ULB

ECOLE
POLYTECHNIQUE
DE BRUXELLES

Paroxysmal Atrial Fibrillation Onset Forecast and Risk Identification During Sinus Rhythm:

A Machine Learning Approach

Thesis presented by Cédric GILON

in fulfilment of the requirements of the
PhD Degree in Engineering Sciences and Technology
("Doctorat en Sciences de l'Ingénieur et Technologie")
Academic year 2023-2024

Supervisor: Professor Hugues BERSINI
Co-supervisor: Professor Stéphane CARLIER

IRIDIA
Institut de Recherches Interdisciplinaires et de
Développements en Intelligence Artificielle

IRIDIA

fnrs
LA LIBERTÉ DE CHERCHER

Paroxysmal Atrial Fibrillation Onset Forecast and Risk Identification During Sinus Rhythm:
A Machine Learning Approach

Cédric GILON

IRIDIA, Université libre de Bruxelles, Belgium

2024

Composition of the jury

Prof. Christine DECAESTECKER, chair
Université libre de Bruxelles, Belgium

Prof. Tom LENAERTS, secretary
Université libre de Bruxelles, Belgium

Prof. Gianluca BONTEMPI
Université libre de Bruxelles, Belgium

Prof. Thierry DUTOIT
Université de Mons, Belgium

Prof. Luc JORDAENS
Erasmus University Rotterdam, Netherlands

Dr. Jean-Marie GRÉGOIRE
Clinical cardiologist
Scientific collaborator, Université libre de Bruxelles, Belgium
Scientific collaborator, Université de Mons, Belgium

Prof. Hugues BERSINI, supervisor
Université libre de Bruxelles, Belgium

Prof. Stéphane CARLIER, co-supervisor
Université de Mons, Belgium

The thesis

Machine learning models can forecast an incoming paroxysmal atrial fibrillation episode moments before its onset. They demonstrate increasing performance as the prediction gets closer to the event.

Moreover, in the comparative analysis of sinus rhythm recordings from patients with and without paroxysmal atrial fibrillation, machine learning models can identify a specific signature of paroxysmal atrial fibrillation within the sinus rhythm.

Summary

The cardiovascular system is a central part of the biological system that constitutes our body. This system ensures that organs and tissues function properly by supplying them with the oxygen and nutrients they need to function and maintain internal balance. Unfortunately, cardiovascular disease is one of the leading causes of death worldwide. It is estimated to be responsible for 17.9 million deaths worldwide, corresponding to 32% of all deaths. Heart attacks and strokes account for 85% of these deaths.

Atrial fibrillation is the most common sustained heart rhythm disorder in adults. Patients with atrial fibrillation have a fivefold increased risk of stroke. This condition affects the rhythmic contractions of the atria, the two upper chambers of the heart, which are disrupted by irregular impulses, resulting in uncontrolled trembling or irregular beating of the muscle. The first stage of the condition is called paroxysmal, in which episodes of atrial fibrillation begin and end spontaneously within 7 days of onset.

In this thesis, we propose a machine learning approach to understand if an incoming paroxysmal atrial fibrillation episode can be forecast moments before its onset. For this purpose, we created a new database composed of electrocardiogram Holter monitoring records from patients with atrial fibrillation. The records were selected retrospectively in three Belgian centres and one Luxembourg centre between 2005 and 2023. Recordings with paroxysmal atrial fibrillation and no major disturbances in signal quality were selected and annotated by a cardiologist and a specialist cardiac nurse. The annotations correspond to the exact time of the QRS complexes of all atrial fibrillation onsets and offsets.

Using this new database, we study the evolution of a prediction made by a machine learning model before the onset of atrial fibrillation. We find that the closer the analysed electrocardiogram window is to the onset of atrial fibrillation, the better the resulting prediction. We then compared several

machine learning models on selected dataset 30-minute electrocardiogram windows close and distant from atrial fibrillation episodes selected from the database. We found that models using heart rate variability and RR intervals performed better compared to models based on the full electrocardiogram signal.

We extended the database with electrocardiogram Holter monitoring records from healthy patients. Using these additional recordings, and building on the previous results, we compare the sinus rhythm windows distant from episodes from patients with and without atrial fibrillation. We show that machine learning models can identify a specific signature of paroxysmal atrial fibrillation within sinus rhythm. The results demonstrate that new machine learning-based strategies could be explored for practical atrial fibrillation screening and treatment.

Keywords

Atrial Fibrillation, Onset Forecast, Prediction, Risk Identification,
Electrocardiogram, RR intervals, Heart Rate Variability
Machine Learning, Deep Learning

Statement

This thesis presents an original work that has never been submitted to the Université libre de Bruxelles or to any other institution for the award of a doctoral degree. Parts of this thesis are based on a number of peer-reviewed articles published in the scientific literature by the author together with his supervisor, co-supervisor and collaborators.

Parts of the state-of-the-art, presented in Chapter 2, are based on:

- Jean-Marie Grégoire, **Cédric Gilon**, Stéphane Carlier, and Hugues Bersini. "Autonomic nervous system assessment using heart rate variability". In: *Acta Cardiologica* 78.6, pp. 648–662 (2023).
- Jean-Marie Grégoire, and **Cédric Gilon**. "Assessing the Autonomic Nervous System using ECG recordings". In: *Asian Hospital & Healthcare Management* 61 (2023).

The first version of the database, presented in Chapter 3, was published and is available online. The description of the materials in this thesis is based on:

- **Cédric Gilon**, Jean-Marie Grégoire, Marianne Mathieu, Stéphane Carlier, and Hugues Bersini. "IRIDIA-AF, a large paroxysmal atrial fibrillation long-term electrocardiogram monitoring database". In: *Scientific Data* 10.1, publisher: Nature Publishing Group, pp. 1–10 (2023).

The experiments and results presented in Chapter 4 are based on:

- **Cédric Gilon**, Jean-Marie Grégoire, Jérome Hellinckx, Stéphane Carlier, and Hugues Bersini. "Reproducibility of machine learning models for paroxysmal atrial fibrillation onset prediction". In: *Computer in Cardiology* (2022).
- Jean-Marie Grégoire, **Cédric Gilon**, Stéphane Carlier, and Hugues Bersini. "Role of the autonomic nervous system and premature atrial

contractions in short-term paroxysmal atrial fibrillation forecasting: Insights from machine learning models". en. In: *Archives of Cardiovascular Diseases* 115.6-7, pp. 377–387 (2022).

- **Cédric Gilon**, Jean-Marie Grégoire, and Hugues Bersini. "Forecast of paroxysmal atrial fibrillation using a deep neural network". In: *2020 International Joint Conference on Neural Networks (IJCNN)* pp. 1–7 (2020).

Parts of the discussion in Chapter 4 and Chapter 5 are based on:

- Jean-Marie Grégoire, **Cédric Gilon**, Stéphane Carlier, and Hugues Bersini. "The potential of artificial intelligence in medical decision making". In: *Revue Medicale de Bruxelles* 43.3, pp. 265–273 (2022).

Acknowledgements

This thesis has been a journey during which I have been fortunate to meet, discuss, debate, benefit from support and collaborate with many people.

First, I would like to thank Hugues, my supervisor, for his support during these years. During all our discussions, you gave me the keys to build this work and helped me find the right path during the rougher times.

I would also like to thank Stéphane, my co-supervisor, who was the medical counterpart, and with Hugues they formed a supervisor duo that allowed me to push this work even further.

This work would not have been possible without Jean-Marie. He was the one who had the first idea for this project. The first recordings were selected from his patients and over the last few years his expertise has enabled us to build up the impressive database presented in this work. Our discussions on Tuesday and Wednesday afternoons allowed us to challenge ideas and encourage each other to make things simple enough for the cardiologist to understand the computer scientist and the computer scientist to understand the cardiologist.

I would also like to thank Marianne, who spent countless hours with Jean-Marie selecting and annotating recordings. Without the work of both of you, this thesis would not be what it is.

I would like to thank the members of the jury for accepting my invitation and for the discussions we had during my private defence, which enabled me to question certain parts of the thesis in order to improve it further. I would especially like to thank Gianluca and Tom, who were also part of my supervisory committee during the years of my thesis. Our annual discussions helped me to keep this work on track and to always come up with new ideas.

From September 2019 to December 2020, I was funded by a C+D partnership between IRIDIA and Procter&Gamble. During this period, I had less time to work on my main subject, but I still had the opportunity to work on machine

learning applications, where I was able to learn new valuable skills.

This work has been supported by the *Fonds de la Recherche Scientifique* - FNRS under FRIA grant n° FC 038733 since December 2020. When I applied for this doctoral fellowship, my application was supported by Ségolène Martin and Shahram Sharif, whom I would like to thank again.

My colleagues at IRIDIA have been invaluable in this work. After the COVID, we had the opportunity to meet again in the lab to discuss and exchange ideas over lunch and in seminars. The IRIDIA'lympics were also part of the highlights. A big thank you to Guillaume and Brice for everything we have shared, laughed and discussed while sitting across from you in our office.

Speaking of COVID, I had the opportunity to work with Carl on a project for FARI. An unexpected friendship grew out of that project and we still get the chance to interact at FARI and for afterworks. Our discussions gave me an inside view and a better understanding of the academic world.

This work has benefited from the results of several master students and interns, who have been able to advance certain areas of my research. These various master theses and internships have forced me to rediscuss my research, to formulate clear objectives, to learn how to support and guide students myself.

This work has benefited from the collaboration of the various cardiologists and hospitals with whom we have built the database: Dr Godart at the CHU Ambroise Paré, Dr Groben at the Centre Hospitalier de Luxembourg, Dr Nguyen at the CHU Brugmann. I would like to thank them all, as our work together has made it possible to further enrich this database. I hope that this project will continue to develop and that other centres will join in.

To my friends and family, with whom I have followed my progress on this thesis over the years. You have helped me to reflect on whether I was going in the right direction. A special thank you to Arnaud, as we started studying computer science together at ESI and since then, our shared curiosity about computers, libraries and new technologies has always pushed us to discover new things. Our specialisations are different now, but that makes our discussions all the more interesting. In addition, our video and board games allow us to spend some timeless moments away from my research bubble.

I would also like to say a special thank you to Yannick. We met during our Masters in Computer Science and discovered each other both professionally and personally. In the second year of my Masters, you talked to me about the idea of doing a PhD, and a few years later I finished mine and yours is well under way. Thank you for all the lunches, coffee breaks and debates that

challenged our ideas and the direction of our research. Thank you again for all the time you have taken to proofread this thesis and discuss the latest changes.

To my parents, Maman et Papa, thank you for awakening and cultivating my curiosity since I was a child. The education, support and love you have given me are the foundations on which I have built the researcher I am today.

Thomas and Alice, my precious siblings, we are so strong together. We have supported each other through difficult times and shared some intense and wonderful moments together. Héloise joined the family some time ago and I have been able to count on her support over the last few years. It makes our moments together even more beautiful.

Finally, the last words are for you, Laura. You have been my biggest supporter all these years and I cannot express how much your motivation has kept me going and enabled me to finally finish this thesis. At times, you were probably more convinced than I was that I would actually finish this thesis. The last year has been particularly challenging, but your motivation, support (and cookies) and love have kept me going. I look forward to discovering where our next journey will take us!

With all my gratitude,
Cédric

28 February 2024

Contents

Acronyms

AC Acceleration

AF Atrial Fibrillation

AFDB MIT-BIH Atrial Fibrillation Database

AFPDB Paroxysmal Atrial Fibrillation Prediction Database

AI Artificial Intelligence

ANN Artificial Neural Network

ANS Autonomic Nervous System

AUC Area Under the Curve

AUPRC Area Under the Precision-Recall Curve

AUROC Area Under the Receiver Operating Characteristic Curve

BIS Bispectral Index

CHL Centre Hospitalier de Luxembourg

CI Confidence Interval

CIED Cardiac Implantable Electronic Device

CNN Convolutionnal Neural Network

CPSC2021 China Physiological Signal Challenge 2021 Database

CSI Cardiac Sympathetic Index

CTM Central Tendency Measure

CVI Cardiac Vagal Index

DC Deceleration

DL Deep Learning

DNN Deep Neural Network

ECG Electrocardiogram

ECGMV ECG Morphology Variability

EEG Electroencephalogram

FFT Fast Fourier Transform

FN False Negative

FNR False Negative Rate

FPR False Positive Rate

GA Genetic Algorithms

HF High Frequencies

HR Heart Rate

HRF Heart Rate Fragmentation

HRV Heart Rate Variability

HRVi HRV Triangular Index

IALS Inverse of the Average Length of the acceleration/deceleration Segments

KNN K-Nearest Neighbours

LF Low Frequencies

LTAFDB Long Term AF Database

ML Machine Learning

NN Normal to Normal

NPV Negative Predictive Value

NSR Normal Sinus Rhythm

PAC Premature Atrial Contraction

PAS Percentage of Alternating Segments

PIP Percentage of Inflection Points

PITP Pill-In-The-Pocket

PNS Parasympathetic Nervous System

PPG PhotoPlethysmoGraphy

PPV Positive Predictive Value

PR Precision-Recall

PRSA Phase-Rectified Signals Average

PSD Power Spectral Density

PSS Percentage of Short Segments

QRS QRS complex

RF Random Forest

RMSSD Root Mean Square of Successive RR interval Differences

ROC Receiver Operating Characteristic

SDNN Standard deviation of NN intervals

SDSD Standard Deviation of Successive RR interval Differences

SNS Sympathetic Nervous System

SODP Second Order Difference Plot

SVM Support Vector Machine

TINN Triangular Interpolation of the NN interval histogram

TN True Negative

TNR True Negative Rate

TP True Positive

TPR True Positive Rate

ULF Ultra Low Frequencies

VHF Very High Frequencies

VLF Very Low Frequencies

XGB XGBoost

Chapter 1

Introduction

The cardiovascular system is a central part of the biological system that constitutes our body. This complex network of heart, blood vessels and blood is responsible for blood circulation. Blood carries oxygen and nutrients to the cells in our body through arteries and removes waste products such as carbon dioxide through veins. This system ensures that organs and tissues function properly by supplying them with the elements they need to function and maintaining internal balance.

At the centre of this system is the heart, a muscular organ that works tirelessly to pump blood throughout the body. Its contractions circulate the blood and are rhythmically controlled by electrical impulses from pacemaker cells. The proper functioning of the heart is therefore essential to sustain life, making it a central issue when discussing cardiovascular health of patients.

The heart is made up of four hollow chambers: two upper chambers, the atria, and two lower chambers, the ventricles. A diagram is shown in Figure 1.1. These chambers are separated by valves and surrounded by the heart wall. This wall is mainly made up of the heart muscle, which is responsible for the rhythmic contraction of the four chambers. The interventricular septum separates the left and right sides of the heart. The right side of the heart receives deoxygenated blood from the body and sends it to the lungs, while the left side receives oxygenated blood from the lungs and sends it to the rest of the body.

Unfortunately, cardiovascular diseases are one of the leading causes of death in the world. It is estimated to be responsible for 17.9 million deaths worldwide, corresponding to 32% of all deaths. Heart attacks and strokes account for 85% of these deaths (WHO 2022). Despite its robustness, the heart is vulnerable to various diseases and malfunctions that can compromise its functioning.

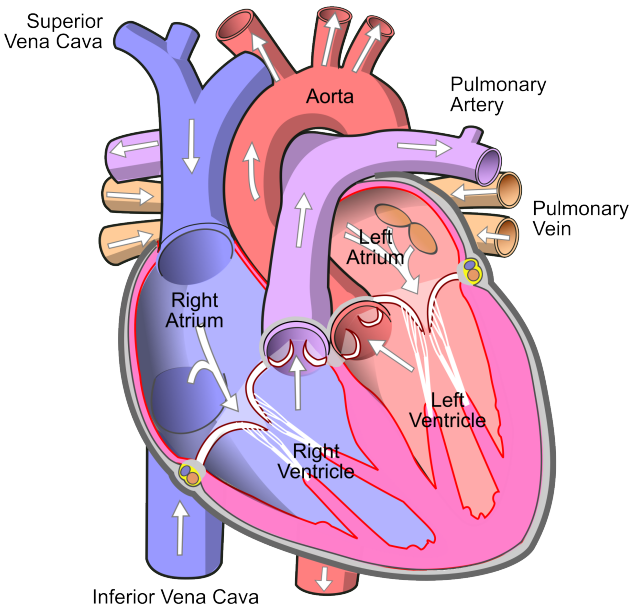


Figure 1.1: Anatomy of the human heart, modified from Wikimedia Commons

These conditions include:

- arrhythmias — abnormal heart rhythms,
- coronary heart disease — restriction of the flow of blood to the heart muscle,
- congenital heart defects — structural problem that are present at birth,
- or heart failure — the muscle is incapable of pumping sufficient blood and meet the blood and oxygen demand of the body.

These diseases can significantly affect the ability of the heart to pump blood efficiently, which can lead to serious health complications. Understanding these diseases, their causes, symptoms, and management is essential to maintaining a healthy heart and preventing potentially life-threatening situations. Researchers are therefore studying them in order to better understand, propose new treatment path and ultimately to treat them better.

In this work, we focus on Atrial Fibrillation (AF), which is the most common sustained heart rhythm disorder among adults. The rhythmic contractions of the atria are disrupted by irregular impulses, resulting in uncontrolled trembling or erratic beating of the muscle. AF it is defined as:

> *a supraventricular tachyarrhythmia with uncoordinated atrial electrical activation and consequently ineffective atrial contraction. Electrocardiographic characteristics of AF include (i) irregularly irregular R-R intervals, when atrioventricular conduction is not impaired, (ii) absence of distinct repeating P waves, and (iii) irregular atrial activations.*
>
> Hindricks et al. (2021)

AF is diagnosed using an Electrocardiogram (ECG), which is the medical test required to diagnose this condition. This disease can often be asymptomatic until revealed by a consequence, such as a stroke. In some cases, AF symptoms such as palpitations, chest pain, fatigue, or dizziness, can be found, but they are not commonly present (Lip et al. 2016). Clinical scores can help physicians to better target patients at risk of AF e.g. CHARGE-AF, or stroke e.g. CHA_2DS_2-VASc (Christophersen et al. 2016). It is important to diagnose AF, as asymptomatic AF patients have a 2-fold increase in risk of mortality compared to symptomatic patients (Boriani et al. 2015).

AF is an evolving disease and multiple stages are defined. The first stage of the disease is called paroxysmal, where AF start and end spontaneously within 7 days after the AF onset. Then, the disease evolves to a permanent stage where the AF crisis lasts more than 7 days. AF episodes can be terminated with a cardioversion, a procedure to restore a regular heart rhythm. It has been observed that 20% to 30% of patients in paroxysmal AF are evolving to a permanent state in the following 4 to 5 years (Al-Khatib et al. 2000; Lakkireddy et al. 2009). Finally, the disease evolves to long-standing persistent AF, where AF is continuous for more than 12 months and a rhythm control strategy is adopted. The term "permanent AF" is used when AF is accepted by both the patient and the physicians. At this stage, no further attempts are made to re-establish sinus rhythm.

The main consequence of this condition is an increased risk of blood pooling in the atria, leading to the formation of blood clots. Patients with AF have a fivefold increased risk of ischaemic stroke (Wolf et al. 1991), i.e. when a blood cloth obstructs a brain blood vessel. It is estimated that 10% to 20% of patients with cryptogenic stroke, i.e. stroke of unknown origin, have paroxysmal AF (Christensen et al. 2014; Sanna et al. 2014). More than 10 000 strokes occur every year in Belgium, with around 85% of strokes labelled as ischaemic stroke (S. Pandya et al. 2011). It is estimated to cost around 393 million euros to the social security system (Wafa et al. 2020).

AF patients also have an increased risk of heart failure, and other cardiac complications (Hindricks et al. 2021). The Framingham Heart Study reported that AF increase risk of death, by 1.5 times for men and 1.9 times for women (Benjamin et al. 1998). In patients with heart failure, AF has also been shown to be an independent predictor of in-hospital mortality and prolonged hospitalization, particularly in the intensive care unit (Rivero-Ayerza et al. 2008).

The global number of patients with AF was estimated at 33.5 millions in 2010 (Chugh et al. 2014). This estimation increase to 60 millions in 2019 (Roth et al. 2020). It is estimated to affect between 1% and 2% of the world population, around 8% of the population over 55 (Krijthe et al. 2013). The prevalence rises to 20% for the population over 80 years old, as the prevalence of the disease increases with the age of the patients (Friberg et al. 2013). In Europe, the lifetime risk of AF is estimated between 1 in 4 and 1 in 3 individuals, with a lower risk for women (Heeringa et al. 2006; Hindricks et al. 2021). The amount of adults aged 55 years and over with AF is projected to double between 2010 and 2060 (Krijthe et al. 2013), therefore the number of patients is likely to continue to grow. In Belgium, the number of AF patient was estimated to 150 000 patients in 2016 (Proietti et al. 2016). Other risk factors in addition to age and male gender include ethnicity, smoking, alcohol intake and obesity, which influence the risk of developing AF (Lip et al. 2016).

The management of AF includes a wide variety of techniques, as simple as lifestyle changes and treatment of comorbidities or the use of drugs such as anticoagulants, surgical procedures such as pulmonary vein isolation using catheter ablation (Marrouche et al. 2018) and electrical shocks (Samuel Lévy et al. 1997). Finally, Cardiac Implantable Electronic Device (CIED) have also been proposed for the management of AF with devices such as implantable pacemakers and atrial defibrillators (Wellens et al. 1998; Cooper et al. 2002). All of these techniques aim to regulate the heart rhythm and ultimately reduce the risks to the patient. Of these, catheter ablation is the most invasive, as it involves creating scars to electrically isolate the pulmonary veins, which have been found to be a major source of ectopic beats and AF impulses (Haïssaguerre et al. 1998).

An ECG is required to medically diagnose AF. This non-invasive test records the heart electrical activity from the patient skin. For AF diagnosis, ECG can be either a 12-lead recording or a single lead recording, showing heartbeats with irregular RR intervals and no discernible repetitive P waves.

ECG have been studied from the end of the 19th century, with Waller (1887) demonstrating the first recording of a human electrocardiogram. In 1902, Einthoven developed the first practical ECG device and assigned the nomenclature used to describe the different waves that make up the ECG. In 1924, he received the Nobel Prize in physiology and medicine for his discovery of the mechanism of the electrocardiogram. He described an ECG as irregular and unequal, which might correspond to the first recording of AF, in his work of 1906 (Einthoven 1906). It was later recognised as AF and described as a common clinical condition (Lewis 1909). Today, an increasing number of ECG are recorded each year, with an annual estimate of more than 300 million ECG recordings (Zhu et al. 2020).

The quality of recording devices has improved over the years and there is now a wide range of devices that can be used for AF screening. The most commonly used devices in clinical practice for AF are the ambulatory long-term Holter monitor and the implantable cardiac monitor (Zimetbaum et al. 2010; Podd et al. 2016). In Belgium, about 270 000 Holter monitoring recordings are performed each year, 75% of which are performed in outpatient clinics (Meeus et al. 2023). The average age of the patients is 64 years. This represents an average annual cost of 18 million euros to the Belgian social security system.

Today, wearables are emerging as a compelling alternative to traditional ECG and Holter monitors, offering a convenient way to track heart rhythm irregularities. Recording can be continuous or intermittent, but with more regular follow-up. Devices such as smartwatches, mobile ECG, connected ECG patches are proposed as alternatives to facilitate AF screening.

In parallel with hardware improvements, researchers have been exploring the use of algorithms and models to automate the interpretation of ECG for the past 25 years (Holst et al. 1999). Machine Learning (ML) approaches have been proposed for automatic ECG interpretation, particularly for AF detection. ML can be distinguished from the classical programming paradigm, in which the programmer defines rules that compute output results from input data. The ML paradigm allows the computer program to learn rules from the combination of data and expected results without being explicitly programmed for the task.

The first use of the term ML is attributed to the work of Samuel (1959). More formally, ML can be defined as:

a computer program is said to learn from experience E with respect to some

> *class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.*
>
> Mitchell (1997)

In this work, we use supervised learning to train ML models, which is the case when both data and annotations are available. Three main consecutive phases can be described.

1. The first phase is model training, where the model learns from data and expected responses. During this training phase, the model adapts to reduce the error on the training data.

2. At the end of training, model performance is evaluated by testing the model on test data, i.e. new data not seen during training. The testing phase validates that the model is able to generalise to new examples.

3. Finally, in the third phase, the trained model is used for inference on the selected task.

To learn insight from the ECG, we feed the ML model with handcrafted features extracted from the ECG data. The definition of the features relies on the technical experts and cardiologists.

Artificial Neural Network (ANN) and Deep Neural Network (DNN) represent an evolution in ML, allowing complex data analysis without the need for manually designed features. It enables autonomous feature extraction and representation learning of the input data. In Artificial Intelligence (AI), ANN form the foundation for models inspired by the structure and function of the human brain, mimicking its interconnected neurons to process information and perform complex tasks. The perceptron, a fundamental concept in neural networks, represents a simplified model of interconnected neurons (McCulloch et al. 1943; Rosenblatt 1958).

The transition from ANN to DNN has been accelerated by advances in both computing power and the availability of big data. This development has enabled the creation of deeper and more powerful models capable of handling complex tasks in different domains. However, the downside of DNN is their reduced interpretability, which poses challenges in understanding and explaining the decision-making process due to their deep and complex internal connections. DNN are often described as black boxes because they encapsulate the internal mechanisms and reasoning, making them difficult to interpret.

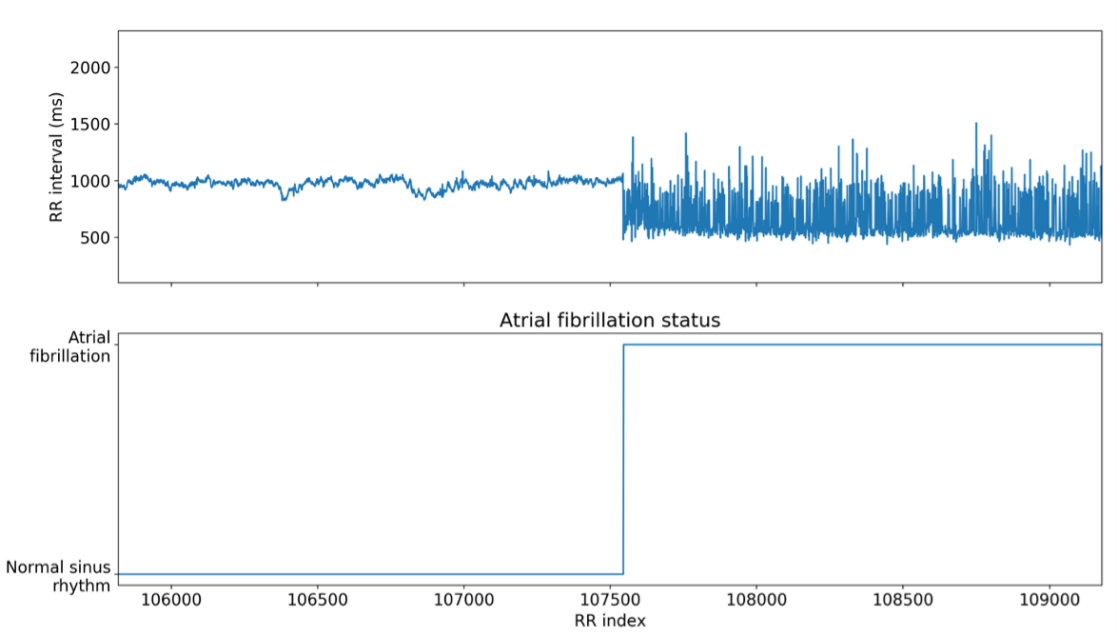


Figure 1.2: Heart rhythm transition from normal before AF onset to irregular after AF onset

The aim of this thesis is to apply the ML approach to the prediction of AF onset and the identification of risk of AF during sinus rhythm. This research is motivated by the potential of these methods to (i) predict early warning signs of AF in the windows preceding AF onset, as presented in Figure 1.2, and (ii) identify the AF signature in sinus rhythm distant from AF. This research opens up improved opportunities for early detection, improved screening and proactive management of this heart condition.

Thesis contribution and structure

The thesis contributes to the field of AF research across several key areas. Beginning with Chapter 2, we review the current state-of-the-art methodologies, including Heart Rate Variability (HRV) analysis, as well as ML models and Deep Learning (DL) models using ECG data presented in the literature for AF detection, onset forecast, and risk identification. Additionally, publicly ECG databases available to researchers are reviewed.

In Chapter 3, we introduce a new large-scale annotated database designed to extend the existing material for AF onset forecast. The validity of these annotations is rigorously tested using ML models.

Chapter 4 focuses on AF onset forecast. First, we explore the reproducibility of existing ML models presented in the literature. Then, we study the evolution of ML methods prediction and performance preceding AF onset. Following those results and using a benchmarking process, we compare the performance of various ML techniques using a diverse range of inputs derived from ECG data. The aim of this in-depth study is to improve the performance and reliability of models for the AF onset forecast.

Continuing the main contributions of this work, Chapter 5 explores the aspect of AF risk identification during sinus rhythm. As for the previous chapter, this chapter presents a benchmark study leveraging ML approaches. We extend the study of model performance by a comparative analysis across different age groups. This analysis provides an overview of AF risk profiles among the distinct groups.

Together, these contributions aim to improve the understanding, prediction and proactive management of AF, with the objective of improving strategies and interventions to improve the heart health of AF patients. The work is concluded in Chapter 6, where we summarise the results achieved within this thesis, while outlining prospective directions for future research following this work.

Chapter 2

State of the art

2.1 Introduction

Heartbeats irregularities and RR intervals variability is a key indicator of Atrial Fibrillation (AF), and can be assessed through Heart Rate Variability (HRV) analysis. In this chapter, we first introduce a selected overview of the scientific literature on HRV and three subdomains: time-domain measurements, frequency-domain measurements and geometric methods. HRV is used as a tool in this thesis to measure of the heart rate changes before or during AF crisis.

Then, we explore the existing databases of Electrocardiogram (ECG) that are publicly available to researchers for AF related research. Database access remains one of the main challenges for AF research and reproducibility. In the last years, a concentration of the research on a limited number of databases have been observed in various Machine Learning (ML) task communities (Koch et al. 2021).

Finally, we define and explore three AF predictions tasks: AF detection, AF onset forecast, and AF risk identification during Normal Sinus Rhythm (NSR). We study ML approaches proposed in the literature.

2.2 Heart rate variability

2.2.1 Human heart and electrocardiogram

The human heart is the central muscle of the cardiovascular system. It ensures the continuous circulation of blood through the body into veins and arteries.

The heart is composed of four hollow chambers : two upper chambers called atria and two lower chambers called ventricles. The heart is divided vertically into right and left sides, each consisting of an atrium and a ventricle, separated by a valve. A second valve is present in each side at the ventricle exit, i.e. the pulmonary valve for the right side and the aortic valve for the left side. A diagram of the heart can be found in the previous chapter as Figure 1.1. The right side of the heart receives deoxygenated blood from the body and expels it to the lungs for re-oxygenation. The left side of the heart receives newly oxygenated blood from the lungs and sends it to the body through the aorta.

Anatomically, the orientation of the left and right sides of the heart is conventionally described in relation to the position of the human body, with the left side of the heart referred to as the area closer to the left side of the body and the right side correspondingly closer to the body. Therefore, the anatomical orientation of the heart in schematic representations or diagrams is typically shown in reverse, with the left side of the heart appearing on the right side of the illustration and vice versa. This intentional inversion matches the perspective of an observer looking at the diagram, allowing for a clearer understanding and correspondence with the actual orientation of the body, despite the visual inversion in the diagrammatic representation.

The contraction of the heart muscles takes place in several phases, all following the rhythm of the sinoatrial node, a group of cells located in the right atrium. These cells are the natural pacemaker of heartbeats. The electrical signal for contraction passes through the cardiac nervous system, first to the atria, then to the atrioventricular node and finally to the ventricles. The heart muscle contracts following the electrical impulse, and each heartbeat follows the same cardiac cycle.

The heart electrical activity can be recorded from the skin of the patient using electrodes. The resulting recording is called an ECG, a graph of the recorded voltage versus the time. For a healthy heartbeat, the signal is called sinus rhythm or NSR, as shown in Figure 2.1. A single heartbeat is made up of three main parts:

- P wave — the contraction of the atria,
- QRS complex — the contraction of the ventricles
- and T wave — the repolarization of the ventricles.

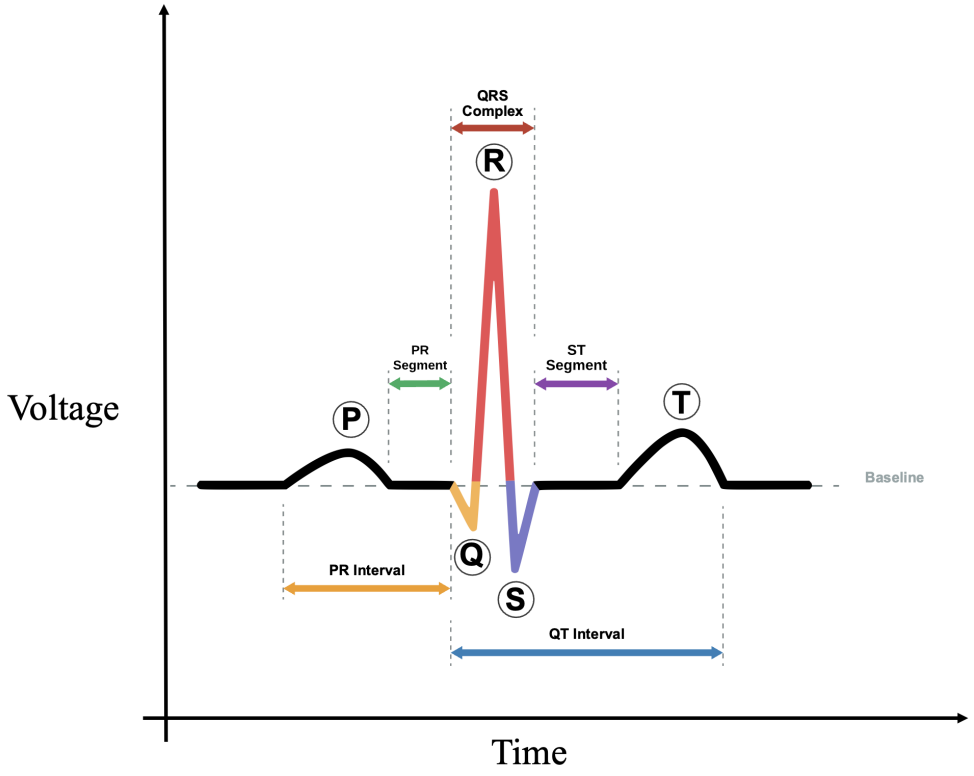


Figure 2.1: Electrocardiogram of one heartbeat in normal sinus rhythm, from Wikimedia Commons

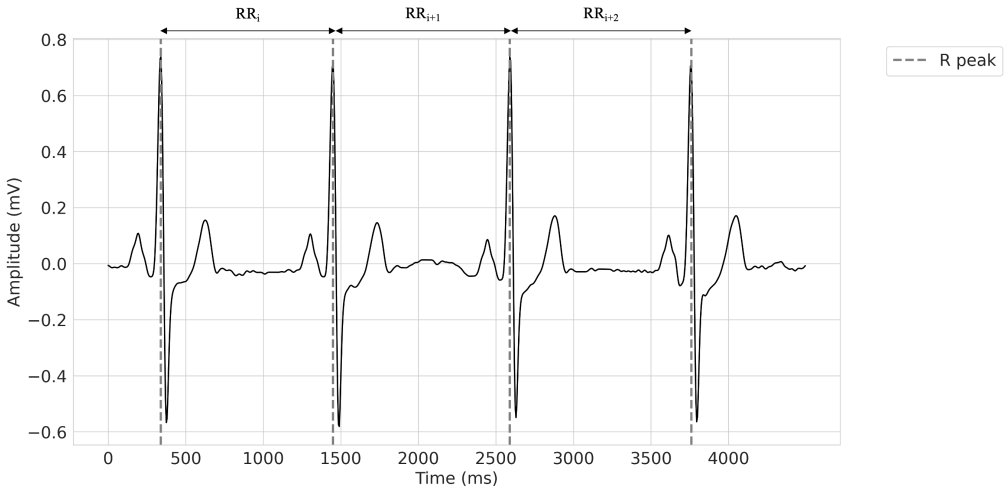


Figure 2.2: RR interval corresponds to the distance between consecutive R peak. Each R peak is used twice to build the RR interval series.

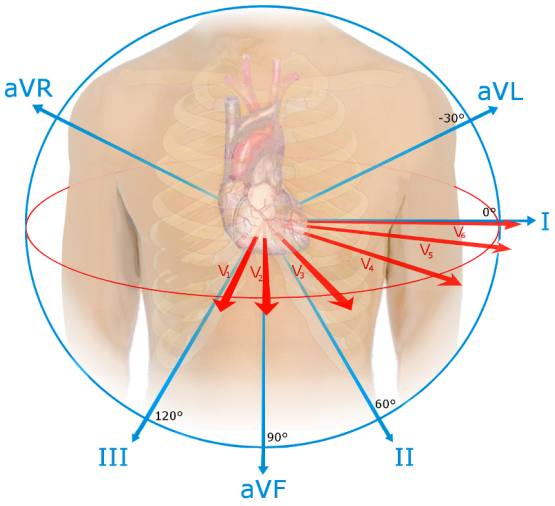


Figure 2.3: 12-lead ECG, composed of 6 red leads in the coronal plane (vertical) and 6 blue leads in the transverse place (horizontal), from Wikimedia Commons

Intervals and segments are measured between the waves and peaks. RR intervals are the time between two consecutive QRS complexes, as presented in Figure 2.2. PR intervals are from the beginning of the P wave to the beginning of the QRS complex. The QT interval follows the PR intervals, from the beginning of the QRS to the end of the T wave. The TQ interval is measured between two beats, from the end of the T wave to the beginning of the next QRS complex.

For the measurements, electrodes are used in pairs to form a lead. The standard recording is a 12-lead from 10 electrodes. The 12 leads provide 12 views of the electrical conduction in the heart, as shown in Figure 2.3. The first 6 leads are in the coronal plane, which is the vertical plane that divides the body into dorsal and ventral sections. These leads can be divided in two categories: (i) the 3 limb leads, lead I, lead II and lead III and (ii) the 3 extended limb leads aVR, aVL and aVF. The precordial leads, labelled V1 to V6, are in the transverse plane, i.e. the horizontal plane dividing the body into superior and inferior sections. The limb leads are bipolar: they use a pair of electrodes to measure the signal. The extended limb leads and the precordial leads are unipolar, using an artificial reference point as a second measuring point.

For AF screening, ambulatory Holter monitorings are used to record the patient heart rhythm during a longer period of 24 hours up to a week. Today, wearables are being proposed as an alternative to standard ECG. PhotoPlethysmoGraphy (PPG) recorders (Pereira et al. 2020; Guo et al. 2021), smart-

watches (Perez et al. 2019; Wasserlauf et al. 2019), smartphones (Proesmans et al. 2019; Freyer et al. 2021; Rizas et al. 2022), wearable mobile ECG (Lopez Perales et al. 2021; Kleiman et al. 2021), chest straps (Hartikainen et al. 2019), and ECG patches (Turakhia et al. 2013; Vijayan et al. 2023) have been studied as tools for AF screening.

The many facets of the cardiac rhythm and cardiac arrhythmias can be analysed using Heart Rate (HR) measurements from long-term ambulatory ECG recordings and HRV measurements. HRV is a measure of the variation in the intervals between successive heartbeats. The study of HR, RR intervals and HRV allows the examination of a cardiovascular state of a patient. They have the advantage of being non-invasive and can be carried over an extended period of time. Additionally, other studies have shown the usefulness of HRV for identifying risks in other diseases, such as coronary artery disease (Goldenberg et al. 2019), heart failure and sudden cardiac death (La Rovere et al. 2003; Sessa et al. 2018), and ventricular tachycardias (Lee et al. 2016).

2.2.2 Autonomic nervous system

Heart rate is regulated by the Autonomic Nervous System (ANS) through the intrinsic cardiac nervous system, and ANS variations can be measured by HRV. The ANS is the subsystem that regulates the involuntary functions of the body, including heart rate, digestion and the immune system. This system is directly under the control of the central nervous system. Within the ANS are the Sympathetic Nervous System (SNS) and the Parasympathetic Nervous System (PNS). They control the same parts of the body and the same general functions, but with opposite effects. While the SNS sends signals that alert the body's systems, the PNS sends signals that relax these systems, using in particular the vagus nerve, which is the main nerve of the PNS. The ANS has been shown to play an important role in the development of atrial and ventricular arrhythmias (Zhu et al. 2019). ANS can be studied by HRV analysis.

2.2.3 Time-domain methods

Time domain measurements are a common method used to study the variability of RR interval time series. During an ECG measurement, arrhythmic events such as Premature Atrial Contraction (PAC), body movement or sensor malfunction can cause artefacts (Citi et al. 2012). PAC are irregular heartbeats that

originate in the atria and cause an early heartbeat before the regular rhythm resumes. Sinus rhythm beats must be distinguished from premature beats, as premature beats distort the results of the calculation of HRV parameters. In addition, R peak artefacts can also affect the HRV measurements, and to ensure the validity of the intervals, only normal R peaks are selected to construct the Normal to Normal (NN) intervals. In scientific ML publications, RR and NN intervals are often used as interchangeable (Tarvainen et al. 2014; Kubios 2023). Moreover, the description of the method in scientific publications does not always explicitly state whether the RR intervals are used as recorded, or whether a correction has been applied and therefore corrected intervals are used.

RR interval variations can be characterised by parameters from several subdomains. The first is the time domain, also known as the statistical domain. A summary of time domain measurements is presented in Table 2.1. Value-based parameters such as mean, median, minimum and maximum values are the first descriptors of the RR interval serie. HRV percentiles such as 20th percentile and 80th percentile are used by Han et al. (2017) and Hovsepian et al. (2015) to detect stress in ECG recordings.

Deviation-based parameters such as Standard deviation of NN intervals (SDNN) reflect the distribution of the components responsible for HRV during the recording period. The conventional measurements are made on 5 minutes recordings and (Shaffer et al. 2017), but ultra-short-term SDNN measurements in 60 seconds recordings have been proposed in the literature (Salahuddin et al. 2007). The regularity of heartbeats can be disturbed by PAC.

Difference-based parameters, such as the Standard Deviation of Successive RR interval Differences (SDSD), the pNN_x or the Root Mean Square of Successive RR interval Differences (RMSSD) are related to deviation-based parameters, i.e. SDNN, but represent shorter-term variability. The pNN_x represents the percentage of RR intervals with a difference higher than x ms. The pNN_{50}, i.e. the percentage of adjacents RR intervals that differ by more than 50 ms, was first proposed by Bigger et al. (1988) and is the most commonly used in the literature.

Deceleration capacity and acceleration capacity

Bauer et al. (2006) proposed a Phase-Rectified Signals Average (PRSA) approach to the analysis of RR interval series. This algorithm allows the separate characterisation of HR Deceleration (DC) and Acceleration (AC) over long

Table 2.1: HRV time-domain measurements

Sud-domain	Measurements	Unit	Description
	HR	bpm	Heart rate
Value-based	MeanNN	ms	Mean value of NN intervals
	MinNN	ms	Minimal value of NN intervals
	Prc_xNN	ms	xth percentile of NN intervals
	MaxNN	ms	Maximal value of NN intervals
	MedianNN	ms	Median value of NN intervals
Deviation-based	SDNN	ms	Standard deviation of NN intervals
	CVNN	%	SDNN divided by MeanNN
Difference-based	RMSSD	ms	Root Mean Square of Successive RR interval Differences
	SDSD	ms	Standard Deviation of Successive NN interval Differences
	CVSD	%	RMSSD divided by MeanNN
	pNN_x	%	Percentage of successive NN intervals that differ from each other by more than x ms
PRSA-based	DC	ms	Decceleration capacity
	DC_{mod}	ms	Modified DC
	DC_k	ms	Modified (Kubios) DC
	AC	ms	Acceleration capacity
	AC_{mod}	ms	Modified AC
	AC_k	ms	Modified (Kubios) AC

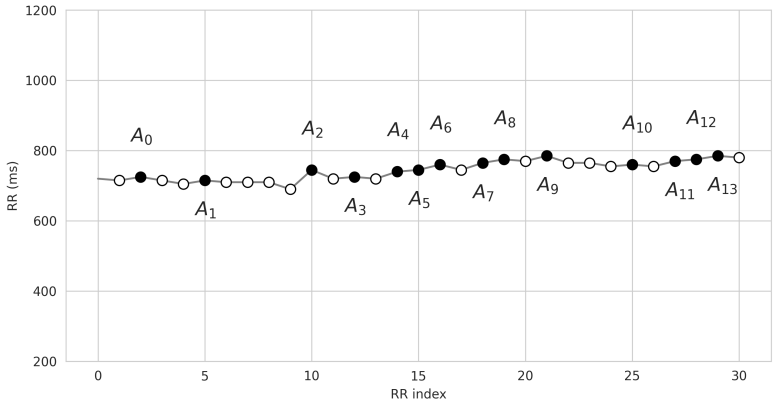
recording times. This method, summarised in Figure 2.4, proposes to align RR interval segments with respect to selected anchor points. The anchor points are selected according to AC or DC in the signal, i.e. RR intervals larger than the previous RR interval for DC and RR intervals smaller than the previous RR interval for AC. Then, windows of interest are defined around these anchor points and windows are stacked. The values obtained from all points i in the windows are averaged to create an average segment. Finally, points of interest X_i are selected around the average anchor point X_0.

DC and AC are computed using Equation (2.1) and Equation (2.2). The equations are similar, but the selection of anchors points is different. An alternative method was later proposed by Nasario-Junior et al. (2014), proposing AC_{mod} and DC_{mod} measurements. Finally, the Kubios HRV analysis software use a third method to produce AC_k and DC_k measurements (Tarvainen et al. 2014).

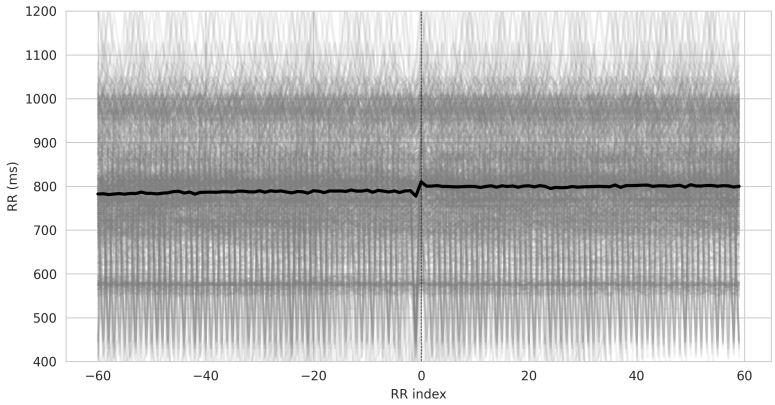
$$\begin{aligned} DC &= \frac{(x_i + x_{i+1}) - (x_{i-1} + x_{x-2})}{4} \\ DC_{mod} &= x_i - x_{i-1} \\ DC_k &= \frac{(x_i + x_{i+1}) - (x_{i-1} + x_{x-2})}{2} \end{aligned} \tag{2.1}$$

$$\begin{aligned} AC &= \frac{(x_i + x_{i+1}) - (x_{i-1} + x_{x-2})}{4} \\ AC_{mod} &= x_i - x_{i-1} \\ AC_k &= \frac{(x_i + x_{i+1}) - (x_{i-1} + x_{x-2})}{2} \end{aligned} \tag{2.2}$$

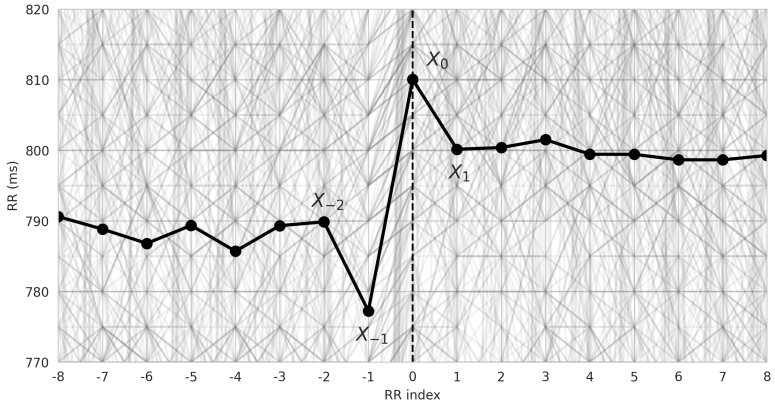
HRV analysis through AC and DC method offers the possibility to analyse periodic behaviours related to HR, which could provide more differentiated information about the autonomic regulation processes of the heart. AC and DC measurements were used successfully for AF detection by Maji et al. (2014). Chen et al. (2018) suggested that AC and DC could distinguish and quantify the roles of SNS and PNS in discriminating AF recurrence after ablation. They showed that DC and AC could discriminate between AF recurrence-free and AF recurrence patients, whereas traditional HRV measures failed to do so. This result was confirmed by Călburean et al. (2021), who found DC capacity allowed for predicting the recurrence of AF after a repeated catheter ablation



(a) Selection of DC anchors points A_i



(b) Selected segments superpositions around anchors points and average DC segment creation



(c) Points of interest in the average DC segment

Figure 2.4: Construction of the DC measurement from the RR interval series. (a) Anchor points A_i selection: RR intervals larger than the previous interval. For AC, anchors points are RR intervals smaller than the previous interval. (b) Average segments construction from the superposition of selected segments. (c) Points of interest X_i selection in the average segment.

procedure. Finally, Pan et al. (2016) investigated the correlation between DC, AC and ANS activity using an experimental model and showed that DC and AC reflect the same aspect of ANS activity and depend exclusively on PNS activity.

2.2.4 Frequency-domain methods

Frequency-domain measurements allow the extraction and analysis of additional information contained in the sequence of RR intervals. Power Spectral Density (PSD) estimation is computed from the interpolated RR interval series. The interpolation is needed to transform the RR intervals into equidistantly sampled series (Berger et al. 1986). In this work, we use linear interpolation and the PSD estimation is computed using Fast Fourier Transform (FFT)-based method (Welch 1967). The RR intervals are divided in overlapping subwindows and the PSD estimation is computed for each subwindow. The results correspond to the average of the spectra of all samples. From the final PSD, the different power bands of the signal can be highlighted and measured.

Autocorrelation methods have also been proposed in the literature and have been compared with FFT-based methods. They were found not to be interchangeable as such, but the normalisation process allows a comparison between the two methods (Fagard et al. 1998; Pichon et al. 2006), as the proportional change between the defined HRV frequency bands can be considered roughly equivalent, regardless of the spectral method used as studied by Heathers (2014). It therefore allows a degree of comparability and interpretability between studies using the two methods. In this work, we use FFT-based methods to estimate frequency-based measurements.

The significance of the different frequency bands obtained in this way needs to be established. One of the keys to understanding ANS activity lies in the greater speed of action of PNS and vagal activity, of the order of 500 milliseconds, which enables it to exert beat-to-beat control. In comparison, the SNS has an adrenergic reactivity with an estimated speed of action of the order of a few seconds. The permanent interactions between the two parts of the ANS complicate the interpretation of the results obtained. HRV frequency-based parameters have previously been used by Pomeranz et al. (1985) and Hayano et al. (2019) to estimate the state of the ANS.

The frequency band for the interpretation of HRV is generally divided in multiple bands between 0.003 and 0.9 Hz, as presented in Table 2.2 (Task Force

Frequency band		Lower bound (Hz)	Higher bound (Hz)
Ultra Low Frequency	(ULF)	0	0.003
Very Low Frequency	(VLF)	0.003	0.04
Low Frequency	(LF)	0.04	0.15
High Frequency	(HF)	0.15	0.4
Very High Frequency	(VHF)	0.4	0.9

Table 2.2: General frequency bands of the HRV

of The European Society of Cardiology and The North American 1996). The total power is measured for each band, by integrating between the lower and higher bounds.

LF and HF bands

The HF and LF bands constitute the majority of the short-term recording signal band, as shown in Figure 2.5. Frequencies between 0.04 Hz and 0.4 Hz are associated with SNS or PNS control mechanisms. The LF band indirectly represents at least part of the SNS activity. It is related to the baro-receptor reflex, the mechanism that keeps the blood pressure in the body at a constant level. Therefore, part of the LF can be linked to PNS and vagal activity, as the vagal nerve mediates this reflex. As a result, the participation of SNS or PNS in the LF band is highly variable, depending on the position and activity of the subjects. This variable participation complicates interpretation and invalidates the fact that LF can be considered solely representative of SNS. Lombardi et al. (2004) describes SNS modulation as characteristic of the majority of AF onsets. The HF band reflects mostly PNS activity. Because the SNS is relatively slow, it cannot generate significant fluctuations at frequencies above 0.15-0.20 Hz. Therefore, all nervous contributions to HR spectra at higher frequencies are essentially of PNS origin. In addition, respiratory activity affects the HF band (Bernston et al. 1997; Shaffer et al. 2017), so the significance of the HF band is questionable in the presence of abnormal respiratory rates, i.e. less than 9 per minute or greater than 24 per minute (Song et al. 2003).

Normalised LF and HF

Normalized Low Frequencies $(LF)_{nu}$ and High Frequencies $(HF)_{nu}$ are computed with respect to the total power limited in the two bands. For short

recordings, LF_{nu} and HF_{nu} are normalized with respect to only the LF and HF band, i.e. from 0.04 Hz to 0.4 Hz. As shown by Burr (2007), there is redundant information between LF_{nu}, HF_{nu} and the LF/HF ratio, as VLF are not taken in account.

For longer recordings, from at least 5 minutes to 24 hours, the normalized values are computed from an extended power spectrum, including the VLF in the selected band. This resolves the redundancies described above. The distribution of spectral values does not follow a normal distribution. A logarithmic scale can be used to obtain a distribution closer to the normal distribution.

LF/HF ratio

The LF/HF ratio has previously been used as an index of SNS-PNS interactions, to study the autonomic nervous system modulations and its impact on the heart rate. Because of the complexity of the relationship between two systems, we should be cautious about drawing conclusions about interactions from this ratio alone (Billman 2013). Expressions such as autonomic modulations (Lombardi 2002) or ANS responsiveness (Malik et al. 2019) have been suggested as alternatives in the literature to describe this ratio.

ULF, VLF and VHF

The literature on the extreme frequency bands of PSD analysis of HRV is scarce. The ULF are associated with the circadian cycle (Shaffer et al. 2014). The Very Low Frequencies (VLF) may correspond to slower rhythms, such as hormonal rhythms and thermoregulation. Chang et al. (2014) suggests that Ultra Low Frequencies (ULF) and VLF could also result from artefacts due to the non-stationarity of the signal over longer periods of time. Very High Frequencies (VHF) are associated with respiration and changes in body tilt.

Bispectrum

The bispectrum measures the phase couplings and interactions between the frequencies in a signal (Pinhas et al. 2004). It can reveal non-linear interactions that are not captured by other methods, such as spectral analysis. The main features extracted from the bispectrum plot are the power averages in each of the regions of interest, as shown in Figure 2.6. The bispectrum plot contains symmetrical regions, between 0.04 Hz and 0.4 Hz, so only a subset is analysed.

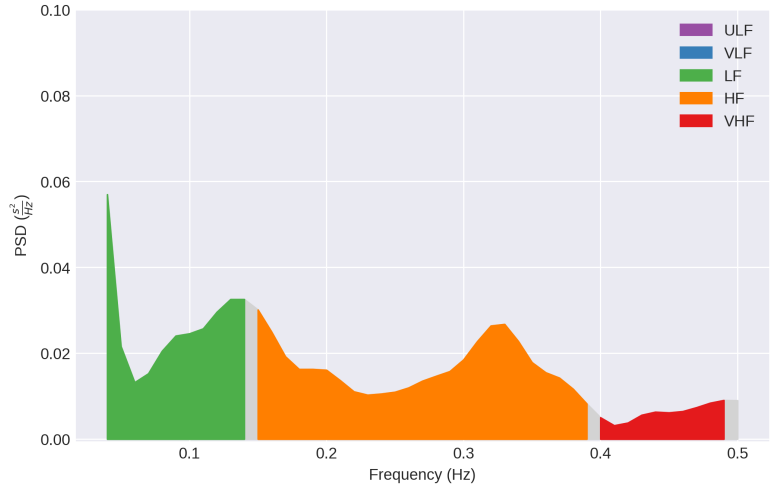


Figure 2.5: HRV PSD estimation of a 5-minute ECG window using FFT-based method

Table 2.3: HRV frequency-domain measurements

Measurements	Unit	Description
LF	ms^2	Low frequency power (0.04 Hz—0.15 Hz)
LF_{nu}	%	Normalized LF
HF	ms^2	High frequency power (0.15 Hz—0.4 Hz)
HF_{nu}	%	Normalized HF
Total power	ms^2	Total power (0 Hz—0.5 Hz)
LF/HF	ratio (%)	Ratio between LF and HF

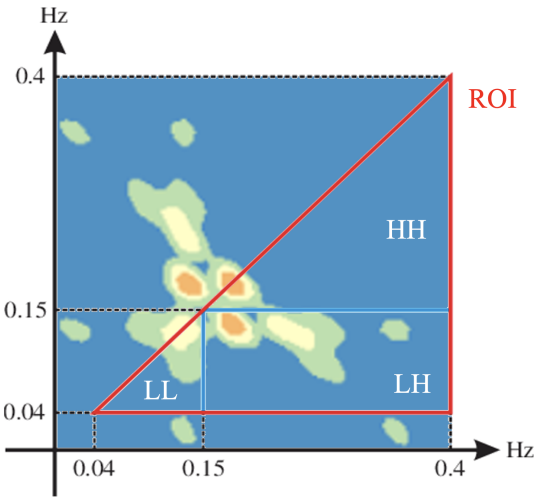


Figure 2.6: Bispectrum plot showing the interaction between the frequencies. The regions of interest (ROI) are interactions between (i) LL: LF and LF, (ii) LH: LF and HF, and (iii) HH: HF and HF

One of the most common use of bispectrum analysis is in anaesthesia monitoring through the Bispectral Index (BIS) score calculated from the patient's Electroencephalogram (EEG) (Mathur et al. 2024). Its use has also been proposed for HRV analysis (Saliu et al. 2002), heart failure detection (Yu et al. 2012) and AF onset forecast (Mohebbi et al. 2012).

2.2.5 Geometric methods

We selected four geometric methods proposed in the literature to analyse HRV: histogram, Poincaré plot, Second Order Difference Plot (SODP) and recurrence plot. The features from these plots are summarized in Table 2.5.

Histogram

The interval sequence can first be converted into histograms, which already allows visualisation of the distribution of HRV. Two features are computed from the RR interval histogram as shown in Figure 2.7. The first is the Triangular Interpolation of the NN interval histogram (TINN) which corresponds to the baseline width of the histogram. The second is the HRV Triangular Index (HRVi) which corresponds to the total number of RR intervals divided by the most present RR values. The total number of bins selected to construct the histogram affects both values. The value of approximately 8 ms, $\frac{1}{128}\ s = 7.8125\ ms$,

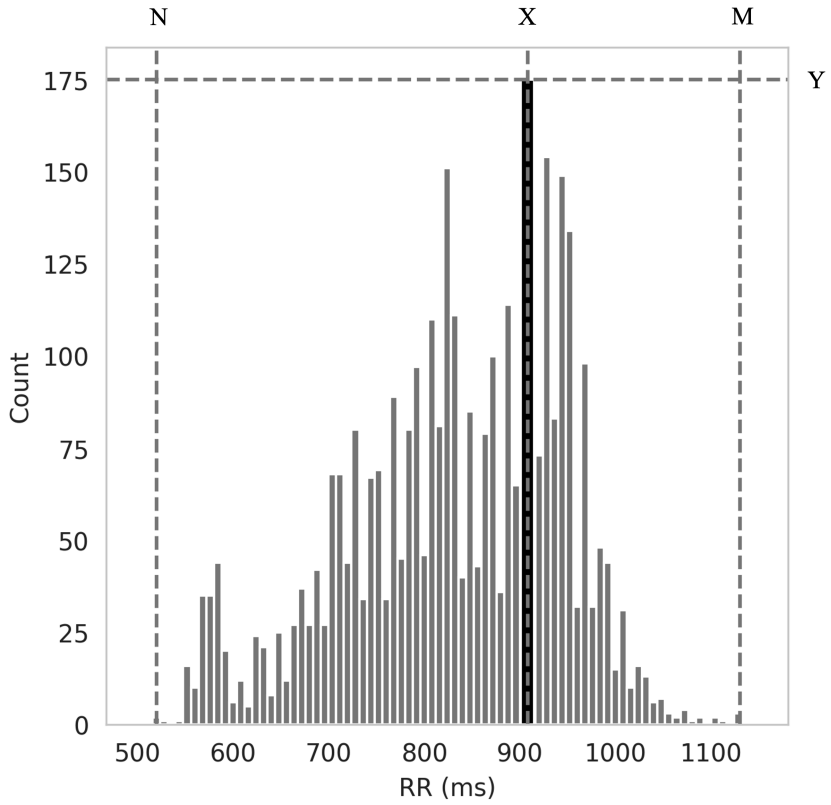


Figure 2.7: Sample distribution of the RR intervals of 1 hour recording, built using 8 ms bins. The most frequent RR interval duration is marked with X and a black color. The corresponding number of RR intervals in the bin is marked as Y. The M and N are marking the corresponding maximal and minimal RR value in the RR intervals. The HRVi correspond to $HRVi = \frac{N_{RR}}{Y}$ where N_{RR} corresponds to the total number of RR intervals. The TINN correspond to the width of the distribution, i.e. $TINN = M - N$.

was previously proposed (Task Force of The European Society of Cardiology and The North American 1996) to correspond to the 128 Hz accuracy of medical equipment at the time of the proposal. In this work, we have chosen to round it up to 8 ms for the construction of the histogram. We could have lowered it to 5 ms as the sampling frequency of the recording is 200 Hz, but we chose to keep 8 ms to correspond to the conventional duration. HRVi has been shown to be a predictor of cardiovascular mortality in patients with AF (Hämmerle et al. 2020).

Poincaré plot

The Poincaré plot is a scatter two-dimensional plot from the RR intervals, as show in Figure 2.8. Each successive pair of RR intervals is plotted as one point where $x = RR_i$ and $y = RR_{i+1}$. Their use has been proposed as been proposed to easily visualize the HRV as a point cloud (Brennan et al. 2001).

The two main measurements are SD1 and SD2. They correspond to the two standard deviations of the point cloud: (i) SD1 perpendicular to the identity axis and (ii) SD2 parallel to the identity axis. SD1 is an index of short-term HRV variations and has been shown to be equivalent to RMSSD from the time domain (Ciccone et al. 2017). On the other hand, SD2 is an index of long-term HRV variation. The ratio SD1/SD2 and the area S are also measured. Finally, Cardiac Sympathetic Index (CSI) and Cardiac Vagal Index (CVI), linked to the PNS, were introduced by Toichi et al. (1997) and are described in Equation (2.3). These measurements are based on

- l - the full length of the point cloud parallel to the identity axis,
- t - the full width of the point cloud, i.e. perpendicular to the identity axis

l and t can be estimated using four times SD1 and SD2. Jeppesen et al. (2014) introduces The modified CSI, with the aim of emphasising the l axis, i.e. long-term variations, proportionally in the CSI calculation.

$$\begin{aligned} t &= 4 * SD1 \\ l &= 4 * SD2 \\ CSI &= l/t \\ CVI &= \log_{10}(l \times t) \\ CSV_{mod} &= l^2/t \end{aligned} \tag{2.3}$$

The usefulness of Poincaré plots for systematic AF screening has been demonstrated by Kisohara et al. (2020), who used them as input to Convolutionnal Neural Network (CNN) models. It also allows visualization of the global aspect of the rhythm, and their automatic recognition is likely to be used to facilitate the analysis of the big data of ECG recordings received by mobile health applications (Lopez Perales et al. 2021). Poincaré plots are also used for other medical signal classification, such as the diagnosis of epileptic seizures using EEG (Goshvarpour et al. 2020).

Quadrant	x_i	y_i	RR relationship	Description
I	DC	DC	$RR_1 < RR_2 < RR_3$	Deceleration of the HR over three successive RR intervals
II	AC	DC	$RR_1 < RR_2 > RR_3$	A long interval is surrounded by two short intervals
III	AC	AC	$RR_1 > RR_2 > RR_3$	Acceleration of the HR over three successive intervals
IV	DC	AC	$RR_1 > RR_2 < RR_3$	A short interval is surrounded by two long intervals

Table 2.4: Four quadrants in the SODP

Second-order difference plot

The SODP is a recurrence plot using the difference of successive pairs of intervals, as described in Equation (2.4) (Babloyantz et al. 1996).

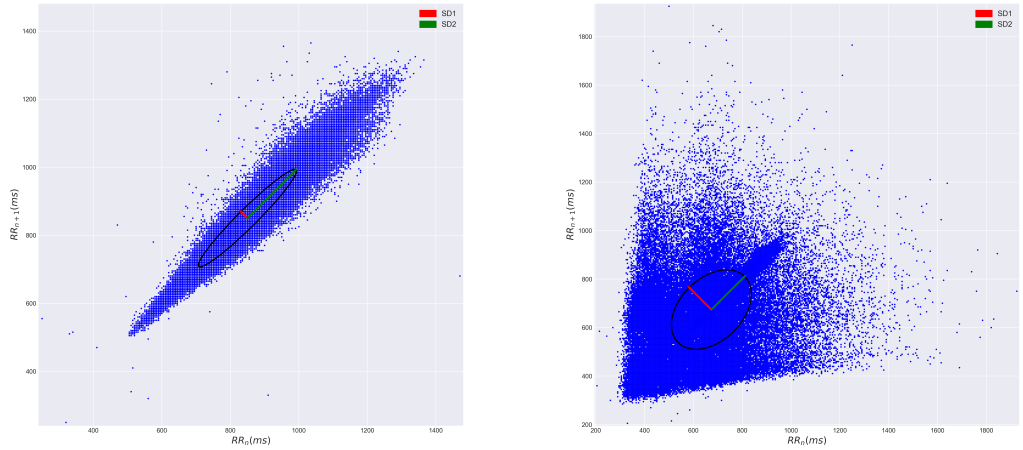
$$\begin{aligned} x_i &= \Delta RR_i = RR_{i+1} - RR_i \\ y_i &= \Delta RR_{i+1} = RR_{i+2} - RR_{i+1} \end{aligned} \tag{2.4}$$

The SODP is divided in four quadrants, as shown in Figure 2.9 and summarized in Table 2.4. A healthy heart has more centred points in an elliptical shape towards quadrants I and III. AF changes the shape of the SODP to point towards quadrants II and IV. The SODP allows the evaluation of the VHF, since it reflects the dynamics over 2 beats, giving the percentage of acceleration, i.e. $RR_{i+1} < RR_i$, and deceleration, i.e. $RR_{i+1} > RR_i$

The Central Tendency Measure (CTM) counts the ratio of points inside a circle of radius r centred on the origin (Thuraisingham 2009; Diao et al. 2022), such as in Equation (2.5). For a given threshold r, more points inside the CTM indicate a more stable heart rate.

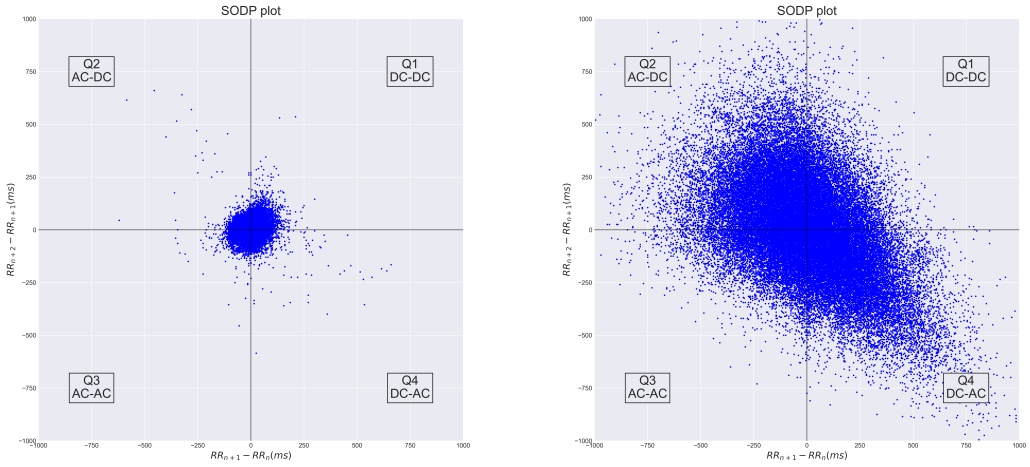
$$\begin{aligned} CTM(r) &= \frac{\sum_{i=1}^{n-2} \delta(RRi)}{n-2} \\ \delta(RRi) &= \begin{cases} 1, \text{ if } \sqrt{\Delta RR_{i+1}^2 + \Delta RR_i^2} < r \\ 0, \text{ otherwise} \end{cases} \end{aligned} \tag{2.5}$$

Other features from SODP are the number of points in each quadrant or



(a) NSR recording (b) AF recording

Figure 2.8: Poincaré plots: (a) NSR patients and (b) AF patients. A larger SD1 can be observed in (b) because AF has a large effect on short-term variability.



(a) NSR recording (b) AF recording

Figure 2.9: SODP plots: (a) patient in NSR and (b) patient in AF. The point cloud in (a) is more centred and in (b) we can observe more points in quadrants II and IV, which show more fluctuating accelerations and decelerations.

in subdivision of quadrants. SODP can be extended to 3 dimensions by using a third ΔRR_{i+2} and computing the CTM feature with spheres instead of circles (Altan et al. 2018; Diao et al. 2022). SODP can measure PNS activity over short durations of recordings, which differs from the AC and DC which measures PNS activity over longer periods of time. The comparison of RR interval differences over multiple data points provides a more complete view of the dynamics than Poincaré plots, which allow discrimination between apparently identical dynamics. Sarkar et al. (2008) proposed to implement SODP in an implantable loop recorder, for AF screening in patients with cryptogenic stroke. It could also provide regular information, e.g. daily, on rhythm and rate control in patients with AF over a long period of time.

Recurrence plot

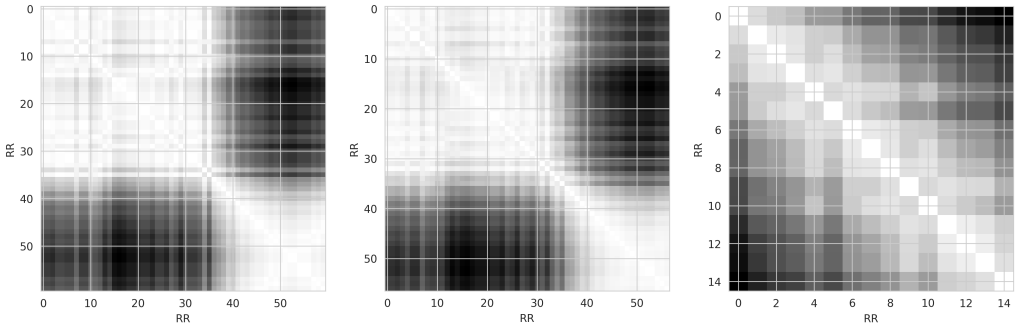
The last geometric method used in this work is the recurrence plot analysis. A recurrence plot is a symmetrical plot of the Euclidean distance between embedding vectors (Eckmann et al. 1987; Marwan et al. 2007). It is constructed using two parameters: m, the embedding dimension, and τ, the embedding lag. From the series of N RR intervals, the embeddings vectors are constructed as described in the matrix Equation (2.6), where each embedding vector corresponds to a row. Each RR_i interval in the matrix corresponds to an RR interval selected from the series of RR intervals using the index, m and τ. The visual effect of varying m and τ can be observed in Figure 2.10

$$X_{emb} = \begin{bmatrix} RR_1 & RR_{1+\tau} & \dots & RR_{1+(m-1)\times\tau} \\ RR_2 & RR_{2+\tau} & \dots & RR_{2+(m-1)\times\tau} \\ \vdots & \vdots & \ddots & \vdots \\ RR_M & RR_{M+\tau} & \dots & RR_N \end{bmatrix} \quad (2.6)$$
$$where\ M = N - (m-1) \times \tau$$

The recurrence plot is built from the embedding matrix, using the Euclidean distance between the embedding vectors, i.e. the rows of the matrix. The recurrence plot is created using Equation (2.7).

$$RP_{ij} = ||X_{emb_i} - X_{emb_j}|| \quad (2.7)$$

The result is a $M \times M$ matrix. If $m = 1$ and $\tau = 1$, the first two rows of the



(a) $m = 1$ $\tau = 1$ (b) $m = 2$ $\tau = 3$ (c) $m = 10$ $\tau = 5$

Figure 2.10: Recurrence plots for the same RR interval window with varying embedding dimension m and embedding lag τ

matrix correspond to the ΔRR used in the SODP plot. Finally, a threshold t can be defined to make the recurrence plot binary, as described in Equation 2.8.

$$RP_{ij} = \begin{cases} 1 \text{ if } ||X_{emb_i} - X_{emb_j}|| < t \\ 0 \text{ if } ||X_{emb_i} - X_{emb_j}|| \geq t \end{cases} \tag{2.8}$$

Recurrence plots were used to measure the variation of PNS and SNS in rats by Dabiré et al. (1998). They used an embedding dimension of $m = 10$ and an embedding delay of $\tau = 1$ to analyse blood pressure measurements. It was also used by Sun et al. (2008) to predict AF offsets using an embedding dimension of $m = 3$ and an embedding delay of $\tau = 70$ on the ECG signal. Ayatollahi et al. (2023) used recurrence plots to detect sleep apnea in ECG. Finally, in a preprint, Gavidia et al. (2023) propose to use recurrence plots to forecast the onset of atrial fibrillation. They propose an embedding dimension $m = 2$ and an embedding delay $\tau = 3$ on the RR intervals and an embedding dimension $m = 5$ and an embedding delay $\tau = 7$ on the ECG.

2.2.6 Heart rate fragmentation

Costa et al. (2018) introduces Heart Rate Fragmentation (HRF) as an additional HRV-based biomarker of cardiovascular risk. HRF has been shown to be a predictor of impaired ANS activity and AF in the healthy general population (Guichard et al. 2022). The four measurements focus on short-term HRV and are summarised in Table 2.6. Percentage of Inflection Points (PIP) represents the percentage of zero crossing points in the RR interval series. The selected points correspond to an RR interval where the preceding difference

Table 2.5: HRV geometric measurements

Sud-domain	Measurements	Unit	Description
Histrogram-based	TINN	ms	Triangular Interpolation of RR intervals
	HRVi	%	HRV triangular index
Poincaré-based	SD1	ms	Standard deviation of the point cloud perpendicular to the identity line
	SD2	ms	Standard deviation of the point cloud parallel to the identity line
	SD1/SD2	%	Ratio between SD1 and SD2
	S	ms^2	Surface of the point cloud
	CSI		Cardiac Sympathetic Index
	CSI_{mod}		Modified Cardiac Sympathetic Index
	CVI		Cardiac Vagal Index (PNS)
SODP-based	SODP Q_i	%	Percentage of point in i quadrant
	CTM_r	%	Percentage of points in a circle of radius r
	SODP origin	%	Percentage of points at origin (0,0)

Table 2.6: HRF measurements

Measurements	Unit	Description
PIP	%	Percentage of Inflection Points
IALS	#RR	Inverse of the Average Length of the acceleration/deceleration Segments
PSS	%	complement of the Percentage of Short Segments
PAS	%	Percentage of RR intervals in Alternation Segments

and the following difference are of opposite sign, i.e. an acceleration followed by a deceleration or a deceleration followed by an acceleration, as described in Equation (2.9).

$$\begin{aligned} &\Delta RR_i \times \Delta RR_{i+1} \leq 0 \\ &\text{where } \Delta RR_i = RR_i - RR_{i-1} \end{aligned} \tag{2.9}$$

PIP are therefore related to the number of points in quadrant II and quadrant IV of the SODP described in Section 2.2.5. The Inverse of the Average Length of the acceleration/deceleration Segments (IALS) counts the inverse of the average RR interval length of the segments between inflection points. The longer the segments, the lower the IALS. For the Percentage of Short Segments (PSS), a short segment is defined as an acceleration or deceleration segment of three or more intervals. Finally, the Percentage of Alternating Segments (PAS) defines alternating segments as $ADAD$ or $DADA$, where A represents an acceleration and D a deceleration. All measurements are defined such that the more fragmented the series, the higher the four measurements, which explains the use of inverse and complement.

2.3 ECG databases

Two categories of publicly available ECG recording databases are described in the literature: (i) short ECG, ranging from 10-second to 60-second recordings, and (ii) long-term recordings, ranging from 30-minute to 24-hour Holter monitoring.

Table 2.7: Comparison of selected short-term publicly available ECG databases, sorted by release date. The duration is indicated per recording.

Name	Year	Patients	Records	Duration	Sample Rate	Leads	Classes
AF challenge 2017 (Clifford et al. 2017)	2017	8528	8528	9 - 61 s	300 Hz	1	4
CU-SPH database (Zheng et al. 2020)	2020	10646	10646	10 s	500 Hz	12	4
PTB-XL (Wagner et al. 2020)	2020	18885	21837	10 s	500 Hz	12	5 (24)
SPH database (Liu et al. 2022)	2022	24666	25770	10 - 60 s	500 Hz	12	11

2.3.1 Short-term ECG databases with AF

There are several large publicly available databases. These databases are composed of records with a duration of approximately 10 seconds to 60 seconds. A selection of databases that have been recently published and used as training material in multiple publications is presented in Table 2.7, where the composition of each database is presented. Except for the `AF challenge 2017` database, these databases have 12-lead recordings and a high sampling frequency of 500 Hz. For a 10-second recording, the file contains 500×10 recorded points for each lead and therefore a 5000×12 matrix for the complete recording, which corresponds to 60 000 values.

The annotations are not similar between the four selected databases. The AF Challenge 2017 database contains 4 classes: NSR records, AF records, other rhythm and noisy records. For the CU-SPH database, 11 rhythm annotations from cardiologists were grouped into 4 groups: NSR, AF, sinus bradycardia and supraventricular tachycardia. PTB-XL annotations are grouped into 5 superclasses, e.g. conduction disturbance or hypertrophy, and 24 subclasses. Finally, in the SPH database, 44 types of cardiologist annotations are grouped into 11 categories. The differences between the annotations can be explained by the fact that each database was created independently to meet different requirements and for different research purposes. Nevertheless, correspondences between databases can be used to group some annotations for cross-database analysis.

2.3.2 Long-term ECG databases with AF

We have identified 4 publicly available long-term recordings with a focus on paroxysmal AF. Table 2.8 presents an overview of the databases. The four databases are available on the Physionet website (Goldberger et al. 2000).

The MIT-BIH Arrhythmia Database (Moody et al. 2001b) consists of 48

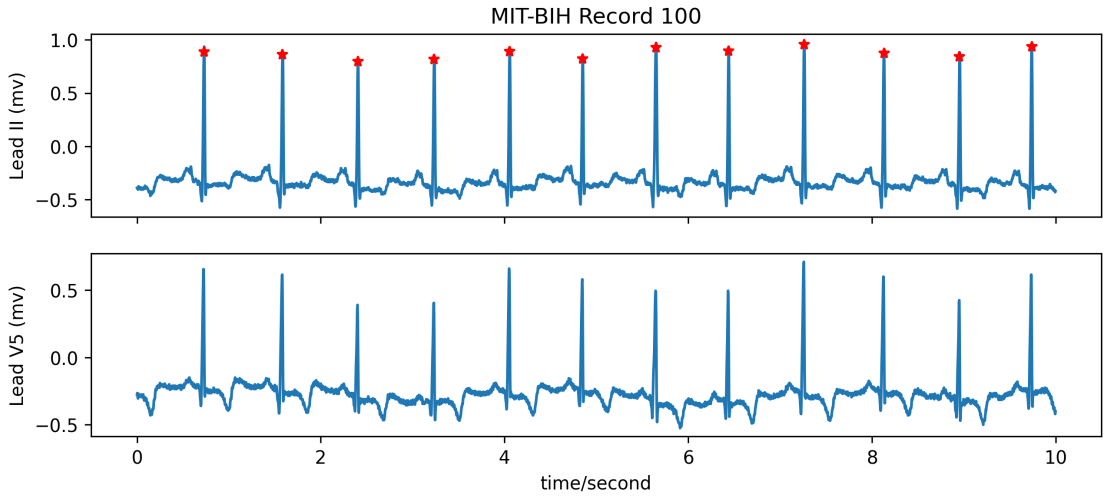


Figure 2.11: 10 seconds of 2-lead ECG from `recording 100` from the MIT-BIH database, with QRS complex annotated

recordings from 47 patients recorded between 1975 and 1979 at Boston Hospital in the United States. The database consists of a mixed population of inpatients and outpatients. Each recording lasts 30 minutes and is recorded at a sampling rate of 360 Hz. Two leads are available and QRS complexes are annotated as shown with stars in Figure 2.11. The first lead is lead II and the second lead can vary between V1, V2, V4 and V5 depending on the recording. The annotations were made by two cardiologists at beat level, i.e. on each QRS annotation.

The MIT-BIH Atrial Fibrillation Database (AFDB) consists of 23 two-channel recordings from patients with AF (Moody et al. 1983). The long-term recordings last 10 hours at a sampling rate of 250 Hz. It contains a total of 299 AF episodes, of which 11 episodes have more than 30 minutes of NSR before the onset of AF and an AF duration of at least 5 minutes or more. The annotations were made manually on the original analogue recordings and then converted into digital files.

The Long Term AF Database (LTAFDB) contains 84 24-hour recordings from 84 patients (Petrutiu et al. 2007). The sampling frequency is 128 Hz. We found 7314 AF episodes present in the database, but only 2 AF have more than 30 minutes of NSR before the AF onset and more than 5 minutes.

Finally, the China Physiological Signal Challenge 2021 Database (CPSC2021) contains 1436 2-lead recordings selected from 3-lead or 12-lead Holter monitoring (Wang et al. 2021). The duration of the recordings is variable, with a mean

Table 2.8: Comparison of publicly available ECG database, sorted by release date. The duration is indicated per record.

Name	Year	Patients	Records	Duration	Sample Rate	Leads
MIT-BIH Arr. DB (Moody et al. 2001b)	1980	47	48	30 min	360 Hz	2
MIT-BIH AFDB (Moody et al. 1983)	1983	23	23	10 h	250 Hz	2
LTAFDB (Petrutiu et al. 2007)	2007	84	84	24 h	128 Hz	2
CPSC2021 (Wang et al. 2021)	2021	105	1436	34 min	200 Hz	2

recording duration of 34 minutes. The shortest recordings last 8 seconds and the longest 6 hours. 229 recordings show evidence of paroxysmal AF, representing 493 AF crises, but only 6 episodes meet the criteria of 30 minutes of NSR before onset and an AF duration of at least 5 minutes.

2.3.3 AF onset forecast databases

Based on the selected database from the previous section, we identified databases that could be used to predict the onset of AF. This required recordings containing paroxysmal AF episodes preceded by a continuous period of sinus rhythm.

The primary database used in the literature is the Paroxysmal Atrial Fibrillation Prediction Database (AFPDB) release for the 2001 Physionet challenge (Moody et al. 2001a). The specific purpose of this challenge was to promote and stimulate research into the forecasting of atrial fibrillation. The database contains 200 recordings from 100 patients, including 53 patients with paroxysmal AF and 47 healthy patients. Each recording lasts 30 minutes. For the healthy patients, the database contains two NSR recordings per patient. For patients with AF, the database contains one NSR recording before the onset of AF and one NSR recording at least 2 hours before any signs of AF. The database also contains a 5-minute continuation recording for all recordings. It follows the end of the NSR recording to confirm the presence or absence of AF. The 53 AF recordings can be used to forecast the onset of AF as they are all 30 minutes long and have the 5-minute continuation recording to prove the presence of AF after the NSR recording.

Other long-term databases contain a larger number of AF crisis, but when we compare using the same selection criteria, we found that the number of AF crisis that can be selected is only a small proportion of the total number of crisis present, as summarised in Table 2.9. The results for the AFDB (Moody

Table 2.9: Comparison of publicly available ECG database for AF onset forecast, sorted by release date. The duration is indicated per recording. The AF* episodes selected have $>$ 30 minutes normal sinus rhythm before the AF onset and $>$ 5 minutes of AF duration after the onset.

Name	Year	Patients	Records	Duration	Leads	AF episodes	AF* episodes
AFPDB (Moody et al. 2001a)	2001	100	200	30 min	2	53	53
MIT-BIH AFDB (Moody et al. 1983)	1983	25	25	30 min	2	299	11
LTAFDB (Petrutiu et al. 2007)	2007	84	84	24 h	2	7358	2
CPSC2021 (Wang et al. 2021)	2021	105	1436	30 min	2	493	6

et al. 1983) is presented in Table 2.10. This database contains a total of 299 AF episodes, but 120 episodes have less than 1 minutes of sinus rhythm before the AF onset and 99 episodes have less than 5 minutes before AF onset. The LTAFDB contains a total of 7314 AF crisis but only 2 corresponds to the selection criteria, as shown in Table 2.11. Finally, the CPSC2021 database contains 493 AF crisis, but Table 2.12 shows that only 6 episodes are matching the selection criteria.

2.4 Machine learning for AF predictions

Multiple AF-related tasks have been investigated in the scientific literature. A graphical summary of the three tasks is presented in Figure 2.12. (i) The primary task involves AF screening, which corresponds to detecting AF presence within an ECG recording. This screening task applies to both short-term and long-term ECG recordings. (ii) The second task focuses on forecasting AF onset, seeking to predict incoming signs of AF episodes during NSR preceding the onset. This type of prediction necessitates the utilization of extended ECG recordings. The model analyses the ECG window preceding AF onset. (iii) Finally, the last task involves identifying the individual risk of developing AF. This task involves a comparative analysis between NSR data obtained from AF patients and that from healthy individuals to find and AF signature.

In this section, we review selected publications about the three tasks, with an in-depth focus on AF onset forecast and AF risk identification. The metrics used to evaluate the model are introduced in Section 2.5.

Table 2.10: Count of the number of AF episodes in the MIT-BIH Atrial Fibrillation database, based on the durations of sinus rhythm before the AF onset

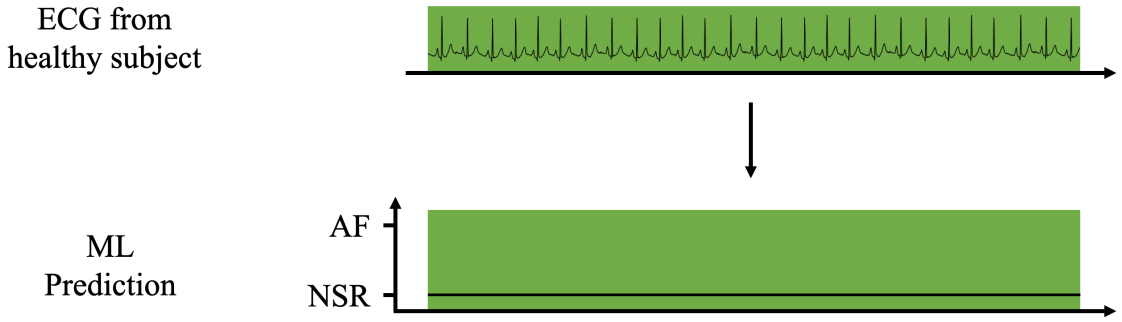
NSR duration before AF (min) \ AF duration (min)	> 0	> 5	> 10	Total
< 1	79	12	29	120
1 - 5	77	9	13	99
5 - 30	34	2	3	39
30 - 60	12	2	0	14
> 60	18	2	7	27
Total	220	27	52	299

Table 2.11: Count of the number of AF episodes in the Long Term Database (Petrutiu et al. 2007), based on the durations of sinus rhythm before the AF onset

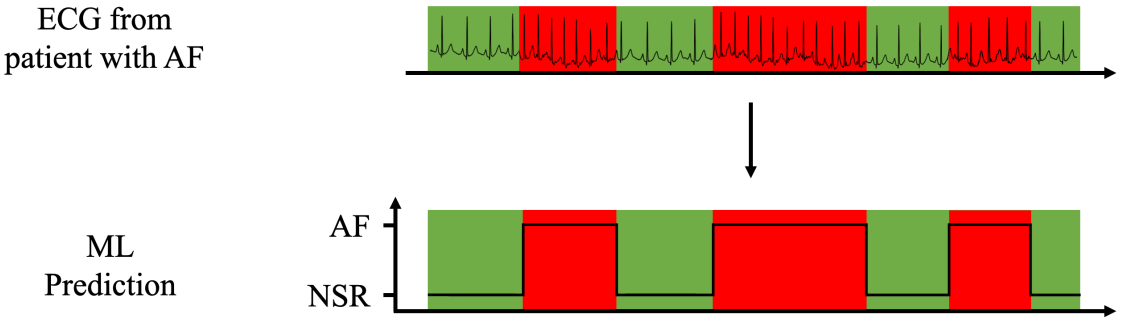
NSR duration before AF (min) \ AF duration (min)	> 0	> 5	> 10	Total
< 1	6217	193	296	6706
1 - 5	476	11	8	495
5 - 30	94	3	20	117
30 - 60	18	0	0	18
> 60	20	1	1	22
Total	6825	208	325	7358

Table 2.12: Count of the number of AF episodes in the CPSC2021 database (Wang et al. 2021), based on the durations of sinus rhythm before the AF onset

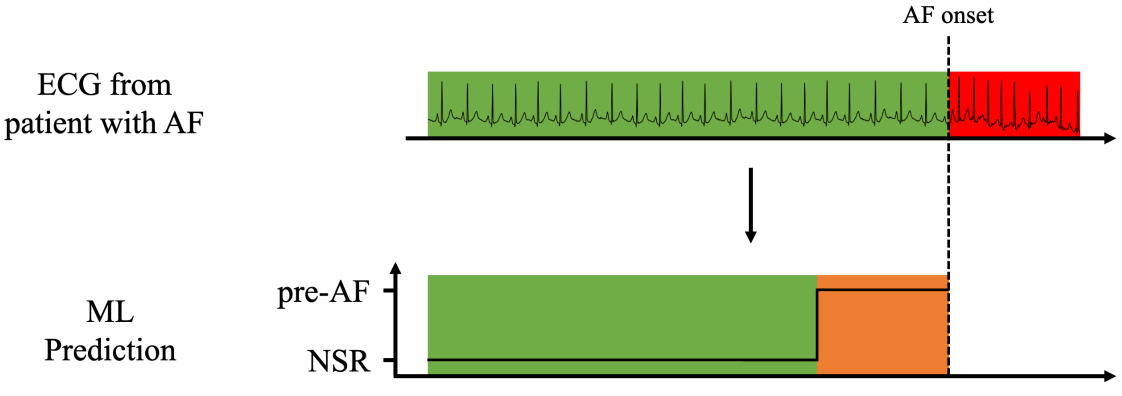
NSR duration before AF (min) \ AF duration (min)	> 0	> 5	> 10	Total
< 1	261	4	5	270
1 - 5	167	1	2	170
5 - 30	26	0	0	26
30 - 60	0	0	0	0
> 60	21	0	6	27
Total	475	5	13	493



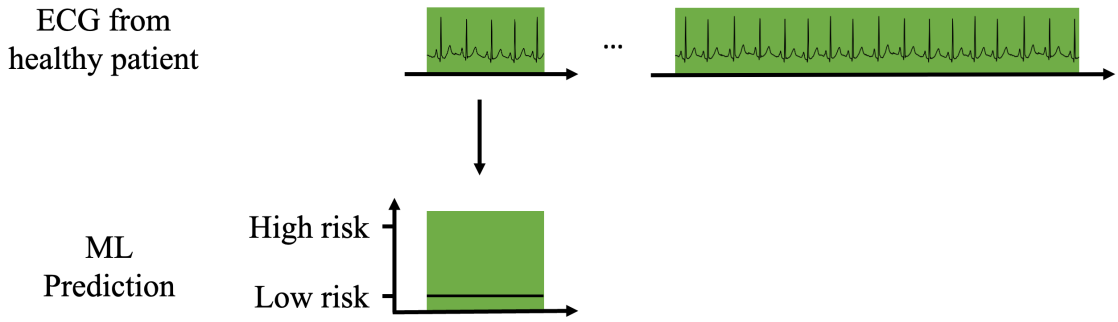
(a) AF detection for healthy patient



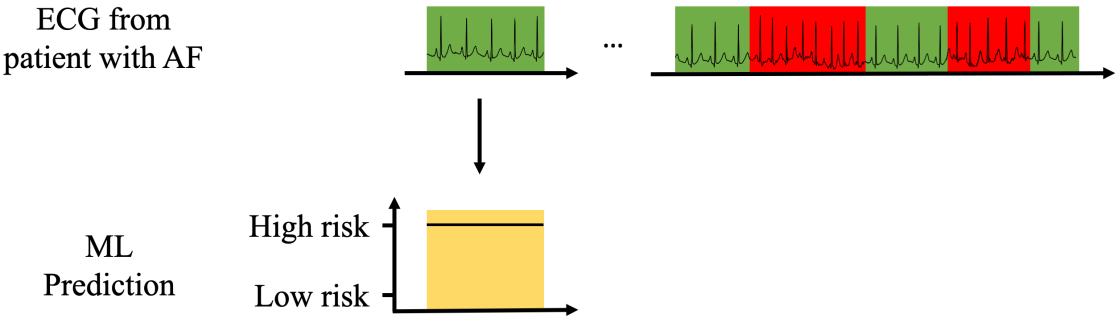
(b) AF detection for AF patient



(c) AF onset forecast



(d) AF risk identification for healthy patient



(e) AF risk identification for AF patient

(f) Legend

Figure 2.12: AF prediction tasks:
(a) and (b) AF detection — the model predicts if an ECG contains AF and where the AF is located,
(c) AF onset forecast — the model predicts incoming AF onset by detection signs in pre-AF window,
(d) and (e) AF risk identification — the models predicts the risk to develop AF, based on NSR recordings

2.4.1 Atrial fibrillation detection

AF detection models predict whether an ECG segment or RR interval window contains AF. Various ML and Deep Learning (DL) methods have been proposed in the literature. The results of these models have achieved a high level of performance, sometimes described as corresponding to cardiologist level as by Hannun et al. (2019). Methods in the literature include the use of Poincaré plots analysis with ML models (Bashar et al. 2021), RR analysis with CNN-RNN (Biton et al. 2023), RR analysis with RNN (Faust et al. 2018), ECG analysis with RNN (Singh et al. 2018) and ECG analysis with CNN (Erdenebayar et al. 2019; Ribeiro et al. 2020).

In particular the recent results from Hannun et al. (2019) can be detailled. They propose a Deep Neural Network (DNN), and in particular a 1D ResNet based CNN (He et al. 2015) composed of 16 blocks, achieving performance of 0.973 (95% Confidence Interval (CI) [0.966–0.980]) Area Under the Receiver Operating Characteristic Curve (AUROC) for the AF class on 1.5 seconds ECG segments. For the training of this model, they used a private database composed of 91 000 ECG records from 53 000 patients, with 12 annotated heart diseases. Their model achieved a comparable or superior Area Under the Curve (AUC) for all the 12 diseases compared to a committee of cardiologists.

Faust et al. (2018) achieved an accuracy of 98.51% on the MIT-BIH Atrial Fibrillation Database (Moody et al. 2001b) using a bidirectional LSTM model (Hochreiter et al. 1997) with RR intervals as input. Equivalent accuracy was achieved by Xia et al. (2018) using 5-second ECG segments. They extracted a 2D spectrogram of the ECG signal using FFT as input to a CNN model. They used the MIT-BIH atrial fibrillation database and the model achieved an accuracy of 98.29%. More recently, Hu et al. (2022) proposed a transformer-based model and achieved an accuracy of 99.23%.

Comparing models on the same medical task has become more challenging, given that ML models are trained and tested on different datasets. This variation in training and testing data introduces complexity in assessing their relative performance and generalisability. However, despite this challenge, a prevailing global trend indicates consistently high performance and predictive value for AF detection across different research results. This highlights the robustness of these models in a clinical context, with opportunities for the development of novel clinical tools.

Table 2.13: PAF challenge 2001 entries using the AFPDB database

Author & Data	Methodology	
	Model	Input
Zong et al. (2001)	Stat. analysis	PAC analysis
Maier et al. (2001)	Stat. analysis	HRV
Chazal et al. (2001)	Stat. analysis	P wave
Yang et al. (2001)	Stat. analysis	HRV
Schreier et al. (2001)	Stat. analysis	P wave
Lynn et al. (2001)	KNN	HRV

2.4.2 Atrial fibrillation onset forecast

In 2001, Physionet launched the PAF Prediction Challenge (Moody et al. 2001a) with the aim of understanding whether early signs of AF onset can be detected in the 30-minute window before the onset of an AF crisis. The idea is to compare 30-minute windows close to AF onset and 30-minute windows distant from AF onset of at least 45 minutes. The database of the challenge is the AFPDB.

For the competition, the proposed methods were mainly based on statistical analysis of the HRV, PAC and P wave. One research team proposed the use of K-Nearest Neighbours (KNN) as a predictive model. The methods proposed in 2001 are summarised in Table 2.13.

The dataset proposed in the challenge has continued to be used in many publications since then, as shown in Table 2.14. The few publications using alternative databases did not make the ECG recordings publicly available. The results presented in the publications using this database are rather optimistic, with accuracy values reaching 90% (Mohebbi et al. 2012; Narin et al. 2018; Boon et al. 2018). However, these results should be treated with caution, as we have shown that some could not be reproduced (Gilon et al. 2022). Therefore, in Table 2.14 we focus on the different methodologies rather than the results and metrics. We can classify the inputs used into four classes: ECG, ECG Morphology Variability (ECGMV), RR and HRV. Most of the recent methods use ML approaches with different models and DNN models.

Table 2.14: Paroxysmal AF onset forecast selected publications

Author & Date	Dataset	Methodology	
		Model	Input
Hickey et al. (2002)	AFPDB	ML (LR)	HRV
Mota et al. (2003)	AFPDB	ML (GA & KNN)	ECGMV
Ros et al. (2004)	AFPDB	ML (KNN)	ECGMV
Thong et al. (2004)	AFPDB	Stat. analysis	PAC
Kikillus et al. (2007)	MIT-BIH	Stat. analysis	HRV
Chesnokov (2008)	AFPDB	ML (SVM) & DL (DNN)	HRV
Pourbabaee et al. (2008)	AFPDB	ML (KNN) & DL (DNN)	ECGMV
Panusittikorn et al. (2010)	AFPDB	ML (KNN)	HRV
Mohebbi et al. (2012)	AFPDB	ML (SVM)	HRV
Costin et al. (2013)	AFPDB	Stat. analysis	HRV & ECGMV
Anwar et al. (2013)	AFPDB	DL (DNN)	HRV & ECGMV
Alcaraz et al. (2015)	private	Stat. analysis	P wave
Martinez et al. (2015)	private	Stat. & ML (DT)	P wave
Boon et al. (2016)	AFPDB	ML (SVM)	HRV
De Giovanni et al. (2017)	AFPDB	ML (SVM)	ECGMV
Ozcan et al. (2017)	AFPDB	ML (GA & KNN)	HRV
Pourbabaee et al. (2018)	AFPDB	CNN	ECG
Bianchi et al. (2018)	AFPDB	ML (KNN & SVM)	ECGMV
Boon et al. (2018)	AFPDB	ML (SVM)	HRV
Ebrahimzadeh et al. (2018)	AFPDB	ML (KNN & SVM) & DL (DNN)	HRV
Narin et al. (2018)	AFPDB	ML (KNN)	HRV
Lee et al. (2018)	AFPDB	ML (RF)	HRV & ECGMV
Cho et al. (2018)	private	DL (CNN)	ECG spectrogram
Aligholipour et al. (2018)	AFPDB	ML (clustering)	HRV
Jalali et al. (2020)	AFPDB	DL (CNN)	ECG spectrogram
Gilon et al. (2020)	IRIDIA-AF v1	DL (CNN-RNN)	RR
Castro et al. (2021)	AFPDB	ML (RF & KNN & SVM)	HRV
Tzou et al. (2021)	private	DL (CNN)	P wave spectrogram
Guo et al. (2021)	private	ML (XGB)	HRV from PPG
Parsi et al. (2021)	AFPDB	ML (SVM)	HRV
Bashar et al. (2021)	AFPDB	ML (SVM & RF)	HRV
Surucu et al. (2021)	AFPDB	DL (CNN)	HRV
Hammer et al. (2022)	MITBIH AF	ML (LR & RF)	HRV & ECGM
Mendez et al. (2022)	AFPDB	DL (CNN)	Poincaré plot
Gavidia et al. (2023)	private	DL (CNN)	Recurrence plot
Rooney et al. (2023)	LTAFDB	DL (CNN+Attention)	ECG

2.4.3 Atrial fibrillation identification

Since 2019, several studies have presented models that are able to identify patients at risk of AF using NSR alone. This opens up the use of ML as an alternative to the risk score in clinical practice. The studies suggest using large databases available in hospital centres with DL models. Unfortunately, most of these databases are still private.

Attia et al. (2019a) proposed a CNN model able to identify patients at risk up to 30-days before the first sign of AF with and AUC of 0.87. The database they used is composed of 180 000 patients and 649 000 10-second records. They used all the NSR records from healthy patients and NSR records 30-days prior to the first AF sign for AF patients. The ECGs are 10-second 12-lead 500 Hz record. The final model is composed of 9 residual blocks, in which each block is composed of 2 sub-blocks. Other studies proposed similar results in 2021 and 2022. Suzuki et al. (2022) and Kim et al. (2022) also used CNN models. Baek et al. (2021) used bi-LSTM mode. Selected state-of-the-art studies are listed in Table 2.15.

All models presented in Table 2.15 are trained using a private database. Data sharing is critical to the reproducibility of research, especially in machine learning, as it allows replication of results and methods across studies, ensuring transparency and scientific rigour in research (Miyakawa 2020). Access to private databases can occasionally be granted on request to the original author, but such limited access can disrupt the continuous flow of research. In addition, Gabelica et al. (2022) analysed the data availability statements made in the literature. The authors found that of the publications that mentioned that the database could be accessed on reasonable request, only 6% of the original authors responded to the data sharing request.

These models can be proposed as an extension or alternative to the clinical scores currently used in medical practice to identify the risk of AF and stroke in different patient groups. The $Chad_2DS_2Vasc$ measures the risk of ischaemic stroke in patients with AF over the next 7-10 years Lip et al. (2010) and Olesen et al. (2011). CHARGE-AF measures the risk of AF over the next 5 years (Alonso et al. 2013). $HARMS_2$-AF has recently been proposed as a superior alternative to CHARGE-AF, with an AUROC of 0.75 for AF detection over the next 5 to 10 years (Segan et al. 2023). Finally, Mr-DASH (Mitrega et al. 2021) and C2HEST (Li et al. 2019) are two specialised identification scores, the first for elderly patients and the second for Asian patients. Noseworthy et al. (2022)

Table 2.15: State-of-the-art studies for the identification of AF patients during sinus rhythm. Access to private* (star) database can be requested from the author.

Author (Date)	Database	Model	Data	Prediction horizon horizon	AUROC
Attia et al. (2019a)	private	CNN	ECG (10 s - 12 leads - 500 Hz)	1 month	0.88
Raghunath et al. (2021)	private*	CNN	ECG (10 s - 12 leads - 500 Hz)	1 year	0.85
Baek et al. (2021)	private	LSTM	ECG (10 s - 12 leads - 500 Hz)	3 months	0.79
Biton et al. (2021)	private*	RF	ECG features & ECG (10 s - 12 leads - 400 Hz)	5 years	0.90
Singh et al. (2022)	private	RF	PAC, age, and sex	2 weeks	0.75
Singh et al. (2022)	private	CNN-LSTM	HR	2 weeks	0.74
Singh et al. (2022)	private	CNN-Attention	ECG (80 beats from 24 hours - 1 lead - 250 Hz)	2 weeks	0.75
Singh et al. (2022)	private	Ensemble	ECG (24 hours - 1 lead - 250 Hz)	2 weeks	0.79
Suzuki et al. (2022)	private*	CNN	ECG (10 s - 12 leads - 500 Hz)	1 month	0.83
Khurshid et al. (2022)	private & UK BioBank	CNN	ECG (10 s - 12 leads - 500 Hz)	5 year	0.82
Myrovali et al. (2023)	private*	P-wave features	RF	7 days	0.96
Hygrell et al. (2023)	private*	CNN	ECG (30 s - 1 lead - 500 Hz)	2 weeks	0.80
Gruwez et al. (2023a)	private	CNN	ECG (10 s - 12-lead - 500 Hz)	1 month	0.87
Yuan et al. (2023)	private	DL	ECG (12-lead)	1 month	0.86
Biton et al. (2023)	private*	CNN-RNN	60 RR	4-years	0.99
Gadaleta et al. (2023)	private*	RNN	10 minutes	2 weeks	0.80
Dupulthys et al. (2023)	private	CNN-Transformer	10 seconds	91 days to 365 days	0.76

have shown that an ML approach outperforms a risk score in a prospective study following patients for 30 days. The model increased AF detection by 10%.

2.5 Performance evaluation

This section discusses the evaluation of the performance of ML models for AF prediction. In the case of perfect prediction model, all healthy recordings are classified as healthy and all AF recordings are classified as AF. In practice, models make misclassifications. In the AF literature, ML model predictions are evaluated using metrics such as AUROC, Area Under the Precision-Recall Curve (AUPRC), accuracy, sensitivity, specificity and F_1 score.

AF tasks are commonly defined as a binary classification problem between two classes: healthy and sick patients. For a given input, such as an ECG, the model predicts the probability of being sick, i.e. having AF. The probability of having AF is a real value between 0 and 1. Given a threshold value, e.g. 0.5, the predictions can be classified into the 4 categories of the confusion matrix, as shown in Figure 2.13. If the predicted probability value is lower than the threshold value, the prediction is classified as healthy, and if the predicted probability value is higher than the threshold value, the prediction is classified as AF. True Positive (TP) represents recordings that were correctly predicted as disease, False Negative (FN) represents recordings that were incorrectly predicted as healthy. True Negative (TN) represents recordings correctly predicted as healthy and TP represents recordings incorrectly predicted as AF.

The accuracy represents the total number of correct predictions for both healthy and AF, by counting the total of TP and TN over the total number of recordings. This corresponds to Equation (2.10). In the case of an unbalanced classification problem, the accuracy score can be misleading: if 99% of the recordings are healthy and the remaining 1% are AF the classifier can always predict recordings as healthy and achieve a 99% accuracy score.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.10}$$

Row-based metrics

The sensitivity, also called recall or True Positive Rate (TPR), is the number of correct predictions for the positive class, i.e. AF patients in this work. It is

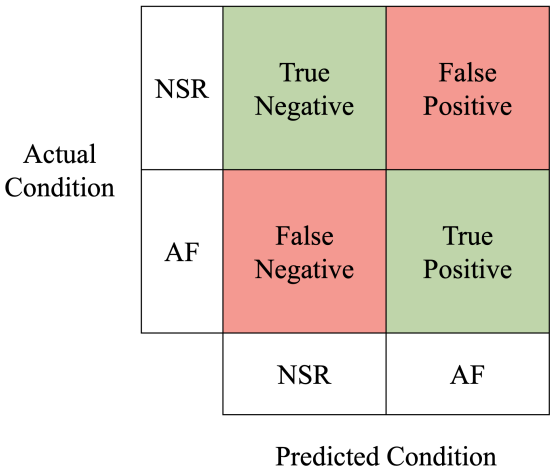


Figure 2.13: Confusion matrix for the binary classification task between NSR and AF. The first row represents the recordings from the healthy patients and the second row represents the recordings from the AF patients. The first column represents the recordings predicted as healthy by the model and the second column represents the recordings predicted as AF by the model.

described in Equation (2.11).

$$Sensitivity = Recall = TPR = \frac{TP}{TP + FN} \tag{2.11}$$

The specificity, also called True Negative Rate (TNR), is the number of correct predictions for the negative class, i.e. healthy patients in the work. It is described in Equation (2.11).

$$Specificity = TNR = \frac{TN}{FP + TN} \tag{2.12}$$

The False Positive Rate (FPR) and the False Negative Rate (FNR) are the complement of the sensitivity and specificity. It counts the number of misclassified recordings in the positive and negative classes.

$$FPR = 1 - TNR = \frac{FP}{FP + TN} \tag{2.13}$$

$$FNR = 1 - TPR = \frac{FN}{FN + TP} \tag{2.14}$$

Column-based metrics

Column-based metrics are based on the predicted class. The precision, also called Positive Predictive Value (PPV), count the number of correct predictions for the predicted positive categories.

$$Precision = PPV = \frac{TP}{FP + TP} \tag{2.15}$$

The Negative Predictive Value (NPV) counts the number of correct predictions for the predicted negative categories.

$$NPV = \frac{TN}{FN + TN} \tag{2.16}$$

Aggregate metrics

The F_1-score is proposed has an aggregation between the positive class row metric and the positive class column metric. It is defined has the harmonic mean of the precision and the recall, as in Equation (2.17).

$$F_1 - score = 2 \times \frac{precision \times recall}{precision + recall} \tag{2.17}$$

Threshold-free metrics

The metrics presented above depend on the choice of threshold t. The most common choice for the threshold is 0.5, as this threshold divides the probability range into two equal parts, i.e. from 0 to 0.5 for the negative class and from 0.5 to 1 for the positive class. Varying the threshold will change the value of the metric.

Threshold-free metrics allow the prediction of models to be evaluated without having to rely on the choice of a threshold. The Receiver Operating Characteristic (ROC) curve (Fawcett 2004; Fawcett 2006) is constructed by plotting the TPR against the FPR for all selected thresholds t between 0 and 1. A random classifier is represented as a straight line between (0,0) and (1,1). The better the classifier, the higher the curves, as shown in Figure 2.14a. The performance of the model is evaluated by the AUC under the ROC curve, i.e. the AUROC, as in Figure 2.14b.

The Precision-Recall (PR) curve is another curve constructed by plotting precision against recall for all selected thresholds t between 0 and 1 (Boyd et al.

2013). The TPR used as y for the PR curve and the recall used as x for the PR curve are the same metric. As for the ROC evaluation, models that produce better predictions have a higher PR curve, as shown in Figure 2.15a. The PR is evaluated using the AUC, as shown in Figure 2.15b. Finally, Saito et al. (2015) showed that PR curves are more informative than ROC curves when using imbalanced databases.

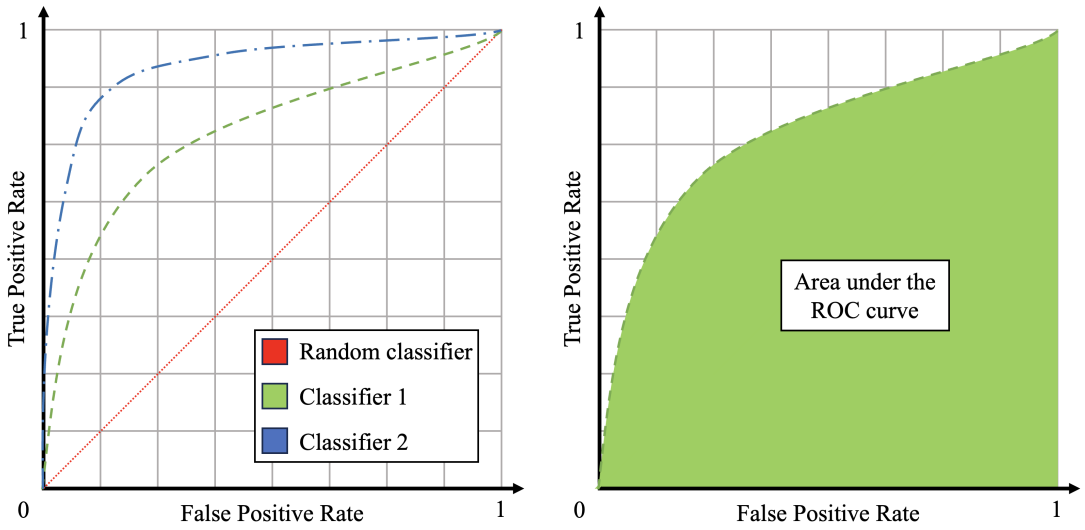
Multi-class metrics

In the case of a multi-class classification with C classes, the confusion matrix can be extended to a $C \times C$ matrix, as in Figure 2.16. The accuracy of the predictions is the sum of the diagonal of the matrix divided by the number of predictions. The metrics presented above in the case of binary classification can be used in this case of multi-class by adopting a one-against-all approach (Tharwat 2020). For each class, the metrics are calculated using the chosen class C_i as the positive class and all other classes as the negative class.

2.6 Summary

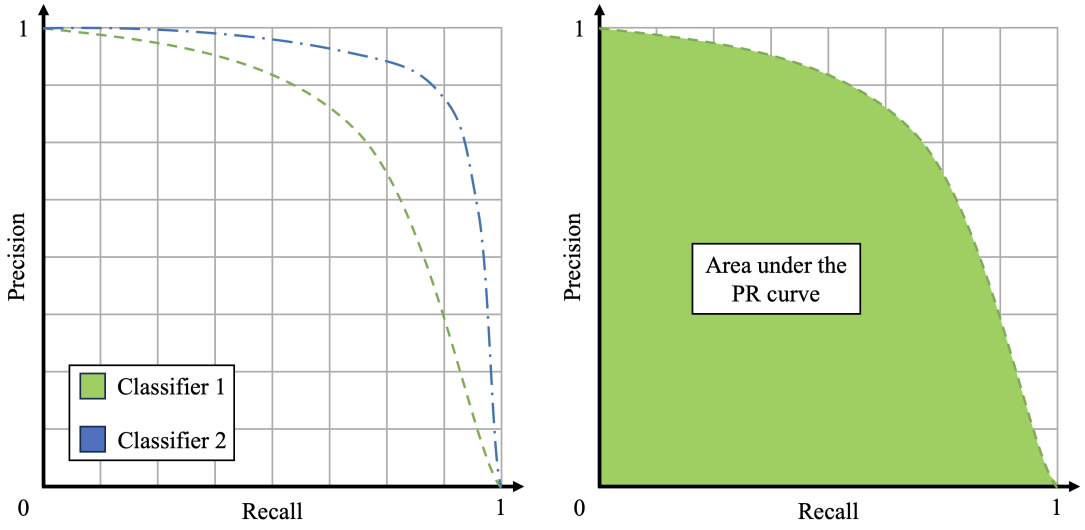
In this chapter we first introduce the electrocardiogram and heart rate measurements. We studied the analysis of HRV through three main domains: the time domain, the frequency domain and the geometric domain. All HRV measurements essentially reflect the modulation of the ANS and its two subsystems, the SNS and the PNS. Long-term measurements such as 24H-SDNN, 24H total spectral power, DC and HRT provide information on the individual ANS state. Spectral measurements taken over short periods, e.g. 5 minutes, provide information on the dynamics of systems that disturb this ANS balance and may be part of the trigger for arrhythmias and premature beats. Geometric methods such as Poincaré plots allow rapid screening for AF and are emerging as an important part of e-cardiology systems. Although mathematical and computational techniques allow manipulation of the ECG signal to extract information and use it in predictive models for individual cardiac risk stratification, its explicability remains difficult. Conclusions about ANS activity from these models must remain cautious.

In this chapter we also reviewed the existing and publicly available ECG database. The database can be divided into two groups: short-term recordings and long-term recordings. In the case of forecasting the AF onset, databases



(a) ROC curve (b) AUROC

Figure 2.14: Receiver Operating Characteristic curve: (a) ROC curves from a random classifier in red and two classifiers (green and blue), and (b) the area under the ROC curve for the green classifier. In (a), the blue classifier (dash and dot) is the best classifier of the three, as is the curve above the green (dash) and red (dot) curves.



(a) PR curve (b) AUPRC

Figure 2.15: Precision-Recall curve: (a) PR curves of two classifiers (green and blue), and (b) the area under the PR curve for the green classifier. In (a), the blue classifier (line and dot) is the best classifier of the two, as the curve is above the green one (line).

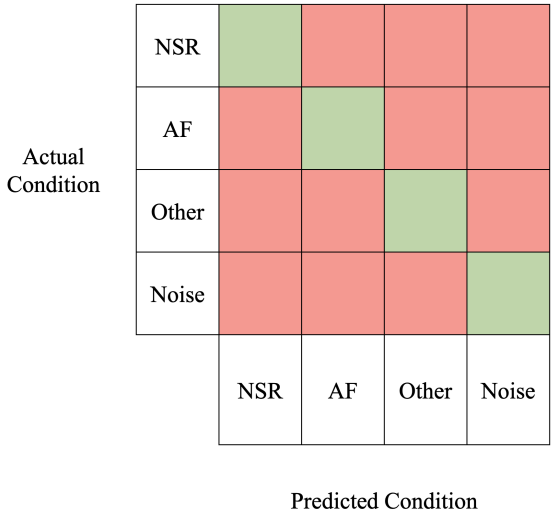


Figure 2.16: Multi-class confusion matrix for ECG classification. In this case, the prediction can belong to 1 of 4 classes: healthy recordings, recordings with AF, recordings with other diseases and noisy recordings. Correct predictions are on the diagonal of the matrix.

containing short-term recordings are of no use, as we want to analyse the period of at least 30 minutes prior to the onset of fibrillation. In the case of long-term databases, we have seen that several databases exist and are publicly available. In the case of these databases, the number of AF episodes preceded by at least 30 minutes of NSR and with AF episode duration of at least 5 minutes is limited.

This review highlights several research needs. The first need is the creation of a new large-scale database of long-term Holter monitoring ECG from the onset of paroxysmal AF. For the forecasting of AF onset, we need to determine the transferability of existing algorithms to a new database. To do this, we first need to assess the reproducibility of previous research. Then we will build on the results and extend the existing models from the three types of prediction to AF onset prediction and AF risk identification.

Chapter 3

Paroxysmal atrial fibrillation Holter monitoring database

3.1 Introduction

In the previous chapter, we have reviewed existing and publicly available database for Atrial Fibrillation (AF) onset forecast and AF identification during Normal Sinus Rhythm (NSR). We have showed that there is no large scale database for AF onset forecast, and the quantity of available recordings is limited. The largest database is the Paroxysmal Atrial Fibrillation Prediction Database (AFPDB) (Moody et al. 2001a) available on Physionet (Goldberger et al. 2000), which contains 53 AF onsets recordings.

For this research, we created and labelled a new Electrocardiogram (ECG) database composed of long-term Holter recordings. Holter recordings are made daily in hospitals, and most of these recordings are stored locally. It is estimated that 300 million ECGs are recorded worldwide every day (Zhu et al. 2020). However, access to these recordings requires compatibility or permissions within the proprietary format. This highlights the need and importance for collaboration between public research and industry, which can only benefit patients.

In the database, AF crisis were manually labelled to detect AF onsets and AF offsets for each AF crisis. Therefore, it could be later used for supervised training of Machine Learning (ML) models. In recent publications related to AF task using ML and Deep Learning (DL) models, database are sometimes kept private by the research teams. An additional objective of this database is to expand the options available to researchers by publishing the database.

We described the collection of data at the centres, the annotation process to create the labels and the final format of the database. The whole process is time-consuming, but the quality of the data and the quality of the annotations have a major impact on the performance of the model.

3.2 IRIDIA-AF database version 1

The first version of the IRIDIA-AF is the result of a mono-centric retrospective study on Holter monitorings from Dr Grégoire outpatient cardiology clinic. The total duration of all recordings in the database represents more than 24 million seconds of recordings in total, which represents 278 days or 6690 hours of Holter monitorings. In total, 388 AF episodes were recorded and annotated, with a total duration of 5 million seconds, which represent 67 days or 1609 hours. It corresponds to 24% of the total duration of the dataset.

3.2.1 Recordings selection and annotation

The ECG signal data was recorded using Microport Spiderview Holter recorders. The data acquisition phases started in January 2006 and ended in August 2017. The recording frequency of the device is 200 Hz, with a precision of 10 µV. Two leads were recorded: lead `I` and lead `II`. The medical analysis and annotations of AF onset and offsets were done using Microport Syneview (version 3.30a). The software was used to view the data, evaluate the quality of the recording and to select the precise time of events in the recording.

This study was approved by the institutional ethics committee Erasme-ULB P2017/413. The request for exemption from consent has been granted by the committee, due to the unrealistic feasibility of obtaining consent given the large number of involved cases and the high probability of being unable to reach numerous patients, and the publication of the anonymous data was allowed.

A total of 167 recordings from 152 patients were selected from the 9568 recordings. The recordings were selected as follows.

1. The database of Holter recordings was reviewed and searched by an experienced cardiac nurse. Holter recordings from patients with a Cardiac Implantable Electronic Device (CIED) were rejected. Holters with persistent or permanent atrial fibrillation or other heart disease were discarded. Holters with low recording quality or excessive noise were rejected. The

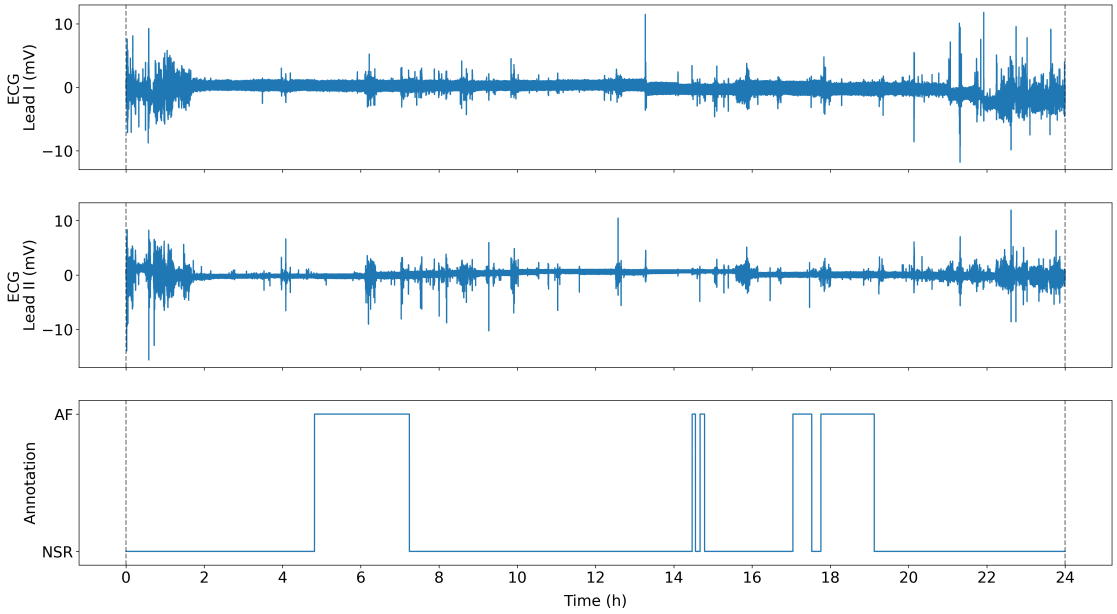


Figure 3.1: ECG and corresponding annotations in `recording_104`

selection of recordings was based on the analyses of the cardiologist and nurse, based on years of experience in Holter reading and interpretation for AF screening.

2. Holters with paroxysmal atrial fibrillation were selected. The selected recordings were reviewed by an experienced cardiologist and cardiac specialist nurse to validate the diagnosis.

3. All the recordings passing the previous validations were annotated. The annotation consists of searching and determining the precise beginning and end of each AF crisis in each recording, as presented in Figure 3.1. The start of the AF crisis corresponds to the first beat in AF, as shown in Figure 3.2. The annotation is positioned on the QRS complex of this first AF beat. The end of the AF crisis corresponds to the first beat in normal sinus rhythm (NSR) after the crisis. The annotation is also positioned on the QRS complex of this first NSR beat. In case of doubt about one event, a second opinion was asked to validate the annotation.

4. The recording was then exported, using the Microport Syneview software, from Microport proprietary format to ISHNE format (Badilini et al. 1998). Each recording was stored along the annotations and RR intervals. The RR intervals files were exported from the automatic QRS annotation by

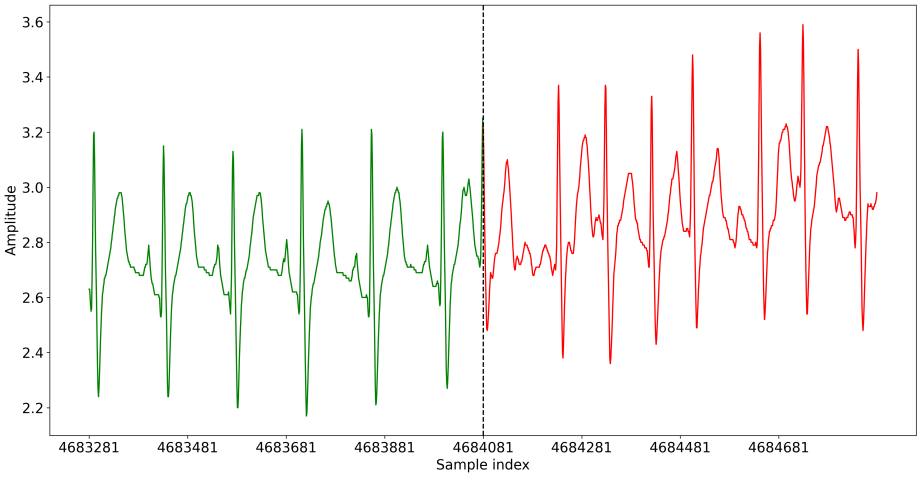


Figure 3.2: AF onset for the first AF crisis in `recording_026`

Microport Syneview software.

5. The labels were checked by a technical expert to validate the correspondence with the waveform data available in the exported ECG signal file. Because the recording frequency is 200 Hz and the annotations were accurate down to the second, the sample index that corresponds to the annotated time may not precisely align with the selected QRS complex. If a difference was found, the label was manually corrected and realign the sample index to correspond precisely to the QRS complex index chosen by the annotators. An example of annotation, label and correction is presented in Figure 3.3.

6. For some recordings, the Holter monitor does not seem to have been stopped just after the electrodes were removed from the patient skin. The end of each recording was visually inspected to determine if *end of record* noise is present. An example of such *end of recording* noise is presented in Figure 3.4, where most of the recording is noise. If *end of recording* noise is present, the recording was trimmed to only contain the interesting data. The RR files were automatically reworked to correspond to the new length of the file.

7. The waveform files and RR files were exported from ISHNE format to HDF5 format.The metadata files were double-checked with annotations to validate the conversion.

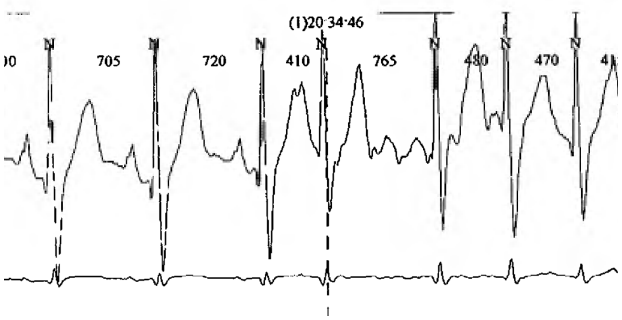
The quality of the waveform file was left as it was recorded by the Microport Holter devices, in order to match real life recordings. Recordings with high levels of noise were discarded by the cardiologist during the selection phase. The sampling frequency of the recordings was left unchanged at 200 Hz. The unique patient identifier and the unique recording identifier were randomly generated. Some patients have multiple Holter recordings, so the recordings are assigned to the same patient identifier. Each recording acquisition date was shifted by a random offset for each patient, as suggested by previous ECG databases (Wagner et al. 2020; Liu et al. 2022), to unlink recordings from the original database in the de-identification process. If there were multiple recordings for a patient, the chronological order of the recordings was maintained. We converted each patient's birthday to their age at the time of recording. Even with these pseudonymisation measures, re-identification remains a risk, as the ECG itself could be used to identify the patient, as suggested by (Guillaudeux et al. 2023).

3.2.2 Results

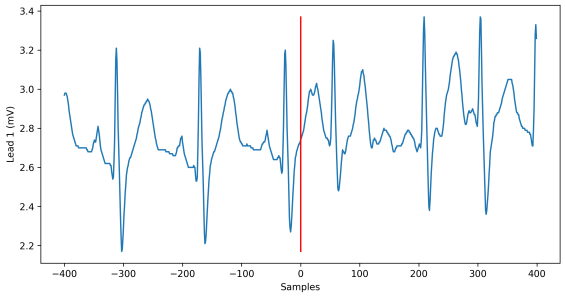
The first version of the database is based on Holter recordings database from Dr Jean-Marie Grégoire outpatient clinic from 2006 to 2017. From the 9568 available recordings, a total of 167 recordings from 152 patients were selected.

3.2.3 Comparison with existing database

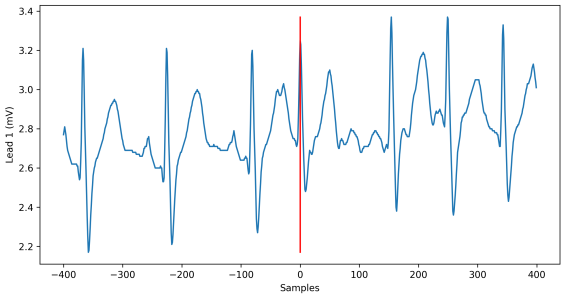
The comparison of the IRIDIA-AF database and selected publicly available database with AF diagnosis is presented in Table 3.1 and Table 3.2. Other databases, such as the PTB-XL (Wagner et al. 2020) or the AF classification challenge 2017 database (Clifford et al. 2017) propose a larger number of patients and number of heart disease diagnosis. IRIDIA-AF database proposes a longer total cumulative recording duration when compared to other publicly available databases with AF. In addition, thanks to the length of the recordings, this database can also be used for other AF related tasks, such as short-term AF onset forecast (Gilon et al. 2020). Other databases, as PTB-XL, cannot be used for AF onset short-term forecast as the recordings are 10-second long and does not include the minutes before AF onsets.



(a) Cardiologist annotation



(b) Converted annotation



(c) Manually corrected annotation

Figure 3.3: Example of annotation correction in `recording_026`. The first annotation (a) is made by the cardiologist. The converted annotation (b) corresponds to the conversion of the annotation time, i.e. time (h:m:s) to the sample index in the file. The sample index in the recording may differ from the annotation made by the cardiologist due to the conversion, as the accuracy of the annotation is limited to 1 second and the recording frequency is 200 Hz. A manual correction (c) is therefore required to realign the annotation of the selected QRS complex in the time series.

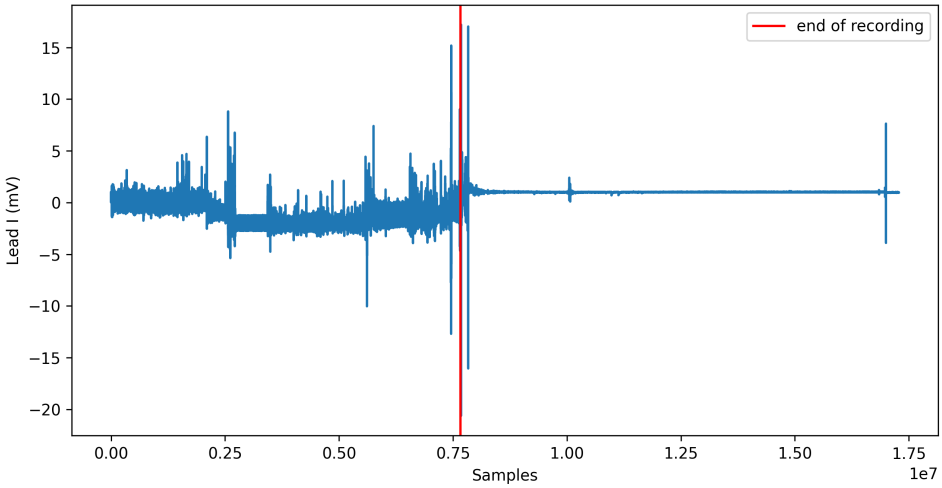


Figure 3.4: ECG recording `record_142` with noisy end after electrode removal

Table 3.1: Comparison of selected publicly available ECG arrhythmia database and IRIDIA-AF

Name		# patients	# recordings	# leads	recording duration (seconds) min	recording duration (seconds) max	# classes
MIT-BIH Arrhythmia	(Moody et al. 2001b)	47	48	2	1800	1800	2
AF classification challenge 2017	(Clifford et al. 2017)	8528	8528	1	9	61	4
SPH dataset	(Liu et al. 2022)	24666	25770	12	10	60	59
CU-SPH dataset	(Zheng et al. 2020)	10646	10646	12	10	10	11
PTB-XL	(Wagner et al. 2020)	18885	21837	12	10	10	71
IRIDIA-AF version 1	(Gilon et al. 2023b)	152	167	2	71 408	345 596	2

Table 3.2: Comparison of selected available samples in publicly available ECG arrhythmia database

Name	Total duration (seconds)	Sampling rate (Hz)	Total samples
MIT-BIH Arrhythmia (Moody et al. 2001b)	86 400	360	31 104 000
AF classification challenge 2017(Clifford et al. 2017)	277 138	300	83 141 400
SPH dataset (Liu et al. 2022)	281 109	500	140 554 500
CU-SPH dataset (Zheng et al. 2020)	106 460	500	53 230 000
PTB-XL 100 Hz (Wagner et al. 2020)	218 370	100	21 837 000
PTB-XL 500 Hz (Wagner et al. 2020)	218 370	500	109 185 000
IRIDIA-AF	24 085 688	200	4 817 137 600

3.2.4 Annotation evaluation

ECG and ECG annotation quality

The quality assessment for the waveform data was done during the data selection process. As stated previously, the data was first validated by an experienced specialist cardiac nurse and then validated again by an experienced cardiologist. Recordings presenting a high level of noise were rejected during this phase. All the AF crisis, starting from the AF onset to the AF termination, were annotated by the cardiologist and reviewed by the specialist cardiac nurse if a second opinion was needed. The labels were then cross-validated during the creation and clean-up of the database, as discussed in the previous sections.

AF detection with ML and DL models for annotation validation

Before the publication of the database, we evaluated the ECG waveform annotations and the RR interval annotations using ML and DL models. The task given to the model is to detect the presence of AF in an ECG window or RR interval window. The first model was trained and tested on the RR intervals and corresponding RR interval annotations. We created a XGBoost (XGB) model and derived Heart Rate Variability (HRV) features from the RR intervals. The model was implemented using the XGBoost and scikit-learn packages. The number of tress was set to 150 with a maximal depth of 5. The HRV features were extracted from the time domain, frequency domains and the Poincaré plot. We used a 10-fold cross-validation with stratification on the patient level, i.e. all the recordings from one patient can only be found in either the train or the test split.

We compared the performance of the XGBoost model, with a DL model. We implemented a 1-dimensional Convolutionnal Neural Network (CNN) using PyTorch, using an input window of 8192 samples corresponding to 40 seconds of ECG. We use a step of 4096 to select the windows for the training set, i.e. 50% overlap between windows and a 8192 step for the testing, i.e. 0% overlap between windows. The model is composed of 9 blocks of CNN with two branches, where the second branch is a skip connection has it has been shown to improve the training and results in deeper models. The model is described in Figure 3.9. CNN models have shown great performance for the classification of 1-dimensional time series and in particular for ECG classification and AF-related tasks, as shown by Attia et al. (2019a). The model was trained dur-

Table 3.3: Comparison of the results for AF detection task using two models: ML model (XGBoost) vs DL model (CNN). The value in parentheses represents the 95% confidence interval. AUROC is the area under the ROC curve. The metrics are computed using a threshold of 0.5.

Model	Input	Window size	AUROC	Accuracy	Sensitivity	Specificity	F1 score
XGBoost	HRV	300 RR	0.967	0.972	0.951	0.983	0.957
	RR	($\approx$ 5 minutes)	(0.950-0.983)	(0.961-0.983)	(0.917-0.984)	(0.975-0.990)	(0.938-0.975)
CNN	ECG	8192 samples	0.995	0.982	0.952	0.992	0.971
		($\approx$ 40 seconds)	(0.990-0.999)	(0.972-0.992)	(0.919-0.985)	(0.988-0.997)	(0.954-0.989)

ing 100 epochs with early stopping with patience of 5 epochs, i.e. the training stops after 5 epochs with no improvement on the validation set. The model was optimized using Adam, using a learning rate of 10^{-4}. The loss function was the binary cross-entropy. The bach size was 32. We used boostraping with 5 repetitions. For each one of the model trainings, a new train-validation-test split was created. Confidence intervals were computed for each metrics across the 5 repetitions. As for the first ML model, the recordings were separated at the patient level to avoid any contamination of the test set. The results of the two models are presented in Table 3.3.

Finally, we evaluated both models on an unseen patient recording. We used a sliding window to create the annotation of the model on the whole recording and compared it visually to the cardiologist annotations. The results for the ML model are presented in Figure 3.5 and the results from the DL model are presented in Figure 3.6. Both models were able to create new annotation corresponding to the cardiologist annotations with the 5 AF episodes present in the recording. It confirms the ability of ML and DL models to be used as a tool for medical decision support.

3.2.5 Database publication

This version 1 of the database is published on Zenodo (Gilon et al. 2023a), with the DOI `10.5281/zenodo.8186845` and is accessible at `https://zenodo.org/doi/10.5281/zenodo.8186845`. A main DOI represents the overall database, and each version of the database is assigned a new specific DOI. The URL to the main DOI has the nice property of redirecting to the latest version of the database.

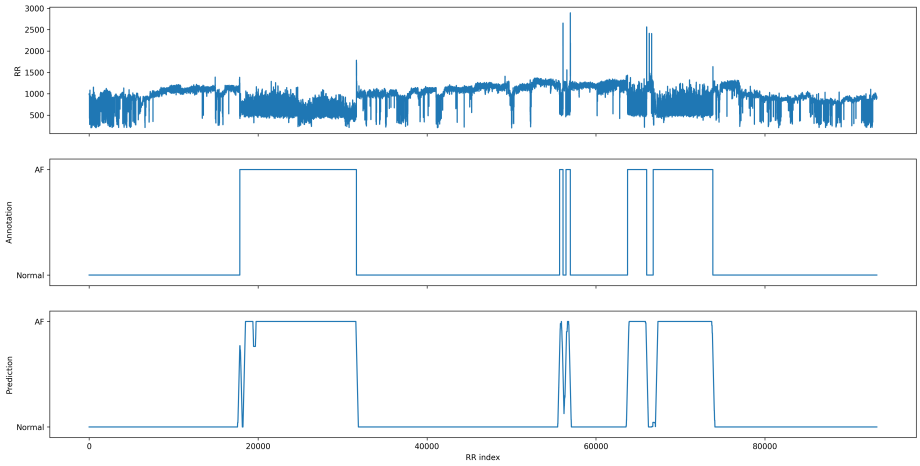


Figure 3.5: Prediction of the ML model on a test recording

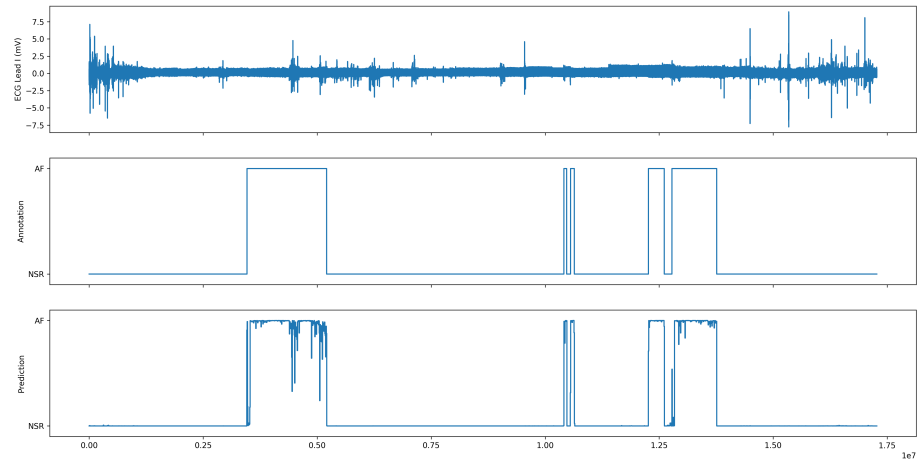


Figure 3.6: Prediction of the DL model a test recording

3.3 IRIDIA-AF database version 2

The version 2 of the IRIDIA-AF database is an extended version of the published IRIDIA-AF version 1. New recordings from the first centre, Dr Grégoire outpatient clinic, were added to the database. In addition, three new centres were added to the database. The three new centres are:

- CHU Ambroise Paré in Mons (Belgium),
- Centre Hospitalier de Luxembourg (Luxembourg),
- CHU Bruggmann in Brussels (Belgium).

The three centres were selected because of the use of the Holter monitors from Microport and corresponding Holter recordings software, Microport Synescope. Indeed, during the creation of the first version, the team developed understanding of the proprietary database structure. It was therefore easier to apply the same process as for Dr Grégoire outpatient clinic to select the patients and corresponding recordings. As for the first version, each paroxysmal AF crisis were annotated.

3.3.1 Dr Grégoire outpatient clinic

We continued the analysis of the recordings in the archive database of the Dr Grégoire outpatient clinic. Figure 3.7a shows the distribution of the 11656 available records. We found 1105 recordings from 2017 to 2018. After selection, 27 new Holter recordings were added to the database.

3.3.2 CHU Ambroise Paré

The retrospective study at the CHU Ambroise Paré in Mons was conducted in collaboration with Dr Pascal Godart and Dr Stéphane Carlier. The study was approved by the hospital ethics committee of the CHU Ambroise Paré. A total of 24 872 Holter recordings were collected between 2010 and 2020, as shown in Figure 3.7b. The ECG signal was recorded using Microport Spiderview Holter recorder at the sample rate of 200 Hz, with two leads: `Lead I` and `Lead II`. A total of 80 Holter recordings from 74 patients were selected, and 130 AF crisis were annotated. In addition, 29 healthy patients with Holter recordings presenting no signs of AF or other cardiac diseases and with a low level of noise

were selected. The selection of recordings is the same for all centres: rejection of patients with CEID, rejection of noisy recordings and rejection of recordings with other diseases.

3.3.3 Centre Hospitalier de Luxembourg

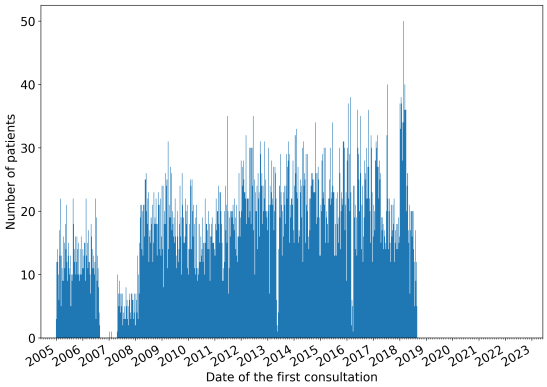
The retrospective study at the Centre Hospitalier de Luxembourg was made in collaboration with Dr Laurent Groben. This study was approved by the *Comité National d'éthique de recherche* from Luxembourg (N° 202101/01). Figure 3.7c presents the Holter recordings database from Centre Hospitalier de Luxembourg (CHL). It contains a total of 29 579 Holter monitorings, recorded from 2006 to 2021. The raw ECG signal data was recorded using Microport Spiderview Holter recorder at a sample rate of 200 Hz. `Lead I` and `Lead II` were recorded. Each year an average of 1848 $\pm$ 394 Holters recordings were recorded in the centre. A total of 250 recordings with paroxysmal AF from 226 patients were selected. It represents a total of 610 annotated AF episodes. In addition, 322 Holter recordings from 315 healthy patients were recorded.

3.3.4 CHU Brugmann

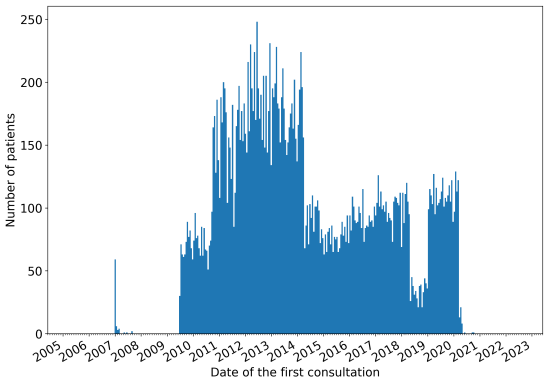
This retrospective study at the CHU Brugmann was made in collaboration wih Dr Thomas Nguyen. This study was approved by the *Comité d'Ethique CHU Brugmann*. Figure 3.7d shows the 29 764 Holter recordings available, recorded from 2007 to 2023. 362 recordings were selected by Dr Nguyen, using the annotations available with the recording. 317 files could be extracted from hospital archives. After re-analysis of recordings by a cardiologist, 113 recordings from 106 patients were selected and a total of 170 AF crisis were annotated.

3.3.5 Results

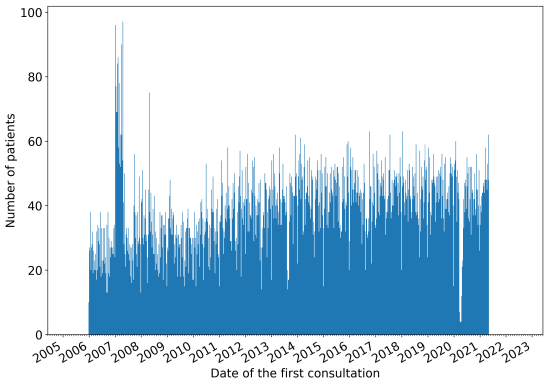
The IRIDIA-AF version 2 database is composed of 988 recordings from 928 patients. This represents only a small fraction, around 1%, of the 95 871 recordings composing the available archives in the centres as show in Figure 3.8. The number of recordings and patients in each centre is presented in Table 3.4. The database contains a total of 1319 AF crisis, as show in Table 3.6. A total of 835 AF have more than 60 minutes of sinus rhythm before the AF onset and more than 10 minutes of AF after the onset, and a total of 964 AF crisis have 30



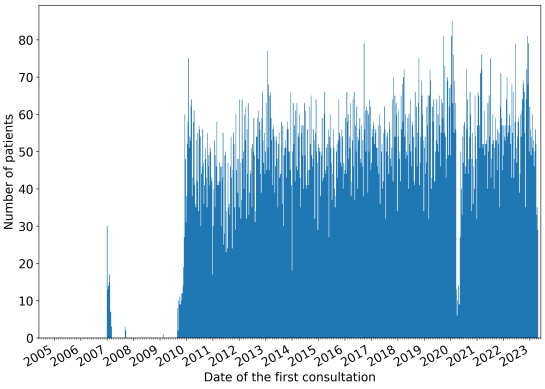
(a) Dr Grégroire outpatient clinic (n = 11 656)



(b) CHU Ambroise Paré (n = 24 872)



(c) CHL Luxembourg (n = 29 579)



(d) CHU Brugmann (n = 29 764)

Figure 3.7: Distribution of the Holter monitoring dates in the four centres: (a) Dr Grégoire outpatient clinic, (b) CHU Ambroise Paré, (c) CHL Luxembourg and (d) CHU Brugmann

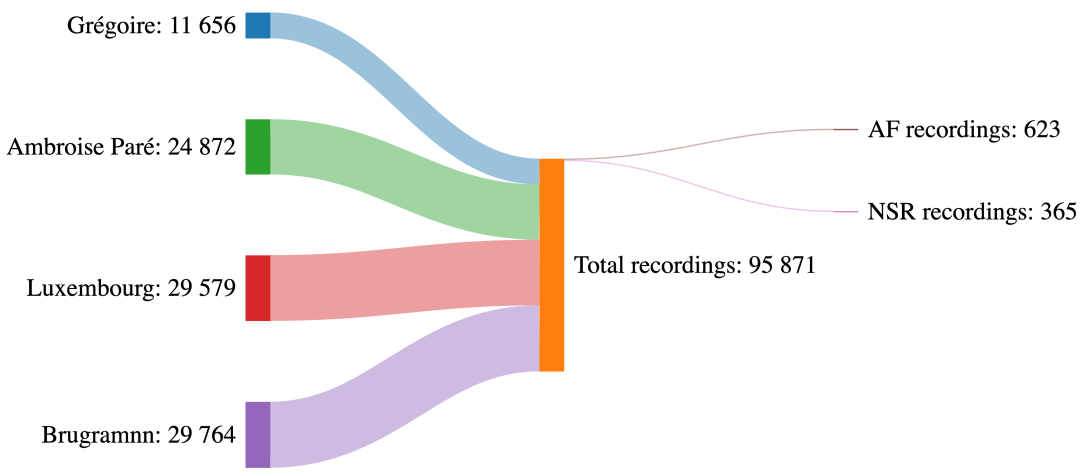


Figure 3.8: Diagram of the database composition

minutes or more sinus rhythm before the AF onset and 5 minutes or more of AF after the onset.

Microport recordings are separated in 24 hours recordings, therefore, the sinus rhythm window previous to the AF or AF crisis can be separated between two sub-recordings. The goal of this new database is to propose a new larger database than existing publicly available database to enable supervised learning for AF onset forecast and AF identification during sinus rhythm.

We can compare this new database with existing publicly available database, as presented in Table 3.7. This new database contains more patients, more recordings and more AF onset windows than the previously existing database by a large margin. This allows supervised training and testing of ML models on a larger scale. The amount of data would also be sufficient for training and testing DL models.

3.4 Database validation using AF detection

AF detection is a daily task in hospitals and outpatient clinics in order to detect and manage AF before one of its consequences, such as a stroke, reveals it (Hin-

Table 3.4: Composition of the IRIDIA-AF database version 2, comparing the number of patients and number of recording in each of the four centres included in the study

				Patients			Recordings			
Centre	City	Country	Years	Selected	AF	NSR	Available	Selected	AF	NSR
Dr Grégoire	Brussels	Belgium	2005–2018	178	164	14	11 656	194	180	14
CHU Ambroise Paré	Mons	Belgium	2007–2020	103	74	29	24 872	109	80	29
CHL	Luxembourg	Luxembourg	2005–2021	541	226	315	29 579	572	250	322
CHU Brugmann	Brussels	Belgium	2007–2023	106	106	0	29 764	113	113	0
Total				928	570	358	95 871	988	623	365

Table 3.5: Count of the number of AF episodes in the IRIDIA-AF database version 1, based on the durations of sinus rhythm before the AF onset and the duration of AF crisis

NSR duration before AF (min) \ AF duration (min)	> 0	> 5	> 10	Total
< 1	8	3	6	17
1–5	19	5	16	40
5–30	17	6	10	33
30–60	5	4	9	18
> 60	32	8	240	280
Total	81	26	281	388

Table 3.6: Count of the number of AF episodes in the IRIDIA-AF database version 2, based on the durations of sinus rhythm before the AF onset and the duration of AF crisis

NSR duration before AF (min) \ AF duration (min)	> 0	> 5	> 10	Total
< 1	8	3	6	17
1–5	30	6	26	62
5–30	60	27	51	138
30–60	23	11	71	105
> 60	115	47	835	997
Total	236	94	989	1319

Table 3.7: Comparison of publicly available ECG database for AF onset forecast with IRIDIA-AF v2, sorted by release date. The duration is indicated per recording. The AF episodes selected have $>$ 30 minutes sinus rhythm before the AF onset and $>$ 5 minutes of AF duration.

Database name	Year	Patients	Recordings	Duration	Sample Rate	Leads	AF episodes
MIT-BIH AFDB (Moody et al. 1983)	1983	25	25	30 min	250 Hz	2	11
AFPDB (Moody et al. 2001a)	2001	100	200	30 min	128 Hz	2	53
LTAFDB (Petrutiu et al. 2007)	2007	84	84	24 h	128 Hz	2	2
IRIDIA-AF v1 (Gilon et al. 2023b)	2023	152	167	24 - 96 h	200 Hz	2	261
IRIDIA-AF v2	2023	928	988	24 - 96 h	200 Hz	2	964

dricks et al. 2021). The detection of AF was used in Section 3.2.4 to validate the usability of the annotations on the IRIDIA-AF version 1. The same process has been applied to the Holter recordings of version 2. It provides a good starting point for a technical validation and usability test of the recordings and annotations contained in the new database.

For this task, we trained and tested multiple ML and DL models on the two types of recording signals, ECG and RR intervals, available for each recording. The RR intervals available in the database are based on the automatic detection made by the Microport Syneview software during the automatic analysis of the Hotler recordings.

AF detection can be modelled as a binary classification task. The models receive an ECG window as input and determine the probability of AF being present in this window. A high probability, close to 1, indicates that the model considers that AF is present in the ECG and a low probability, close to 0, indicates that the model considers that the ECG window contains only sinus rhythm.

3.4.1 Methods

AF detection using HRV parameters

All recordings in the database, from both AF and healthy patients, were divided into overlapping RR interval windows using a sliding window method. We selected all windows of size 300 RR intervals, with a step of 100 RR intervals. This means that two successive windows have 200 RR in common. The duration in seconds of each window is therefore variable, as it depend on the heart rate in the window. It corresponds to a 3-minute window when the heart rate is 100 bpm and a 5-minute window when the heart rate is 60 bpm. Out-

Table 3.8: Comparison of selected publicly available ECG arrhythmia database and IRIDIA-AF

Database name	# patients	# recordings	# leads	recording duration (seconds)		# classes
				min	max	
MIT-BIH Arrhythmia (Moody et al. 2001b)	47	48	2	1800	1800	2
AF classification challenge 2017 (Clifford et al. 2017)	8528	8528	1	9	61	4
SPH dataset (Liu et al. 2022)	24666	25770	12	10	60	59
CU-SPH dataset (Zheng et al. 2020)	10646	10646	12	10	10	11
PTB-XL (Wagner et al. 2020)	18885	21837	12	10	10	71
IRIDIA-AF v1 (Gilon et al. 2023b)	152	167	2	71 408	345 596	2
IRIDIA-AF v2	928	988	2	41 969	345 596	2

Table 3.9: Comparison of seconds and sample per lead in IRIDIA-AF with selected publicly available ECG arrhythmia database

Database name	Total duration (seconds)	Sampling rate (Hz) (Hz)	Total samples (Samples per lead)
MIT-BIH Arrhythmia (Moody et al. 2001b)	86 400	360	31 104 000
AF classification challenge 2017 (Clifford et al. 2017)	277 138	300	83 141 400
SPH dataset (Liu et al. 2022)	281 109	500	140 554 500
CU-SPH dataset (Zheng et al. 2020)	106 460	500	53 230 000
PTB-XL 100 Hz (Wagner et al. 2020)	218 370	100	21 837 000
PTB-XL 500 Hz (Wagner et al. 2020)	218 370	500	109 185 000
IRIDIA-AF v1 (Gilon et al. 2023b)	24 085 688	200	4 817 137 600
IRIDIA-AF v2	100 655 332	200	20 131 094 524

liers in the window, i.e. RR intervals less than 200 ms and greater than 4000 ms, were removed and linear interpolation was used to replace the missing RR intervals. This was applied to 0.01% of the total RR intervals. HRV parameters were calculated on the window. Parameters were extracted from time domain, frequency domain, Poincaré plot, Second Order Difference Plot (SODP) plot and heart rate fragmentation.

Four different ML models have been trained to classify HRV parameters from RR interval window. We first trained a logistic regression as baseline. Then, we selected three ML models and trained them on the same training split. The models are Decision Tree, Random Forest (RF), XGB. The decision tree and the random forest were implemented using scikit-learn package. The random forest was implemented using 200 trees with a depth of 10. The XGB model was implemented using xgboost library and the number of trees was set to 100 with a maximum depth of 5.

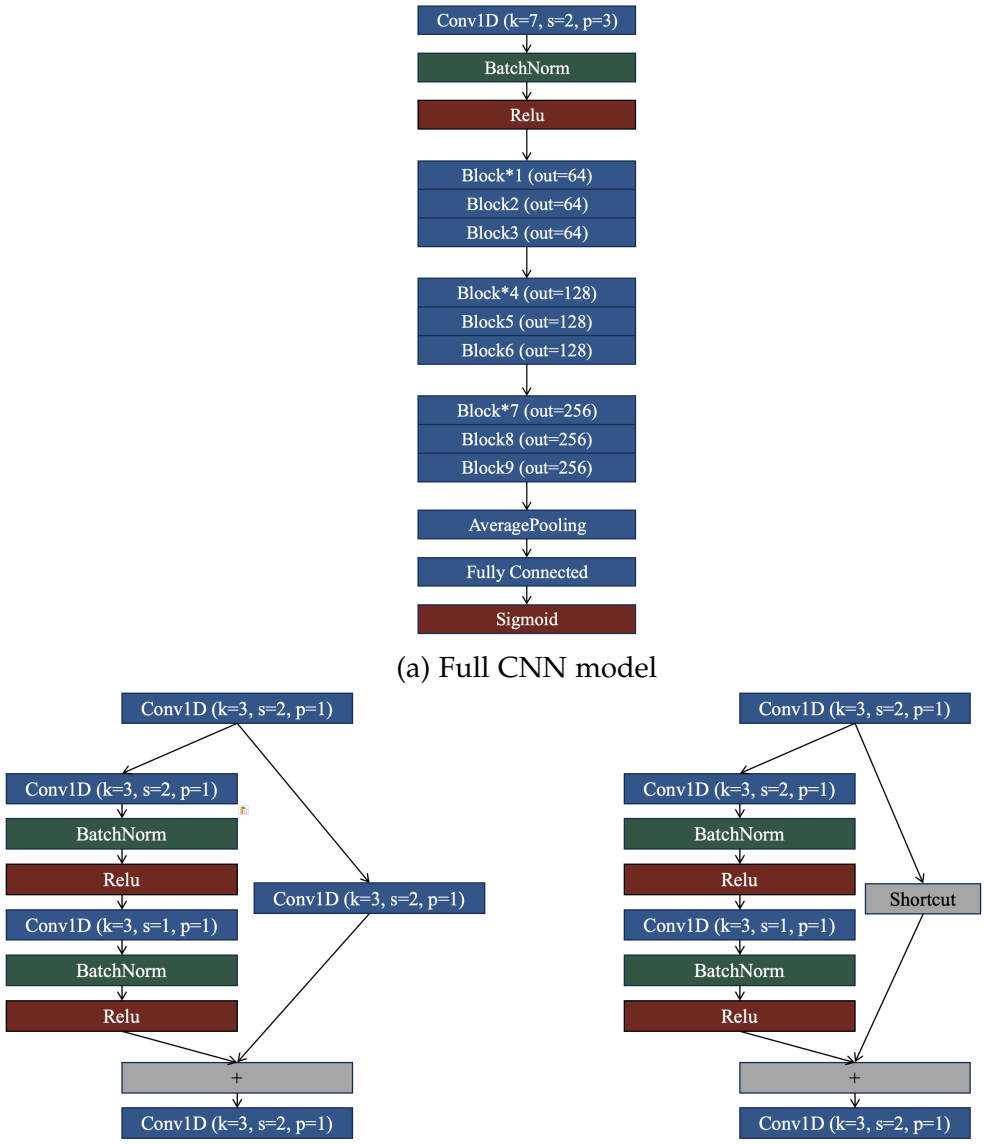
The performance of the models were compared across 10-fold cross-validation at patient level. The same splits were used across all models to obtain comparable metrics.

AF detection using ECG

For the second class of model, as for IRIDIA-AF version 1 annotation evaluation, we used a 1D CNN model, with an input size of 8192 samples corresponding to 40 seconds of ECG. We used a step of 8192 to select windows, i.e. there is no overlapping between windows.

The model is using the same architecture and is composed of 9 blocks of CNN with two branches, where the second branch is a skip connection. 1-dimensional CNN models have shown great performance for ECG classification and AF related tasks, as show by Attia et al. (2019a). The CNN model is composed of 9 blocks, with shortcut, as shown in Figure 3.9. The final prediction is the results of the sigmoid activation of the output from the last fully connected layer.

The model is created in PyTorch (Paszke et al. 2019), we used Adam (Kingma et al. 2017) as optimizer, and the binary cross-entropy loss function. We used a learning rate of 10^{-4} and a batch size of 128. The model was trained up to 100 epochs, with an early stopping strategy after 3 epochs if the validation loss did not decrease. After two epochs with no increase, the learning rate was divided by 2.



(a) Full CNN model

(b) Block*

(c) Block

Figure 3.9: Architecture of the CNN model used for AF identification. The model architect (a) is composed of an input block, followed by 9 blocks and a final classification. The 9 blocks are divided into 3 groups, where the first block is a Block* (b) and the next two are Blocks (c).

Cross-validation methods

To evaluate the generalization of the predictions made by a trained ML model, two evaluation methods were used: temporal cross-validation and spatial cross-validation. For each of the two methods, the database is separated into two parts: a training part and a test part. For some models, a validation part can also be created for use during training. The separation is carried out at patient level to avoid any data leakage between the training and the testing set. A patient and Holter recordings can only be contained in either the training or testing split. This separation process is repeated multiple times to determine the stability of the results.

For the temporal cross-validation, the recordings are sorted by recording date and separated into n splits, as suggested by Cerqueira et al. (2020). The first split contains the oldest recordings and the last split contains the most recent recordings, as shown in Figure 3.10. During the annotation process, we had the opportunity to visually analyse all the recordings, and we found no difference in quality between the first and last recordings, as the recording devices and the analysis software were the same for across all four centres. We wanted to keep the number of splits in the training set constant, with one split used as a test split and another as a validation split if the model required it. In this paper, the DL models use a validation split to stop training early. If a patient has multiple recordings, the first recording date is used and all recordings from that patient are placed in the same split.

The database consists of recordings from four centres. Spatial cross-validation uses this feature to divide the database into a test split containing recordings from one centre and a training split containing the remaining three centres. For models using a validation split during the training phase, the remaining three centres are divided proportionally into a training split and a validation split, as shown in Figure 3.11. Due to the variable number of recordings in each centre, we decided to use the first option and split the recordings from the three centres between training and validation. This is equivalent to using an ML model in clinical practice in the first three hospitals and testing the performance of the model in a fourth hospital.

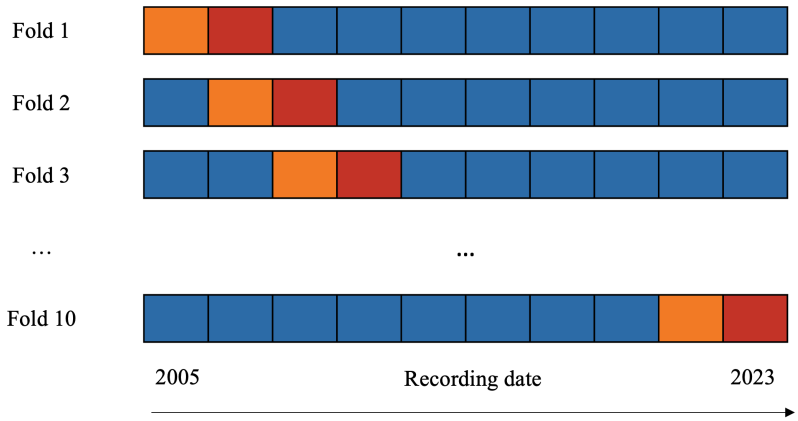


Figure 3.10: Temporal cross-validation. The data set is divided into a training set (blue), a validation set (orange) and a test set (red). The validation set is used in particular by the DL model to stop training early.

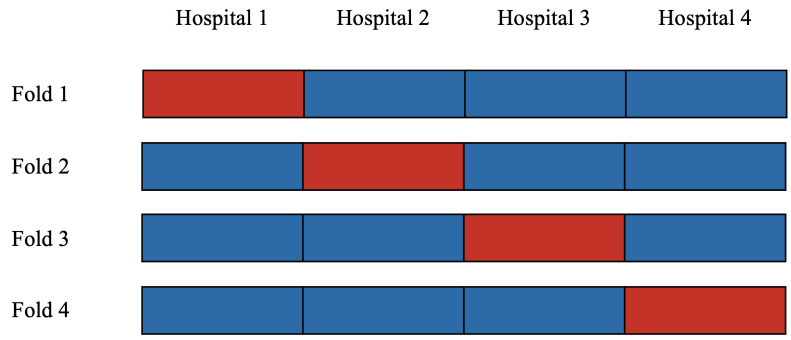


Figure 3.11: Spacial cross-validation: the blue splits are used for training and the red split for testing

Table 3.10: Evaluation metrics on the 10-fold temporal cross-validation using ML models

Model	AUROC	AUPRC
Logistic Regression	0.957 (0.943–0.971)	0.772 (0.726–0.818)
Decision Tree	0.947 (0.937–0.957)	0.926 (0.911–0.942)
Random Forest	0.990 (0.984–0.997)	0.963 (0.937–0.989)
XGBoost	0.990 (0.983–0.996)	0.960 (0.936–0.984)

Table 3.11: Evaluation metrics on the 10-fold temporal cross-validation using CNN model

Model	Input size	Leads	AUROC	AUPRC
CNN	8192	I	0.987 (0.982-0.993)	0.948 (0.920-0.975)
CNN	8192	I&II	0.988 (0.983-0.993)	0.954 (0.928-0.980)

3.4.2 Results

Temporal cross-validation

We evaluated ML and DL models using temporal split on the IRIDIA-AF v2 database. The results are presented in Table 3.10 for the ML models and in Table 3.11 for the DL models. We found that RF and XGBoost achieve the best average performance with an Area Under the Receiver Operating Characteristic Curve (AUROC) of 0.990. Compare to the evaluation on IRIDIA-AF v1 where the CNN model achieved better results than the XGB model, the tendency is now reversed.

Inter-hospital cross-validation

As a second evaluation method, we compared the performance of ML models using inter-hospital cross-validation on the four centres composing the IRIDIA-AF v2 database. The results are presented in Table 3.12 for the ML models. We found that RF and XGB achieve similar AUROC performance, but the XGB model achieved better Area Under the Precision-Recall Curve (AUPRC) on three of the four centres. The best results were achieved on recordings from Dr Grégoire outpatient clinic, and the lowest results were achieved when the model were tested on recordings from CHU Ambroise Paré.

Table 3.12: Performance of the classifier on the hospital

	XGBoost		Random Forest	
Hospital	AUROC	AUPRC	AUROC	AUPRC
Dr Grégoire	0.999	0.998	0.996	0.964
CHU Ambroise Paré	0.956	0.960	0.953	0.958
CHL Luxembourg	0.976	0.844	0.974	0.865
CHU Brugmann	0.983	0.980	0.980	0.979

Complete recording prediction

AF detection algorithm could help to reduce cardiologist Holter monitoring recordings reading and annotation work. If an algorithm is proven to have at least the same level of annotation accuracies as a trained clinical technician or cardiologist, this model could be used to make an initial reading and annotations of the recordings.

The algorithm using HRV features computed from RR intervals has a high level of accuracy on a single window, but these results should be extended to correspond to the whole recording.

To validate the model-based approach on the complete recording, we selected a random patient and trained a new RF model on the HRV feature dataset, excluding the recordings of the selected patient. Then, using a sliding window over the entire recording, we compute the mean AF detection probability. This is done by taking the mean of all predictions from all windows containing the RR. The corresponding prediction is shown in Figure 3.12.

Using the same patient, we used the trained model from the fold in which the selected patient is in the test split. We compute a new prediction every 256 samples and computing the mean of the result for each sample. The results are presented in Figure 3.13 showing predictions corresponding to cardiologist annotations.

3.5 Summary

From a state-of-the-art analysis, we have highlighted the need for a new database for the task of AF onset forecast, as the publicly available databases for this task are limited. In this chapter, we presented a new publicly available database consisting of two versions: IRIDIA-AF version 1 and version 2.

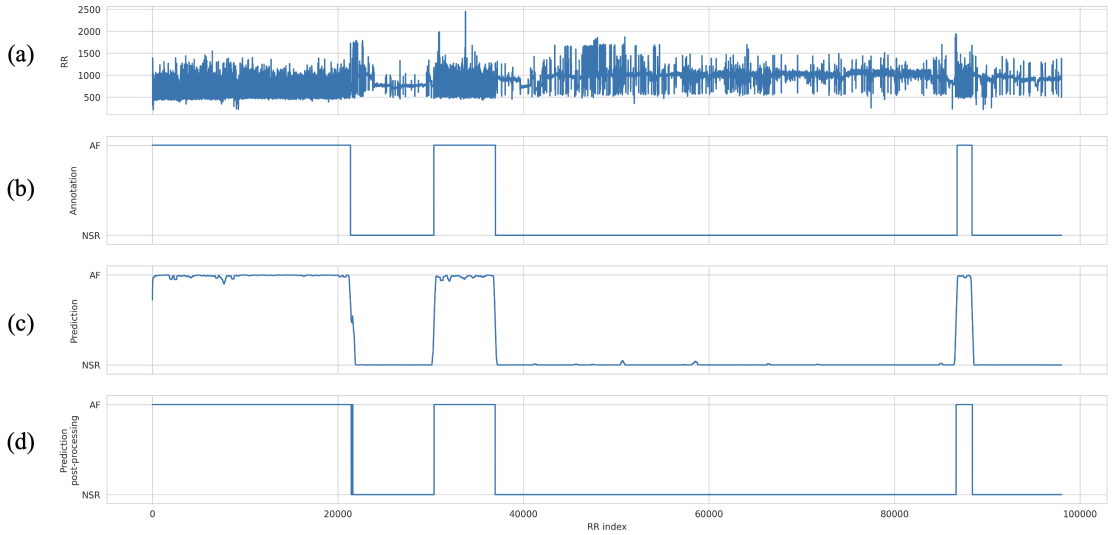


Figure 3.12: ML model prediction of AF presence in a Holter recording. The ECG intervals (a) are used by as input by the CNN model. Using the cardiologist annotations (b) of AF onset and AF offset, AF crisis can be labelled. The mean AF presence prediction (c) is computed for all RR interval. Using a 0.5 threshold, the final prediction is presented in (d).

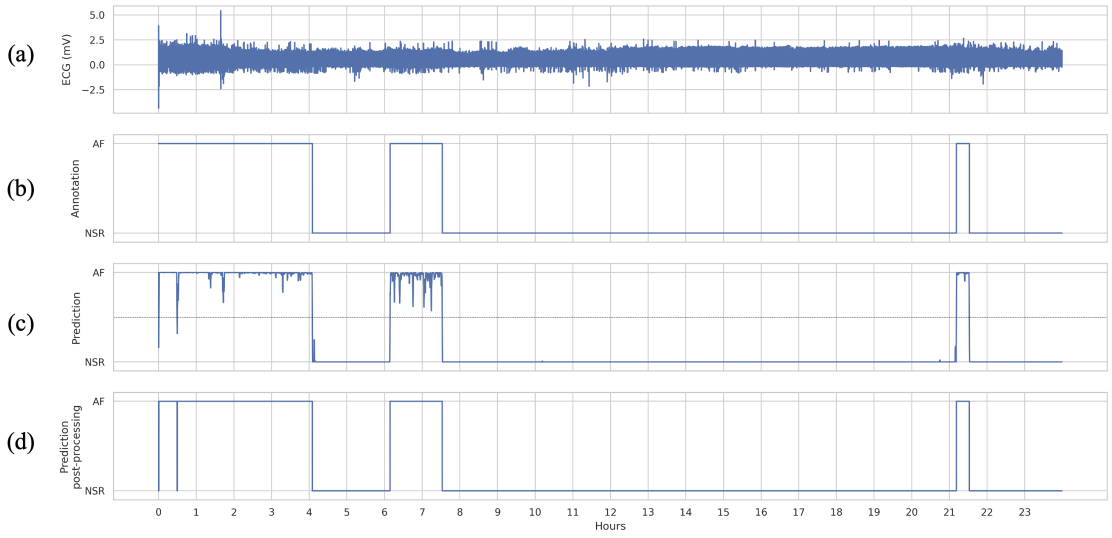


Figure 3.13: CNN model prediction of AF presence in a Holter recording. The RR intervals (a) are used to compute the HRV features for each window. Using the cardiologist annotations (b) of AF onset and AF offset, AF crisis can be labelled. The mean AF presence prediction (c) is computed for all ECG windows. Using a 0.5 threshold, the final prediction is presented in (d).

IRIDIA-AF version 1 has already been published online on Zenodo and is available to the research community. This database contains 167 Holter recordings from 157 patients in an outpatient clinic. The second version, IRIDIA-AF version 2, consists of 988 recordings from 928 patients. Three new hospitals were added to the database, bringing the total number of centres to 4. All AF onsets and offsets were annotated in the recordings, with a total of 964 AF crises with AF lasting more than 5 minutes and 30 minutes or more of NSR before AF onset. To the best of our knowledge, this database is the largest publicly available database of annotated recordings for AF research.

To validate the annotations, we evaluated the performance of the ML and DL models on both IRIDIA-AF version 1 and version 2. In both cases, the models showed impressive performance in AF detection, with results consistent with those previously obtained by machine learning and deep learning models using alternative databases in the literature.

In this context, and in line with the high performance of the models, the detection of AF can be considered an easy task. The extension of the database towards a version 3 should be further developed by including new centres and other cardiovascular diseases. Automatic selection of high quality recordings and rejection of low quality recordings could also simplify the process. For versions 1 and 2, we relied on the expertise of a cardiologist and a cardiac nurse to make this selection, but defined selection criteria could be created based on the existing database as noise ratio or p-wave, qrs and t-wave detection ratio. The annotations should be redefined to include four labels: (i) AF, (ii) NSR, (iii) other diseases and (iv) noisy parts. This could help to understand the generalisation of model performance. Automated annotation of atrial fibrillation in Holter monitoring can significantly improve clinical accessibility and interpretation, speeding up analysis and enhancing the use of such AF screening monitoring in clinical settings.

Chapter 4

Paroxysmal atrial fibrillation onset forecast

4.1 Introduction

The main research question of this thesis is to understand if Atrial Fibrillation (AF) onset can be forecast, particularly using Machine Learning (ML) and Deep Learning (DL) models. The underlying question is to understand if these models are able to extract meaningful information from the Normal Sinus Rhythm (NSR) preceding the AF onset. The comparison of larger DL models with more classical ML models was made possible by the use of the novel database introduced in Chapter 3, as existing public databases do not contain enough recordings. Previous studies in this research area have mainly used the Paroxysmal Atrial Fibrillation Prediction Database (AFPDB), which is significantly smaller than IRIDIA-AF. In the literature, the AF onset forecast task is presented as a binary classification problem. To achieve good performance, two classes of Electrocardiogram (ECG) windows should be distinguished: (i) ECG windows close to the AF onset, in the pre-AF windows, and (ii) ECG windows distant from the AF onset.

Most published results are based on the AFPDB. However, it is important to note that this database has a limited number of AF onset recordings. Although the ML models presented in the existing literature have reported high performance on this task, it may be premature to consider the task as completely solved. In this chapter, we begin our investigation by assessing the performance of state-of-the-art models reimplemented according to the methodology described in their respective publication. The aim is to assess the level of re-

producibility of the scientific literature on the prediction of AF onset.

We then discuss the evolution of ECG waves, intervals and complexes prior to AF onset to understand if changes can be observed in the ECG. We then compare the evolution of the performance of a ML model in classifying 5-minute windows close to and distant from AF onsets as the windows close to the AF onset are moved further away. Finally, we compare the performance of several ML models using a variety of input representations in classifying selected 30-minute windows close to AF onsets and 30-minute windows at least 2 hours away from AF onsets. Using the model which achieved the best average metrics, we analyse the predictions on complete recordings.

4.2 State-of-the-art reproducibility

Reproducibility of published work is not always straightforward, and can be even more difficult without access to the original code, or even impossible without access to the dataset used in the study. Recently, ML-based methods have shown great results in various areas, including AF onset prediction. Reviews have shown that ML publications contain errors or missing information in the methodology. In addition, data leakage and overfitting from train to test split can also be found in publications, making the results difficult or sometimes impossible to reproduce (Vandewiele et al. 2021; Shim et al. 2021). It is necessary to provide an independent data split to test the model to ensure a good generalisation of the prediction (Walsh et al. 2021). We have selected, analysed and reproduced the method of 3 publications on AF onset prediction using the AFPDB to understand whether the results can be reproduced.

4.2.1 PAF Prediction Challenge Database

The AFPDB (Moody et al. 2001a) is the database of the three selected studies. This dataset is composed of 200 ECG recordings from 100 patients, with 2 recordings per patient. The data and labels are also available on the Physionet website (Goldberger et al. 2000). Each recording is 30 minutes long and the sampling frequency is 128 Hz. For AF patients, one recording is preceding an AF and the other is distant from any AF sign, with at least 45 minutes before and 30 minutes after. For healthy patients, the two recordings are NSR.

For the 2001 Physionet Challenge, the dataset was split into a train set and a hidden test set and later published as a whole. The challenge train set consists

Table 4.1: Selected models for AF onset prediction

Authors	Model	Window	Features
Mohebbi et al. (2012)	SVM	30	non-linear frequency-domain bispectrum
Boon et al. (2018)	SVM	5	time-domain bispectrum frequency-domain non-linear
Narin et al. (2018)	KNN	5	time-domain frequency-domain

of 50 patients with 25 healthy patients and 25 AF patients, resulting in a total of 100 healthy recordings, 50 NSR recordings from AF patients and 50 pre-AF recordings. The test set consists of 50 patients with 22 healthy and 28 AF patients, resulting in a total of 44 healthy recordings, 28 AF NSR recordings and 28 pre-AF recordings. If both healthy and AF NSR recordings are grouped together, the data set consists of 53 pre-AF recordings and 147 recordings far from any AF sign.

4.2.2 Materials and methods

We searched the literature for publications on AF onset forecast. 10 challenge entries and 6 resulting publications were presented at the Computing in Cardiology 2001 conference. We found more than 30 publications on this topic, published after the 2001 conference and the release of the database. We selected 3 ML-based methods using Heart Rate Variability (HRV) features, based on the total number of citations at the time of the selection. The methods are summarised in Table 4.1.

For each selected publication, we reproduce the methodology as presented in each publication. We also create several alternative scenarios by varying the dataset selection, dataset split, HRV features computation, and model parameters. Each scenario was run 1000 times and the average accuracy, sensitivity and specificity were reported with a 95% confidence interval.

Model SVM-30

A Support Vector Machine (SVM) classifier is proposed by Mohebbi et al. (2012). They used the full 30-minute window of the signal as a single window. After preprocessing, they used QRS complex (QRS) detection to create the HRV signal. They computed frequency domain features (Low Frequencies (LF) and High Frequencies (HF)), bispectrum features, Poincaré plot features (SD1, SD2, SD1/SD2) and sample entropy. They used the first train-test split from the challenge, but restricted the dataset to the 53 AF patients. In total, 106 recordings were used, 50 ECGs for training and 56 ECGs for testing. The best average results were obtained using an SVM with $C = 1000$ and $\gamma = 3.6$.

We implemented several models using the 30 minute window. The first with C=1000 and γ=3.6 as presented in the methodology, and an alternative second SVM with C chosen in $[0.1, 1, 10, 100, 1000, 10000]$ and γ chosen in $[10, 3.6, 1, 0.1, 0.01, 0.001, 0.0001]$. We used the initial train-test split as in the publication. We also tested whether the use of feature standardisation improved the results.

Model SVM-5

Boon et al. (2018) proposed an SVM classifier using HRV features computed from 5-minute windows. HRV is extracted from ECGs using QRS detection. Features from the time domain, frequency domain and bispectrum are used. They used Genetic Algorithms (GA) to select features and the final set still consists of temporal features (NN50, pNN50), non-linear features (entropy, SD2), frequency domain features and bispectrum features. They used 10-fold cross-validation to analyse their performance on the 106 ECGs from the 53 AF patients.

We created an SVM model with HRV features extracted from 5-minute windows, using a GA to select the features. We used the variable C and γ as in the SVM-30 model. We tested two types of window selection: only the last five minutes of the recordings or all 5-minute windows available from the 30-minute window (with 50% overlap). We tested two datasets: the first with all AF patients and the second with the whole dataset, using 10-fold CV at the patient level.

Table 4.2: Reported results for AF onset forecast in the original publications

Authors	Accuracy	Sensitivity	Specificity
Mohebbi et al. (2012)	-	96.30%	93.10%
Boon et al. (2018)	87.7%	86.8%	88.7%
Narin et al. (2018)	90.0%	92.0%	88.0%

Model KNN

Narin et al. (2018) used a K-Nearest Neighbours (KNN) model. The 30-minute recordings are divided into 5-minute windows with 50% overlap. They used all the training data except the *n27* recording. In total, there are 74 NSR recordings and 25 AF recordings. To compute their classifier performance, they used 10-fold CV. They extracted HRV from the ECG signal and used temporal, frequency and non-linear features. To select features, they used a GA where each feature usage is encoded as a bit. They present several models with results for different dataset splits, feature selection and k value for KNN models. The best model uses $k = 3$ and 5 features (Root Mean Square of Successive RR interval Differences (RMSSD), LF, VLF and total power).

We trained a KNN model with HRV features from the 5-minute RR interval window and selected HRV features using a GA. We tested three dataset splits: the train split from the challenge using only AF patients, the train split from the challenge using all patients, and finally the full dataset. We used two types of 10-fold cross-validation, either at the recording level or at the patient level, i.e. the two recordings of a patient should be included in the same split. We could not verify whether the same constraint was applied in the original methodology.

4.2.3 Results

For all 3 methods, we were unable to reproduce the results reported in the publications. The performance of the reproduced models and extended models were lower than reported. The reproduced results are presented in Table 4.2.

Model SVM-30

For the model from Mohebbi et al. (2012) we obtained a sensitivity of 78.57% at the cost of a low specificity of 10.71%. The results are shown in Table 4.3 for

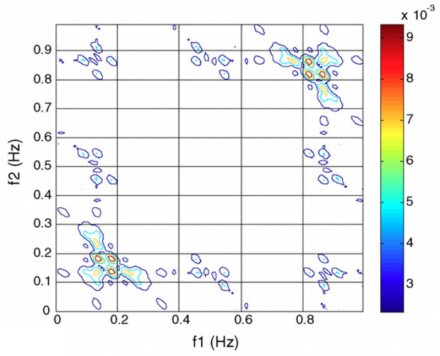


Figure 4.1: Figure 4a from Mohebbi et al. (2012)

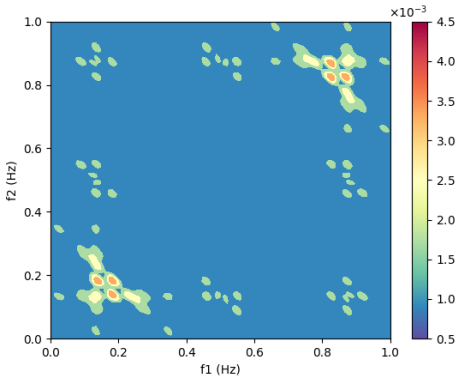


Figure 4.2: Reproduction

Figure 4.3: Reproduction of the biamplitude contour plot for a 30-minute window from the p03 recording of the AFPDB

both fixed and variable parameters. We were able to reproduce the bispectral plot presented in the publication as shown in Figure 4.3 to validate our method. The model seems to overfit on one class when no standardisation is used. The results were better using varying C and γ.

Model SVM-5

For the method of (Boon et al. 2018) we achieved a top accuracy of 74.45% corresponding to a low specify of 17.85% and a high specificity of 94.82%. The results with the different scenarios and data splits are shown in Table 4.5. The results with the whole data set were better in terms of accuracy, but at the cost of a much lower sensitivity.

Model KNN

For the model of (Narin et al. 2018) we obtained 75.62% accuracy for KNN with $k = 3$, with a sensitivity of 40.80% and a specificity of 87.44%. The results are shown in Table 4.6. The distribution of sensitivity and specificity for each fold is shown in Figure 4.4. The results using splits are better at the recording level, but could be related to data leakage between the training and test sets. Indeed, in this case, the model could be tested on the HRV of a patient already present in the training data.

Table 4.3: Reproduced results for SVM-30 model with $C = 1000$ and $\gamma = 3.6$. Results are presented with 95% confidence interval.

Dataset train	Dataset test	Patient	# ECG	Norm.	Accuracy (%)	Sensitivity (%)	Specificity (%)
train	test	AF	50	no	50.0 (50.0-50.0)	100.0 (100.0-100.0)	0.0 (0.0-0.0)
train	test	AF	50	yes	44.64 (44.64-44.64)	78.57 (78.57-78.57)	10.71 (10.71-10.71)
train	test	AF+NSR	108	no	72.0 (72.0-72.0)	0.0 (0.0-0.0)	100.0 (100.0-100.0)
train	test	AF+NSR	108	yes	70.0 (70.0-70.0)	0.0 (0.0-0.0)	97.22 (97.22-97.22)
train	test	AF	50	no	50.0 (50.0-50.0)	100.0 (100.0-100.0)	0.0 (0.0-0.0)
train	test	AF	50	yes	47.92 (47.58-48.26)	59.28 (57.72-60.83)	36.56 (34.8-38.32)
train	test	AF+NSR	108	no	72.0 (72.0-72.0)	0.0 (0.0-0.0)	100.0 (100.0-100.0)
train	test	AF+NSR	108	yes	69.67 (69.46-69.87)	5.81 (5.35-6.27)	94.5 (94.06-94.95)

Table 4.4: Reproduced results for SVM-30 model with variable C and γ. Results are presented with 95% confidence interval.

Dataset train	Dataset test	Patient	# ECG	Norm.	Accuracy (%)	Sensitivity (%)	Specificity (%)
train	test	AF	50	no	50.0 (50.0-50.0)	100.0 (100.0-100.0)	0.0 (0.0-0.0)
train	test	AF	50	yes	47.92 (47.58-48.26)	59.28 (57.72-60.83)	36.56 (34.8-38.32)
train	test	AF+NSR	108	no	72.0 (72.0-72.0)	0.0 (0.0-0.0)	100.0 (100.0-100.0)
train	test	AF+NSR	108	yes	69.67 (69.46-69.87)	5.81 (5.35-6.27)	94.5 (94.06-94.95)

Table 4.5: Reproduced results for SVM-5 model. Results are presented with 95% confidence interval.

Dataset	Windows	Patient	# ECG	CV	Accuracy (%)	Sensitivity (%)	Specificity (%)
train+test	last	AF	106	patients	53.17 (52.87-53.47)	56.89 (56.37-57.4)	50.91 (50.17-51.64)
train+test	last	AF+NSR	200	patients	72.33 (72.16-72.5)	10.06 (9.45-10.67)	94.79 (94.39-95.19)
train+test	all	AF	106	patients	62.84 (62.46-63.22)	61.33 (60.45-62.21)	64.39 (63.83-64.96)
train+test	all	AF+NSR	200	patients	74.45 (74.35-74.54)	17.95 (16.83-19.07)	94.82 (94.43-95.2)

Table 4.6: Reproduced results for KNN model with k=3. Results are presented with 95% confidence interval.

Dataset	Patient	# ECGs	CV	Accuracy (%)	Sensitivity (%)	Specificity (%)
train	AF	50	patients	58.11 (58.11-58.11)	60.61 (60.61-60.61)	55.61 (55.61-55.61)
train	AF	50	recording	65.07 (65.01-65.13)	65.12 (65.03-65.2)	65.21 (65.13-65.28)
train	AF+NSR	100	patients	69.58 (69.53-69.64)	33.34 (33.2-33.48)	82.06 (82.0-82.11)
train	AF+NSR	100	recordings	75.62 (75.58-75.65)	40.80 (40.71-40.89)	87.44 (87.4-87.47)
train+test	AF+NSR	200	patients	64.45 (64.41-64.48)	28.31 (28.22-28.39)	77.65 (77.6-77.69)
train+test	AF+NSR	200	recordings	70.03 (70.01-70.06)	32.65 (32.59-32.72)	83.63 (83.6-83.66)

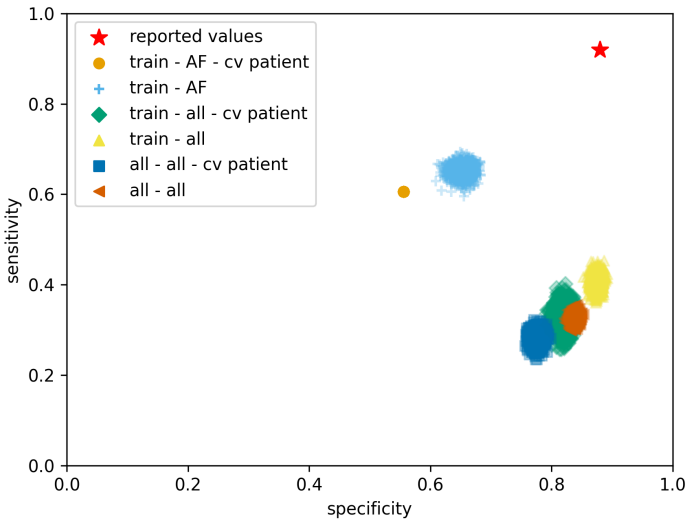


Figure 4.4: Sensitivity and specificity for KNN with k=3 in the different scenarios

Table 4.7: Comparison of reported results and reproduced results for AF onset forecast

Authors	Reported		Reproduced	
	Sensitivity	Specificity	Sensitivity	Specificity
Mohebbi et al. (2012)	96.30%	93.10%	59.28%	36.56%
Boon et al. (2018)	86.8%	88.7%	61.33%	64.39%
Narin et al. (2018)	92.0%	88.0%	65.12%	65.21%

4.2.4 Discussion

Published methodologies achieving overly optimistic and non-reproducible results may be the cause of rejection of ML methods by clinicians, as discussed by Shah et al. (2022). We reviewed three publications proposing ML models for AF onset forecast. We reproduced the methodologies presented in the publications. Our results did not correspond with those reported by the authors as shown in Table 4.7, with large differences. The results also highlight the class imbalance issue in this database, as the model is able to achieve a 72% accuracy with a 0% sensitivity and a 100% specificity, meaning that the ML model was not able to find useful information from the data. No AF episode has been detected, as the model is always predicting recordings in the NSR class.

In addition, some methodological questions remain unanswered. We could

not verify whether the same constraints were applied in the original methodologies regarding the cross-validation between patient. This could have produce a data leakage in the original publications. We did not receive answers when we contacted the authors. In particular, we should know whether the authors pre-processed or normalised the features before using them in the model. Another question concerns the selection of data, as if only a subset of the data was selected and used by the model, this could also help to explain the differences between the results reported in the articles and those reproduced. Researchers are now proposing frameworks to help authors include all the necessary material and methodological information to better reproduce their work (Walsh et al. 2021). We think a good step forward would be to open source the code created by the author to retrain the exact same model.

4.3 Evolution of the ECG before AF onset

This section examines the evolution of the waves, complexes and intervals that make up the ECG heartbeats before the onset of AF. The aim is to understand whether underlying patterns can be detected in the evolution of the ECG. Although we were not able to reproduce the results presented in existing publications, some models showed both sensitivity and specificity of over 65% for their predictions. These models were able to find useful information in the HRV data for the prediction of recordings preceding AF onset. Based on these results, we examined the evolution of HRV features and ECG waves, complexes and intervals before AF onset to understand if we could observe an evolution in the ECG signal.

4.3.1 Method

We used the IRIDIA-AF v2 database, as presented in Chapter 3. We selected all the AF crisis onsets with 30 minutes of NSR before the AF onset of an AF lasting at least 5 minutes, as originally proposed in the AFPDB. Baseline wander was removed using a high-pass filter (Kher 2019; Makowski et al. 2021) and power line interference were removed using a notch filter at 50 Hz. We checked the recording to ensure that the ECG was not inverted and, if it was, the trace was inverted along the y-axis QRS complexes and R peaks were detected using the Pan-Tomkins algorithm (Pan et al. 1985).

The selected NSR window is analysed using a 5-minute sliding window with 30 second increments, as the conventional duration for short-term recording analysis is 5 minutes (Task Force of The European Society of Cardiology and The North American 1996; Shaffer et al. 2017). For each window, each heartbeat is analysed separately, using the positions of the corresponding QRS complex as anchor points to divide the sliding window into heartbeats. Invalid measurements were excluded from the analysis. The onset, peak and offset of the P wave were located in the 200 ms before the R peak, as the P-R interval is typically 200 ms. The T wave was located in the 320 ms after the end of the QRS complex (Feher 2012; Padsalgikar 2017). We detected the boundaries using signal processing and calculated the corresponding features for each wave and complex.

For the P wave, we compute the wave duration, the amplitude and the total area between the onset and offset. For the QRS complex, we computed the duration, the amplitude, and area under the QRS complex. In addition, we computed the same value for the R peak only. For the T wave, we computed the duration, amplitude, and area between T wave onset and T wave offset. We also compute the QT interval, RR interval, and TQ intervals. For the RR interval and TQ interval, the duration is computed between the current beat and the previous beat.

Finally, we plotted the features evolutions from -30 minutes, i.e. -1800 seconds before AF to 0, i.e. the AF onset. We use the mean and 95% Confidence Interval (CI) to represent each time window. Corrected QT intervals were calculated using Bazett Equation (4.1) (Bazett 1920; Taran et al. 1947). This correction removes the effect of the heart rate on the QT duration, allowing the QT duration to be compared at different heart rates. We used the same equation for the TQ interval correction in Equation (4.2) and for the P wave duration correction in Equation (4.3).

$$QT_c = \frac{QT}{\sqrt{RR}} \tag{4.1}$$

$$TQ_c = \frac{TQ}{\sqrt{RR}} \tag{4.2}$$

$$P_c = \frac{P}{\sqrt{RR}} \tag{4.3}$$

As a second analysis, we studied the evolution of the HRV parameters. We

selected 5-minute sliding windows with a step of 30 seconds. For each window we detected the QRS complexes, created the RR interval series and calculated the selected HRV features. In particular, we analysed the evolution of the mean RR interval, Standard deviation of NN intervals (SDNN), RMSSD and pNN50 in the time domain features. For the frequency domain, we analysed the LF and HF evolution. Finally, for Poincaré, we analysed SD1 and SD2. The complete description of HRV features is available in Appendix A.

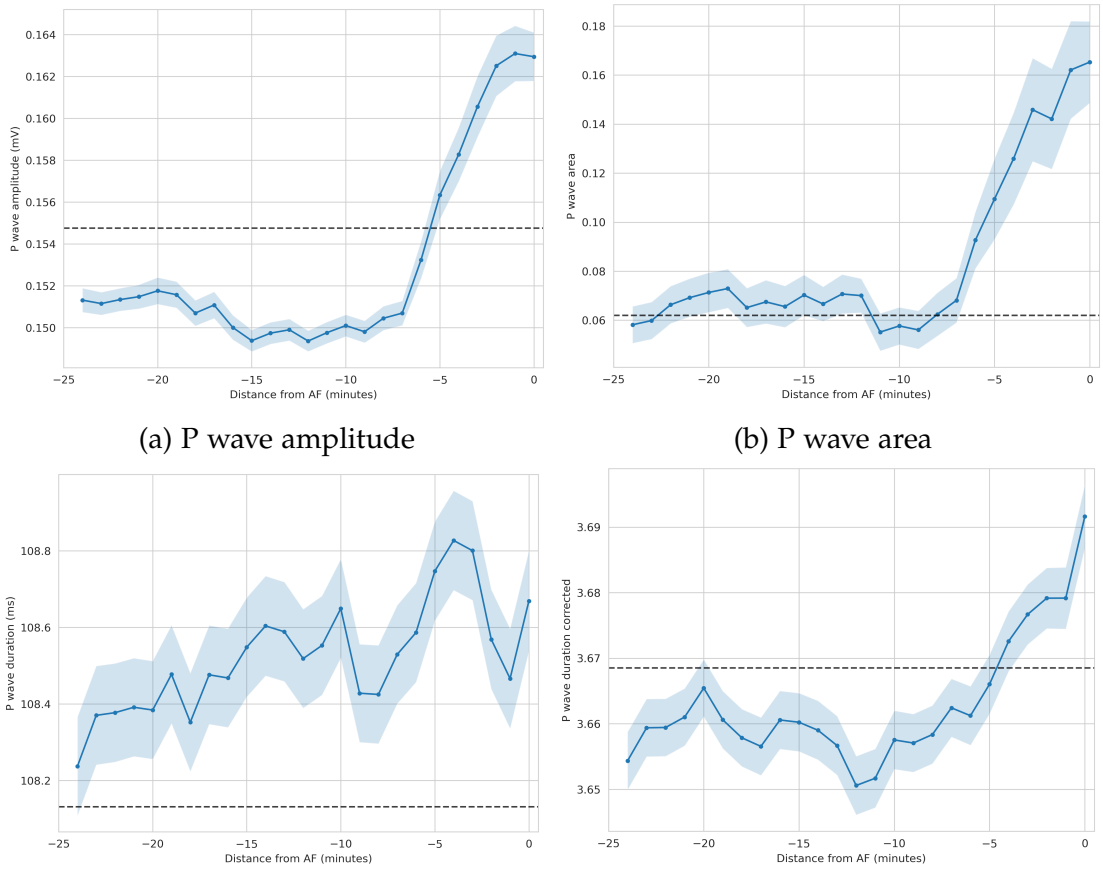
4.3.2 Results

ECG waves complexes and intervals

Among the 623 patients with AF, we identified 964 AF episodes that met the selection criteria. This corresponds to AF in 521 patients in 570 recordings. The results were grouped into Figure 4.5 for the P waves, Figure 4.6 for the QRS complexes and R waves, Figure 4.7 for the T waves, Figure 4.8 for the QT and TQ intervals and finally Figure 4.10 for the RR intervals and ectopic beats. The T wave duration has a remarkably low confidence interval around the mean at AF onset.

HRV features

The selected parameters were calculated from the corresponding 964 AF onsets: mean RR, SDNN, RMSSD, pnn50, Poincaré SD1, Poincaré SD2, LF and HF. The time domain characteristics are shown in Figure 4.13. We found that the mean RR is decreasing, which corresponds to the results we previously obtained for heart rate in Figure 4.12, i.e. an increasing heart rate means shorter RR intervals. We found an increase in mean heart rate of 2 beats per minute just before the onset of atrial fibrillation compared to 30 minutes before. Long-term variability and short-term variability also increase before the onset of AF. Long-term variability was measured mainly using SDNN and pNN_{50} and the short-term variability using RMSSD. The LF results show an increasing trend, but it should be noted that the range of confidence intervals increases in the 5 minutes before the onset of AF. In this period it is much larger than for the other measures, between 5 minutes and 1 hour before AF.



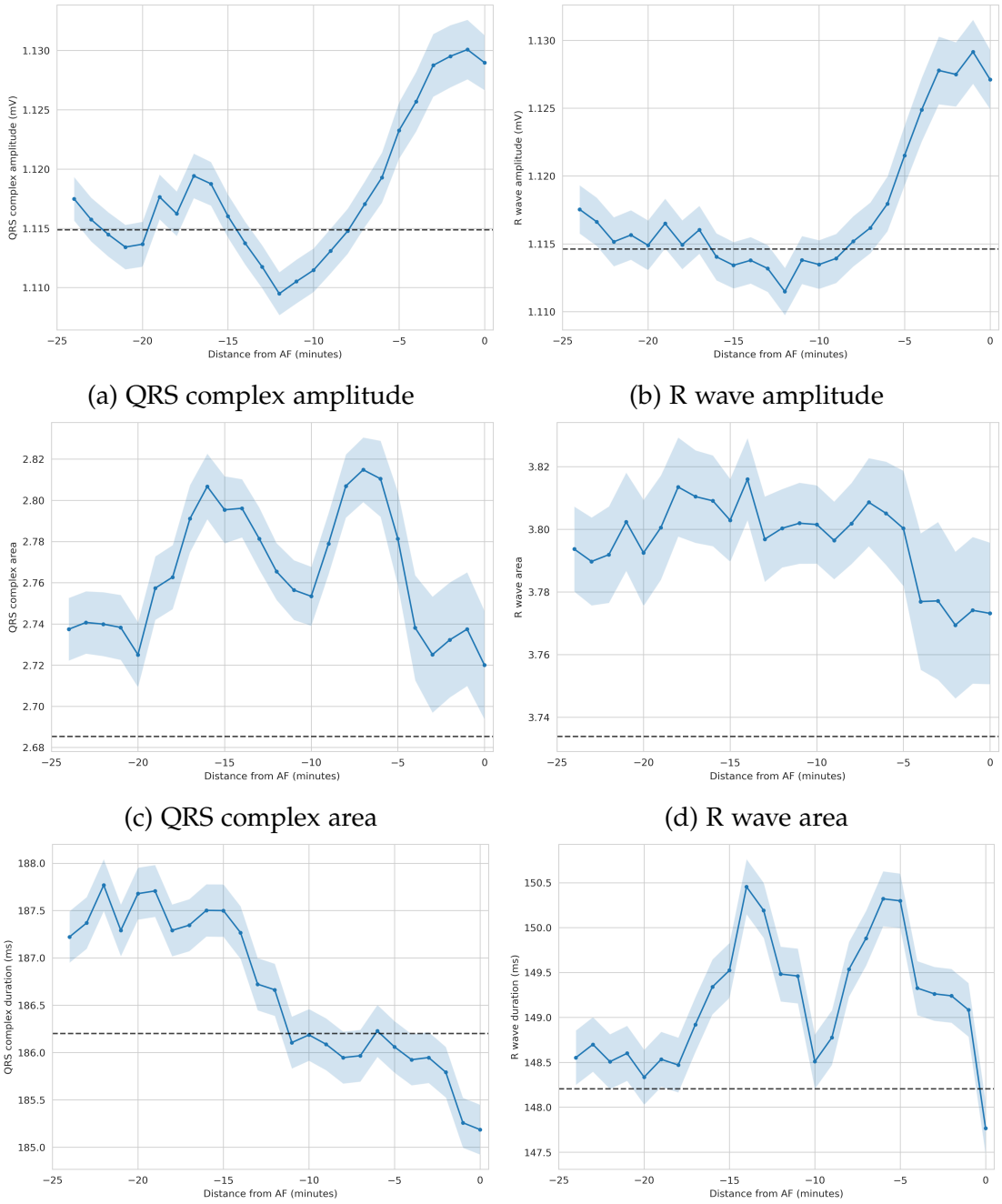
(a) P wave amplitude

(b) P wave area

(c) P wave duration

(d) P wave corrected duration

Figure 4.5: Evolution of P wave before AF onset. The analysis was performed using a 5-minute sliding window and a step of 30 seconds. The main line corresponds to the mean value of all selected windows. The 95% confidence interval is displayed around the line. The baseline (black dotted line) represents the average value of the features in the 2-hour period preceding the analysed 30-minute window.



(a) QRS complex amplitude

(b) R wave amplitude

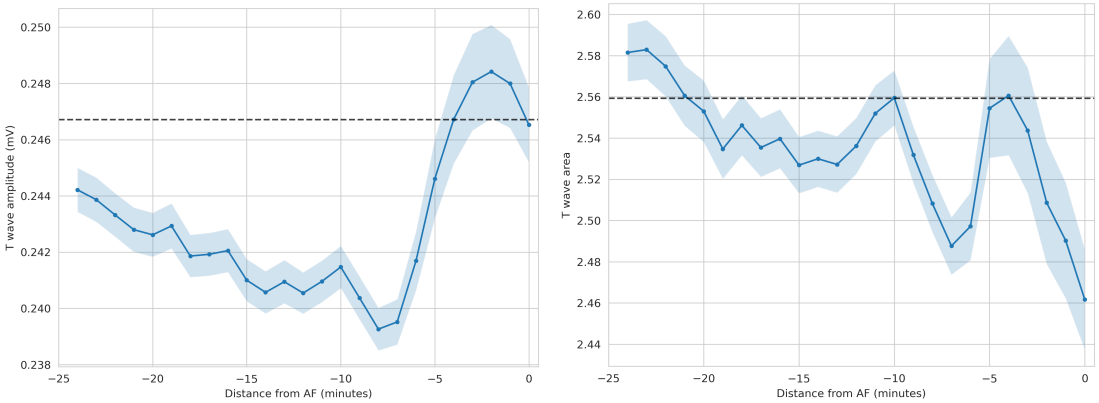
(c) QRS complex area

(d) R wave area

(e) QRS complex duration

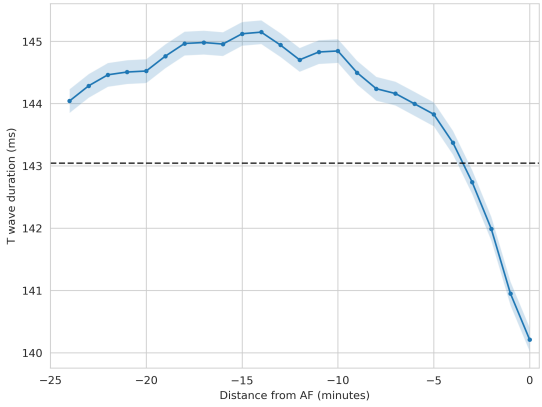
(f) R wave duration

Figure 4.6: Evolution of QRS complex and R wave before AF onset. The analysis was performed using a 5-minute sliding window and a step of 30 seconds. The main line corresponds to the mean value of all selected windows. The 95% confidence interval is displayed around the line.



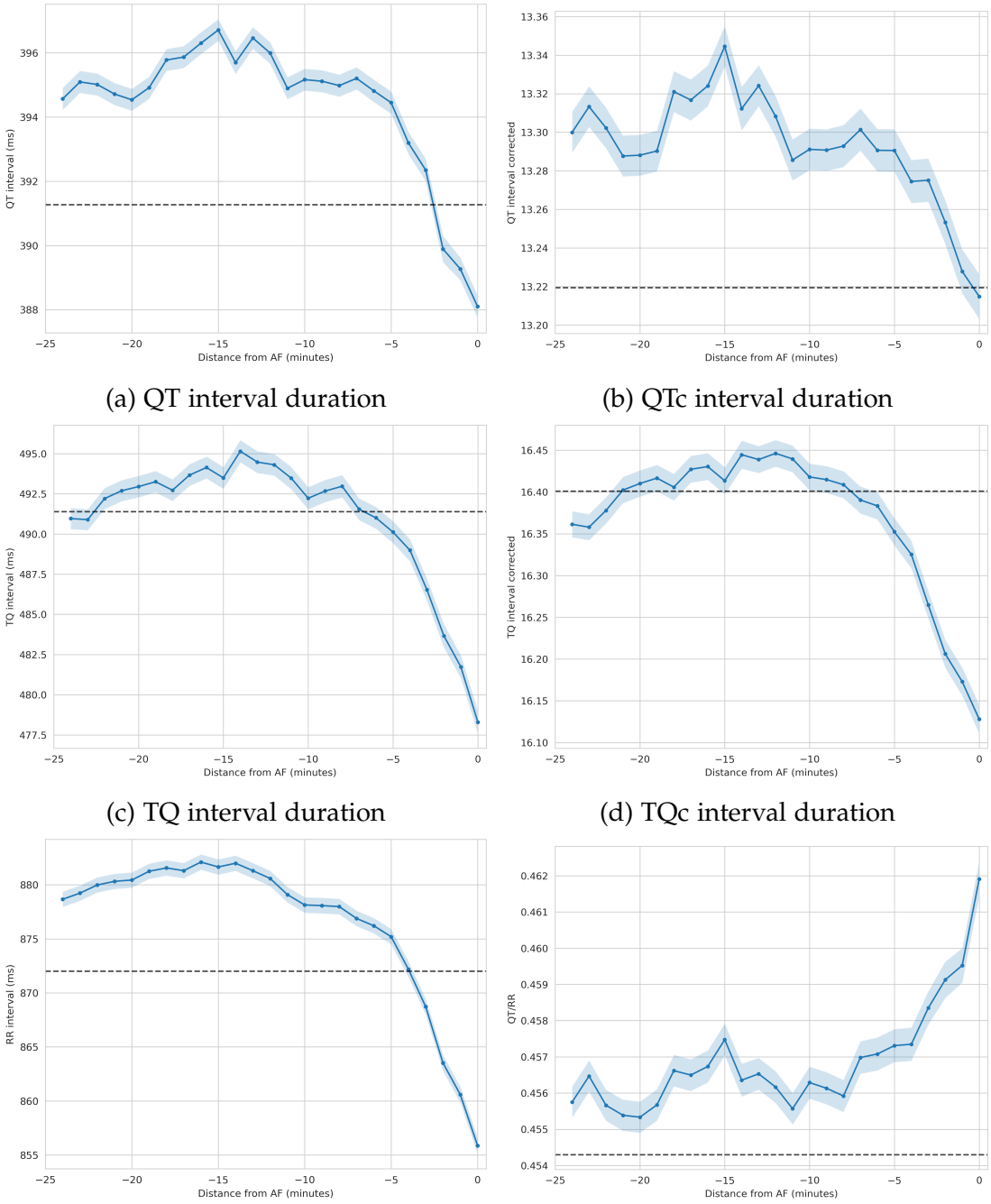
(a) T wave amplitdude

(b) T wave area



(c) T wave duration

Figure 4.7: Evolution of T wave before AF onset. The analysis was performed using a 5-minute sliding window and a step of 30 seconds. The main line corresponds to the mean value of all selected windows. The 95% confidence interval is displayed around the line.



(a) QT interval duration

(b) QTc interval duration

(c) TQ interval duration

(d) TQc interval duration

(e) RR interval duration

(f) QT-RR

Figure 4.8: Evolution of QT and TQ intervals before AF onset. The analysis was performed using a 5-minute sliding window and a step of 30 seconds. The main line corresponds to the mean value of all selected windows. The 95% confidence interval is displayed around the line.

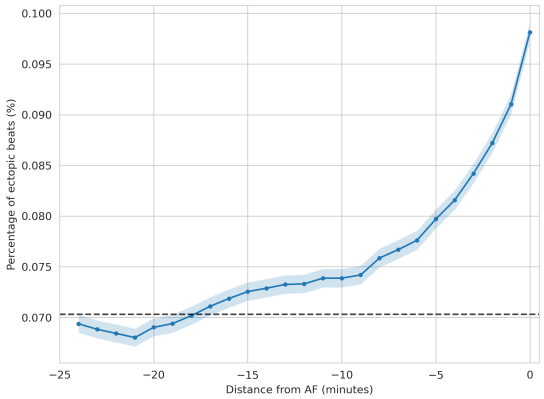


Figure 4.9: Percentage of ectopic beats

Figure 4.10: Evolution of number of ectopic beats before AF onset. The analysis was performed using a 5-minute sliding window and a step of 30 seconds. The main line corresponds to the mean value of all selected windows. The 95% confidence interval is displayed around the line.

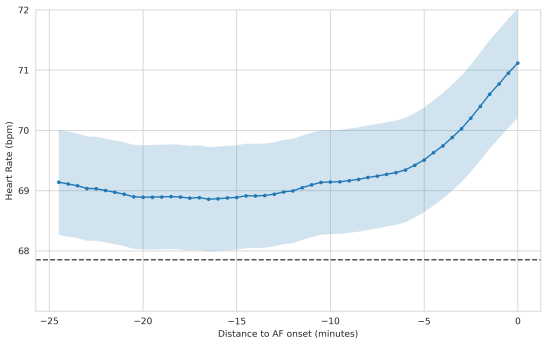
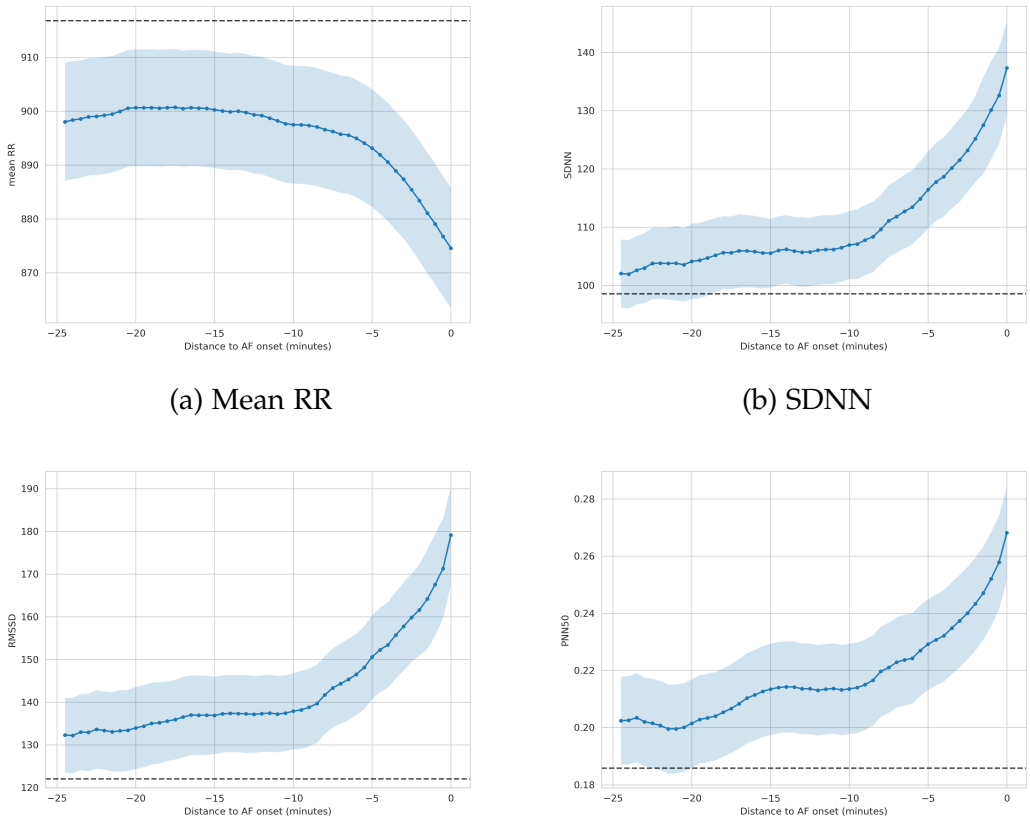


Figure 4.11: Heart rate

Figure 4.12: Evolution of heart rate before AF onset. The analysis was performed using a 5-minute sliding window and a step of 30 seconds. The main line corresponds to the mean value of all selected windows. The 95% confidence interval is displayed around the line.



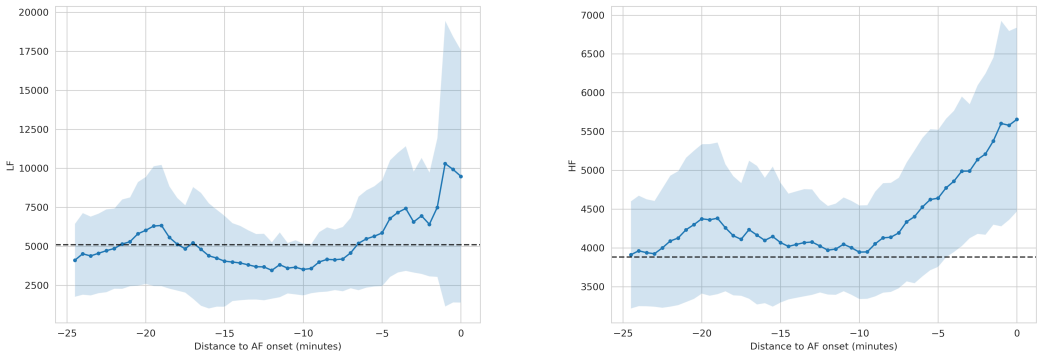
(a) Mean RR

(b) SDNN

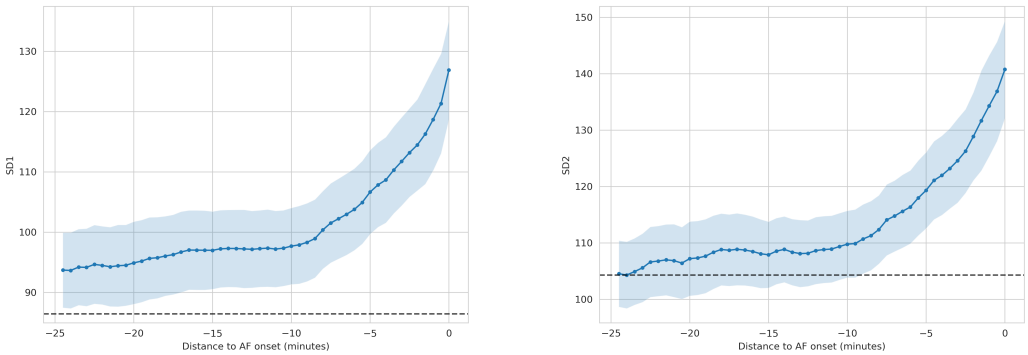
(c) RMSSD

(d) pNN50

Figure 4.13: Evolution of HRV features before AF onset. The analysis was performed using a 5-minute sliding window and a step of 30 seconds. The main line corresponds to the mean value of all selected windows. The 95% confidence interval is displayed around the line.



(a) LF (b) HF



(c) Poincaré SD1 (d) Poincaré SD2

Figure 4.14: Evolution of HRV features before AF onset. The analysis was performed using a 5-minute sliding window and a step of 30 seconds. The main line corresponds to the mean value of all selected windows. The 95% confidence interval is displayed around the line.

4.3.3 Discussion

We found that there are moving trends in the ECG waves, complexes and intervals as well as in the HRV parameters in the ECG windows preceding the AF onset. The P wave evolves before the onset of AF, as shown in Figure 4.5. It was previously shown that a relationship was found between AF risk and P wave duration (Sebestyén et al. 2016; Hari et al. 2018). We found an increasing trend for P-wave amplitude, area under the P-wave and P-wave corrected duration, with final values above the baselines. Both area and amplitude start to increase around 400 seconds before the onset of AF. The P-wave duration shows an increasing trend line without a clear increase in the last minutes, similarly to the other parameters. It has been shown that the P-wave duration is increased in stroke patients compared to healthy patients (Dewland et al. 2013; Deschamps et al. 2023).

Coumel et al. (1994) describe that any heart disease alters the QT dynamics and Batchvarov et al. (2002) found that the QT-RR has a high inter-patient variability but a low intra-patient variability, i.e. a high stability. In Figure 4.8f we found (i) that the QT-RR varies before this AF onset and (ii) that the increasing trend is similar between all recordings, with values above the baseline. We also observed an increase in high frequencies above the baseline, as we found previously in the analysis of IRIDIA-AF v1 (Grégoire et al. 2022). This could be related to an increase in Parasympathetic Nervous System (PNS) activity.

The development of ectopic beats before the onset of AF has previously been studied by Vikman et al. (1999). We confirmed the findings of increasing trends before the onset of AF, from 7% of the heartbeats to 10% of the heartbeats. Waktare et al. (2001) also found that ectopic beats were more frequent when they preceded an episode of AF. Premature Atrial Contraction (PAC) are related to the trigger described in the Coumel triangle. The triangle defines three components required for AF onset: a substrate, a modulator and finally a trigger (Coumel 1994; Rebecchi et al. 2023). The substrate is the anatomical and electrophysiological support that allows the arrhythmia to persist. The trigger is the electrical element that is likely to cause the arrhythmia by activating the arrhythmogenic substrate. Finally, the Autonomic Nervous System (ANS) acts as a modulator and uses the first two factors by increasing the sensitivity of the arrhythmogenic substrate, thereby facilitating the onset of the AF.

4.4 Evolution of predictions before AF onset

Forecasting the onset of AF using ML is also a more complex task than detecting AF. The idea is to detect whether there are signs of incoming AF in the NSR that precede the onset of AF. As this has never been shown to be possible in real life conditions, there is currently no developed treatment. We can think of several research directions for the future of AF onset forecasting, one being a *pill in the pocket* strategy if AF can be detected a few hours before its onset. Indeed, the time required for a pill to take effect is not immediate. The other direction could be the implementation of an effective overdrive algorithm in the Cardiac Implantable Electronic Device (CIED), which could benefit from the results of an initial AF onset prediction algorithm and be activated as a second step. The last direction we can think of is stimulation of the parasympathetic and vagal systems. We have shown that parasympathetic activity increases before the AF onset. Stimulating it could reduce its activity and allow a return to a normal state. (Kharbanda et al. 2022).

4.4.1 Materials and methods

As a first step in analysing the performance of ML models in forecasting AF, we compared the performance of the model in a binary classification of windows close to AF compared to windows distant from AF. In this experiment, we examined the performance of the model when the windows defined as pre-AF are moving away from AF onset.

Holter recordings in the IRIDIA-AF database can be divided into three categories: AF, pre-AF and inter-AF. AF crises start with the onset of AF and end with the offset of AF. It should be noted that Holter monitoring recordings are long-term recordings and do not always contain the beginning and end of each crisis. For some AF crises, the AF onset happened before the start of Holter monitoring. For other AF episodes, the AF offset happened after the end of monitoring. Therefore, not all annotated AF crises in the database could be used AF onset forecast.

The pre-AF is the selected window preceding the onset of AF. Pre-AF windows were defined as 30-minute windows in the AFPBD database (Moody et al. 2001a). We used a sliding 5-minute sub-window, because short-term HRV measurements are usually based on 5-minute recordings. (Shaffer et al. 2017). Inter-AF windows correspond the residual portion of the recordings that is

neither AF nor pre-AF. Inter-AF are windows between AF episodes and are defined as 30 minute windows separated by at least 2 hours from any AF sign in the AFPBD database (Moody et al. 2001a). This distance to AF onset can be selected as an additional parameter. If the post-NSR distance is greater than the pre-AF window size, there is a gap between the NSR and pre-AF windows.

We created a dataset for AF onset forecast using IRIDIA-AF v2 database, described in Chapter 3. Only patients with AF were included in the dataset, and only AF crises lasting more than 5 minutes were selected. We selected evolving pre-AF windows, starting with pre-AF windows just before the onset of AF up to pre-AF windows 1 hour before the onset of AF.

We selected all distance from 0 up to 1 hour, i.e. just before the AF onset to 1 hour before the AF onset. We used a step of 30 seconds, for a total of 120 selected distances. This is shown in Figure 4.15. For each distance, we selected all windows between the selected distance d and 5 minutes after, i.e. 300 seconds corresponding to $d - 300$. We added an extension of 30 seconds, as data augmentation, to allows the selection of additional windows for each distance. Therefore, for each distance d, the windows are selected between d and $d - 300 - 30$, with a duration of 300 and a step of 1. The selection of window boundaries is presented in Table 4.8.

The dataset was balanced with randomly selected windows at least 2 hours away from any AF signs before the start of the windows and 2 hours after the end of the windows. For each distance, we created 10 datasets and for each dataset we used a 10-fold cross validation at the patient level, for a total of 100 experiments per distance. IN total, it represents 12 000 experiments. We ensure that each fold was composed of the same ratio of pre-AF and NSR windows.

As a first baseline model from the results we obtained on AF detection during database validation, we chose to use a Random Forest (RF) model (Breiman 2001). A random forest is an ensemble learning method that fits a number of decision tree classifiers and uses averaging of the individual predictions to improve the global predictive accuracy of the model and control over-fitting. For this experiment, we set the number of trees to 100, with a maximal depth of 10.

From each 5-minute window, we selected up to 300 corresponding RR intervals. From the RR intervals, we compute HRV parameters from time-domains, geometrical plots and frequency-domain. The complete feature input consists of:

- time domain features: mean heart rate, SDNN, RMSSD, Standard Devia-

Distances			AF windows		
Index	Start	End	Index	Start	End
0	0	-329	w_0	0	-300
			w_1	-1	-301
			w_2	-2	-302
			...		
			w_{29}	-29	-329
1	-30	-359	w_0	-30	-360
			...		
			w_{29}	-59	-359
	...			...	
119	-3570	-3900	w_0	-3570	-3870
			...		
			w_{29}	-3599	-3899

Table 4.8: Selection of window boundaries for the evolution of prediction before AF onset. For each experiments from 0 to

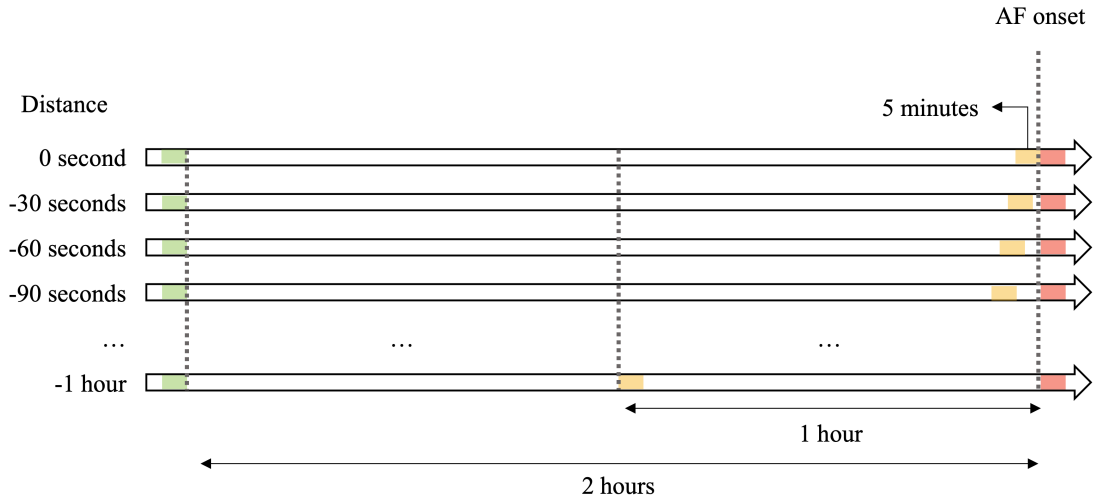


Figure 4.15: Selection of windows for AF onset forecast prediction evolution. The Holter recording is divided into three types of windows: AF (red), pre-AF (yellow) and inter-AF (green). We successively construct new datasets with increasing pre-AF distance, starting from AF windows just before AF onset to windows 1 hour before onset. Inter-AF windows are chosen to be 2 hours away from any AF sign. Each dataset correspond to a pre-AF distance and is composed of 5-minute pre-AF windows (yellow) and 5-minute inter-AF windows (green).

tion of Successive RR interval Differences (SDSD), CVNN, CVSD, pNN10, pNN20, pNN50, minNN, maxNN, medianNN, prc20NN, prc80NN, Triangular Interpolation of the NN interval histogram (TINN), and HRV Triangular Index (HRVi);

- frequency domain features: total power, power LF band, power HF band, their normalized values, and LF/HF ratio;
- Poincaré plot features: SD1, SD2, and SD1/SD2 ratio, Cardiac Sympathetic Index (CSI), Cardiac Vagal Index (CVI) and CVI modified;
- Second Order Difference Plot (SODP) features: number of ΔRR in Q1 to Q4, Central Tendency Measure $(\text{CTM})_{20}$, CTM_{50}, and CTM_{100};
- Acceleration (AC), Deceleration (DC), AC modified, DC modified, AC_k and DC_k;
- Heart Rate Fragmentation (HRF) features: Percentage of Inflection Points (PIP), Inverse of the Average Length of the acceleration/deceleration Segments (IALS), Percentage of Short Segments (PSS), and Percentage of Alternating Segments (PAS).

For each distance, the performance was evaluated using the Area Under the Receiver Operating Characteristic Curve (AUROC) and the Area Under the Precision-Recall Curve (AUPRC) for the binary classification between pre-AF at distance d and NSR windows. We also evaluate the predictions using threshold-based metrics, choosing a threshold of 0.5.

4.4.2 Results

Using the RF model with HRV features computed from the RR intervals, we found that the model is able to make the distinction between pre-AF and NSR windows. We also found a gradual increase in AUROC performance when the pre-AF windows are closer to the AF onset, up to a performance of 0.714 AUROC (95% CI 0.692–0.735) and AUPRC 0.697 (95% CI 0.671–0.724). For the evaluation 1 hour before the AF onset, the model achieved a lower performance, with an AUROC of 0.562 (95% CI 0.539–0.586) and an AUPRC of 0.555 (95% CI 0.534–0.576). The complete results are presented in Appendix A in Tables A.1 to A.3.

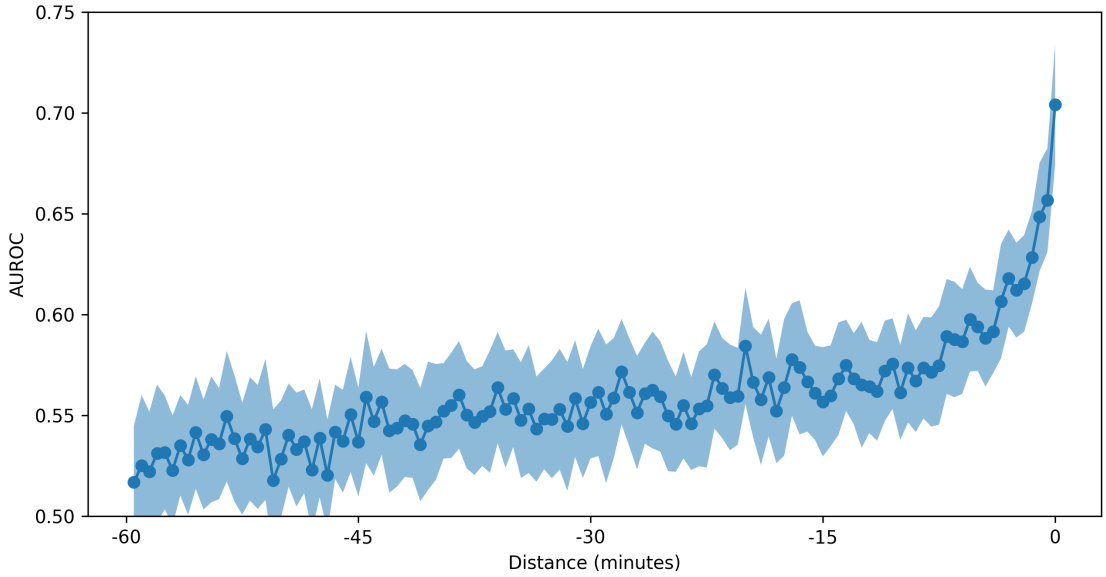


Figure 4.16: Evolution of AUROC performance for balanced binary classification between NSR windows close to AF and NSR windows distant from AF using a gradient boosted tree model. Each point represent the AUROC mean for a binary classification task between NSR windows at the distance d and randomly selected NSR windows distant at 2h of any AF sign.

Using a threshold of 0.5, we observed the same trend for the following metrics:

- accuracy increase from 53.6% (51.8–55.4) to 66.9% (65.0–68.8),
- sensitivity increase from 41.7% (38.5–44.8) to 63.6% (59.3–68.0),
- specificity from 65.5% (63.0–68.0) to 70.1% (66.6–73.6),
- PPV from 54.7% (52.4–57.0) to 68.1% (66.0–70.3),
- NPV from 52.9% (51.4–54.5) to 66.0% (63.6–68.5),
- and F1-score from 47.2% (44.6–49.9) to 65.6% (63.1–68.1).

4.5 Comparison of models for AF onset forecast

From the results in the previous section, we have shown that the sinus rhythm before AF onset contains meaningful information about the oncoming AF onset. The ML model is able to identify this information and is able to distinguish

NSR windows close to AF onset and NSR windows distant from any signs. In the last years, DL model have shown their great capacity to generalize prediction in various tasks, including medical ones (Hannun et al. 2019). In this section, we compare the performance of various ML and DL models using various inputs in the AF onset forecast task.

4.5.1 Materials and methods

Window selection

We selected all the recordings with AF from the IRIDIA-AF v2 database. In those recordings, we selected all the AF onsets with at least 30 minutes of NSR before the AF onset and AF crisis with a duration of at least 5 minutes. From those selected onsets, we selected the 30 minutes of recordings before the AF, as highlighted in yellow in Figure 4.17. For the NSR window selection, if the 3 hours before the AF onset are in sinus rhythm, we selected the 30-minute window between 2 hours and 2 hours 30 minutes before the AF onset, as highlighted in green in Figure 4.17. If the NSR duration before the AF onset is not sufficient, we search for another window in the recording with at least 1 hour of NSR before the window and 2 hours of NSR after the window.

The ECG data from the selected windows can be analysed in 4 different ways. The most straightforward is the analysis of the raw ECG signal. The ECG signal can then be decomposed, and the signal morphology analysed. Next, the series of RR intervals can be analysed. Finally, the HRV parameters can be calculated from the RR intervals and analysed. For each model, the selected 30-minute windows are split into variables sub-window size. This extends the data available for the training, but we should note that windows closer to AF are contains more signs about the incoming AF onset, as discussed in the previous section.

Other data augmentation techniques such as the random concatenation of beats, ECG window squeezing or dilatation are proposed in the literature (Liu et al. 2020). It can increase the diversity in the data set, but it disrupts the order of the beats and the morphology of the signal. We decided to limit ourselves to window slicing enhancement.

For all models, the dataset was divided into 10 folds using temporal cross-validation. The 10 folds were used to evaluate each model, using AUROC and AUPRC. For DL models, a validation set is also used during model training. In

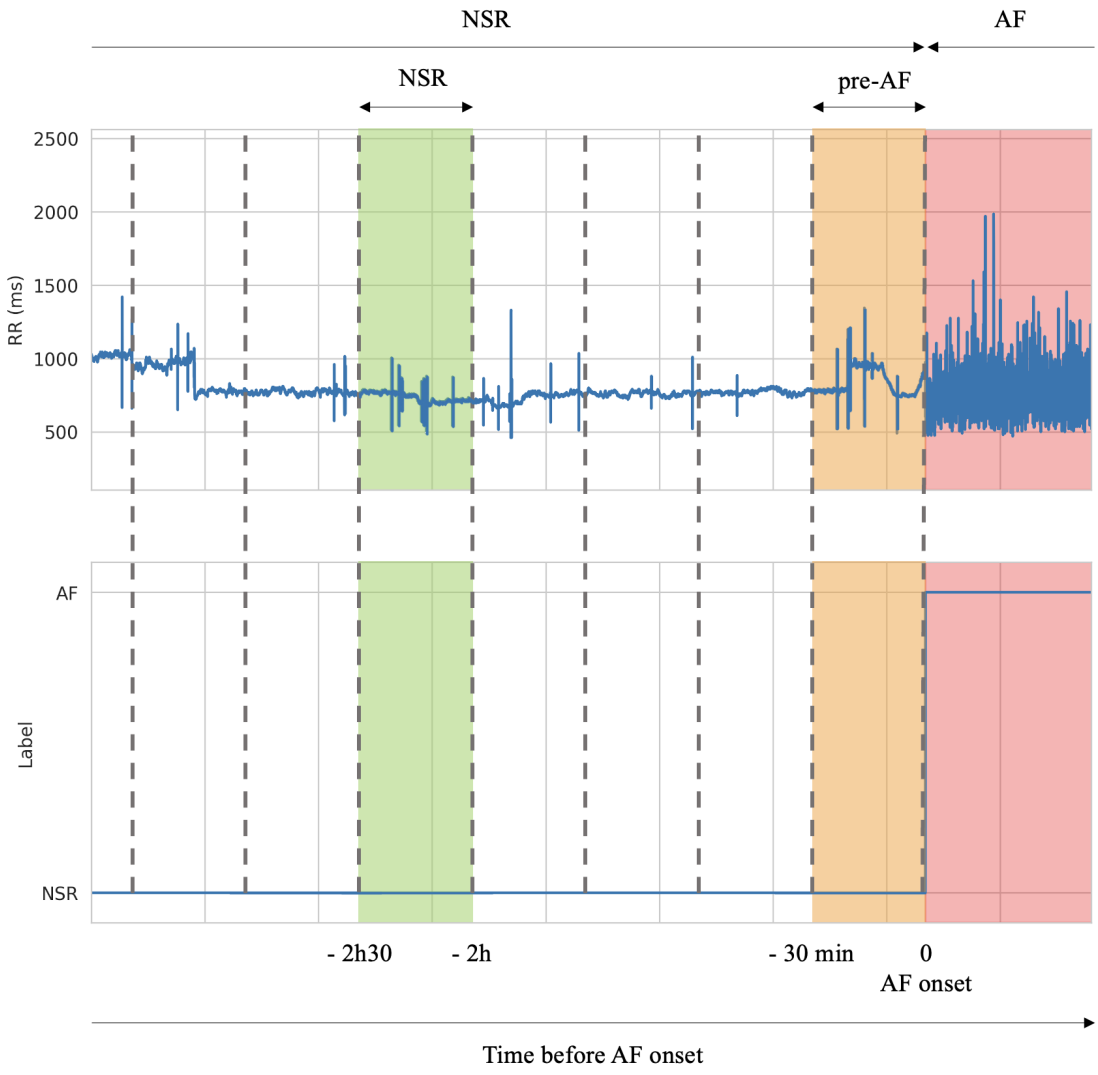


Figure 4.17: Window selection for AF onset forecast. The AF onset is located at 0 on the x-axis and is preceded by NSR and followed by AF (red). Two types of windows are selected for the binary classification task of AF onset forecast: pre-AF windows (yellow), which correspond to the 30-minute window preceding the AF (red), and NSR windows (green), which correspond to 30-minute window distant from at least 2 hours from any AF sign.

this case, another split is used as validation. If a patient has multiple recordings, the date of the first recording is used and all recordings of the patient are in the same split.

Models selection

We want to compare the performance of various ML and DL models using various inputs in the AF onset forecast task. We distinguish 4 classes of models, grouped by their input type. The four input types are: (i) ECG, (ii) ECG morphology, (iii) RR intervals, and (iv) HRV parameters. The ECG and RR intervals are used by Deep Neural Network (DNN) models to create the predictions. The ECG Morphology Variability (ECGMV) features and HRV features are used by ML models to create the features. The relation between input types and their use by the models is presented in Figure 4.18.

ECG model

The most straightforward input is the use of the ECG as is. We choose to use two types of Convolutionnal Neural Network (CNN) models. The first model is a CNN, with 3 blocks, where each block is composed of a convolution layer, a batch normalization layer, and finally a max pooling layer. A ReLU activation function is used between the batch normalization and the max pooling. The final prediction is obtained by fully connected layer and a sigmoid function, with the features from the CNN layers as input. The model is depicted in Figure 4.19.

The second selected model is a deeper CNN with 9 blocks. Each block consists of a convolution layer, followed by a batch normalisation, a second convolution layer and a second batch normalisation. We added a shortcut connection in the blocks to help the network propagate the features during the forward pass and the gradient during the backward pass. This type of ResNet base network is used in the literature for AF risk identification, as in the research from Attia et al. (2019a). The model is described in Figure 4.20.

Baseline wander was removed using a high-pass filter (Kher 2019; Makowski et al. 2021) and power line interference were removed using a notch filter at 50 Hz. We checked the recording to ensure that the ECG was not inverted and, if it was, the trace was inverted along the y-axis. As the full 30-minute window could not be used by the model, we selected 3 sub-window input sizes: 10 seconds, 30 seconds and 60 seconds. The sampling rate of the

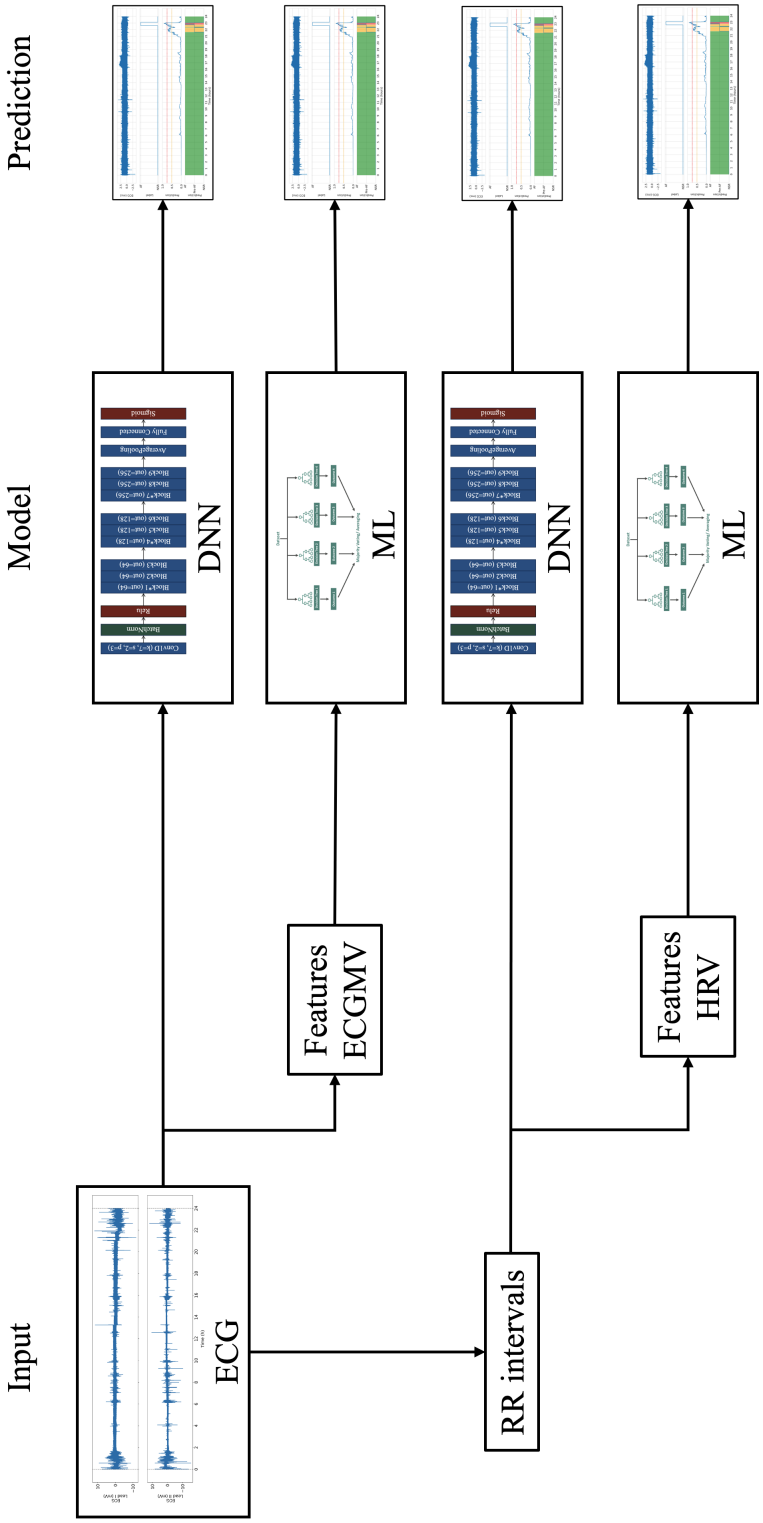


Figure 4.18: Prediction pipelines for the 4 input types. The ECG is the base input type. The ECG is used to create the ECGMV features and the RR intervals. From the RR intervals, the HRV features are computed. ECG and RR intervals are used by DNN models. ECGMV features and HRV features are used by ML models.

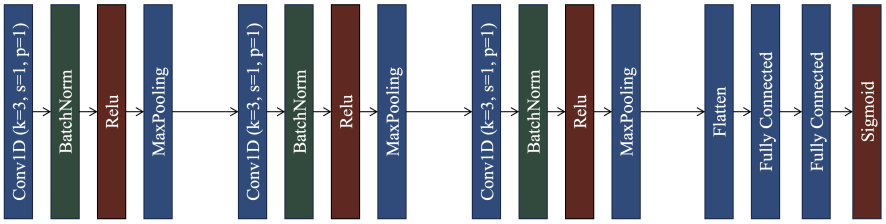
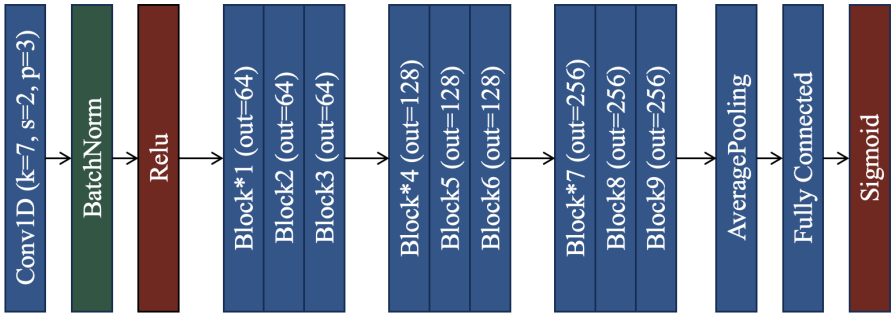
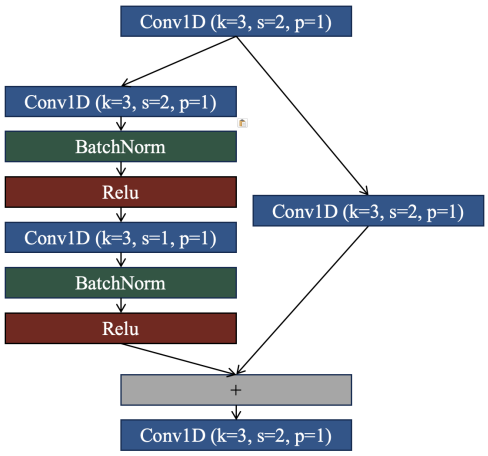


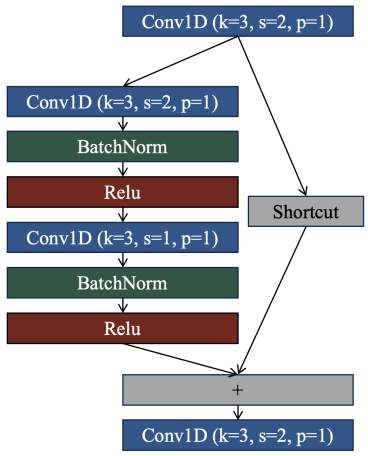
Figure 4.19: Selected architecture of the CNN model



(a) Selected architecture of the ResNet model



(b) Block*



(c) Block

Figure 4.20: Architecture of the CNN model used for AF identification. The model architect (a) is composed of an input block, followed by 9 blocks and a final classification. The 9 blocks are divided into 3 groups, where the first block is a Block* (b) and the next two are Blocks (c).

signal is 200 Hz, therefore the input window size is 2000 for the 10-second window, 6000 for the 30-second window and 12000 for the 60-second window.

For the two models, we use a batch size of 32 and an initial learning rate of 10^{-5}. We used the binary cross entropy loss function and the model weights were optimised using Adam (Kingma et al. 2017). A total of 100 epochs were allowed for each fold, with an early stopping after 5 epochs if the performance of the model did not improve on the validation split. The same validation set is used to reduce the learning rate after 2 epochs if no improvement is observed. The model is built in PyTorch (Paszke et al. 2019).

ECG morphology based model

As a second type of input, ECGMV was used to analyse the recordings. QRS complexes are detected and individual beats are analysed to find the P wave onset, peak and offset, Q, R and S peaks, R onset and R offset, S peak and finally T onset, peak and offset.

Using these points of interest in the signal, we computed multiple features and intervals as follows. For the P wave, QRS complex, R wave and T wave, we calculated the duration, amplitude and area. The QT, TQ and RR intervals were calculated. Corrected intervals were calculated using the RR interval duration. The corrected intervals are QTc and TQc. The corrected P-wave duration was also calculated. Using the 18 features for all beats, we calculated the mean, median, first quartile, third quartile and standard deviation. Finally, we also included the heart rate, PAC count and PAC percentage in each window. In total, 92 features were computed for each window. Additionally, we studied the pairwise correlations between these 92 features using the Spearman correlation coefficient. We found that some features were highly correlation, with a coefficient above 0.9. The complete correlation matrix is available in Appendix A in Figure A.12.

Two ML models were trained and tested: RF (Breiman 2001) and XGBoost (XGB) (Chen et al. 2016). Like RF, XGB uses ensemble decision trees, but with boosted decision trees. This is an ensemble learning technique in which decision trees are sequentially added to correct the errors of previous models to achieve high predictive accuracy. For both models, we set the number of trees to 200 and the maximum depth of the trees to 10. As we found that some features were highly correlated, both models were trained on the full dataset and on a selected dataset in which one feature from each highly correlated pair

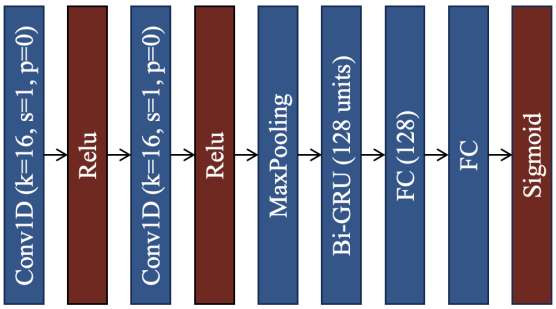


Figure 4.21: Architecture of the CNN-RNN model used for AF identification

was removed from the dataset. The high correlation was defined as an absolute correlation value above 0.9. The feature selection was computed on the training set and used to select the features available in the test set.

RR intervals model

Using the detected QRS complexes we construct the RR interval series. This series can be used as input for a DL model. We trained 3 models using an input window size of 300 RR. The first two models are CNN with a similar architecture to those used for the ECG signal: a simple CNN and a ResNet-based model. We also tested the performance of a CNN-RNN used in our previous work (Gilon et al. 2020). The three models were implemented using PyTorch.

We used a batch size of 32 and an initial learning rate of 10^{-5}. We used the binary cross entropy loss function and Adam as the optimisation algorithm (Kingma et al. 2017). A total of 100 epochs were allowed for each fold, with an early termination after 5 epochs if the performance of the model did not improve on the validation split. If no improvement was observed after 2 successive epochs, the learning rate was divided by 2.

The CNN-RNN model is presented in Figure 4.21. The first two convolutional layers are used to create features. ReLU layers are used after each convolution. Next, a GRU receives its inputs from the CNN. We used a bidirectional layer to create features in both directions in the time series: from start to end and from end to start. The features are flattened and the final prediction is computed using a fully connected layer and a sigmoid.

RR intervals recurrence plot model

Gavidia et al. (2023) proposed a new approach using recurrence plots constructed from RR intervals, described in Section 2.2.5. We used a methodology similar to the one proposed in the publication. We used a 60-second window to find the QRS complex and generate the RR interval series, with a 30 seconds overlap between windows. From the RR intervals, we created the recurrence plot using the following parameters: an embedding size of 3 and a lag of 2. The recurrence plot was expanded to an image size of 224×224.

We used a CNN model to classify the images. We reproduce the EfficientNet based model as proposed by Tan et al. (2021). This model is based on using uniform scaling of all widths and depths of the convolution.

In the methodology described by Gavidia et al. (2023), a CNN model was used for the classification of recurrence plots. Specifically, we reproduced the model based on EfficientNet. This architecture has been proposed as an efficient alternative to other CNN architectures. In this case the width, depth and resolution of the network are scaled uniformly. This approach allows the model to achieve better performance while being computationally efficient compared to other architectures. In the original paper by Tan et al. (2021), they achieved better performance on a benchmark database while using a smaller architecture and therefore faster training time.

Compared to the other DL models used in this work, this model uses a 2D input of size 224×224, rather than a 1D input. As with the model using raw RR intervals, a total of 100 epochs were used for each fold, with an early stopping after 5 epochs if the performance of the model did not improve on the validation split. The same validation performance is used to reduce the learning rate after 2 epochs if no improvement is observed. The batch size was 32. The model is implemented using Keras (Chollet et al. 2015) and TensorFlow (Abadi et al. 2016).

HRV models

The final class of models are those that use HRV features. We use features from the time domain, geometric plots and the frequency domain as defined in Section 4.4.1. As for ECGMV features, we studied the pairwise correlations between these features using the Spearman correlation coefficient. We found that some features were highly correlation, with a coefficient above 0.9. The complete correlation matrix is available in Appendix A in Figure A.11.

We compared two ML ensemble decision tree models: Random Forest and XGBoost. The models were using 200 trees with a maximal depth of 10. We compared 4 sub-window sizes: 60 seconds as suggested by Levasseur et al. (2022), 85 seconds as suggested by Kisohara et al. (2020), 5 minutes and the complete 30 minutes. As for the ECGMV features, we found that some features were highly correlated. We applied the same method for both models using HRV. The models were trained on the full dataset and on a selected dataset in which one feature from each highly correlated pair was removed from the dataset. The high correlation was defined as an absolute correlation value above 0.9. The feature selection was computed on the training set and used to select the features available in the test set.

4.5.2 Results

From the 623 recordings with annotated AF onset, we selected the 900 AF onsets with at least 5 minutes of AF and 30 minutes of NSR before AF onset in the same file. The remaining 64 AF onsets with at least 5 minutes of AF and at least 30 minutes of NSR before AF onset were excluded because the NSR window overlapped on two recording days and therefore included the calibration phase. Of these, 538 AF onsets had 3 hours of NSR before AF onset, and 301 NSR windows were selected elsewhere in the recordings. For 61 recordings, we could not find an NSR window that met the criteria.

The benchmark is presented in Table 4.9 for the full-window models, Table 4.10 for the window-level evaluation and Table 4.11 for the record-level evaluation. The window-level evaluation compares the model to individual sub-windows of the 30-minute window. The record level evaluation averages all predictions from all sub-windows into a single evaluation for each recording. The threshold-based metrics for the two models with the best average metrics are shown in Tables A.5 and A.6.

Overall, the best average results were obtained by the XGB model using HRV parameters from the full 30-minute windows, with an AUROC of 0.643 and an AUPRC of 0.634. It should be noted that the confidence intervals with the RF model using HRV overlap to a large extent. We also found that using a selected feature set in which the correlation between features was used did not improve the results obtained by the models.

Using a threshold of 0.5, the predictions across all folds have an accuracy of 0.608, a sensitivity of 0.706, a specificity of 0.504, a Positive Predictive Value

(PPV) of 0.604, a Negative Predictive Value (NPV) of 0.615 and an F1 score of 0.651. If we choose a threshold of 0.3, the sensitivity increases to 0.933 but at the cost of a lower specificity of 0.175. The results with different thresholds between 0 and 1 are shown in the Table A.4 of the Appendix A. In the context of clinical practice, thresholds should be chosen individually for each patient based on their previous recordings. However, further testing of the validity of the threshold should be performed as the threshold should be chosen using a validation set to understand if the previous recordings are sufficient to allow the correct choice of threshold.

When analysing the results of the models using smaller windows, the RF using ECGMV achieved the best average AUROC performance of 0.579 and an AUPRC of 0.577 when evaluated at the window level. The confidence intervals were computed using the test metrics for each test set over the 10 folds. The best average performance was achieved by the CNN using RR intervals when comparing AUPRC, as the CNN achieved 0.585. When evaluated at the recording level, the best average performance for both AUROC and AUPRC was achieved by the CNN model, with an AUROC of 0.604 and an AUPRC of 0.595. The confidence intervals of the results overlap to a large extent.

To understand whether the models using the full 30-minute window were statistically significantly better, we compute the pairwise p-value for both AUROC and AUPRC using Mann-Whitney U test. The p-values were computed between these four models and the CNN, which achieved the best average in the benchmark. The results are shown in Table 4.12. We found that except for the AUPRC results between the XGB model using the HRV features from the 30-minute window and the CNN using 300 RR intervals, all the p-values were above 0.05 and therefore the difference in AUROC and AUPRC was not statistically significant.

4.5.3 Predictions analysis

Following these results, we studied the evolution of the prediction over a complete 24-hour Holter monitoring. For each patient, we retrained a new XGB model on the entire database, except for the patient which is used for testing. The model used 200 trees with a maximum depth of 10. As Holter monitoring consists of AF and NSR periods, the model was trained on labels composed of NSR windows and AF windows. The pre-AF windows were assigned the same label as the AF windows.

Table 4.9: Model performance benchmark for 10-fold cross-validation with a balanced dataset of 30-minute windows. The models use the full 30 minutes. For each model, we tested two types of input dataset: (i) a full dataset containing all the features and (ii) a selected dataset in which one features from each highly correlated feature pair was removed.

Input	Model	Input size	Features	AUROC	AUPRC
HRV	RF	30 min	all	0.640 (0.608-0.671)	0.631 (0.599-0.662)
	RF	30 min	selected	0.627 (0.595-0.659)	0.615 (0.579-0.652)
	XGB	30 min	all	**0.643 (0.609-0.677)**	**0.634 (0.594-0.674)**
	XGB	30 min	selected	0.617 (0.589-0.645)	0.610 (0.575-0.646)
ECGMV	RF	30 min	all	0.627 (0.591-0.663)	0.600 (0.565-0.635)
	RF	30 min	selected	0.616 (0.576-0.656)	0.603 (0.559-0.646)
	XGB	30 min	all	0.625 (0.591-0.659)	0.603 (0.562-0.643)
	XGB	30 min	selected	0.621 (0.584-0.659)	0.606 (0.565-0.647)

For training, we selected a 60-minute NSR window, 15-minute pre-AF windows and a 10-minute AF window from the onset of the AF crisis, with at least 60 minutes of NSR before the onset of AF and at least 10 minutes of AF. The problem is formulated as a binary classification, with the NSR window as the negative class and the pre-AF and AF windows as the positive class. The XGB model using HRV was chosen as it performed best with the full windows in the previous section and is faster to train than the CNN model, as a new model has to be trained for each patient.

Model predictions are made using a sliding window. This is done by moving a window of fixed size across the entire recording to average the performance of the model over each window of a recording. This is equivalent to a sequential testing on the entire recording. The predictions were then evaluated visually, as shown in Figure 4.22. We found that for some recordings the model reacts as expected: for the majority of the recording the prediction value is low, then before the onset of AF the prediction values increase up to the arbitrary selected threshold of 0.5. During the AF crisis, the predicted values increase again and exceed the second arbitrarily chosen threshold of 0.75. As for the benchmark, the choice of threshold will be most important if the model is to be used in a clinical setting. This threshold should be chosen from the previous recording of the patient to determine the best value based on the medical decision to aim for a high detection rate of episodes and a high risk of false alarms or a lower detection rate and a lower risk of false alarms but an increased risk

Table 4.10: Model performance comparison for 10-fold cross-validation with a balanced data set of 30-minute windows. Models are evaluated at the window level.

Input	Model	Input size	AUROC	AUPRC
		60 s	0.567 (0.535-0.598)	0.571 (0.539-0.603)
HRV	RF	85 s	0.572 (0.540-0.604)	0.575 (0.543-0.607)
		5 min	0.570 (0.535-0.605)	0.573 (0.545-0.602)
		60 s	0.558 (0.527-0.590)	0.567 (0.533-0.601)
HRV	XGB	85 s	0.563 (0.531-0.595)	0.569 (0.536-0.603)
		5 min	0.559 (0.527-0.591)	0.567 (0.540-0.595)
	CNN-RNN	300 RR	0.563 (0.540-0.586)	0.577 (0.555-0.599)
RR	CNN	300 RR	0.574 (0.553-0.595)	**0.585 (0.559-0.611)**
	ResNet	300 RR	0.545 (0.526-0.563)	0.567 (0.543-0.592)
RR RP	EfficientNet	60 s	0.553 (0.533-0.574)	0.565 (0.543-0.587)
		60 s	0.567 (0.539-0.594)	0.572 (0.538-0.606)
ECGMV	RF	85 s	0.573 (0.546-0.599)	0.575 (0.544-0.607)
		5 min	**0.579 (0.552-0.606)**	0.577 (0.543-0.611)
		60 s	0.565 (0.533-0.596)	0.569 (0.529-0.608)
ECGMV	XGB	85 s	0.560 (0.531-0.589)	0.569 (0.531-0.606)
		5 min	0.565 (0.537-0.592)	0.566 (0.533-0.599)
		10 s	0.541 (0.527-0.555)	0.545 (0.533-0.557)
ECG	CNN	30 s	0.534 (0.521-0.547)	0.540 (0.527-0.552)
		60 s	0.528 (0.516-0.539)	0.534 (0.523-0.545)
		10 s	0.522 (0.504-0.540)	0.533 (0.517-0.550)
ECG	ResNet	30 s	0.543 (0.516-0.570)	0.554 (0.520-0.589)
		60 s	0.541 (0.513-0.569)	0.551 (0.512-0.590)

Table 4.11: Model performance comparison for 10-fold cross-validation using a balanced dataset of 30-minute windows. Models are evaluated at the episode level using the mean prediction aggregated across all windows.

Input	Model	Input size	AUROC	AUPRC
HRV	RF	60 s	0.591 (0.548-0.635)	0.589 (0.547-0.631)
		85 s	0.597 (0.555-0.639)	0.593 (0.551-0.635)
		5 min	0.585 (0.543-0.628)	0.581 (0.543-0.619)
HRV	XGB	60 s	0.589 (0.543-0.635)	0.593 (0.546-0.640)
		85 s	0.594 (0.548-0.640)	0.593 (0.546-0.640)
		5 min	0.580 (0.538-0.622)	0.578 (0.541-0.615)
RR	CNN-RNN	300 RR	0.597 (0.559-0.636)	0.592 (0.559-0.626)
	CNN	300 RR	**0.604 (0.582-0.625)**	**0.595 (0.569-0.622)**
	ResNet	300 RR	0.591 (0.553-0.629)	0.593 (0.557-0.628)
RR RP	EfficientNet	60 s	0.588 (0.559-0.618)	0.589 (0.560-0.617)
ECGMV	RF	60 s	0.583 (0.547-0.618)	0.584 (0.545-0.623)
		85 s	0.590 (0.557-0.623)	0.587 (0.551-0.623)
		5 min	0.589 (0.557-0.621)	0.585 (0.547-0.623)
ECGMV	XGB	60 s	0.583 (0.542-0.624)	0.585 (0.538-0.631)
		85 s	0.577 (0.538-0.616)	0.581 (0.538-0.623)
		5 min	0.577 (0.544-0.610)	0.576 (0.541-0.611)
ECG	CNN	10 s	0.563 (0.540-0.586)	0.565 (0.547-0.584)
		30 s	0.558 (0.535-0.581)	0.565 (0.548-0.582)
		60 s	0.543 (0.522-0.564)	0.551 (0.536-0.566)
ECG	ResNet	10 s	0.529 (0.508-0.549)	0.539 (0.518-0.559)
		30 s	0.550 (0.521-0.579)	0.561 (0.523-0.598)
		60 s	0.546 (0.515-0.577)	0.554 (0.515-0.594)

Model A			Model B			p-value	
Input	Model	Size	Input	Model	Size	AUROC	AUPRC
HRV	RF	30 min	HRV	XGB	30 min	0.791	0.677
HRV	RF	30 min	ECGMV	RF	30 min	0.520	0.185
HRV	XGB	30 min	ECGMV	XGB	30 min	0.344	0.384
HRV	RF	30 min	RR	CNN	300 RR	0.075	0.088
HRV	XGB	30 min	RR	CNN	300 RR	0.064	0.045

Table 4.12: P-value between models using the full 30-minute window and the CNN which obtained the best average AUROC and AUPRC in the model comparison benchmark.

of missing the first AF.

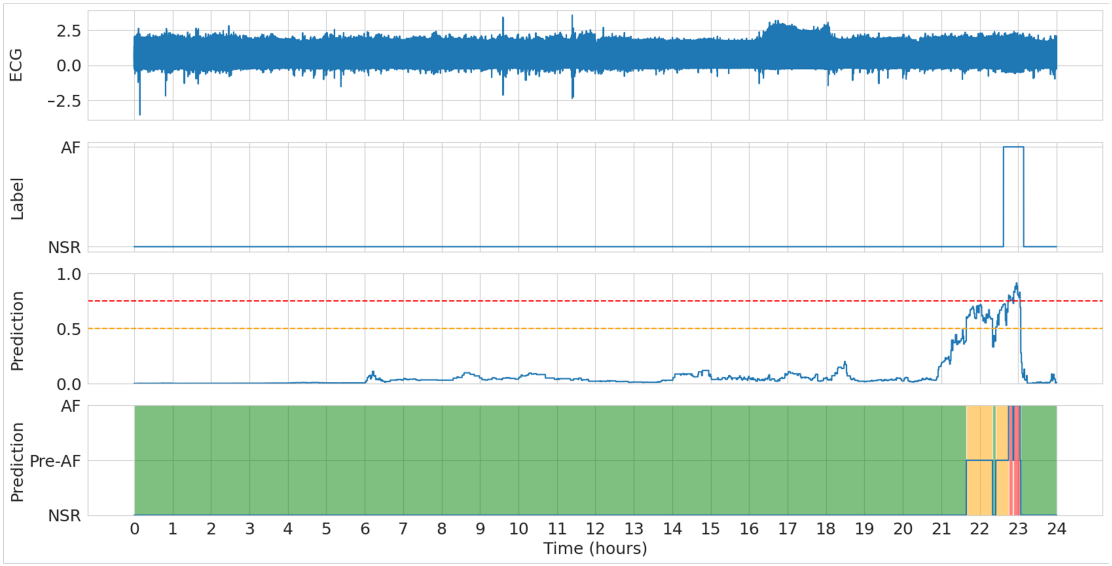
It should be noted that this was not the case for all recordings. For some recordings the predictions contained primarily false alarms, while for others the predicted values did not respond at all during the pre-AF windows. We found no differences in the signals examined that could visually explain the differences in prediction.

It should also be noted that the choice of individual thresholds is central to the performance of the model when assessing performance using the threshold base metrics. In practice, these thresholds should be chosen in consultation with medical professionals to enable the best therapeutic strategy for the patient. In addition, machine learning based methods could guide the choice of these thresholds by assessing the level of risk for a patient. The analysis of model performance using different thresholds should be further analysed in future work.

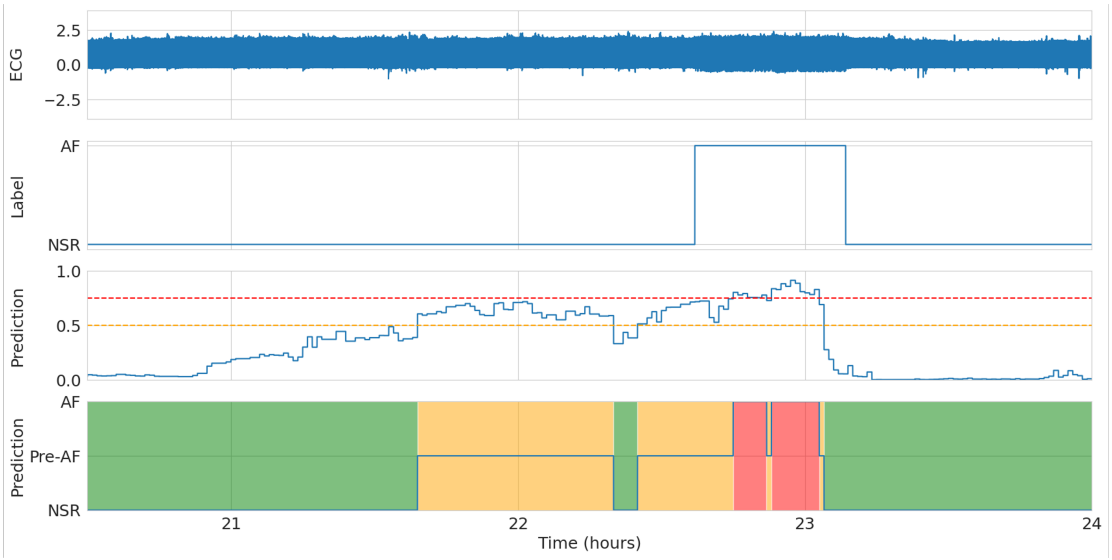
4.6 Discussion

In this study, we found that an ML model is able to detect changes in the sinus rhythm ECG before the onset of AF. We first examined the evolution of ECG and HRV before the onset of AF. We found that both the ECG and HRV varied significantly before the onset of AF. Some parameters increase and others decrease, but changes are observed in most parameters.

We then showed that the closer the prediction is to the onset of AF, the better the prediction performance of the ML model. This correlates with the results obtained for ECG and HRV evolution, as predictions are better closer to



(a) Prediction on the complete 24-hour recording using XGB model with HRV features computed on a 5-minute sliding window



(b) Focus on the AF crisis

Figure 4.22: AF onset forecast prediction on the recording 589 from IRIDIA-AF v2. The model use a 5-minute sliding window and the prediction are concatenated over the whole recording. For (a) and (b), the first row correspond to the raw ECG of the patient, i.e. the input. The second row correspond to the raw prediction of the model, i.e. the concatenation prediction of all windows. The third row corresponds to the decision based on the two threshold: green means (NSR), yellow is the first warning (pre-AF), red is the second warning (AF).

AF onset.

In the benchmarking process to compare the performance of the models, we first compare the performance of two ML models using HRV features and two ML models using ECGMV. The HRV features and ECGMV features were computed on the full 30-minute window. The best performance was achieved by the XGB model using HRV features with an AUROC of 0.640 (0.608-0.671) and an AUPRC of 0.631 (0.599-0.662). The results of this model were almost identical to the performance of the RF model. Both models with HRV features performed better than the two models with ECGMV features.

The features ranking for both the random forest and XGBoost models are compared for each of the 10-fold and then ranked across the folds. For the RF model, the feature importance is computed using the average decrease in impurity resulting from splitting a node using that feature across all decision trees in the forest (Breiman et al. 1984). For the XGBoost model, feature importance is computed using a gain, which corresponds to the sum of the weighted feature contributions from each node split in all decision trees in the forest (Chen et al. 2016). While both impurity reduction and gain are measures of improvement in model performance, they are computed differently. Impurity reduction focuses more on the purity of subsets within decision trees, while gain focuses more on the overall improvement in prediction accuracy across the ensemble of trees in boosting algorithms. However, in both cases, lower impurity and higher gain indicate better model performance.

The 5 features achieving the highest mean rank across the 10 folds are shown in Figure 4.23 and the box plots are shown in Appendix A in Figure A.13. For both models, the top 3 features are the CVSD, the Central Tendency Measure $(CTM)_{100}$ in the Second Order Difference Plot (SODP) and the CVNN. The CVSD corresponds to the short-term variability and the CVNN to the long-term variability. For the pre-AF windows, the mean RMSSD is higher and the mean RR is lower, hence the CVSD is higher. The same is true for the CVNN, as the SDNN is higher for the pre-AF windows. In both cases, the pre-AF windows have higher variability and a lower mean RR, which means that the heart rate is increasing towards AF. The CVNN was previously used as a single feature for AF detection by (Tateno et al. 2001) and is used by Gavidia et al. (2023) as a threshold for the selection of pre-AF windows.

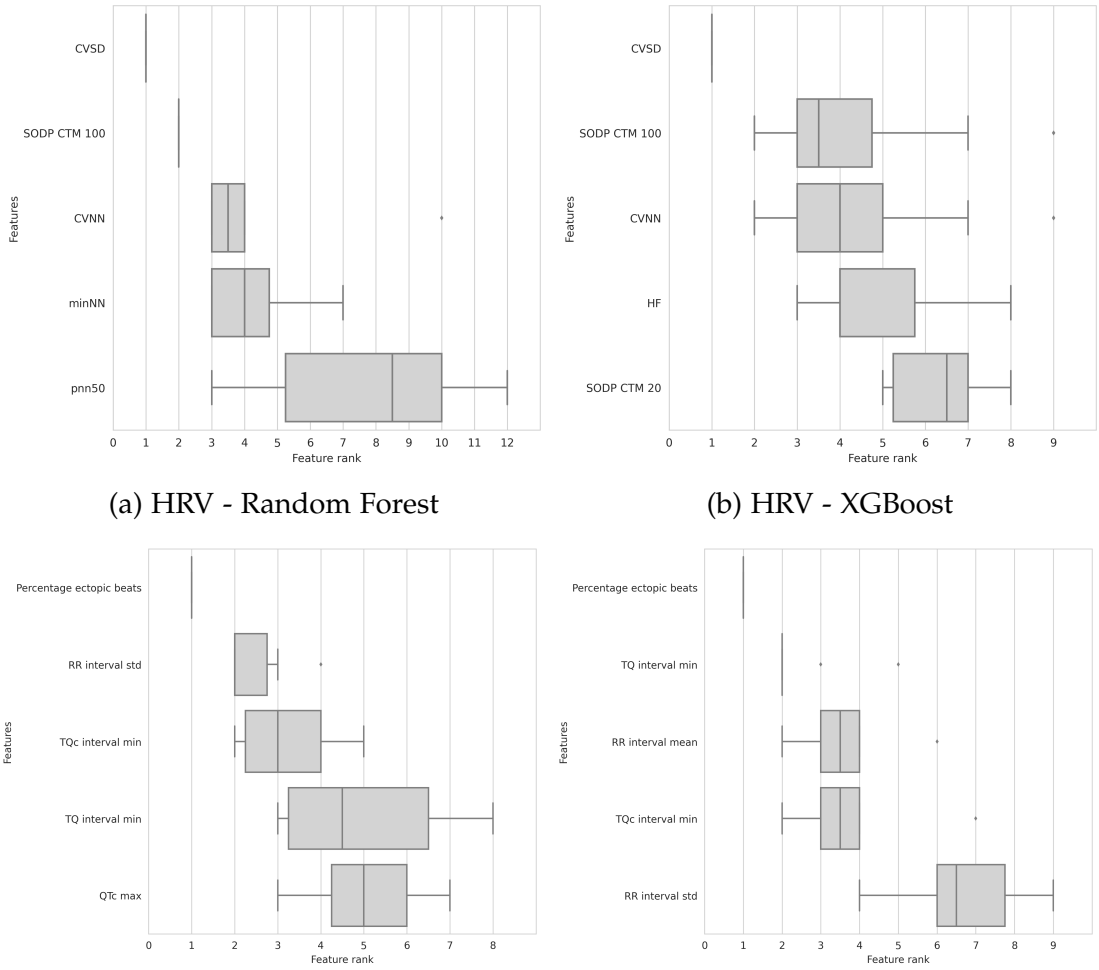
The CTM_{100} is lower for pre-AF windows, further supporting the idea that short-term variability is higher during pre-AF windows. The next two features are different between the two models: the RF use the minimum RR and the

pnn_{50} the most, and the XGB use the HF and the SODP CTM_{20}. An increase in HF activity has been shown in Figure 4.14b and corresponds to an increase in PNS activity. SODP CTM_{20} and pnn_{50} also indicate an increase in short-term variability. Finally, the minimum RR value is lower for the pre-AF windows, which correlate with the increase in Heart Rate (HR).

The top 5 ECGMV features across the 10-folds for the RF and XGB are shown in Figure 4.23 and box plots are shown in Appendix A in Figure A.14. For both models, the percentage of ectopic beats is the most meaningful feature over the 10-fold range. Both models use the standard deviation of the RR, i.e. the SDNN in HRV analysis, which correlates with the CVNN as discussed for the previous models. The minimum values of TQ interval and TQ_c interval are both lower for pre-AF windows. The TQ interval encapsulates the P wave, which has been shown to increase in duration as AF approaches. The decrease in TQ may therefore be more related to the heartbeats being closer together and therefore the increase in HR, rather than the change in P wave. In addition, we measured the changes in the T-P segment from the end of the T wave to the beginning of the P wave and found that the duration of the segment decreases before the onset of AF, as shown in Figure 4.24. Comparing the distribution, we found that the T-P segments are shorter for pre-AF windows compared to NSR windows.

After evaluating the four models using features from the full 30-minute window, we evaluate the models using shorter windows. It should be noted that this benchmark has been constructed with an emphasis on testing a larger number of models and inputs more broadly, rather than deep optimisation and hyper-parameter search for a single specific model.

Interestingly, models based on RR and HRV outperformed those using ECG and ECGMV. Furthermore, DL models using the raw ECG as an input achieved the lowest overall scores. Using the same CNN architecture with RR or ECG as input, a 4% difference in AUROC and a 3% difference in AUPRC were observed. The initial experiments used ML models with HRV features as the first baseline, and it was expected that using the full ECG would improve performance. However, this hypothesis proved to be incorrect. In addition, the P-wave features were in the lower part of the ranking when analysing the full ECG-MV ranking. This also suggests that the meaningful information is contained in short-term and long-term variability, such as RR intervals and HRV, rather than in P-wave variability. Another hypothesis is that detection of the QRS complex is more robust than detection of other waves on a Holter. The



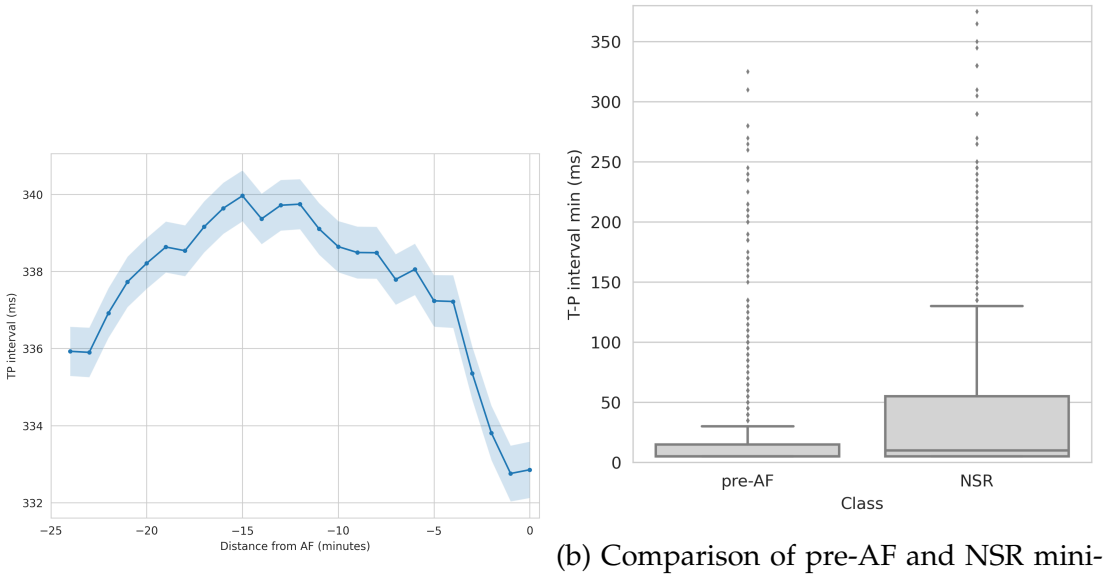
(a) HRV - Random Forest

(b) HRV - XGBoost

(c) ECGMV - Random Forest

(d) ECGMV - XGBoost

Figure 4.23: Comparison of top ranked features. HRV features used by Random Forest (a) and XBGoost (b) and ECGMV features used by Random Forest (c) and XGBoost (d)



(a) Evolution of T-P interval

(b) Comparison of pre-AF and NSR minimum T-P interval

Figure 4.24: T-P interval before AF onset

recording lasts for 24 hours and, despite the cleaning of the signal, it still contains a lot of noise. QRS complexes are therefore more representative of the heart rhythm than the full ECG. Further research into reducing the noise in Holter monitorings could improve the performance of DNN models.

In the context of medical applications, the choice of threshold values is not straightforward when moving from area-based metrics to threshold-based metrics and making clinical decisions. A lower sensitivity may miss episodes and a lower specificity AF may give false alarms. In this case, we could aim for higher sensitivity to predict all episodes, but with a required continuous time above the threshold before triggering an alarm to limit false alarms. However, the current level of performance is not yet sufficient for practical use in clinical practice.

These results also allow us to consider the direction of application of these models. Indeed, if the models that perform best are also those that require the least resources, it is possible to consider implementing them in the cardiac implantable electrical devices. The development of the database and the evaluation of different models and inputs were necessary to reach this conclusion. The input requirements are also lower: if the full ECG is not needed, PhotoPlethysmoGraphy (PPG) could be sufficient to capture the changes in

variability, as proposed by Guo et al. (2021) and Gruwez et al. (2023b). In this work, the authors were able to identify an AF signature in sinus rhythm using the PPG signal.

Studies have shown that flecainide can be used as an antiarrhythmic drug for recovery from AF to sinus rhythm (Echt et al. 2020). Pill-In-The-Pocket (PITP) strategies have been studied as a solution for AF (Reiffel et al. 2023). If we can achieve performance as in the predictive evaluation in Figure 4.22, where the model starts to react 1 to 2 hours before AF onset, the use of PITP for AF management is achievable. Further research is needed to address the variations in problem definition, such as binary versus ternary classification and class balanced versus unbalanced settings, as these may more closely reflect real-world AF scenarios.

4.7 Summary

In this chapter, we investigated the performance of ML models for forecasting AF onset. First, we reproduced the three methodologies described in the literature to understand if the results could be reproduced. We were not able to reproduce the results obtained with the described methodology and variations.

Then, using the new database created and presented in the Chapter 3, we studied the evolution of ECG waves, ECG intervals and HRV parameters before the onset of AF. We found significant variations in several parameters of both ECG and HRV before the onset of AF, reinforcing the fact that changes in the ECG and heart rhythm occur before the onset of AF. Building on these results, we investigated the evolution of the predictive performance of an ML model when comparing NSR windows close to AF and NSR windows far from AF. We varied the distance between the pre-AF and AF onset windows and found that performance increased when the pre-AF windows were closer to AF onset, with the highest scores of 0.714 AUROC and 0.697 AUPRC.

We compared the performance of different ML and DL models using different inputs such as ECG, ECGMV, RR and HRV. We found that models using RR and HRV performed better than models using ECG. This highlights the fact that most of the information for predicting the onset of AF is contained in the variability of the heart rate. This could also be related to the fact that even if we increase the size of our database, there is still not enough data available for DL training, or that this data, although selected, still contains too much noise to

be used in its raw state. The best models were the XGBoost models using HRV parameters with an AUROC of 0.643 and an AUPRC of 0.634. We analysed the key features used by the model and found that short-term variability features were the most important.

Finally, we tested the performance of this model on 24-hour recordings. We visually analysed the prediction and found that for some recordings the predictions were perfectly accurate, i.e. predicting NSR for most of the recording, starting to increase the risk score 1 hour before AF onset and correctly predicting AF. We also found that the models under-predicted and over-predicted multiple AF events, with recordings containing AF alarms for the entire recording and recordings where the model did not detect any AF events. These results correlate with our benchmark results comparing the four classes of models. It supports the view that these models are not yet ready for practical clinical use in predicting the onset of AF.

Chapter 5

Paroxysmal atrial fibrillation risk identification during sinus rhythm

5.1 Introduction

Atrial Fibrillation (AF) patients have an increased risk of ischaemic stroke. Early detection enables the disease to be treated with anticoagulant therapy or rhythm control therapy (Svennberg et al. 2021; Kirchhof et al. 2023). Risk scores such as CHARGE-AF (Alonso et al. 2013) or the $HARMS_2$-AF (Segan et al. 2023) are used for AF screening, i.e. the prediction of the individual risk of AF in the global population. Other related scores, as the CHA_2DS_2-VASc (Lip et al. 2010) are used to predict the individual risk of stroke.

In medical practice, a high risk score is not sufficient to establish the diagnosis of AF. This diagnosis requires Electrocardiogram (ECG) documentation, which includes either a standard 12-lead ECG or a 30-second single-lead ECG with no detectable P waves and irregular RR intervals (Hindricks et al. 2021). For AF screening, long-term ECG, such as Holter monitoring, is used to record heartbeats over a longer period of 24 hours to a week. This is because longer recordings are more likely to capture the required 30-second AF crisis sample.

Although no clinical application has yet been developed, recent research results suggest that Machine Learning (ML) can reliably identity the presence of an AF electrocardiographic signature in Normal Sinus Rhythm (NSR) ECG (Attia et al. 2019a). This identification is made weeks before the first recorded AF event, without using the medical parameters of the patient. The results of studies, presented in the Chapter 2, are obtained from Deep Learning (DL) models analysing 12-lead 10-second ECG. However, even if we can better identify pa-

tients at risk, preventive treatments other than improved screening, lifestyle changes and dietary measures are not obvious because we do not know how many days, weeks or months it will take for AF to develop. Identifying short-term risk, within a window of less than a day, could make it possible to develop predictive models for paroxysmal AF. An efficient model for short-term prediction of AF may prove to be an effective prevention tool for initiating preventive measures. This may allow screening methods to be optimised, thereby reducing the workload of cardiologists. These preventive measures may also include changes in the patient lifestyle and, with a longer-term vision, open up new avenues of treatment research, such as the use of a Pill-In-The-Pocket (PITP) strategy (Reiffel et al. 2023).

In this chapter, we investigate the performance of ML models and DL models to identify the risk of AF. For this, we evaluate the model performance to identify the AF signature in the first hour of NSR from 24-hour ECG Holter monitorings, where AF can be found in the next hours of the recording. We compare these recordings with recordings from NSR patients, i.e. patients with no signs of AF or other cardiac diseases in the recording.

5.2 Materials and method

The state-of-the-art methods presented in the literature use DL models trained to detect AF in 12-lead ECG. Unfortunately, the different databases used are not publicly available. Therefore, we constructed a new database of Holter monitorings composed of records from AF patients and NSR patients. This database, IRIDIA-AF v2, is presented in depth in Chapter 3. IRIDIA-AF v2 is the results of a retrospective multicentre study. The database is composed of recordings from 4 centres: 3 hospitals and 1 outpatient clinic. The 4 centres used 2-lead Microport Spiderview digital recorders, with a recording sample rate of 200 Hz. The recordings last between 24 and 96 hours.

We reviewed all available recordings and selected recordings according to the following inclusion criteria. For patients with AF, the patient should be 35 years or older. In addition, the selected recordings should contain at least one crisis of paroxysmal AF. We excluded Cardiac Implantable Electronic Device (CIED) patients and recordings with persistent and permanent AF or other heart diseases. Automated analysis was performed using Microport Synescope version 3.30a software. Automatic correction of the recording was performed to

label the ECG complexes and to eliminate artefacts, under and over detection of complexes by the Synescope software. For NSR patients, we selected all patients in the database.

All selected AF recordings were analysed to determine the exact beginning and end of each paroxysmal AF episode and the corresponding complexes. This allows a precise analysis of the transition from NSR to AF episodes. Each recording contains both NSR and one or more AF episodes. We selected patients with AF episodes lasting more than 30 seconds and recordings without evidence of AF or other heart disease from patients without CIED. These patients and their recordings are referred to as NSR.

5.2.1 Recordings selection

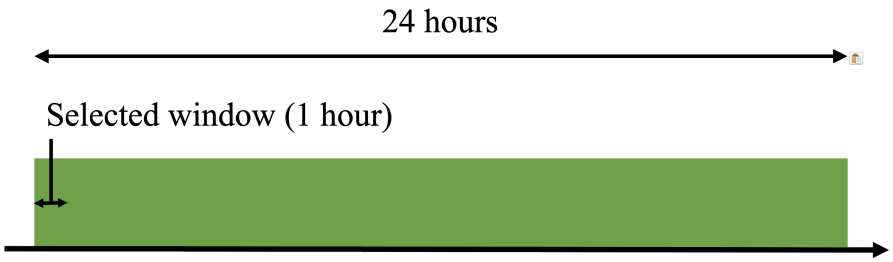
From the recordings of AF patients in the database, we select the first hour for those with no evidence of AF in the first two hours of the recordings and presence of AF in the next 22 hours. The first 10 minutes were skipped to remove any noise caused by the start of the recording. For recordings spanning more than one day, the second and subsequent days were only selected if the last hour of the previous day did not contain AF and the first two hours did not contain signs of AF.

For NSR recordings, we select the first hour of the recording. To increase the data available for model training and to increase the variability of the data of healthy patients, we include the 12th hour, which is the hour in the middle of the recording. The Figure 5.1 summarises the selection of windows for both AF and NSR patients.

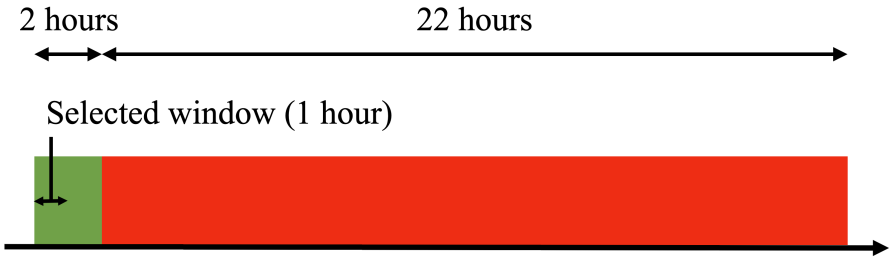
5.2.2 Models comparison using temporal cross validation

ECG-based models

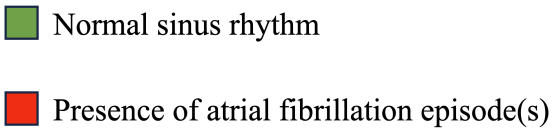
We compared the performance of several models using cross-validation. The first model is a Deep Neural Network (DNN), we implemented a 1-dimensional ResNet Convolutionnal Neural Network (CNN) using PyTorch (Paszke et al. 2019). This kind of model is used in several state-of-the-art studies and was first proposed in the study of Attia et al. (2019a). We chose to use lead I to correspond to wearable applications where only 1 lead is measured and available to the model. We tested several input sizes and evaluated the performance



(a) Healthy subject, i.e. patient without atrial fibrillation in the recording



(b) Patient with atrial fibrillation after the first hour in the recording



(c) Legend

Figure 5.1: Selection of ECG windows of interest from Holter monitoring for AF risk identification. Selected windows are highlighted in blue or yellow.
(a) For healthy subjects, the first hour is selected.
(b) For patients with AF, the window is selected if there is no evidence of AF in the first two hours of the recording and evidence of AF in the following 22 hours.

of the CNN models using a variable input size of 10 seconds, 30 seconds, 60 seconds and 5 minutes.

The CNN model is composed of 9 blocks, composed of two lanes: a two-convolution lane and a shortcut lane, as shown in Figure 3.9. The final prediction is the result of the sigmoid activation of the output from the last fully connected layer. We used Adam (Kingma et al. 2017) as optimizer, and the binary cross-entropy as loss function. We used a learning rate of 10^{-4} and a batch size of 32. The model was trained up to 100 epochs, with an early stopping strategy after 3 epochs if the validation loss did not decrease.

The second model is a CNN-RNN model, as proposed by Biton et al. (2023). The model is presented in Figure 4.21. It is composed of two CNN layers followed by a bidirectional GRU layer and two fully connected layers. The sigmoid function is used for the final prediction. The model is also created in PyTorch (Paszke et al. 2019), we used the same optimizer, loss function, learning rate, and batch size. The model was also trained up to 100 epochs, with an early stopping strategy after 3 epochs if the validation loss did not decrease.

HRV-based models

We compared the performance of DL models with a ML approach, using Heart Rate Variability (HRV) parameters measures from RR intervals. The first model is a Random Forest (RF) classifier. We set the number of estimators to 200 with a maximum depth to 10. The second model is a gradient-boosted decision trees XGBoost (XGB) model (Chen et al. 2016). We used the same limitations as for the RF model, with a limit of 200 trees and a maximal tree depth of 10.

The two classifiers were evaluated on multiple windows size, i.e. 60 seconds as proposed by Levasseur et al. (2022), 85 seconds as proposed by Kisohara et al. (2020), 5 minutes and the complete 1 hour. The two models used short-term and long-term HRV and Heart Rate Fragmentation (HRF) parameters presented in Chapter 2. The input is composed of the following parameters:

- time domain features: mean heart rate, Standard deviation of NN intervals (SDNN), Root Mean Square of Successive RR interval Differences (RMSSD), Standard Deviation of Successive RR interval Differences (SDSD), CVNN, CVSD, pNN10, pNN20, pNN50, minNN, maxNN, medianNN, prc20NN, prc80NN, Triangular Interpolation of the NN interval histogram (TINN), and HRV Triangular Index (HRVi);

- frequency domain features: total power, power Low Frequencies (LF) band, power High Frequencies (HF) band, their normalized values, and LF/HF ratio;

- Poincaré plot features: SD1, SD2, and SD1/SD2 ratio, Cardiac Sympathetic Index (CSI), Cardiac Vagal Index (CVI) and CVI modified;

- Second Order Difference Plot (SODP) features: number of ΔRR in Q1 to Q4, Central Tendency Measure $(CTM)_{20}$, CTM_{50}, and CTM_{100};

- Acceleration (AC), Deceleration (DC), AC modified, DC modified, AC_k and DC_k;

- HRF features: Percentage of Inflection Points (PIP), Inverse of the Average Length of the acceleration/deceleration Segments (IALS), Percentage of Short Segments (PSS), and Percentage of Alternating Segments (PAS).

Evaluation using cross-validation

The models were evaluated using temporal 10-fold cross-validation at patient level. The recordings were ordered using their recording date, and separated in 10 groups. Each of the 10 group is used in turn as the test set, and the remaining 9 groups as the train set, i.e. 90%-10% train-test ratio. If a validation procedure is useful during the model training, a validation set is also set aside and the model is trained on the remaining 8 groups, i.e. 80%-10%-10% train-validation-test ratio. We should note that the separation is done at the patient level to avoid any data contamination between the train and the test set. If a patient has more than one recording, the first recording of the patient is used a reference date and all the recordings from this patient are in the same group.

We evaluated performance using Area Under the Receiver Operating Characteristic Curve (AUROC) and Area Under the Precision-Recall Curve (AUPRC). We also used threshold-based metrics: accuracy, sensitivity, specificity, Positive Predictive Value (PPV), Negative Predictive Value (NPV) and F1 score, as described in Chapter 2. For models using windows smaller than 1 hour, we first evaluated at the window level, then aggregated all windows for a single recording and evaluated the model at the recording level.

5.2.3 Inter-hospital cross-validation

As a second performance evaluation, we trained the model who obtained the best average score from the model benchmark, using the four centres in turn as testing group. The model is trained on the recordings from the first 3 centres and evaluated on the last centres. If a validation procedure is useful during the model training, a validation set is randomly chosen from the three centres and correspond to 10% of the recordings.

5.2.4 Age-group comparison

Finally, we measured the performance of our models according to different age groups. The recordings are divided into 5 groups according to the age of the patient at the time of recording: patients under 60, patients between 60 and 70, patients between 70 and 80 and patients over 80. A patient with multiple recordings could be found in more than one group. Using the model who obtained the best average score from the models benchmark, we train the model using 10-fold cross validation.

5.3 Results

5.3.1 Database and window selection

A total of 95 871 recordings were collected. After analysing the recordings, 879 were selected from 833 patients. 475 recordings are from 514 AF patients and 365 NSR recordings are from 358 patients, as summarized in Table 5.1. In total, we selected 1390 1-hour windows, 654 AF windows and 736 NSR windows. The CHU Luxembourg contribute the most to the number of recordings. We should note that CHU Brugmann does not have any NSR recordings, therefore the performance evaluation is not applicable. The 60-70 and 70-80 age groups contains the most AF recordings. For NSR recordings, the patients less than 60 are the largest group.

Among all included patients, the mean age is 62.6 years $\pm$ 16.4, ranging from 7 to 99 years, on the date of the Holter monitoring. The mean age of AF patients is 70.2 years $\pm$ 11.4, from 35 years to 99 years. The age distribution is presented in Figure 5.2. The CHA_2DS_2-VASc score is 2.9 $\pm$ 1.7. This score is ranging from 1 to 9. This score is computed from the clinical measurements

Table 5.1: Number of selected patients, recordings, and windows for all age groups and all centres

Group		Patient			Recording			1-hour Window		
		All	AF	NSR	All	AF	NSR	All	AF	NSR
All		833	475	358	879	514	365	1390	654	736
Centre	Dr Grégoire	152	138	14	164	150	14	260	232	28
	CHU Ambroise Paré	94	65	29	100	71	29	143	85	58
	CHL Luxembourg	501	186	315	525	203	322	897	247	650
	CHU Brugmann	86	86	0	90	90	0	90	90	0
Age	<60	255	78	177	266	86	180	476	116	360
	60–70	238	142	96	252	154	98	397	199	198
	70–80	210	145	65	222	155	67	337	199	138
	>80	136	116	20	139	119	20	180	140	40

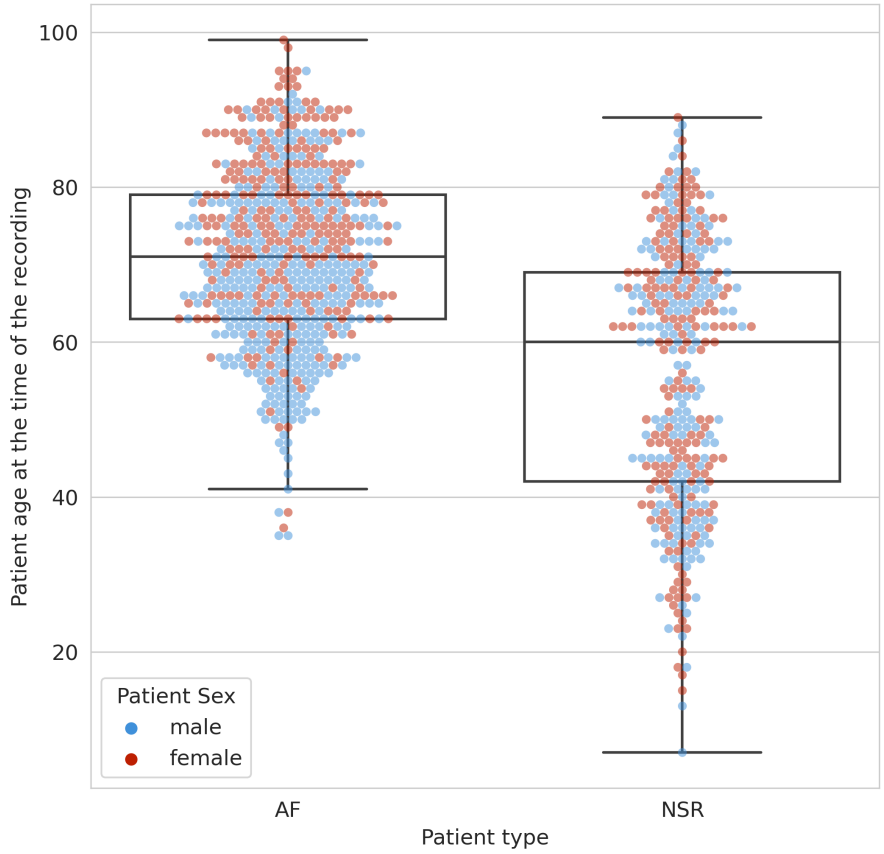


Figure 5.2: Comparison of patient age distribution between AF patients and NSR patients

from 2 centres: Dr Grégoire outpatient clinic and CHU Ambroise Paré. For the other two centres, CHL Luxembourg and CHU Brugmann, we have not yet been able to obtain the clinical data following the work required to achieve the same result as in the other two centres.

Recordings from female and male patients are nearly balanced, as shown in Table 5.2, with a lower ratio in the number of recordings from AF patients, with 44% recordings from female and 56% from male. The mean age is higher for women than for man recordings.

5.3.2 CNN window size comparison

We trained CNN models on the first 8 folds, using the 9th fold as validation and the 10th fold as testing. We compared the performance of the models with different input window sizes. We found that the CNN models performed best with an input window size of 30 seconds, both at the window level and

Table 5.2: Patients age comparison between AF and NSR groups

	Patient			Recording		Age (year)			
Type	Sex	#	%	#	%	Mean	σ	Min	Max
All	All	833		879		62.6	16.4	7	99
	Female	394	47.2	413	47.0	64.0	17.3	15	99
	Male	440	52.8	468	53.0	61.1	15.3	7	95
AF	All	475		514		70.2	11.4	35	99
	Female	208	43.8	225	43.8	74.4	10.6	36	99
	Male	267	56.2	289	56.2	66.9	10.9	35	95
NSR	All	358		365		55.8	17.1	7	89
	Female	186	52.0	188	51.1	56.1	17.3	15	89
	Male	173	48 .0	179	48.9	55.2	16.9	7	88

Table 5.3: Performance comparison for varying windows input size at window level, testing on the last fold

Model	Window size		Windows		Recordings	
	Seconds	Samples	AUROC	AUPRC	AUROC	AUPRC
CNN	10	2048	0.805	0.769	0.845	0.899
	30	6144	**0.822**	**0.784**	**0.856**	**0.912**
	60	12 288	0.814	0.780	0.831	0.889
	300	60 000	0.796	0.763	0.822	0.876

at the recording level, for the AUROC and the AUPRC. The performance are shown in Table 5.3. We found a parabolic tendency in the performance for AF detection using RR intervals, similar to the tendency found by Kisohara et al. (2020), i.e. low and high window sizes achieved lower results and the best results were achieved by a window size in between.

We chose to use a 30-second input window for both the CNN and CNN-RNN models in the benchmark model comparison. For all input sizes, we observed an improvement when the model was evaluated at the recording level rather than at the window level.

5.3.3 Model comparison

We compared the performance of RF and XGB models with two DL models: a CNN model and CNN-RNN model. The best average performance was achieved by the XGB model on the recording evaluation, with an AUROC of

0.919 (0.879-0.958) and an AUPRC of 0.919 (0.879-0.958). The confidence intervals were computed using the test metrics for each test set over the 10 folds.

Using a threshold of 0.5, this corresponds to an accuracy of 84.5 % (81.2–87.8), a sensitivity of 83.0 % (79.5–86.4), a specificity of 86.6 % (79.3–93.9), a precision of 90.2 % (85.5–94.9), a NPV of 78.4 % (74.7–82.1), and a F1-score of 86.2 % (83.5–89.0). An extended table for threshold-based metrics is presented in Table B.1 from Appendix B.

For both RF and XGB models, we observed a performance increase for both AUROC and AUPRC when using longer windows. For the evaluation at the recording level, 5 minutes windows gave the best results, outperforming 1-hour windows.

We observed the same performance increase between windows evaluation and recording evaluation for the DNN models, but the performance is lower than for those from the RF and XGB. The CNN model performs better than the CNN-RNN model using the same 30-second input window. We observed that both models overfit on the training set after one epoch and the performance did not improve after this first step. Lowering the learning rate did not improve the results.

5.3.4 Inter-hospital cross-validation

Using the XBG models on HRV features from 5 minute ECG windows, we evaluate the performance of the model in inter-hospital validation. As there are no NSR patients for CHU Brugmann, AUROC and AUPRC cannot be calculated for this group. The remaining results are all above 0.860, which corresponds to the performance achieved in the benchmarking of the models. The best performance was achieved for the Dr Grégoire outpatient clinic. The lowest performance was obtained on Centre Hospitalier de Luxembourg (CHL) patients. This can be explained by the fact that the highest proportion of patients in the database are from the CHL.

5.3.5 Age-group comparison

Using the XBG models on HRV features from 5-minute ECG windows, we evaluate the performance of the model using the four selected age group. In Table 5.1, the < 60 group has the most recordings. The > 80 group only has 20 NSR records.

Table 5.4: Models performance evaluation at the window level and recording level

Input	Model	Window size (seconds)	Windows		Recording	
			AUROC	AUPRC	AUROC	AUPRC
HRV	RF	60	0.831 (0.774-0.889)	0.829 (0.778-0.880)	0.906 (0.857-0.956)	0.940 (0.909-0.972)
		85	0.840 (0.785-0.896)	0.839 (0.789-0.888)	0.910 (0.864-0.955)	0.943 (0.914-0.971)
		300	0.866 (0.818-0.913)	0.863 (0.818-0.908)	0.917 (0.877-0.956)	0.947 (0.923-0.971)
		3600	0.881 (0.829-0.933)	0.875 (0.819-0.931)	0.899 (0.852-0.946)	0.935 (0.904-0.965)
HRV	XGB	60	0.832 (0.774-0.890)	0.831 (0.777-0.884)	0.908 (0.859-0.957)	0.941 (0.910-0.972)
		85	0.840 (0.784-0.897)	0.840 (0.788-0.892)	0.912 (0.866-0.958)	0.944 (0.916-0.973)
		300	0.864 (0.815-0.913)	0.864 (0.818-0.909)	**0.919 (0.879-0.958)**	**0.949 (0.925-0.972)**
		3600	**0.884 (0.830-0.937)**	**0.880 (0.823-0.937)**	0.901 (0.852-0.950)	0.937 (0.905-0.968)
ECG	CNN-RNN	30	0.754 (0.701-0.807)	0.696 (0.624-0.769)	0.800 (0.750-0.849)	0.834 (0.787-0.881)
ECG	CNN	30	0.809 (0.780-0.838)	0.760 (0.724-0.796)	0.846 (0.816-0.875)	0.881 (0.858-0.903)

Table 5.5: Performance using spacial fold. For CHU Brugmann AUROC and AUPRC are not applicable, as there is no NSR recording in the database for this hospital, the AUPRC value is 1.

Hospital	Windows		Recordings	
	AUROC	AUPRC	AUROC	AUPRC
Dr Grégoire	0.942	0.993	0.965	0.997
CHU Ambroise Paré	0.880	0.910	0.923	0.966
CHL Luxembourg	0.860	0.737	0.894	0.871
CHU Brugmann	N/A	1.000	N/A	1.000

Results are presented in Table 5.6. The best performance was achieved in the > 80 group with an AUROC of 0.983 and an AUPRC of 0.997. The AUPRC is lower for the < 60 group as the prevalence is lower for this group, but surprisingly the AUROC is the lowest for the 70-80 group for both window size. The threshold-based metrics are presented in Table 5.6. The accuracy reflects the performance of the AUROC and AUPRC, with a lower performance for the 70–80 group. Sensitivity and PPV values are increasing with age, with a 98% sensitivity and 94% PPV for the patients aged 80 and older.

Building on these results and the results from Singh et al. (2022), we extended the results presented in the benchmark presented in Table 5.4 to include the age of the patient at the time of recording and the sex of the patient. In addition, we calculated the pairwise correlations between the HRV features using the Spearman correlation coefficient. As for the forecast of AF onset, we found that some features were highly correlated, with a coefficient greater than 0.9. The full correlation matrix is available in Appendix B in Figure B.1. For each experiment, we run it once with all features and once with selected features, i.e. one feature was removed from each highly correlated pair. The results are shown in Table 5.8. The average model score improved with the addition of the two parameters, to 0.92 AUROC and 0.95 AUPRC for the RF using HRV features from the 300-second window. Future work should explore the use of additional clinical variables to improve model performance.

5.4 Discussion

We found that a ML model is able to identify a AF signature in Holter recordings from AF patients. The ML model was able to identify HRV parameters

Table 5.6: Performance of the XGBoost model using HRV parameters computed on 5 minutes and 1-hour window. Validation was done using temporal cross-validation.

Window size	Age group	Windows		Recordings	
		AUROC	AUPRC	AUROC	AUPRC
HRV 5 minutes	< 60	0.850 (0.799-0.901)	0.688 (0.588-0.789)	0.904 (0.850-0.959)	0.866 (0.796-0.937)
	60–70	0.844 (0.810-0.877)	0.844 (0.800-0.887)	0.933 (0.891-0.975)	0.957 (0.929-0.985)
	70–80	0.784 (0.720-0.848)	0.848 (0.798-0.898)	0.845 (0.770-0.920)	0.934 (0.901-0.967)
	≥ 80	0.893 (0.857-0.929)	0.965 (0.953-0.977)	0.983 (0.962-1.004)	0.997 (0.994-0.999)
HRV 1 hour	< 60	0.888 (0.835-0.941)	0.745 (0.646-0.844)	0.898 (0.846-0.949)	0.846 (0.768-0.924)
	60–70	0.877 (0.845-0.909)	0.872 (0.826-0.919)	0.911 (0.869-0.952)	0.948 (0.924-0.972)
	70–80	0.840 (0.778-0.902)	0.878 (0.823-0.932)	0.860 (0.779-0.941)	0.943 (0.908-0.977)
	≥ 80	0.907 (0.861-0.952)	0.970 (0.951-0.990)	0.963 (0.923-1.004)	0.994 (0.988-0.999)

Table 5.7: Metrics for recording

Age group	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	F1 (95% CI)
< 60	0.835 (0.772-0.897)	0.582 (0.415-0.749)	0.956 (0.910-1.001)	0.882 (0.764-0.999)	0.834 (0.774-0.895)	0.668 (0.516-0.820)
60-70	0.849 (0.792-0.907)	0.813 (0.729-0.897)	0.906 (0.836-0.975)	0.935 (0.894-0.976)	0.771 (0.679-0.862)	0.866 (0.811-0.920)
70-80	0.770 (0.685-0.855)	0.845 (0.769-0.920)	0.595 (0.417-0.774)	0.833 (0.768-0.899)	0.625 (0.472-0.779)	0.836 (0.775-0.898)
> 80	0.935 (0.906-0.964)	0.984 (0.960-1.008)	0.650 (0.409-0.891)	0.946 (0.910-0.982)	0.926 (0.813-1.039)	0.963 (0.947-0.979)

Table 5.8: Comparison of model performance between HRV features and HRV features with age and sex. Features with an asterisk (*) indicate selected features, where one feature was removed from each highly correlated pair.

Input	Model	Window size (seconds)	Windows		Recording	
			AUROC	AUPRC	AUROC	AUPRC
HRV	RF	300	0.866 (0.818-0.913)	0.863 (0.818-0.908)	0.917 (0.877-0.956)	0.947 (0.923-0.971)
		3600	0.881 (0.829-0.933)	0.875 (0.819-0.931)	0.899 (0.852-0.946)	0.935 (0.904-0.965)
HRV*	RF	300	0.884 (0.828-0.939)	0.880 (0.823-0.937)	**0.927 (0.884-0.971)**	**0.954 (0.928-0.980)**
		3600	0.892 (0.837-0.947)	0.884 (0.821-0.947)	0.906 (0.856-0.957)	0.938 (0.904-0.973)
HRV+A+S	RF	300	0.882 (0.827-0.937)	0.881 (0.824-0.938)	0.926 (0.882-0.971)	0.954 (0.926-0.981)
		3600	0.896 (0.845-0.947)	0.895 (0.842-0.949)	0.909 (0.860-0.957)	0.942 (0.910-0.974)
HRV+A+S*	RF	300	0.884 (0.828-0.939)	0.880 (0.823-0.937)	**0.927 (0.884-0.971)**	**0.954 (0.928-0.980)**
		3600	0.892 (0.837-0.947)	0.884 (0.821-0.947)	0.906 (0.856-0.957)	0.938 (0.904-0.973)
HRV	XGB	300	0.864 (0.815-0.913)	0.864 (0.818-0.909)	0.919 (0.879-0.958)	0.949 (0.925-0.972)
		3600	0.884 (0.830-0.937)	0.880 (0.823-0.937)	0.901 (0.852-0.950)	0.937 (0.905-0.968)
HRV*	XGB	300	0.877 (0.827-0.927)	0.869 (0.818-0.920)	0.915 (0.878-0.952)	0.944 (0.923-0.966)
		3600	0.896 (0.841-0.951)	0.890 (0.829-0.950)	0.911 (0.859-0.963)	0.940 (0.902-0.977)
HRV+A+S	XGB	300	0.879 (0.830-0.928)	0.874 (0.824-0.925)	0.917 (0.879-0.954)	0.946 (0.924-0.968)
		3600	**0.904 (0.853-0.955)**	**0.903 (0.848-0.958)**	0.917 (0.869-0.966)	0.948 (0.916-0.981)
HRV+A+S*	XGB	300	0.877 (0.827-0.927)	0.869 (0.818-0.920)	0.915 (0.878-0.952)	0.944 (0.923-0.966)
		3600	0.896 (0.841-0.951)	0.890 (0.829-0.950)	0.911 (0.859-0.963)	0.940 (0.902-0.977)

that allowed classification between AF and NSR recordings. We also found that a DL model is able to find an AF signature in an ECG recording. In the performance analysis, the ML model performs better than the DL model when evaluated either at the window level or at the recording level, i.e. when the evaluation is done by aggregating all windows from a single recording. Nevertheless, using a 5-minute window, the model achieves an AUROC of 0.866 and an AUPRC of 0.863.

When compared to the state of the art, such as the results from Attia et al. (2019a), we obtained lower results with a comparable DL model, but higher results with the XGB model. Our hypothesis is that DNN are not able to generalise properly due to the limited size of the database, as all windows are selected from a total of 1390 unique 1-hour windows.

Attia et al. (2019a) used a database composed of 649 931 normal sinus rhythm ECGs from 180 922 patients. During this study, we wondered what the results of the studies proposed above would have been if the variability parameters had been used, but existing research on AF identification is difficult to extend due to data access limitations.

The results of Singh et al. (2018) are interesting to compare with this study, as they proposed a model based only on Premature Atrial Contraction (PAC), sex and age of the patient, and not only on ECG analysis with DL. The model achieved an AUROC of 0.74, and the ECG was able to add valuable information to improve the results. The age and sex of the patient can be inferred from the ECG by DNN (Attia et al. 2019b). We have shown that the results of the models did improve when the age and sex was added as input value to the model. Part of the DNN prediction may be related to these results, but in this study the DNN models performed worse than the ML models using HRV parameters. In the future, the effect of other clinical parameters such as hypertension, obesity, diabetes, hyperthyroidism should also be investigated to understand if other risk factors could help the model to better understand the context of the ECG and therefore make better predictions for the identification of these patients.

5.4.1 Features importance analysis

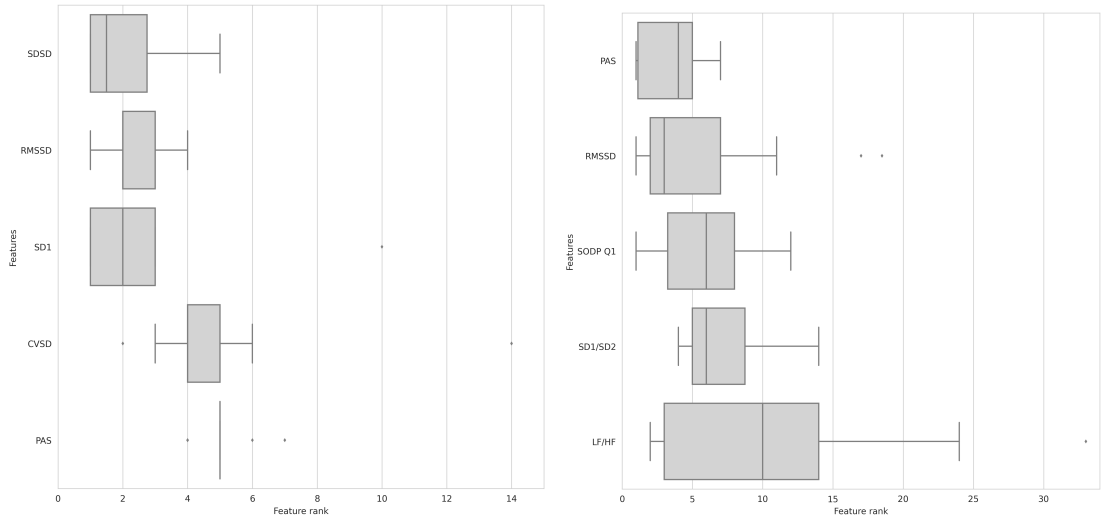
We have analysed the features used by the XGB model to understand which are the most important parameters for AF identification. The analysis is performed using multiple methods and only the top 5 features are shown. The full feature ranking is presented in Appendix B.

For the first analysis, Figure 5.3a, we analysed the gain of each feature, i.e. the relative contribution of the feature to the model prediction, which is determined by computing the contribution of all features across all trees in the model, as defined by Chen et al. (2016). A higher gain value indicates that it is more important for generating a prediction. The results show that the SDSD, RMSSD and SD1 from the Poincaré plot were the most useful parameters over the 10-fold cross-validation. The three parameters described the short-term variability, more precisely the single-beat variability (Brennan et al. 2001).

These parameters are also found in the second analysis, i.e. the aggregation of the first method with two alternative analyses: (i) the frequency, i.e. the number of times a feature is used in a tree, and (ii) the coverage, i.e. the number of HRV windows affected by this feature. The result is presented in Figure 5.3b. The SDSD is replaced by the PAS as the first ranked feature, but this parameter also measures short-term variability using alternating segments. As a reminder, it measures the percentage of short segments, defined as $ADAD$ or $DADA$, where A is an acceleration as $\Delta RR < 0$ and D is a deceleration as $\Delta RR > 0$. NSR recordings are associated with less short-term variability, but as a healthy heart is not a perfect metronome, there is always variability in the heartbeats (Shaffer et al. 2014).

The features highlighted by the two previous evaluation did not yield exactly the same results, but the same underlying meaning to highlight the short-term RR variability. We used the Shapley values as a third method, to examine the features used by the models for the predictions. The SHAP score described the contribution of each features in the model output. It is computed using a subset and permutation-based and subset method, as proposed by Lundberg et al. (2020). As shown in Figure 5.4, similar features to the features importance analysis using the 3 methodologies can be found in the SHAP results. The RMSSD, PAS and SODP Q1 are the top 3 features. The next two features, the SODP CTM 100 and the SD1/SD2 ratio, are also ranked in the two previous methods.

Using the beeswarm plot, presented in Figure 5.4b, we can use the values of the features to explain the impact of these features on the model prediction. A higher RMSSD or PAS, i.e. more short-term variability, tends to lean the prediction towards AF. AF recordings seems to have less ΔRR in the Q1 of the Second Order Difference Plot (SODP). Indeed, alternating segments are classified in Q2 or Q4. The distinction for CTM measure is not clear. The SD1/SD2 ratio is lower for NSR recordings, which mean that the cloud point



(a) Gain method (b) Combination of the 3 methods

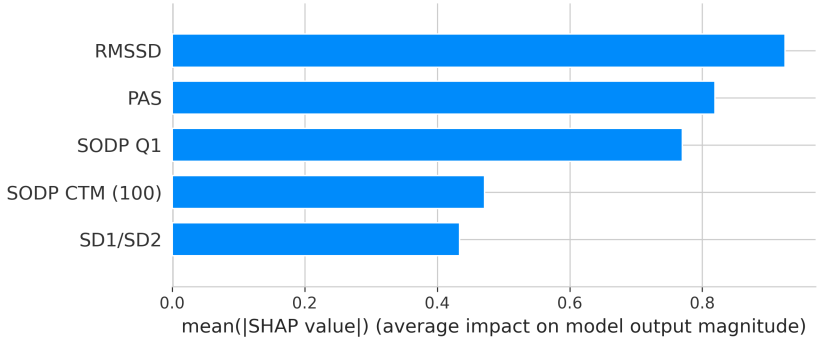
Figure 5.3: Metrics ranks based on either gain analysis or the 3 methods available for the XGB model. The metrics are ranked based on the trained model for each fold and then classified according to the average rank across all folds.

is smaller than for AF recordings.

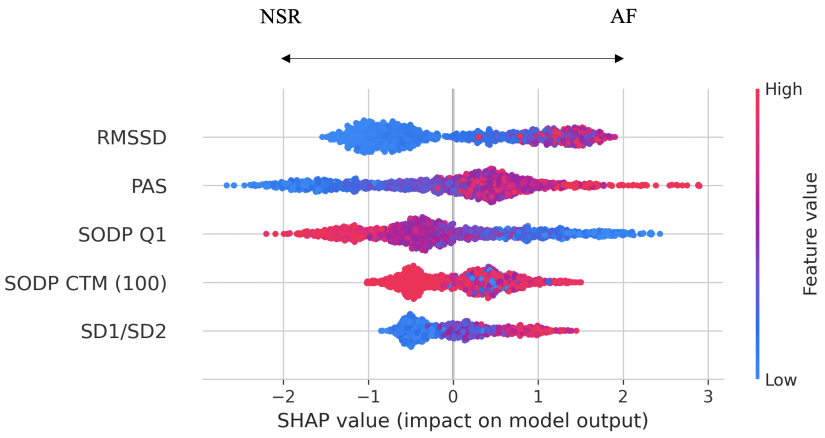
The SHAP score has been extended to explain the gradient activation of the DNN. In future work, this SHAP score could also be used to analyse the score obtained by CNN and understand which part of the ECG the model is using to make the prediction. This could provide a feedback loop from the internal representation of the DNN to the clinician.

5.4.2 Clinical implication

The development of new Artificial Intelligence (AI) models could have clinical implications for both inpatient and outpatient settings. Holter monitoring is still widely used for AF screening in Belgium and worldwide. If the performance of a model convinces health authorities and medical professionals to use it for AF screening, the workload for cardiologists, nurses and technicians could be reduced. In the next few years, the reimbursement system could be questioned to include these new technologies using short-term recordings. In addition, AI models have now started to be used in clinical practice for AF detection in 1-lead wearables (Tison et al. 2018; Perez et al. 2019). If this use could be extended to identify the AF signature in the NSR recording, it could



(a) Importance of top 5 features



(b) Beeswarm plot

Figure 5.4: SHAP analysis results (a) top 5 features ranking and (b) beeswarm plot. Each row of the beeswarm plot consists of one point for each prediction. The y position of a point is related to the influence of the feature on the prediction, i.e. a prediction towards the positive (right) side indicates that the features influenced the prediction towards AF, and a prediction towards the negative (left) side indicates that the features influenced the prediction towards NSR. Next, the colour of each point indicates whether the feature value is high (red) or low (blue). The relationship between the position of the points and the colour helps us to understand the impact of the feature on the prediction. E.g. for the first feature, RMSSD, the points on the right are more red and the points on the left are more blue. This means that if the recordings contain more RMSSD (red dots), the recording contains more short-term variability and therefore the model is more likely to predict AF.

help to detect the disease earlier.

Once accepted by medical professionals, the next challenge will be to get older generations to adopt and use these new technologies. Indeed, older generations are more difficult to convince, and unfortunately the prevalence of AF is highest in these patients.

The positive predictive value obtained by our retrospective study from our sample is very good for patients over 60 years of age and excellent for patients older than 80, with a PPV value of 95% for patients aged 80 years and older. These are precisely the patients most at risk of developing AF, and Sekelj et al. (2021) validated in a prospective study the usefulness of an ML approach, achieving a high PPV in older patients. The sensitivity of our models is better in older patients and specificity is better in younger patients, which is ultimately good news. The low specificity for this group can be linked to the low number of patients of that age with normal Holter recordings. Alternative measurements may also be useful in this population. Yan et al. (2020) demonstrated the possibility of detecting AF using non-contact facial PhotoPlethysmoGraphy (PPG). This type of measurement could be extended to identify AF risk in selected populations.

Conversely, the low sensitivity in patients younger than 60 years can be interpreted as a result of the lower number of AF recordings in this age group. The performance of the model in patients younger than 60 years should be extended in the future, as in IRIDIA-AF version 2 some NSR recordings are from patients younger than 35 years and even children, with the lowest age being 7 years. This was related to the limited number of recordings as we wanted to keep the most recordings, but in practice this could introduce a bias into the model as such young patients will not be screened for AF in everyday clinical practice. In the future, we should limit this class to 35 to 60 years to match the age range present in the selection of AF patients and to have a similar age distribution between AF and NSR patients.

On a larger scale, ML models could be used for improved analysis of the Holter monitoring database. In this study, we used temporal 10-fold cross-validation. In this case, a model trained on the first nine folds and tested on the last fold corresponds to a real-world application where a model is trained on existing data and tested on new data. In this case, the model could be used as an adjunct to existing risk scores and cardiologist assessment. A patient could have a 1-hour Holter and the models could predict the risk score of having atrial fibrillation in the next 24 hours. If the score is high, the patient

continues with the 24-hour Holter. If the score is low, the full Holter is not required, as the likelihood of finding AF is low. As with the task of forecasting the onset of AF presented in the previous chapter, the choice of threshold to discriminate between low and high risk patients is critical and should be based on discussions with cardiologists to better understand the need for the medical application and the impact of false positive and false negative predictions. A prospective study would be required to further develop and validate this methodology.

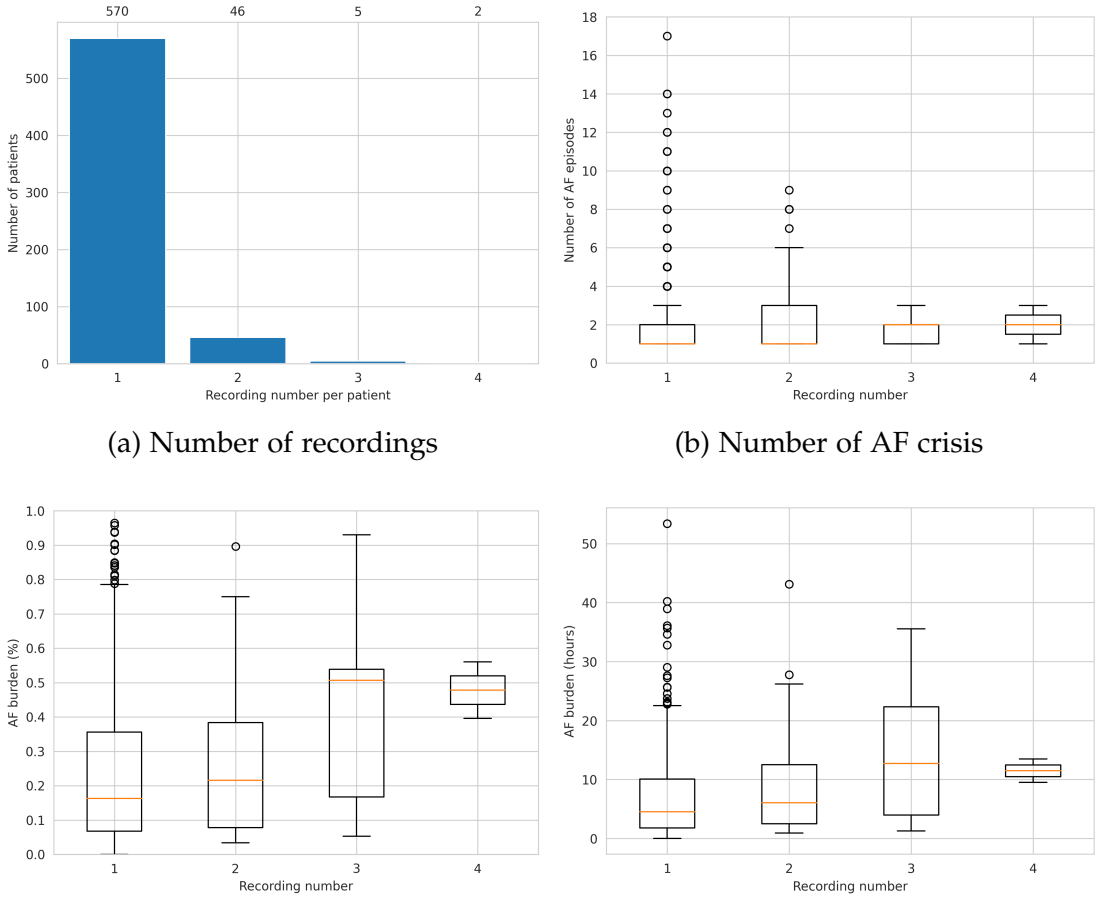
Finally, we investigated if Holter recordings had already an impact in cardiology practice, as reported by the IRIDIA-AF version 2 database. We measured the AF burden evolution in our database, which correspond to the percentage of the recording the patient spent in AF compared to NSR. The results are presented in Figure 5.5. AF burden does not seem to decrease between the first and the following AF recordings, and in fact the tendency is that the mean AF burden is higher in hours and in ratio. The extreme value in the number of AF episodes are no longer present. The patients seem to have less but longer episodes in the following Holter recordings.

5.5 Summary

In this chapter, we investigate the performance of ML models for the identification of patients who are at risk of developing AF episodes in the next 22 hours. We selected 1-hour windows at the start of the record to understand if this windows contains information about the incoming AF crisis.

We found that the use of a 1-hour ECG recording is already sufficient to give a good predictive value, especially in older patients. We compare the performance obtained by 4 different ML and DL techniques, namely the RF, XGB, CNN and CNN-RNN. The analysis of the XGB model suggests that the important information enabling AF prediction is contained in the HRV and in particular in the short-term variability.

This opens up perspectives that can be used by cardiology monitoring techniques and wearables. Prospective studies are needed to confirm the encouraging potential offered by these results.



(a) Number of recordings

(b) Number of AF crisis

(c) AF burden ratio AF/NSR

(d) AF burden in hours

Figure 5.5: Evolution of the AF burden in follow-up Holter monitorings

Chapter 6

Conclusions and perspectives

Conclusions

The objective of this thesis is twofold: to forecast the onset of Atrial Fibrillation (AF) episodes and to identify the risk of AF in sinus rhythm. To forecast AF onset, Electrocardiogram (ECG) windows close to the onset are compared with ECG windows distant from the onset. The objective is to determine whether Machine Learning (ML) models can differentiate between the two types of windows and to obtain a trained model capable of detecting early signs of AF. To identify AF risk, we compare normal sinus rhythm ECG windows from patients with and without AF. Our aim is to determine whether ML models can distinguish between the two types of windows to better identify patients at risk of AF. We focus on the use of ML models, using 1-dimensional ECG data as input. We also extend the representation of the ECG to ECG Morphology Variability (ECGMV), RR and Heart Rate Variability (HRV). This thesis relies on the hypothesis that ECG data prior to AF onsets and ECG sinus rhythm data from AF patients contain information about AF.

Following the introduction in Chapter 1, we discuss the state-of-the-art in Chapter 2. First, we introduce selected HRV measurements, with a particular focus on geometrical measurements. We then review the existing publicly available ECG databases related to AF and show that there is a need for new publicly available materials for the task of AF onset forecast. The final part of the state-of-the-art chapter introduces the three types of AF predictions in the literature, AF detection, AF onset forecast and AF identification during sinus rhythm, with a particular and extensive focus on the AF onset forecast and AF identification. For AF onset forecast, the results tend to use more HRV feature

extraction with different ML models, while Deep Learning (DL) models were used more frequently in the results for AF risk identification.

Based on the results of the state-of-the-art review, we have created a new database of long-term electrocardiograms in paroxysmal AF. The database is presented in Chapter 3. We present version 1, which consists of 167 recordings from 152 AF patients selected retrospectively in an outpatient clinic. This version of the database has been published on Zenodo. We then present the extended database: IRIDIA-AF version 2. This second version consists of 988 records from 928 patients. This version contains recordings from patients with and without AF. The total duration of the database is 100 million seconds, making it, to our knowledge, the largest open-access database of long-term atrial fibrillation recordings. A total of 964 AF episodes were annotated by a cardiologist and a cardiac nurse. This database construction work was necessary before the contributions in AF onset forecast and AF risk identification could be built.

Chapter 4 presents our contribution to AF onset forecast. First, we discuss the reproducibility of the models presented in the state-of-the-art. We selected three published ML approaches and reproduced the described methods. We found that the results were not reproducible. We then studied the ECG at AF onsets available in the IRIDIA-AF database. Furthermore, we found that both ECG morphology and HRV varied significantly before AF onset, suggesting that these features may be suitable for AF onset forecast. We evaluate the performance of a Random Forest (RF) model using HRV features from 5-minute pre-AF windows, i.e. close to AF onset, and HRV features from 5 minutes away from AF onset. We found that the closer the window to the AF onset, the better the results, supporting that the ECG closer to the AF onset contains information about the incoming crisis. At the onset, the RF model achieves an Area Under the Receiver Operating Characteristic Curve (AUROC) of 0.714 and an Area Under the Precision-Recall Curve (AUPRC) of 0.697.

Finally, we evaluated various ML and DL models in an extensive experimental benchmark based on selected 30-minute pre-AF and 30-minute Normal Sinus Rhythm (NSR) windows. The NSR windows were separated by at least 2 hours from any AF onset. We evaluated four classes of models using four types of input: ECG, ECGMV, RR and HRV. Models using RR and HRV gave the best results, suggesting that most of the information for predicting AF is contained in heart rate variability. The Convolutionnal Neural Network (CNN) model using RR intervals achieves the best performance with an AUROC of

0.604 and an AUPRC of 0.595. We evaluate the features selected by the model to construct a feedback loop back to cardiology. We create the HRV features from cardiological knowledge and use ML to understand meaningful features for AF onset prediction.

In Chapter 5 we present our contribution to the identification of AF risk in sinus rhythm. We selected recordings from AF patients with no evidence of AF in the first two hours and all recordings from healthy patients. We compare the performance of the ML model using HRV and the DL model using ECG. The best model achieved an AUROC of 0.90 and an AUPRC of 0.94. In the feature importance analysis, we found that the short-term features contain the most information for the ML model.

Limitations and future work

The results of this study are based on limited Holter data from a mixed patient population. Some of these patients were prescribed antiarrhythmic therapy. This introduces a relative inhomogeneity in the sample, but on the other hand it is more representative of a real clinical situation that our model will have to deal with.

The Holter recordings labelled NSR are those without rhythm disturbances, which does not necessarily mean that there is no associated pathology, as clinical characteristics were not available for all subjects, introducing some bias. Patients were not selected for the study, but were selected retrospectively from hospital archives. However, it is likely that the results would have been even better if healthy volunteers had been recruited for the study. Bias could have an important effect on model performance. In this work, we did not investigate the effect of other potential biases such as race and gender.

Dataset expansion and publication

We have released version 1 of the database. Version 2 has already been completed with the addition of about 750 records and 3 new centres. This second version should be published so that other researchers can build new results and models. The database should be promoted through challenges such as those proposed on Kaggle or at scientific conferences.

A version 3 of the database could be the next step and there are several directions. The first is to continue to work with the centres to start prospec-

tive studies. The second is to recruit new centres where further retrospective studies could be conducted. Holter annotation is a time-consuming task and automatic annotation and semi-supervised learning are methods that should be explored in the future. This will require collaboration between academics and Holter monitoring companies. Finally, the use of synthetic data as suggested by Guillaudeux et al. (2023) could be explored.

Towards better model performance

The results we obtained for forecasting the onset of AF were surprising and disappointing. From the first benchmark with a limited number of patients and only RR intervals, to the latest results with an expanded database and different models with different inputs, we have not seen an increase in performance. One hypothesis about these results is that we still do not have enough data to fully train a DL model. We are now in some intermediate data, which is still not enough.

We should continue to explore new models and layers in DL models, such as Attention (Vaswani et al. 2017), which has shown great performance in healthcare applications in recent years (Hu et al. 2022; Nguyen et al. 2022). We should also continue to explore data representation, such as spectrogram analysis using 2D CNN. HRV parameters could be further explored, as new parameters can always be added as input to the model. We can think of the P-P variability, recurrence plots (Marwan et al. 2007) or the heart rate n-variability proposed by (Liu et al. 2019). The effect of ECG pre-processing should also be measured, and other pre-processing steps and filters should be tested. Time of day may also have an effect on these parameters. This was not considered in this work, but should be further investigated.

The models could be trained in alternative ways. A first pre-training step using AF detection could be used to make an initial optimisation of the weights. We could then use this pre-trained model to retrain the model for AF onset prediction and AF detection to understand if this improves model performance.

Another direction is to move from general models to patient models. In this work, a model is trained on selected patients and tested on unseen patients. We should test the performance of training a model on selected patients and fine-tuning the model on the first recordings or AF episodes of that patient to understand if a specialised model achieves better performance.

Prospective study, proof of concept and applications

The results we obtained for the identification of AF are encouraging, as the results we obtained for the identification of the AF signature in sinus rhythm using ML are higher than those obtained with the risk score in a large cohort. Prospective studies should be conducted to understand whether ML models could identify AF in a larger and more general population and improve AF screening. For this, we can rely on the rapid development of new connected Holter monitoring devices that allow analysis of results during recording. This allows a better and continuous workflow for patient management.

If the sensitivity is high enough to detect all cases and reduce the workload of analysing recordings from patients without AF, we must be careful not to lower the threshold too much for any reason, especially financial reasons. A stroke is more expensive to society than anticoagulant treatment for AF. On a macro level, there will always be false negatives in the statistics, but on a micro level, a stroke can be life-threatening and life-changing for a patient.

Perspectives

The three AF-related tasks, corresponding scores and therapeutic strategies are summarised in Table 6.1. For AF detection, we can assume that humans achieve the perfect score because they define the annotation used as ground truth by the models. We have shown in Chapter 3 that ML models achieve high performance for AF detection.

AF onset forecast is difficult and even impossible for humans, hence the 0% score, but ML models achieve a score of 71%, as shown in Chapter 4. The pill-in-pocket strategy should also be considered, as it can help optimise the medication taken by the patient.

Finally, risk scores, such as CHARGE-AF, allow cardiologists to identify patients at risk of AF. We have shown in Chapter 5 that ML models are able to achieve up to 90% performance, which opens the way for better AF screening using smart Holter monitors, smartwatches and wearables. They could make this screening strategy more easily accessible to the general population.

Ultimately, ML predictions can be used as an extension of the risk score to assist and support the work of the cardiologist. ML is unlikely to completely replace the cardiologist, but cardiologists using ML tools could work more effectively and some specific workloads, such as ECG annotation, could be

Task	Prediction score		Therapeutic strategies
	Human	ML	
AF detection	100%	99%	Treatment, e.g. anticoagulant
AF onset forecast	0%	70%	PITP Forecasting algorithm in CIED
AF risk identification	75%	90%	Screening optimization Lifestyle changes

Table 6.1: AF-related tasks with corresponding approximate scores and therapeutic strategies and perspectives

drastically reduced.

It is important to note that while human errors and ML errors may be similar, the key difference is that ML errors are consistent unless the model is corrected and retrained. It is therefore crucial to maintain human oversight to prevent persistent and potentially biased false predictions. The medical field should prioritise medicine and utilise technology as a tool rather than becoming overly reliant on it.

The decision to use ML models is not always straightforward. The construction of meaningful features requires technical knowledge of the problem and can therefore be a barrier in some cases, which is not necessary for Deep Neural Network (DNN). However, DL models require significantly more data and computing power to optimise their weights and to be able to learn from the data. Depending on the requirements of the solution, one may be preferred over the other. Energy consumption should also be considered. The footprint for training and running DL models has increased in recent years as the models have grown in size (Vries 2023).

In this work, ML proved to be as effective as DL models, but the results could not reflect the final clinical applicability. Nevertheless, this is a positive result in the long run, as these models have a lower resource requirement and simpler inputs, such as PhotoPlethysmoGraphy (PPG) proposed by wearables, are sufficient for a high quality prediction. In addition, understanding the use of features by the model can be a significant advantage for the acceptance of new ML-based tools, rather than DL-based methods where the internal representation of the data is not clear. In this work, we have shown that the DL models are not always the magic solution that gives the best results.

AF screening in the general population using ML predictions from smartwatches and wearables could mean that there are many false positives, as the prevalence in the general population is lower compared to selected population, e.g. CIED patients or older patients. This should be avoided as it could lead to mistrust from healthcare professionals (Reyna et al. 2022; Shah et al. 2022). Instead, older patients and especially those with Cardiac Implantable Electronic Device (CIED) have a higher prevalence of AF. They could therefore be a better focus, as they are the perfect beneficiaries of an AF identification algorithm. In addition, forecasting the onset of AF could be facilitated if a device is already in place, as it can deliver a cardiac overdrive sequence as a treatment. We hope that this work and our contributions will help to improve the results of the models over the next years, so that one day they can be used in practical clinical applications to help and treat patients.

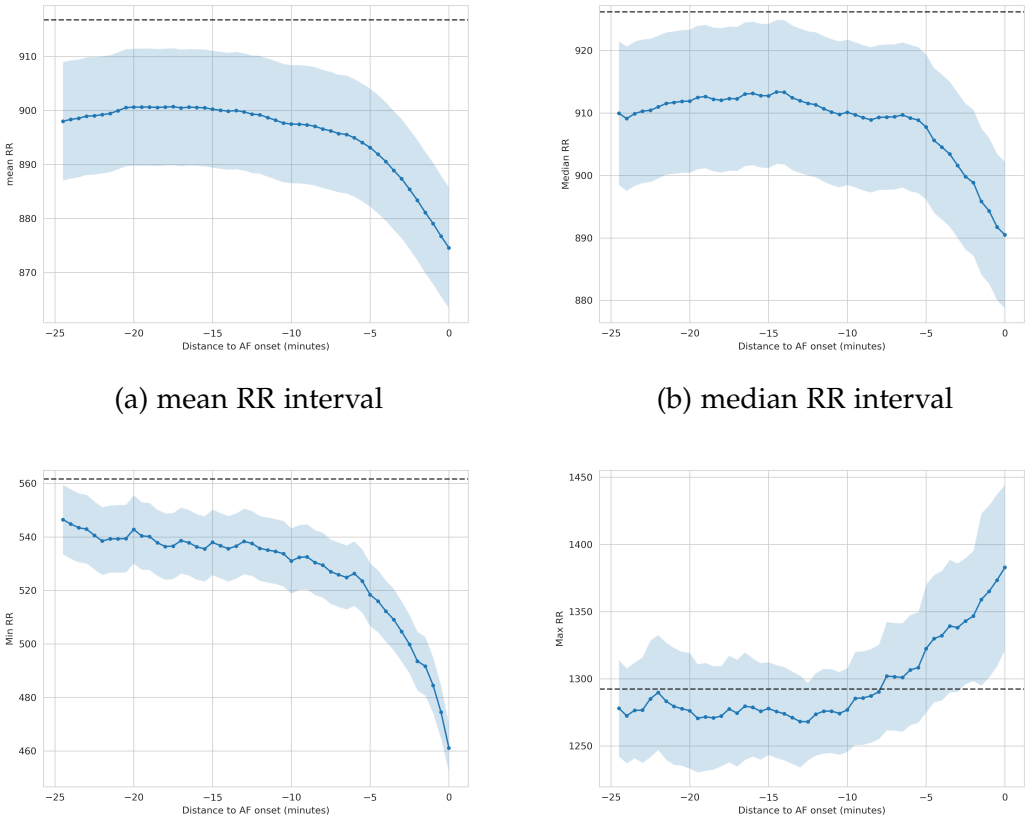
Appendix A

Atrial fibrillation onset forecast

A.1 HRV evolution before AF onset

In this section, we present the evolution of HRV features before the onset of paroxysmal AF, as only selected features are presented in Chapter 4. The features are grouped into the following categories.

- HRV temporal features: Figure A.1, Figure A.3 and Figure A.2.
- HRV histogram features: Figure A.4
- Poincaré plot features: Figure A.5
- Second Order Difference Plot (SODP) features: Figure A.6 and Figure A.7
- Acceleration (AC) and Deceleration (DC) features: Figure A.8
- Heart Rate Fragmentation (HRF) features: Figure A.9
- HRV frequency analysis features: Figure A.10



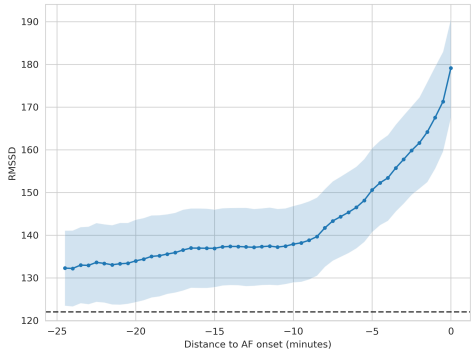
(a) mean RR interval

(b) median RR interval

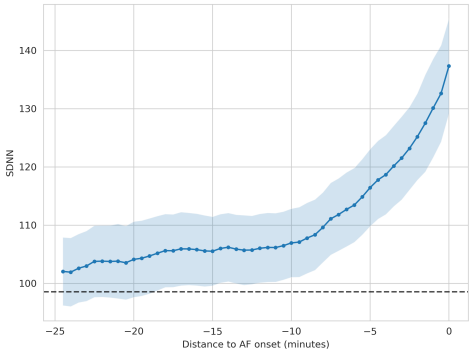
(c) min RR interval

(d) max RR interval

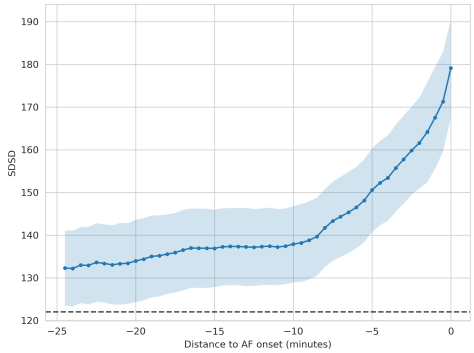
Figure A.1: Evolution of HRV temporal features before AF onset (1/3). The analysis was performed with a sliding window of 5 minutes and a step of 30 seconds. The main line corresponds to the mean of all selected windows. The 95% confidence interval is shown around the line. The baseline (black dotted line) represents the average value of the features in the 2-hour period preceding the analysed 30-minute window.



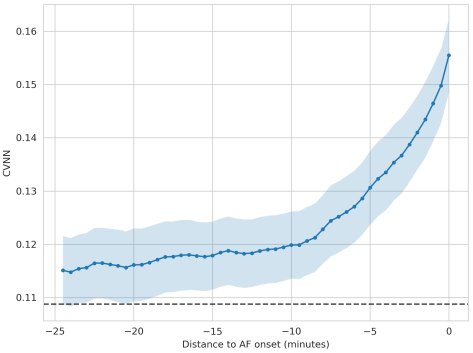
(a) RMSSD



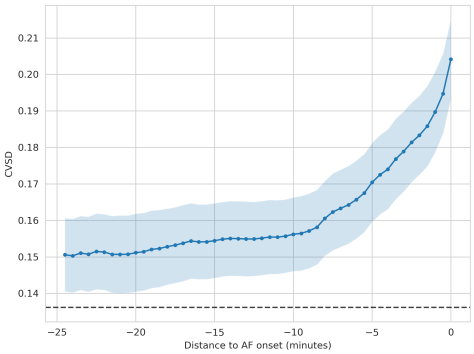
(b) SDNN



(c) SDSD

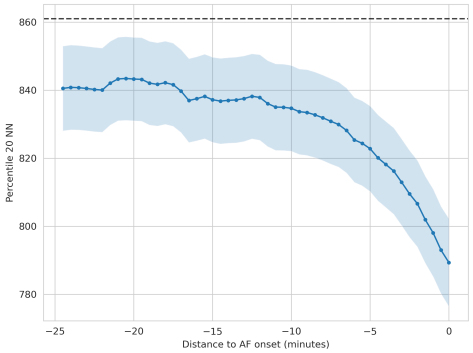


(d) CVNN

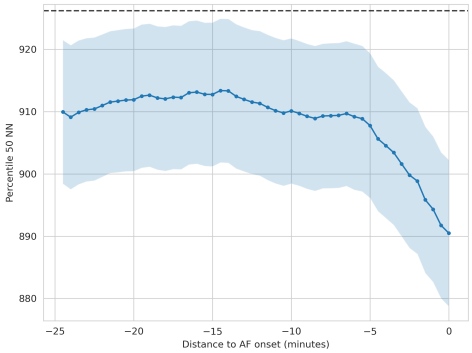


(e) CVSD

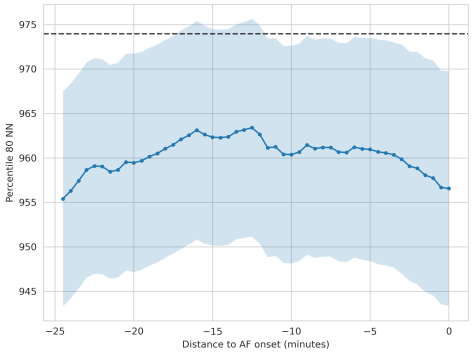
Figure A.2: Evolution of HRV temporal features before AF onset (2/3). The analysis was performed with a sliding window of 5 minutes and a step of 30 seconds. The main line corresponds to the mean of all selected windows. The 95% confidence interval is shown around the line. The baseline (black dotted line) represents the average value of the features in the 2-hour period preceding the analysed 30-minute window.



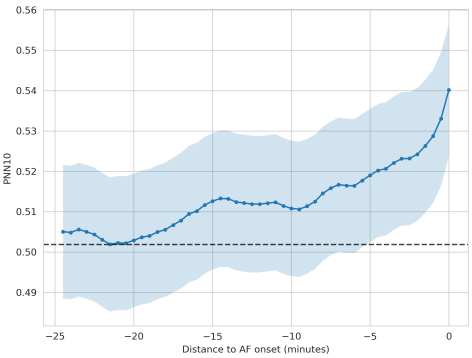
(a) RR intervals 20% percentile



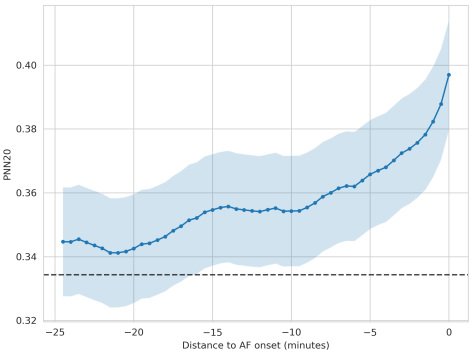
(b) RR intervals 50% percentile



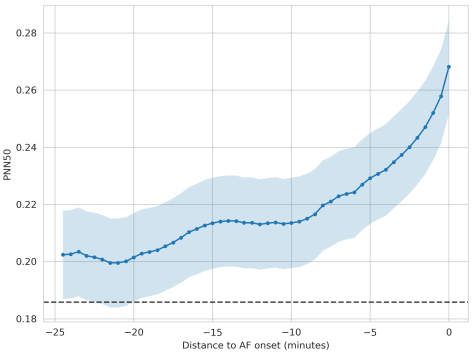
(c) RR intervals 80% percentile



(d) pNN10

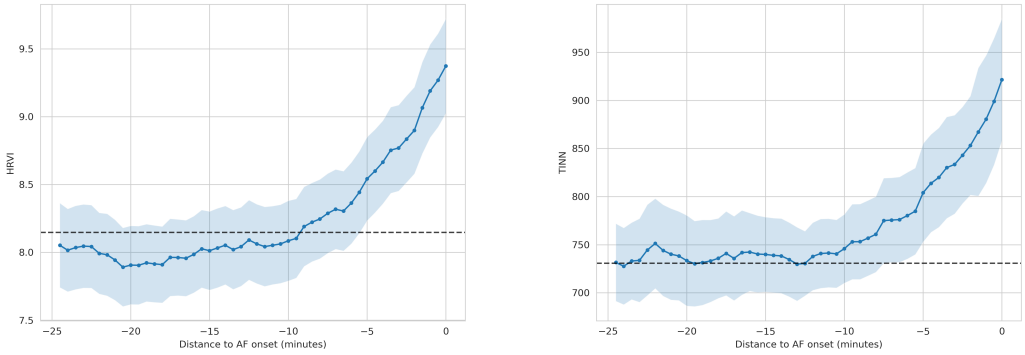


(e) pNN20



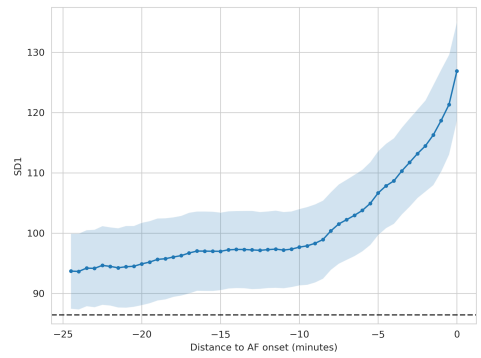
(f) pNN50

Figure A.3: Evolution of HRV temporal features before AF onset (3/3). The analysis was performed with a sliding window of 5 minutes and a step of 30 seconds. The main line corresponds to the mean of all selected windows. The 95% confidence interval is shown around the line. The baseline (black dotted line) represents the average value of the features in the 2-hour period preceding the analysed 30-minute window.

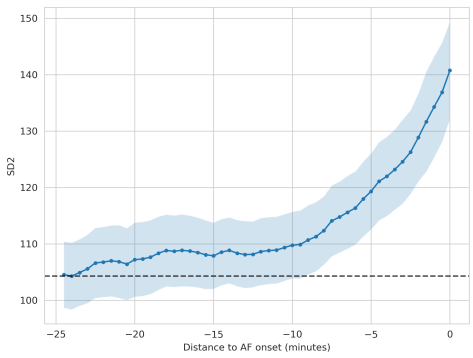


(a) HRVi (b) TINN

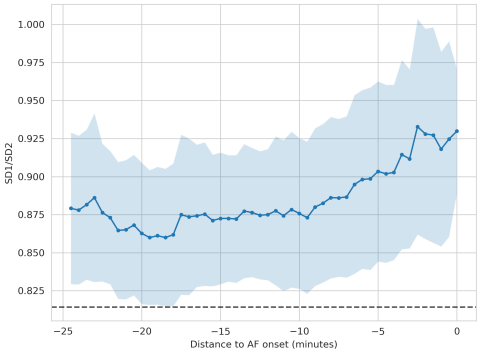
Figure A.4: Evolution of HRV histogram features before AF onset. The analysis was performed with a sliding window of 5 minutes and a step of 30 seconds. The main line corresponds to the mean of all selected windows. The 95% confidence interval is shown around the line. The baseline (black dotted line) represents the average value of the features in the 2-hour period preceding the analysed 30-minute window.



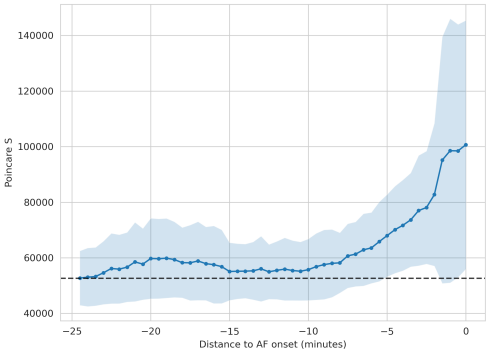
(a) Poincaré SD1



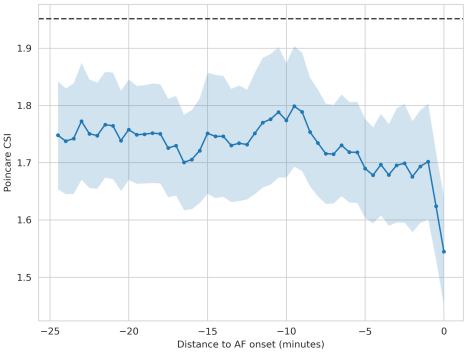
(b) Poincaré SD2



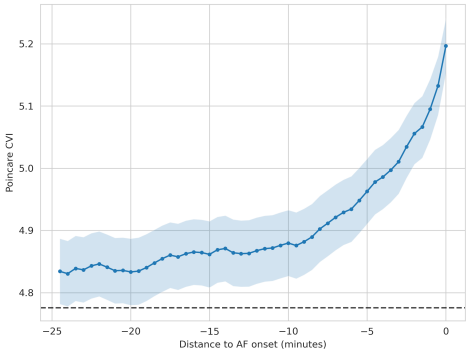
(c) Poincaré SD1/SD2



(d) Poincaré Surface

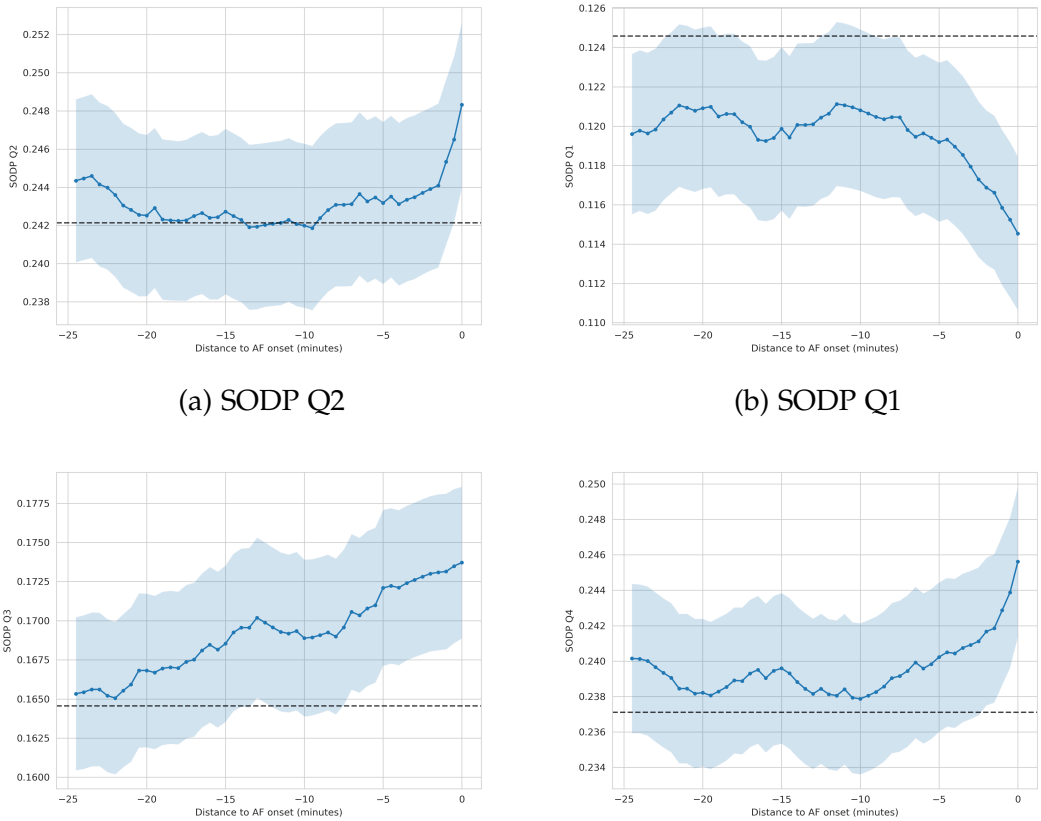


(e) Poincaré Cardiac Sympathetic Index (CSI)



(f) Poincaré Cardiac Vagal Index (CVI) Parasympathetic Nervous System

Figure A.5: Evolution of HRV Poincaré plot features before AF onset. The analysis was performed with a sliding window of 5 minutes and a step of 30 seconds. The main line corresponds to the mean of all selected windows. The 95% confidence interval is shown around the line. The baseline (black dotted line) represents the average value of the features in the 2-hour period preceding the analysed 30-minute window.



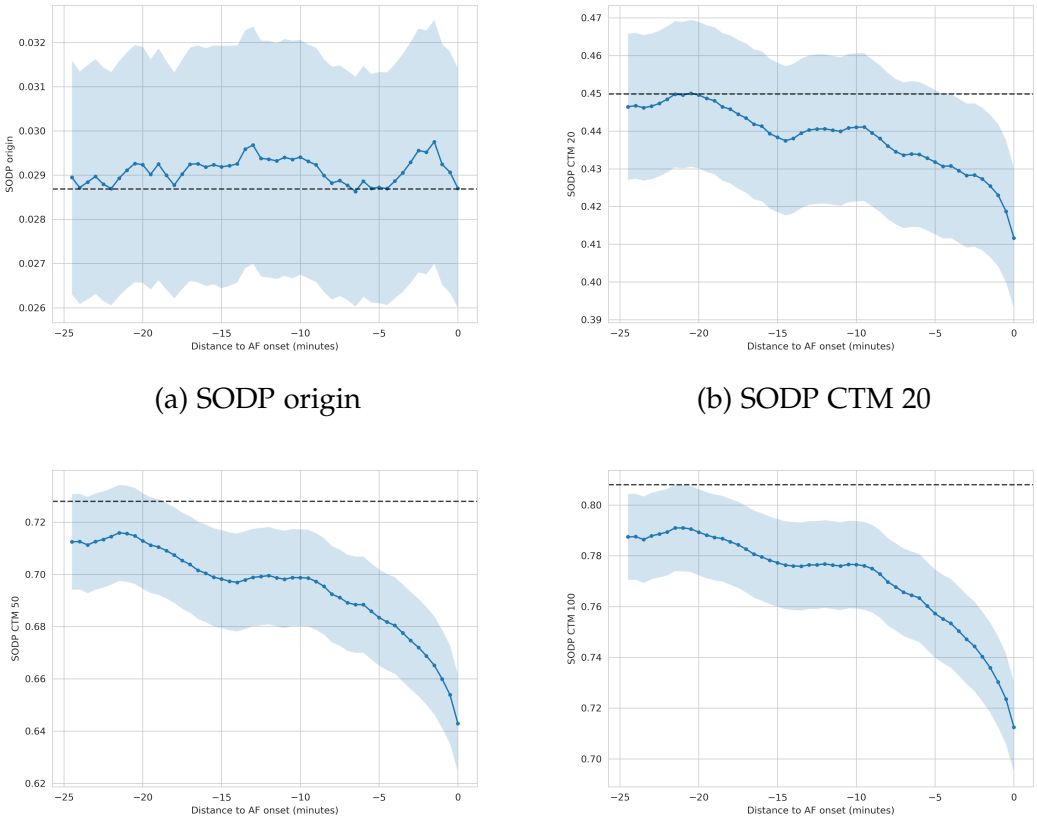
(a) SODP Q2

(b) SODP Q1

(c) SODP Q3

(d) SODP Q4

Figure A.6: Evolution of HRV SODP features before AF onset The analysis was performed with a sliding window of 5 minutes and a step of 30 seconds. The main line corresponds to the mean of all selected windows. The 95% confidence interval is shown around the line. The baseline (black dotted line) represents the average value of the features in the 2-hour period preceding the analysed 30-minute window.



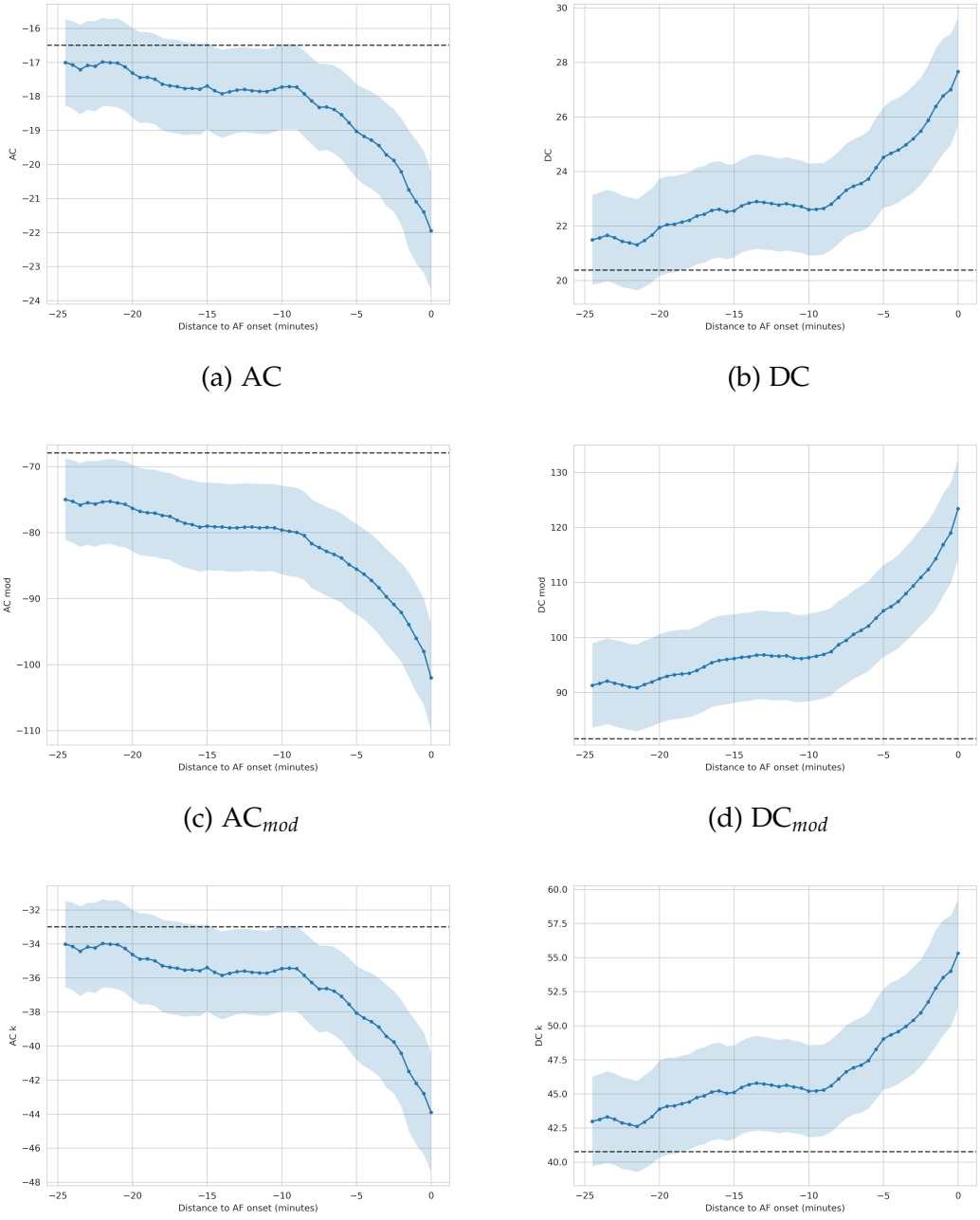
(a) SODP origin

(b) SODP CTM 20

(c) SODP CTM 50

(d) SODP CTM 100

Figure A.7: Evolution of HRV SODP quadrant features before AF onset. The analysis was performed with a sliding window of 5 minutes and a step of 30 seconds. The main line corresponds to the mean of all selected windows. The 95% confidence interval is shown around the line. The baseline (black dotted line) represents the average value of the features in the 2-hour period preceding the analysed 30-minute window.

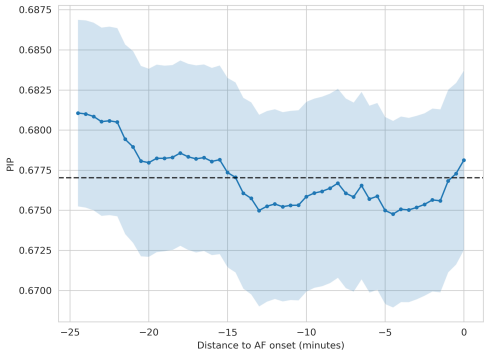


(a) AC (b) DC

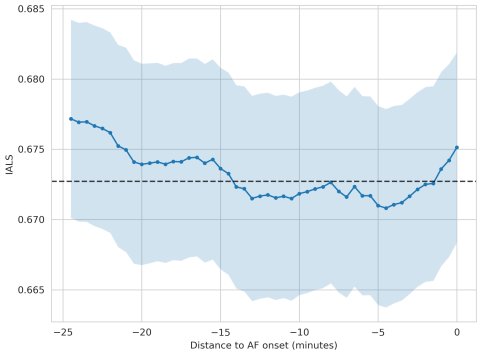
(c) AC_{mod} (d) DC_{mod}

(e) AC_k (f) DC_k

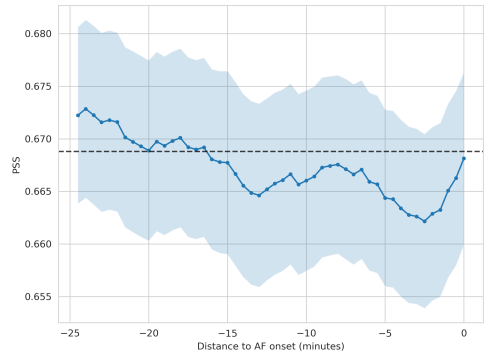
Figure A.8: Evolution of HRV AC and DC features before AF onset The analysis was performed with a sliding window of 5 minutes and a step of 30 seconds. The main line corresponds to the mean of all selected windows. The 95% confidence interval is shown around the line. The baseline (black dotted line) represents the average value of the features in the 2-hour period preceding the analysed 30-minute window.



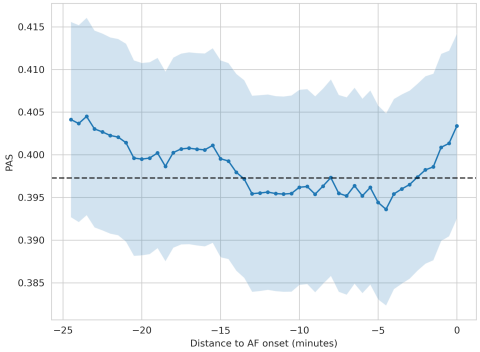
(a) Percentage of Inflection Points of the RR intervals series (PIP)



(b) Inverse of the Average Length of the acceleration/deceleration Segments (IALS)

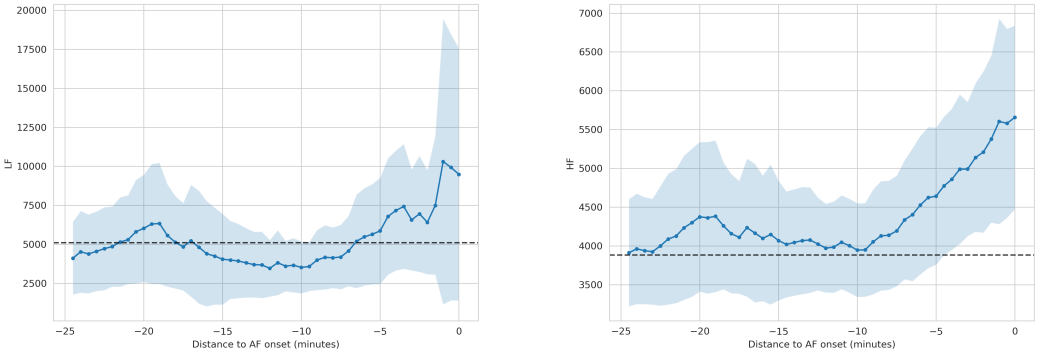


(c) Percentage of short segments (PSS)



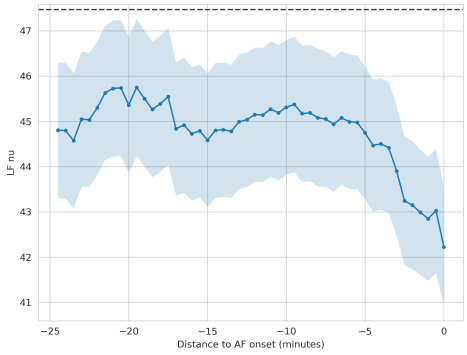
(d) Percentage of RR intervals in Alternation Segments (PAS)

Figure A.9: Evolution of HRF features before AF onset. The analysis was performed with a sliding window of 5 minutes and a step of 30 seconds. The main line corresponds to the mean of all selected windows. The 95% confidence interval is shown around the line. The baseline (black dotted line) represents the average value of the features in the 2-hour period preceding the analysed 30-minute window.

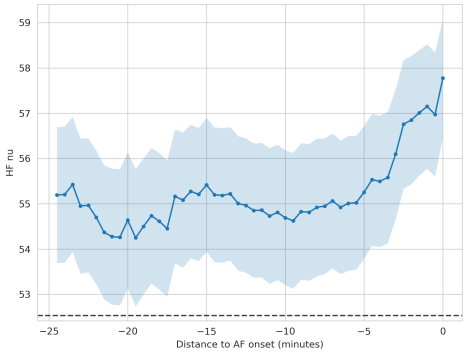


(a) LF

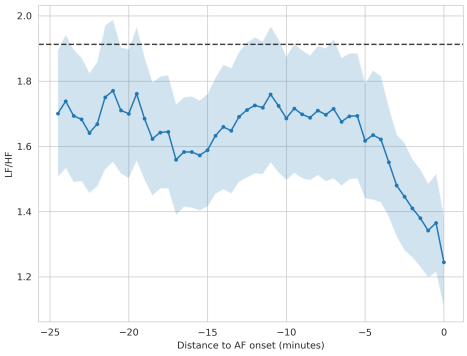
(b) HF



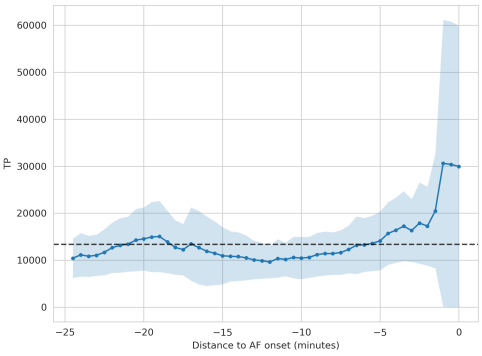
(c) LFnu



(d) HFnu



(e) LF/HF



(f) Total Power

Figure A.10: Evolution of HRV frequency features before AF onset. The analysis was performed with a sliding window of 5 minutes and a step of 30 seconds. The main line corresponds to the mean of all selected windows. The 95% confidence interval is shown around the line. The baseline (black dotted line) represents the average value of the features in the 2-hour period preceding the analysed 30-minute window.

A.2 Features correlation

In this section, we present the correlations matrices for AF onset forecast. The first matrix, Figure A.11, is the correlation matrix for HRV parameters, the second matrix, Figure A.12 is the correlation matrix for ECGMV parameters.

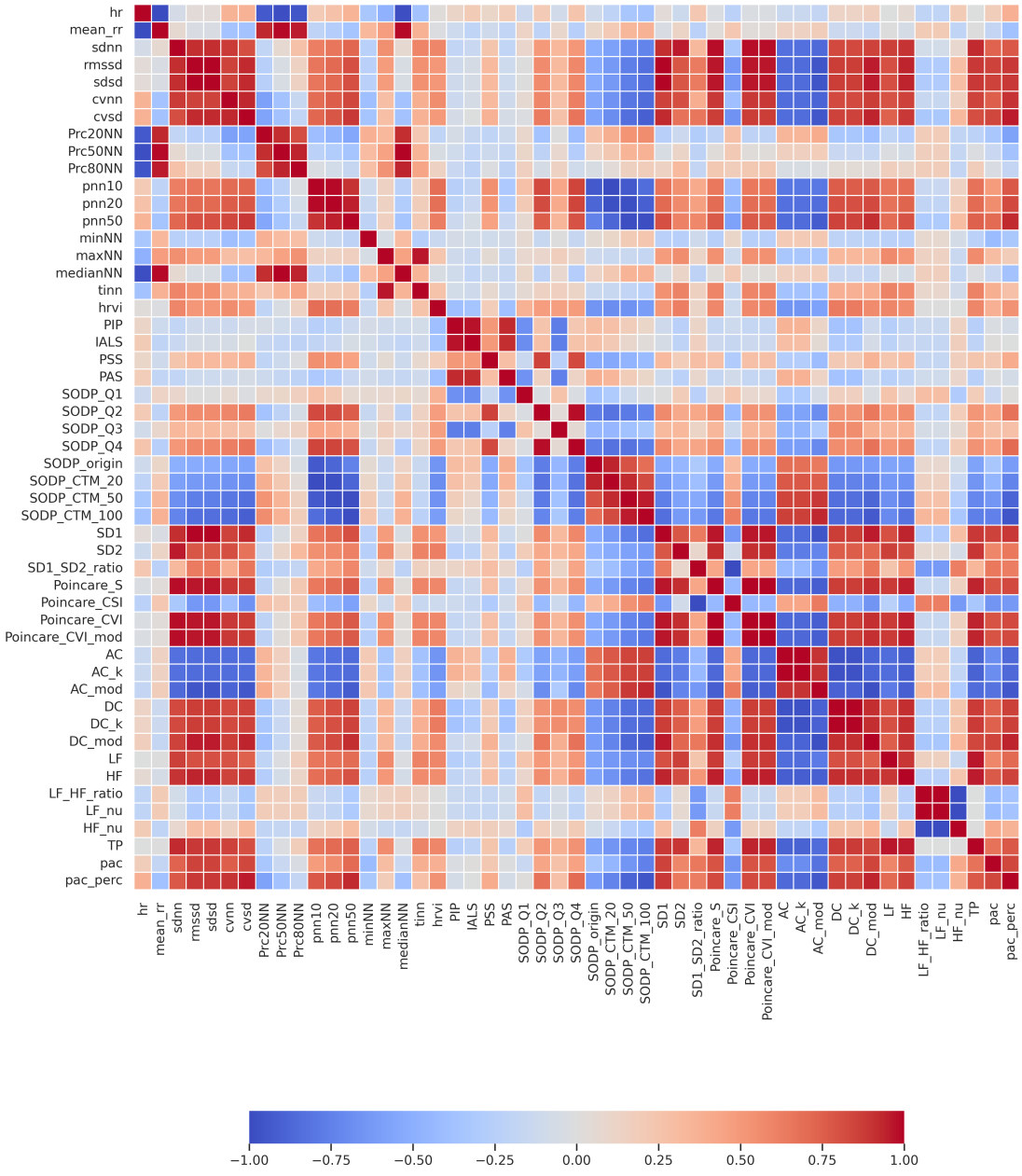


Figure A.11: Pairwise correlation matrix between HRV features used for AF onset forecast. Correlation coefficients were computed using Spearman correlation coefficient.

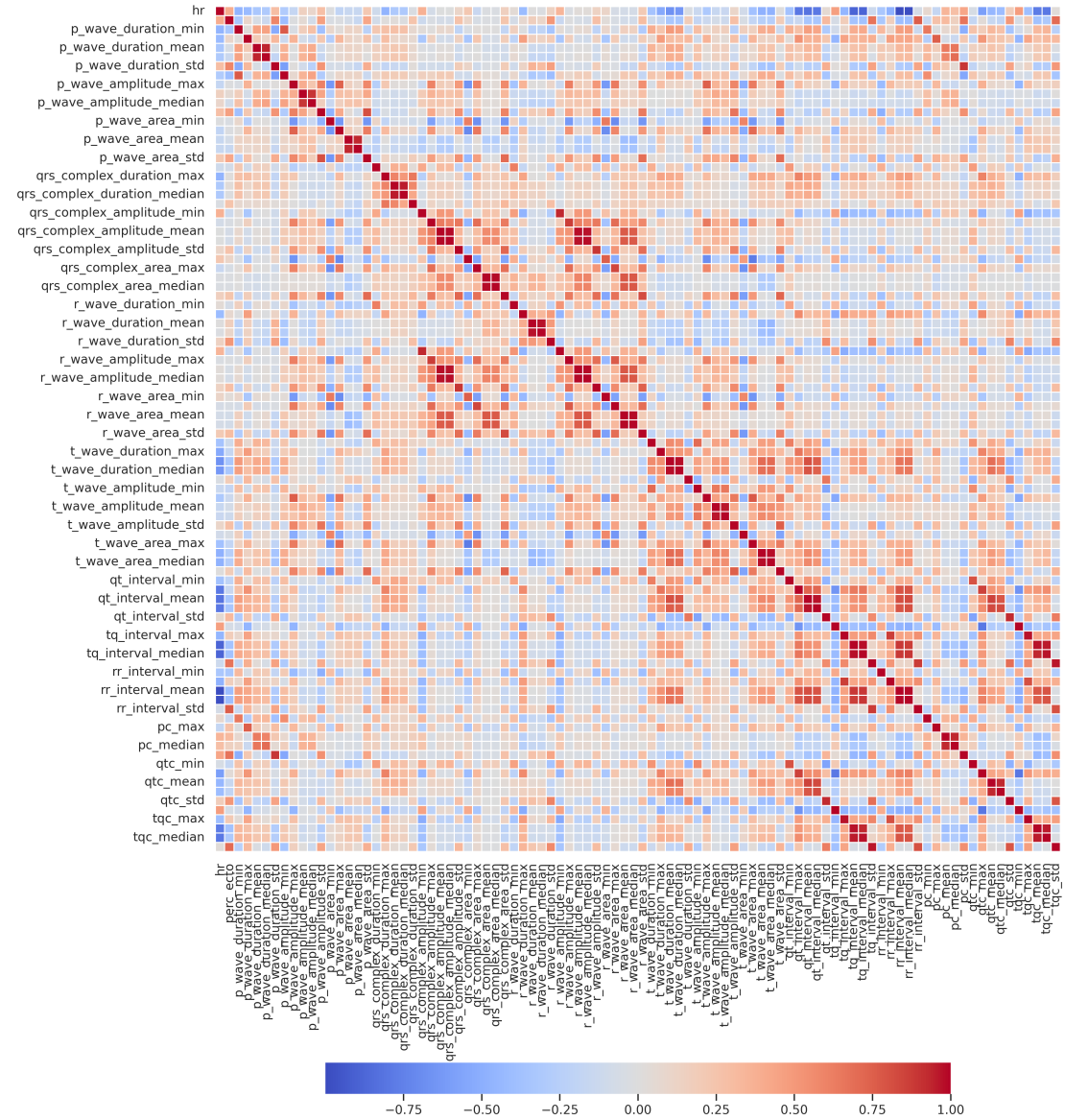


Figure A.12: Pairwise correlation matrix between ECGMV features used for AF onset forecast. Correlation coefficients were computed using Spearman correlation coefficient.

A.3 AF onset forecast evolution using HRV

The AUROC metric used in Figure 4.16 is presented in Tables A.1 to A.3. The AUPRC is added as a second metric, showing the same trend.

Table A.1: Results of model predictions between pre-AF windows and NSR windows. Each line represents the mean results of 10 repetitions of 10-fold cross-validation, for each repetition a new dataset was created by randomly selecting NSR windows to balance the dataset with pre-AF windows. Results for experiments between 0 and 20 minutes before AF onset.

Selected pre-AF windows ending (seconds before AF onset)		AUROC	AUPRC
0	-30	0.714 (0.692–0.735)	0.697 (0.671–0.724)
-30	-60	0.693 (0.668–0.717)	0.675 (0.649–0.701)
-60	-90	0.681 (0.660–0.701)	0.663 (0.641–0.686)
-90	-120	0.665 (0.649–0.682)	0.648 (0.631–0.665)
-120	-150	0.664 (0.644–0.685)	0.647 (0.624–0.669)
-150	-180	0.658 (0.640–0.675)	0.637 (0.616–0.657)
-180	-210	0.649 (0.632–0.667)	0.631 (0.612–0.650)
-210	-240	0.640 (0.622–0.659)	0.622 (0.600–0.644)
-240	-270	0.635 (0.621–0.650)	0.615 (0.597–0.633)
-270	-300	0.634 (0.618–0.650)	0.617 (0.597–0.637)
-300	-330	0.631 (0.615–0.647)	0.615 (0.595–0.634)
-330	-360	0.628 (0.610–0.645)	0.611 (0.590–0.632)
-360	-390	0.630 (0.611–0.648)	0.617 (0.595–0.639)
-390	-420	0.631 (0.611–0.652)	0.619 (0.597–0.640)
-420	-450	0.623 (0.604–0.643)	0.607 (0.589–0.626)
-450	-480	0.623 (0.606–0.640)	0.610 (0.591–0.629)
-480	-510	0.616 (0.598–0.635)	0.601 (0.581–0.622)
-510	-540	0.610 (0.592–0.629)	0.596 (0.576–0.615)
-540	-570	0.612 (0.594–0.630)	0.602 (0.584–0.621)
-570	-600	0.607 (0.583–0.631)	0.597 (0.577–0.617)
-600	-630	0.610 (0.588–0.633)	0.598 (0.577–0.618)
-630	-660	0.608 (0.586–0.630)	0.589 (0.568–0.611)
-660	-690	0.603 (0.581–0.625)	0.586 (0.563–0.609)
-690	-720	0.604 (0.580–0.629)	0.587 (0.561–0.613)
-720	-750	0.613 (0.588–0.638)	0.594 (0.570–0.619)
-750	-780	0.607 (0.583–0.632)	0.583 (0.560–0.606)
-780	-810	0.615 (0.591–0.639)	0.594 (0.570–0.617)
-810	-840	0.610 (0.587–0.633)	0.595 (0.572–0.617)
-840	-870	0.611 (0.589–0.632)	0.595 (0.574–0.617)
-870	-900	0.607 (0.585–0.630)	0.590 (0.568–0.612)
-900	-930	0.609 (0.587–0.631)	0.590 (0.567–0.613)
-930	-960	0.610 (0.583–0.637)	0.587 (0.559–0.616)
-960	-990	0.616 (0.591–0.641)	0.590 (0.561–0.618)
-990	-1020	0.618 (0.591–0.645)	0.597 (0.568–0.626)
-1020	-1050	0.607 (0.581–0.634)	0.587 (0.559–0.614)
-1050	-1080	0.599 (0.571–0.628)	0.585 (0.556–0.613)
-1080	-1110	0.605 (0.579–0.631)	0.588 (0.564–0.612)
-1110	-1140	0.608 (0.582–0.634)	0.586 (0.560–0.612)
-1140	-1170	0.608 (0.583–0.633)	0.587 (0.561–0.612)
-1170	-1200	0.616 (0.595–0.636)	0.587 (0.566–0.607)

Table A.2: Results of model predictions between pre-AF windows and NSR windows (continued 2/3). Results for experiments between 20 and 40 minutes before AF onset.

Selected windows ending (seconds before AF onset)		AUROC	AUPRC
-1200	-1230	0.617 (0.596–0.638)	0.588 (0.568–0.608)
-1230	-1260	0.609 (0.591–0.627)	0.582 (0.563–0.602)
-1260	-1290	0.608 (0.591–0.625)	0.583 (0.565–0.601)
-1290	-1320	0.610 (0.591–0.629)	0.581 (0.561–0.601)
-1320	-1350	0.607 (0.587–0.626)	0.582 (0.560–0.604)
-1350	-1380	0.603 (0.581–0.625)	0.581 (0.559–0.603)
-1380	-1410	0.600 (0.577–0.623)	0.578 (0.556–0.600)
-1410	-1440	0.593 (0.570–0.615)	0.571 (0.549–0.593)
-1440	-1470	0.594 (0.572–0.616)	0.574 (0.553–0.595)
-1470	-1500	0.594 (0.573–0.616)	0.577 (0.558–0.596)
-1500	-1530	0.600 (0.577–0.623)	0.583 (0.562–0.604)
-1530	-1560	0.595 (0.572–0.618)	0.580 (0.558–0.602)
-1560	-1590	0.596 (0.569–0.623)	0.585 (0.560–0.610)
-1590	-1620	0.598 (0.571–0.626)	0.584 (0.555–0.612)
-1620	-1650	0.613 (0.585–0.641)	0.597 (0.567–0.626)
-1650	-1680	0.620 (0.602–0.637)	0.603 (0.581–0.625)
-1680	-1710	0.616 (0.591–0.640)	0.598 (0.573–0.623)
-1710	-1740	0.600 (0.566–0.633)	0.577 (0.546–0.608)
-1740	-1770	0.600 (0.566–0.633)	0.578 (0.548–0.607)
-1770	-1800	0.605 (0.578–0.631)	0.586 (0.560–0.613)
-1800	-1830	0.602 (0.580–0.623)	0.582 (0.559–0.604)
-1830	-1860	0.595 (0.575–0.615)	0.582 (0.563–0.601)
-1860	-1890	0.597 (0.574–0.620)	0.577 (0.555–0.600)
-1890	-1920	0.595 (0.571–0.619)	0.577 (0.552–0.602)
-1920	-1950	0.591 (0.573–0.609)	0.576 (0.557–0.595)
-1950	-1980	0.584 (0.548–0.620)	0.568 (0.535–0.601)
-1980	-2010	0.586 (0.559–0.613)	0.571 (0.547–0.595)
-2010	-2040	0.582 (0.556–0.609)	0.563 (0.542–0.585)
-2040	-2070	0.590 (0.559–0.621)	0.567 (0.538–0.595)
-2070	-2100	0.585 (0.564–0.607)	0.566 (0.542–0.590)
-2100	-2130	0.592 (0.570–0.613)	0.569 (0.546–0.592)
-2130	-2160	0.588 (0.567–0.609)	0.568 (0.549–0.587)
-2160	-2190	0.592 (0.561–0.622)	0.574 (0.547–0.602)
-2190	-2220	0.589 (0.565–0.613)	0.573 (0.553–0.592)
-2220	-2250	0.589 (0.568–0.610)	0.571 (0.554–0.589)
-2250	-2280	0.579 (0.549–0.610)	0.566 (0.540–0.592)
-2280	-2310	0.594 (0.566–0.621)	0.577 (0.554–0.599)
-2310	-2340	0.595 (0.572–0.618)	0.575 (0.551–0.600)
-2340	-2370	0.587 (0.566–0.607)	0.562 (0.546–0.579)
-2370	-2400	0.586 (0.566–0.605)	0.564 (0.544–0.585)

Table A.3: Results of models predictions between pre-AF windows and NSR windows (continued 3/3). Results for experiments between 40 and 60 minutes before AF onset.

Selected windows ending (seconds before AF onset)		AUROC	AUPRC
-2400	-2430	0.587 (0.564–0.611)	0.566 (0.545–0.587)
-2430	-2460	0.587 (0.556–0.618)	0.567 (0.541–0.593)
-2460	-2490	0.588 (0.568–0.608)	0.562 (0.543–0.580)
-2490	-2520	0.592 (0.562–0.622)	0.566 (0.541–0.591)
-2520	-2550	0.583 (0.566–0.600)	0.560 (0.542–0.578)
-2550	-2580	0.583 (0.564–0.602)	0.562 (0.545–0.580)
-2580	-2610	0.588 (0.571–0.604)	0.570 (0.551–0.588)
-2610	-2640	0.583 (0.566–0.600)	0.567 (0.549–0.585)
-2640	-2670	0.590 (0.567–0.612)	0.570 (0.545–0.594)
-2670	-2700	0.584 (0.556–0.612)	0.568 (0.540–0.596)
-2700	-2730	0.584 (0.561–0.608)	0.566 (0.543–0.590)
-2730	-2760	0.575 (0.552–0.598)	0.562 (0.541–0.583)
-2760	-2790	0.571 (0.546–0.596)	0.556 (0.531–0.582)
-2790	-2820	0.571 (0.552–0.590)	0.556 (0.535–0.576)
-2820	-2850	0.566 (0.540–0.592)	0.549 (0.525–0.573)
-2850	-2880	0.573 (0.547–0.599)	0.555 (0.531–0.580)
-2880	-2910	0.576 (0.555–0.597)	0.563 (0.540–0.586)
-2910	-2940	0.574 (0.551–0.597)	0.557 (0.534–0.579)
-2940	-2970	0.578 (0.556–0.600)	0.560 (0.537–0.584)
-2970	-3000	0.576 (0.553–0.599)	0.559 (0.535–0.583)
-3000	-3030	0.576 (0.553–0.598)	0.561 (0.537–0.585)
-3030	-3060	0.567 (0.531–0.604)	0.550 (0.518–0.583)
-3060	-3090	0.579 (0.554–0.603)	0.558 (0.534–0.582)
-3090	-3120	0.576 (0.550–0.601)	0.556 (0.536–0.577)
-3120	-3150	0.570 (0.542–0.598)	0.552 (0.525–0.578)
-3150	-3180	0.561 (0.535–0.586)	0.545 (0.523–0.568)
-3180	-3210	0.574 (0.544–0.604)	0.552 (0.524–0.581)
-3210	-3240	0.579 (0.551–0.607)	0.564 (0.540–0.587)
-3240	-3270	0.568 (0.543–0.594)	0.549 (0.526–0.572)
-3270	-3300	0.569 (0.545–0.592)	0.548 (0.526–0.570)
-3300	-3330	0.566 (0.545–0.588)	0.549 (0.532–0.565)
-3330	-3360	0.566 (0.546–0.587)	0.549 (0.529–0.568)
-3360	-3390	0.561 (0.535–0.587)	0.547 (0.526–0.567)
-3390	-3420	0.561 (0.543–0.580)	0.544 (0.529–0.560)
-3420	-3450	0.564 (0.544–0.584)	0.544 (0.523–0.564)
-3450	-3480	0.567 (0.542–0.592)	0.551 (0.527–0.575)
-3480	-3510	0.558 (0.526–0.590)	0.544 (0.518–0.571)
-3510	-3540	0.565 (0.543–0.586)	0.550 (0.530–0.571)
-3540	-3570	0.556 (0.531–0.580)	0.545 (0.524–0.566)
-3570	-3600	0.562 (0.539–0.586)	0.555 (0.534–0.576)

A.4 Model benchmark: threshold-based metrics

Threshold based metrics for:

- XGB model using HRV features computed from the 30-minute window - Table A.4,
- CNN model using 300 RR as input - Table A.5,
- RF model using HRV features computed from the 5-minute window - Table A.6.

Table A.4: Metrics for the XGB model using HRV features computed from the 30-minute window. Evaluation is at episode level.

Threshold	Accuracy	Sensitivity	Specificity	PPV	NPV	F1-score
0.1	0.518	1.000	0.000	0.518	nan	0.682
0.2	0.531	0.987	0.043	0.525	0.750	0.685
0.3	0.568	0.933	0.175	0.548	0.710	0.691
0.4	0.606	0.846	0.349	0.582	0.678	0.690
0.5	0.608	0.706	0.504	0.604	0.615	0.651
0.6	0.587	0.466	0.716	0.638	0.555	0.538
0.7	0.524	0.167	0.907	0.658	0.504	0.266
0.8	0.486	0.016	0.992	0.667	0.484	0.030
0.9	0.482	0.000	1.000	nan	0.482	nan

Table A.5: Metrics for the CNN model using 300 RR as input. Evaluation is at episode level.

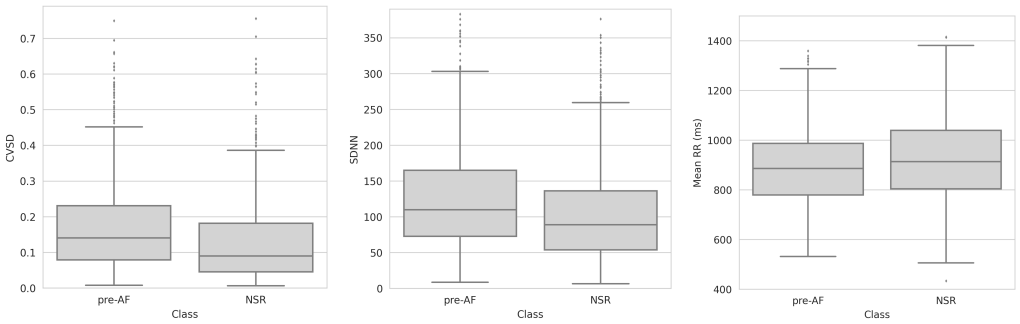
Threshold	Accuracy	Sensitivity	Specificity	PPV	NPV	F1-score
0.1	0.516	0.998	0.000	0.517	0.000	0.681
0.2	0.516	0.993	0.005	0.517	0.400	0.680
0.3	0.518	0.986	0.015	0.518	0.500	0.679
0.4	0.548	0.922	0.147	0.537	0.637	0.679
0.5	0.589	0.620	0.555	0.599	0.577	0.610
0.6	0.538	0.289	0.806	0.615	0.514	0.393
0.7	0.494	0.088	0.930	0.572	0.487	0.152
0.8	0.486	0.029	0.977	0.578	0.484	0.055
0.9	0.484	0.008	0.994	0.583	0.483	0.015

Table A.6: Metrics for the RF model using HRV features computed from the 5-minute window. Evaluation is at episode level.

Threshold	Accuracy	Sensitivity	Specificity	PPV	NPV	F1-score
0.1	0.518	1.000	0.000	0.518	nan	0.682
0.2	0.517	0.999	0.000	0.517	0.000	0.682
0.3	0.516	0.989	0.008	0.517	0.412	0.679
0.4	0.538	0.930	0.118	0.531	0.611	0.676
0.5	0.568	0.688	0.440	0.568	0.568	0.622
0.6	0.527	0.276	0.797	0.593	0.506	0.376
0.7	0.485	0.026	0.977	0.548	0.483	0.049
0.8	0.483	0.004	0.996	0.571	0.483	0.009
0.9	0.482	0.001	0.999	0.500	0.482	0.002

A.5 Top ranked parameters for AF onset forecast

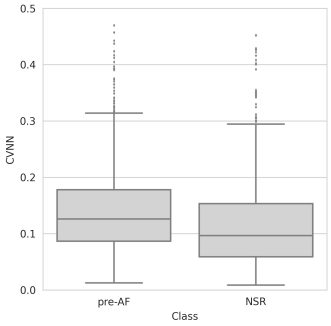
In Figure A.13 and Figure A.14 we present the top features selected by the RF and XGB models for forecasting AF onset.



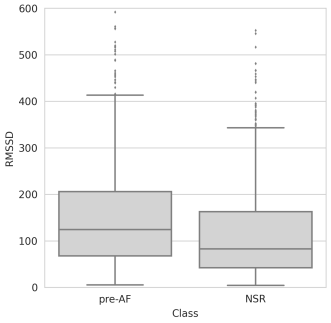
(a) CVSD

(b) SDNN

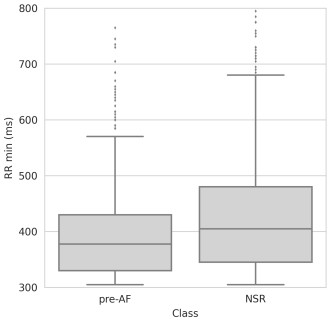
(c) Mean RR



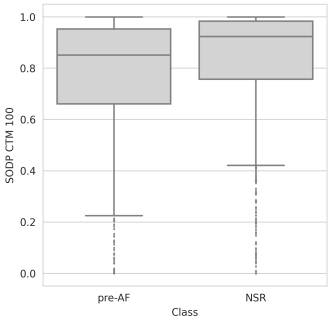
(d) CVNN



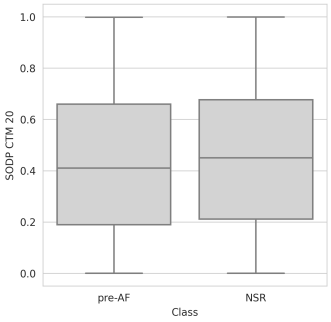
(e) RMSSD



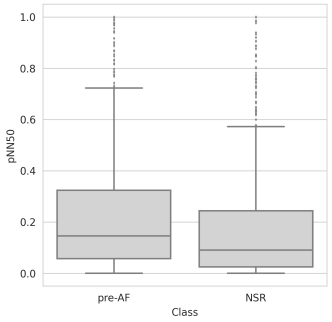
(f) Min RR



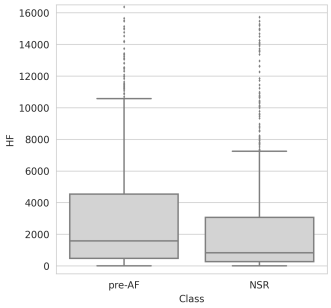
(g) SODP CTM 100



(h) SODP CTM 20

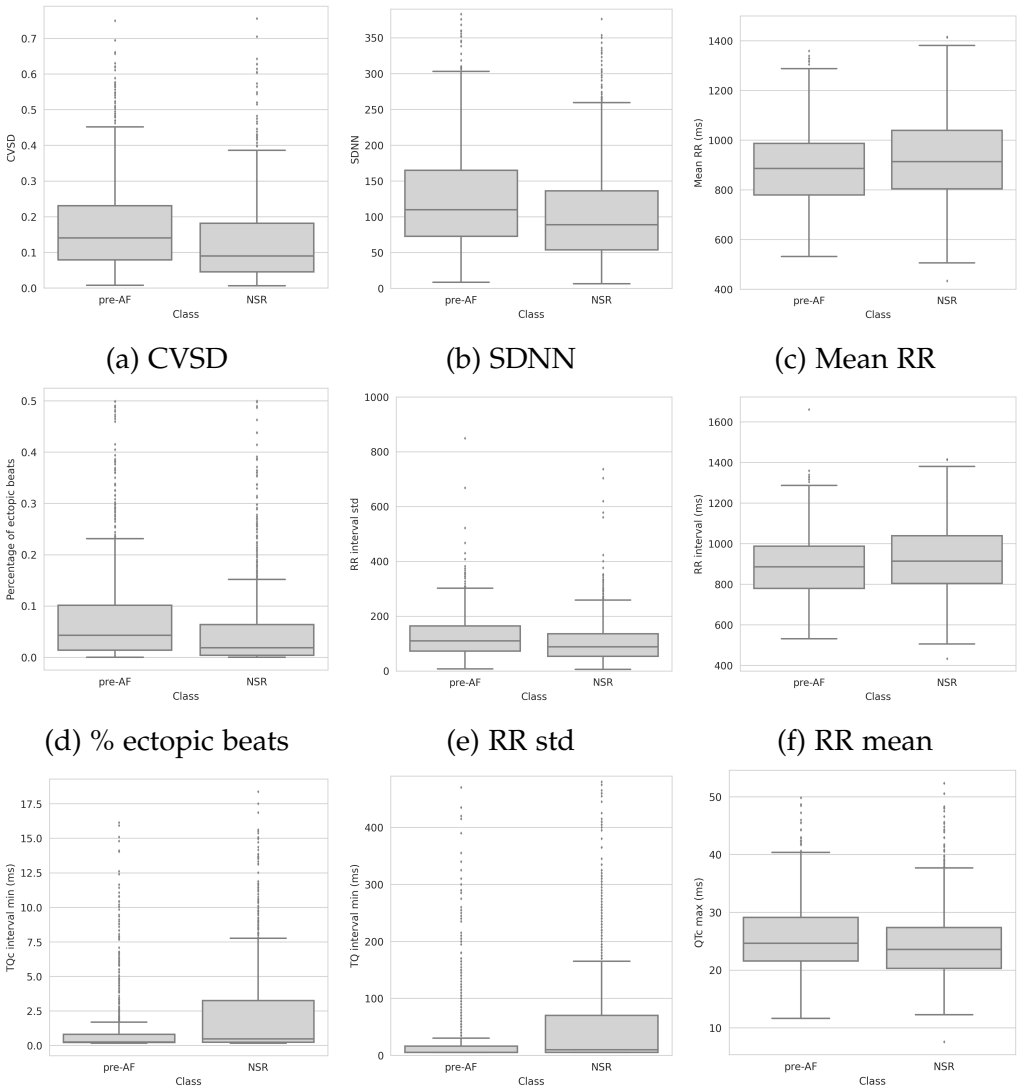


(i) pNN50



(j) HF

Figure A.13: Top ranked HRV features for AF onset forecast (1/2)



(a) CVSD (b) SDNN (c) Mean RR

(d) % ectopic beats (e) RR std (f) RR mean

(g) SODP CTM 100 (h) SODP CTM 20 (i) pNN50

Figure A.14: Top ranked ECG features for AF onset forecast (2/2)

Appendix B

Atrial fibrillation identification

B.1 Model benchmark: threshold-based metrics

Table B.1: Metrics for XGB model using HRV parameters from 5-minute ECG window. The evaluation is done at the recording level.

Threshold	Accuracy	Sensitivity	Specificity	PPV	NPV	F1-score
0.1	0.585	0.998	0.003	0.585	0.500	0.738
0.2	0.702	0.982	0.307	0.666	0.926	0.794
0.3	0.791	0.938	0.584	0.760	0.869	0.840
0.4	0.840	0.893	0.764	0.842	0.835	0.867
0.5	0.844	0.829	0.866	0.897	0.782	0.861
0.6	0.801	0.695	0.951	0.952	0.688	0.803
0.7	0.737	0.566	0.978	0.973	0.616	0.716
0.8	0.646	0.399	0.995	0.990	0.540	0.569
0.9	0.478	0.107	1.000	1.000	0.443	0.193

Table B.2: Metrics for XGB model using HRV parameters from 30-minute ECG window. The evaluation is done at the recording level.

Threshold	Accuracy	Sensitivity	Specificity	PPV	NPV	F1-score
0.1	0.772	0.934	0.545	0.743	0.854	0.828
0.2	0.797	0.905	0.647	0.783	0.828	0.839
0.3	0.812	0.879	0.718	0.814	0.809	0.846
0.4	0.817	0.844	0.778	0.843	0.780	0.844
0.5	0.827	0.813	0.847	0.882	0.763	0.846
0.6	0.824	0.778	0.888	0.907	0.740	0.838
0.7	0.815	0.739	0.921	0.929	0.715	0.823
0.8	0.792	0.683	0.945	0.946	0.679	0.793
0.9	0.760	0.611	0.970	0.966	0.639	0.749

B.2 Features correlation

In this section, we present the correlation matrix for AF risk identification. The matrix, Figure A.11, is the correlation matrix for HRV parameters for 1-hour windows.

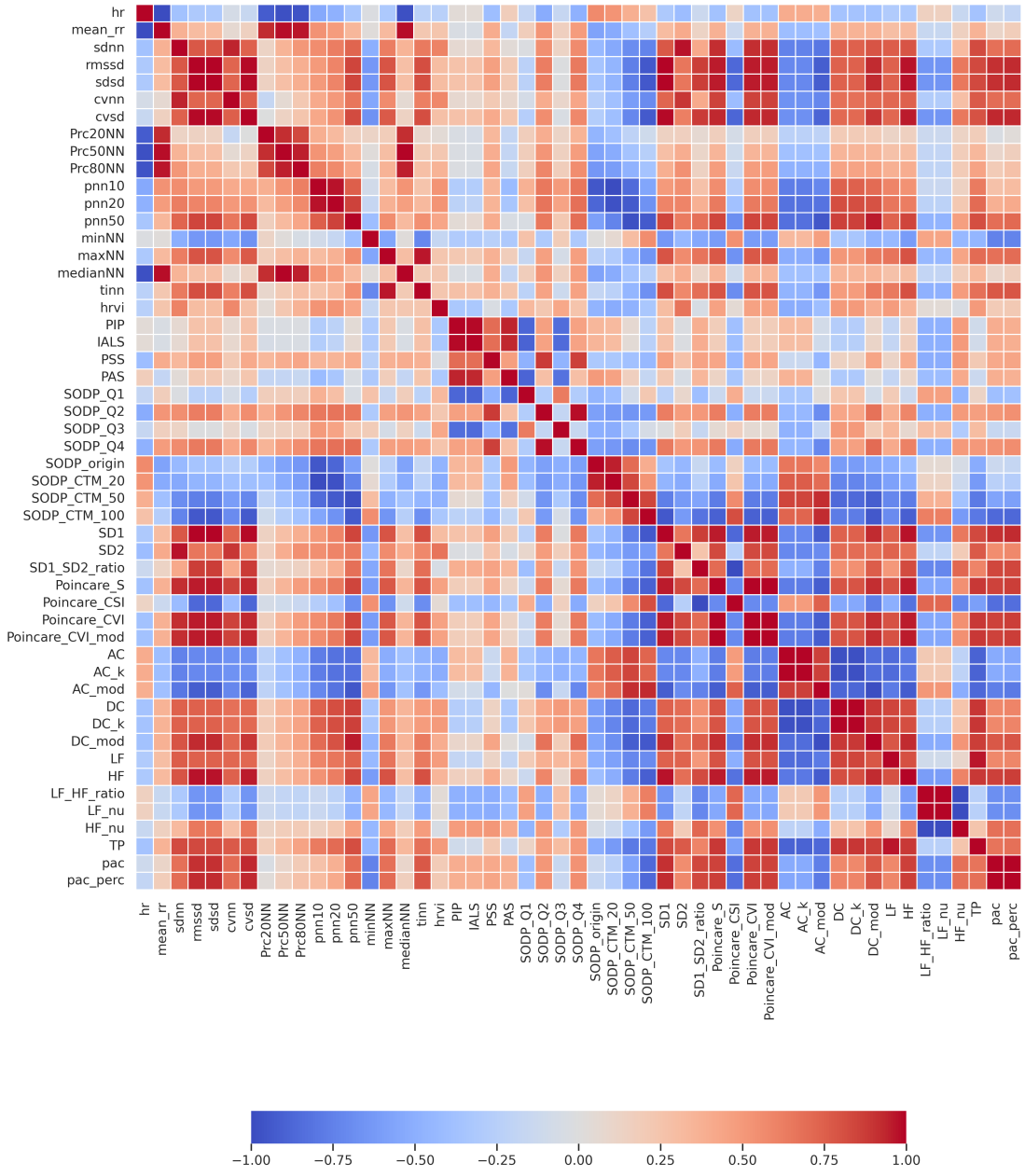


Figure B.1: Pairwise correlation matrix between HRV features used for AF risk identification. Correlation coefficients were computed using Spearman correlation coefficient.

Appendix C

Choice of hyperparameters

In this thesis, mainly in Chapters 3 to 5, the training of Machine Learning (ML) and Deep Learning (DL) algorithms required the choice of hyperparameters. In this appendix, we have grouped the hyperparameters to provide a summary of the hyperparameters across the work.

All the experiments in this work were done using Python 3.10 and 3.11. The experiments are listed below with their main hyperparameters for the Electrocardiogram (ECG) window selection and model. The random seed chosen in experiments is 42, and is used to seed Python random, NumPy, Scikit-learn, PyTorch and TensorFlow libraries.

We evaluated the models using cross-validation. The reported metrics were computed on all test sets from each fold. The 95% confidence intervals were computed on the test values.

C.1 IRIDIA-AF version 1 annotation evaluation

In Section 3.2.4, we evaluated the performance of a XGBoost (XGB) model and a Convolutionnal Neural Network (CNN) for the AF detection on the IRIDIA-AF version 1 annotations.

XGB

Window selection

- Database: IRIDIA-AF v1
- Window size: 300 RR

- Training window step: 100 RR (200 RR overlap)
- Testing window step: 10 RR (290 RR overlap)

XGB model

- XGBoost and Scikit-learn python library
- Objective function: binary logistic
- Number of trees: 150
- Maximum depth: 5

Evaluation

10-fold cross validation at patient level. Each fold is composed of a train set and a test set.

CNN

Window selection

- Database: IRIDIA-AF v1
- Window size: 8192 samples
- Training window step: 4096 samples
- Testing window step: 8192 samples

CNN model

- PyTorch library
- Architecture: Figure 3.9
- Epochs: 100
- Early stopping patience: 5 epochs
- Adam optimizer
- Learning rate: 10^{-4}

- Batch size: 32
- Binary cross-entropy loss function

Evaluation

Boostraping with 5 repetitions. Each repetition is composed of a train set, a validation set and a test set.

C.2 IRIDIA-AF version 2 annotation evaluation

In Section 3.4, we evaluated the performance of four ML models and a CNN for the AF detection on the IRIDIA-AF version 2 annotations.

Models using HRV

Windows selection

- Database: IRIDIA-AF v2
- Window size: 300 RR
- Training window step: 100 RR

Logistic Regression

- Scikit-learn library

Decision Tree

- Scikit-learn library
- No max depth

Random Forest

- Scikit-learn library
- Number of trees: 200
- Maximum depth: 10

XGBoost

- XGBoost and Scikit-learn library
- Number of trees: 100
- Maximum depth: 5

Evaluation

The models were evaluated using 10-fold cross-validation at patient level, with patient sorted by recording date. Random Forest (RF) and XGB were also evaluated using a 4-fold spacial cross-validation, with the four centres. For both cross-validation, each fold is composed of a train and a test split.

CNN

Window selection

- Database: IRIDIA-AF v2
- Window size: 8192 samples
- Window step: 8192 samples

CNN model

- PyTorch library
- Architecture Figure 3.9
- Epochs: 100 (with 1500 steps per epoch)
- Early stopping patience: 3 epochs
- Adam optimizer
- Learning rate: 10^{-4}
- Learning rate decay patience: 2 epochs
- Batch size: 128
- Binary cross-entropy loss function

Evaluation

The model was evaluated using 10-fold cross-validation at patient level, with patient sorted by recording date. Each fold is composed of a train set, a validation set for early stopping and a test split.

C.3 Evolution of predictions before AF onset

In Section 4.4, we evaluated the performance of a RF model in the binary classification between windows close to the AF onset and distant from the AF onset.

Window selection

- Database: IRIDIA-AF v2
- Window size: 5-minute ECG window
- pre-AF window distance: 0 to 1 hour
- pre-AF window distance step: 30 seconds
- NSR window distance from AF: 2 hours

RF Model

- Scikit-learn library
- Number of trees: 100
- Maximum depth: 10

Evaluation

For each distance, the model was evaluated using 10 repetitions for each 10-fold cross-validation at patient level. Each fold is composed of a train set and a test split.

C.4 Comparison of models for AF onset forecast

In Section 4.5, we evaluated multiple ML and DL models on four types of input: ECG, ECGMV, RR intervals and HRV.

Window selection

- Database: IRIDIA-AF v2
- Window size: 30 minutes
- AF window distance: 0
- NSR window distance: 2 hours

ECG model

CNN and ResNet models

- PyTorch library
- Architecture CNN Figure 4.19
- Architecture ResNet Figure 4.20
- Epochs: 100
- Early stopping patience: 5 epochs
- Adam optimizer
- Learning rate: 10^{-5}
- Learning rate decay patience: 2 epochs
- Batch size: 32
- Binary cross-entropy loss function

Evaluation

Temporal 10-fold cross-validation at patient level. Each fold is composed of a train set, a validation set for the early stopping and a test set.

C.4.1 ECGMV model

RF model

- Scikit-learn library
- Number of trees: 200
- Maximum depth: 10

XGB

- XGBoost and Scikit-learn library
- Number of trees: 200
- Maximum depth: 10

Evaluation

Temporal 10-fold cross-validation at patient level. Each fold is composed of a train set and a test set.

RR interval model

CNN, ResNet and CNN-RNN models

- PyTorch library
- Architecture CNN Figure 4.19
- Architecture ResNet Figure 4.20
- Architecture CNN-RNN Figure 4.21
- Epochs: 100
- Early stopping patience: 5 epochs
- Adam optimizer
- Learning rate: 10^{-5}
- Learning rate decay patience: 2 epochs

- Batch size: 32
- Binary cross-entropy loss function

EfficientNet models

- Window size: 1 minute
- Window step: 30 seconds
- Embedding dimension: 3
- Embedding lag: 2
- Image: 224×224
- Keras and Tensorflow library
- Architecture EfficientNetV2S
- Epochs: 100
- Early stopping patience: 5 epochs
- Adam optimizer
- Learning rate: 10^{-4}
- Learning rate decay patience: 2 epochs
- Batch size: 32
- Binary cross-entropy loss function

Evaluation

Temporal 10-fold cross-validation at patient level. Each fold is composed of a train set, a validation set for the early stopping and a test set.

HRV

RF model

- Scikit-learn library
- Number of trees: 200
- Maximum depth: 10

XGB

- XGBoost and Scikit-learn library
- Number of trees: 200
- Maximum depth: 10

Evaluation

Temporal 10-fold cross-validation at patient level. Each fold is composed of a train set and a test set.

C.5 AF onset forecast on complete recording

In Section 4.5.3, we evaluated the performance of a XGB model on complete recording. A new model is trained for each patient in the dataset. The train set is composed of all other patient in the database.

Window selection

- Database: IRIDIA-AF v2
- Window size: 5 minutes
- Window step: 30 seconds

XGB

- XGBoost and Scikit-learn library
- Number of trees: 200

- Maximum depth: 10

C.6 AF identification

In Section 5.2, we evaluated the performance of ML and DL model for AF identification.

HRV models

Window selection

- Database: IRIDIA-AF v2
- Window size: 5 minutes
- Window step: 5 minutes

Random Forest

- Scikit-learn library
- Number of trees: 200
- Maximum depth: 10

XGB

- XGBoost and Scikit-learn library
- Number of trees: 200
- Maximum depth: 10

Evaluation

Temporal 10-fold cross-validation at patient level. Each fold is composed of a train set and a test set.

ECG Model

Window selection

- Window size: 6144 samples
- Window step: 2048 samples

CNN-RNN and CNN models

- PyTorch library
- Architecture CNN Figure 3.9
- Architecture CNN-RNN Figure 4.21
- Epochs: 100
- Early stopping patience: 3 epochs
- Adam optimizer
- Learning rate: 10^{-4}
- Learning rate decay patience: 2 epochs
- Batch size: 32
- Binary cross-entropy loss function

Evaluation

Temporal 10-fold cross-validation at patient level. Each fold is composed of a train set, a validation set for the early stopping and a test set.

Appendix D

Scientific Communications During the Thesis

Peer-reviewed publications in conference proceedings

Cédric Gilon, Jean-Marie Grégoire, Jérome Hellinckx, Stéphane Carlier, and Hugues Bersini. "Reproducibility of machine learning models for paroxysmal atrial fibrillation onset prediction". In: *Computer in Cardiology* (2022).

Cédric Gilon, Jean-Marie Grégoire, and Hugues Bersini. "Forecast of paroxysmal atrial fibrillation using a deep neural network". In: *2020 International Joint Conference on Neural Networks (IJCNN)* pp. 1–7 (2020).

Peer-reviewed journal publications

Cédric Gilon, Jean-Marie Grégoire, Marianne Mathieu, Stéphane Carlier, and Hugues Bersini. "IRIDIA-AF, a large paroxysmal atrial fibrillation long-term electrocardiogram monitoring database". In: *Scientific Data* 10.1, publisher: Nature Publishing Group, pp. 1–10 (2023).

Jean-Marie Grégoire, and **Cédric Gilon**. "Assessing the Autonomic Nervous System using ECG recordings". In: *Asian Hospital & Healthcare Management* 61 (2023).

Jean-Marie Grégoire, **Cédric Gilon**, Stéphane Carlier, and Hugues Bersini. "Autonomic nervous system assessment using heart rate variability". In: *Acta Cardiologica* 78.6, pp. 648–662 (2023).

Jean-Marie Grégoire, **Cédric Gilon**, Stéphane Carlier, and Hugues Bersini. "The potential of artificial intelligence in medical decision making". In: *Revue Medicale de Bruxelles* 43.3, pp. 265–273 (2022).

Jean-Marie Grégoire, **Cédric Gilon**, Stéphane Carlier, and Hugues Bersini. "Role of the autonomic nervous system and premature atrial contractions in short-term paroxysmal atrial fibrillation forecasting: Insights from machine learning models". en. In: *Archives of Cardiovascular Diseases* 115.6-7, pp. 377–387 (2022).

Database

Cédric Gilon, Jean-Marie Grégoire, Marianne Mathieu, Stéphane Carlier, and Hugues Bersini. (2023). IRIDIA-AF, a large paroxysmal atrial fibrillation long-term electrocardiogram monitoring database (1.0.1) [Data set]. Zenodo. `https://doi.org/10.5281/zenodo.8186845`

Software

Cédric Gilon. Codebase for "IRIDIA-AF, a large paroxysmal atrial fibrillation long-term electrocardiogram monitoring database". `https://github.com/cedricgilon/iridia-af`

Cédric Gilon. Codebase for "Reproducibility of machine learning models for paroxysmal atrial fibrillation onset prediction". `https://github.com/cedricgilon/paf-challenge-reproducibility`

Cédric Gilon. Codebase for "Forecast of paroxysmal atrial fibrillation using a deep neural network". `https://github.com/cedricgilon/af-forecast-dnn`

Communications during conferences with scientific selection committee

2024

Cédric Gilon, Jean-Marie Grégoire, Hugues Bersini, Laurent Groben, Thomas Nguyen, Pascal Godart, and Stéphane Carlier. "Deep learning to identify patients who will present an AF onset within 24 hours". Journées Européennes de la Société Française de Cardiologie, Paris, France (2024)

2023

Cédric Gilon, Jean-Marie Grégoire, Jérome Hellinckx, Stéphane Carlier, and Hugues Bersini. "Reproducibility of machine learning models for paroxysmal atrial fibrillation onset forecast". Journées Européennes de la Société Française de Cardiologie, Paris, France (2023)

2022

Cédric Gilon, Jean-Marie Grégoire, Adriana Sibru, Hugues Bersini, and Stéphane Carlier. "Short-term atrial fibrillation onset forecast using geometrical features with machine learning". Belgian Heart Rhythm Meeting (2022)

Other communications

General audience (non peer-review) journal publication

Jean-Marie Grégoire, **Cédric Gilon**, Hugues Bersini, and Stéphane Carlier. "Médecine de précision et cardiologie prédictive: l'exemple de la fibrillation auriculaire", In: *Vaisseaux, Coeur, Poumons*, Vol 28, N°7, p10-15 (2023).

Jean-Marie Grégoire, **Cédric Gilon**, Hugues Bersini, and Stéphane Carlier. "Intelligence artificielle et ECG : revue critique", In: *Journal de cardiologie* pp. 19-30 (2021).

Seminars and oral presentations

Cédric Gilon, and Guillaume Levasseur. "Deep and recurrent neural networks: two real-world use cases" Current trends in AI, ULB, Brussels, Belgium, (2023)

Cédric Gilon. "Paroxysmal atrial fibrillation detection and onset forecast using machine learning". IRIDIA Seminars, ULB, Brussels, Belgium, (2022).

Guillaume Levasseur, and **Cédric Gilon**. "Deep and recurrent neural networks: two real-world use cases". Current trends in AI, ULB, Brussels, Belgium, (2022).

Cédric Gilon. "Paroxysmal atrial fibrillation diagnosis and forecast with Deep Neural Networks". PhD AI Seminars, BeCentral, Belgium, (2019).

List of Figures

List of Tables

Bibliography

Abadi, M. et al. (Nov. 2016). "TensorFlow: a system for large-scale machine learning". In: *Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation*, pp. 265–283.

Alcaraz, R., Martínez, A., and Rieta, J. J. (Apr. 2015). "Role of the P-wave high frequency energy and duration as noninvasive cardiovascular predictors of paroxysmal atrial fibrillation". en. In: *Computer Methods and Programs in Biomedicine* 119.2, pp. 110–119. DOI: 10.1016/j.cmpb.2015.01.006.

Aligholipour, O. and Kuntalp, M. (Apr. 2018). "Clustering of Paroxysmal Atrial Fibrillation (PAF) and non-PAF subjects based on arrhythmia-free records". en. In: *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*. Istanbul: IEEE, pp. 1–5. DOI: 10.1109/EBBT.2018.8391425.

Alonso, A. et al. (Mar. 2013). "Simple risk model predicts incidence of atrial fibrillation in a racially and geographically diverse population: the CHARGE-AF consortium". eng. In: *Journal of the American Heart Association* 2.2, e000102. DOI: 10.1161/JAHA.112.000102.

Altan, G., Kutlu, Y., Pekmezci, A. Ö., and Nural, S. (Aug. 2018). "Deep learning with 3D-second order difference plot on respiratory sounds". en. In: *Biomedical Signal Processing and Control* 45, pp. 58–69. DOI: 10.1016/j.bspc.2018.05.014.

Anwar, F. A. and Al Raddady, F. (Dec. 2013). "ANN-based Classifier for PAF Prediction". en. In: *International Journal of Computer Applications* 83.17, pp. 7–13. DOI: 10.5120/14667-2811.

Attia, Z. I., Noseworthy, P. A., Lopez-Jimenez, F., Asirvatham, S. J., Deshmukh, A. J., Gersh, B. J., Carter, R. E., Yao, X., Rabinstein, A. A., Erickson, B. J., Kapa, S., and Friedman, P. A. (Sept. 2019a). "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. DOI: 10.1016/S0140-6736(19)31721-0.

Attia, Z. I., Friedman, P. A., Noseworthy, P. A., Lopez-Jimenez, F., Ladewig, D. J., Satam, G., Pellikka, P. A., Munger, T. M., Asirvatham, S. J., Scott, C. G., Carter, R. E., and Kapa, S. (2019b). "Age and Sex Estimation Using Artificial Intelligence from Standard 12-Lead ECGs". In: *Circulation: Arrhythmia and Electrophysiology* 12.9, pp. 1–11. DOI: 10.1161/CIRCEP.119.007284.

Ayatollahi, A., Afrakhteh, S., Soltani, F., and Saleh, E. (Apr. 2023). "Sleep apnea detection from ECG signal using deep CNN-based structures". en. In: *Evolving Systems* 14.2, pp. 191–206. DOI: 10.1007/s12530-022-09445-1.

Babloyantz, A. and Maurer, P. (Sept. 1996). "A graphical representation of local correlations in time series — Assessment of cardiac dynamics". en. In: *Physics Letters A* 221.1-2, pp. 43–55. DOI: 10.1016/0375-9601(96)00506-3.

Badilini, F. and ISHNE Standard Output Format Task Force (July 1998). "The ISHNE Holter Standard Output File Format". en. In: *Annals of Noninvasive Electrocardiology* 3.3, pp. 263–266. DOI: 10.1111/j.1542-474X.1998.tb00353.x.

Baek, Y.-S., Lee, S.-C., Choi, W., and Kim, D.-H. (Dec. 2021). "A new deep learning algorithm of 12-lead electrocardiogram for identifying atrial fibrillation during sinus rhythm". en. In: *Scientific Reports* 11.1, p. 12818. DOI: 10.1038/s41598-021-92172-5.

Bashar, S. K., Han, D., Zieneddin, F., Ding, E., Fitzgibbons, T. P., Walkey, A. J., McManus, D. D., Javidi, B., and Chon, K. H. (Feb. 2021). "Novel Density Poincaré Plot Based Machine Learning Method to Detect Atrial Fibrillation from Premature Atrial/Ventricular Contractions". In: *IEEE transactions on bio-medical engineering* 68.2, pp. 448–460. DOI: 10.1109/TBME.2020.3004310.

Batchvarov, V. N., Ghuran, A., Smetana, P., Hnatkova, K., Harries, M., Dilaveris, P., Camm, A. J., and Malik, M. (June 2002). "QT-RR relationship in healthy subjects exhibits substantial intersubject variability and high intrasubject stability". en. In: *American Journal of Physiology-Heart and Circulatory Physiology* 282.6, H2356–H2363. DOI: 10.1152/ajpheart.00860.2001.

Bauer, A., Kantelhardt, J. W., Barthel, P., Schneider, R., Mäkikallio, T., Ulm, K., Hnatkova, K., Schömig, A., Huikuri, H., Bunde, A., Malik, M., and Schmidt, G. (May 2006). "Deceleration capacity of heart rate as a predictor of mortality after myocardial infarction: cohort study". In: *The Lancet* 367.9523. Publisher: Elsevier, pp. 1674–1681. DOI: 10.1016/S0140-6736(06)68735-7.

Bazett, H. (1920). "An Analysis of the Time Relationships of Electrocardiograms". In: *Annals of Noninvasive Electrocardiology* 2, pp. 177–194. DOI: 10.1111/j.1542-474X.1997.tb00325.x.

Benjamin, E. J., Wolf, P. A., D'Agostino, R. B., Silbershatz, H., Kannel, W. B., and Levy, D. (Sept. 1998). "Impact of atrial fibrillation on the risk of death: the Framingham Heart Study". eng. In: *Circulation* 98.10, pp. 946–952. DOI: 10.1161/01.cir.98.10.946.

Berger, R. D., Akselrod, S., Gordon, D., and Cohen, R. J. (Sept. 1986). "An Efficient Algorithm for Spectral Analysis of Heart Rate Variability". en. In: *IEEE Transactions on Biomedical Engineering* BME-33.9, pp. 900–904. DOI: 10.1109/TBME.1986.325789.

Bernston, G. G., Bigger, J. T., Eckberg, D. L., Grossman, P., Kaufman, P. G., Malik, M., Nagaraja, H. N., Porges, S. W., Saul, J. P., Stone, P. H., and Molen, M. W. van der (1997). "Heart rate variability: Origins, methods, and interpretive caveats". en. In: *Psychophysiology* 34. DOI: 10.1111/j.1469-8986.1997.tb02140.x.

Bianchi, F. M., Livi, L., Ferrante, A., Milosevic, J., and Malek, M. (Apr. 2018). *Time series kernel similarities for predicting Paroxysmal Atrial Fibrillation from ECGs*. en. arXiv:1801.06845 [cs, eess, stat].

Bigger, J. T., Kleiger, R. E., Fleiss, J. L., Rolnitzky, L. M., Steinman, R. C., and Miller, J. P. (Feb. 1988). "Components of heart rate variability measured during healing of acute myocardial infarction". eng. In: *The American Journal of Cardiology* 61.4, pp. 208–215. DOI: 10.1016/0002-9149(88)90917-4.

Billman, G. E. (2013). "The LF/HF ratio does not accurately measure cardiac sympatho-vagal balance". eng. In: *Frontiers in Physiology* 4, p. 26. DOI: 10.3389/fphys.2013.00026.

Biton, S., Aldhafeeri, M., Marcusohn, E., Tsutsui, K., Szwagier, T., Elias, A., Oster, J., Sellal, J. M., Suleiman, M., and Behar, J. A. (Mar. 2023). "Generalizable and robust deep learning algorithm for atrial fibrillation diagnosis across geography, ages and sexes". en. In: *npj Digital Medicine* 6.1, pp. 1–10. DOI: 10.1038/s41746-023-00791-1.

Biton, S., Gendelman, S., Ribeiro, A. H., Miana, G., Moreira, C., Ribeiro, A. L. P., and Behar, J. A. (Dec. 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. DOI: 10.1093/ehjdh/ztab071.

Boon, K., Khalil-Hani, M., Malarvili, M., and Sia, C. (Oct. 2016). "Paroxysmal atrial fibrillation prediction method with shorter HRV sequences". en. In: *Computer Methods and Programs in Biomedicine* 134, pp. 187–196. DOI: 10.1016/j.cmpb.2016.07.016.

Boon, K., Khalil-Hani, M., and Malarvili, M. (Jan. 2018). "Paroxysmal atrial fibrillation prediction based on HRV analysis and non-dominated sorting genetic algorithm III". en. In: *Computer Methods and Programs in Biomedicine* 153, pp. 171–184. DOI: 10.1016/j.cmpb.2017.10.012.

Boriani, G., Laroche, C., Diemberger, I., Fantecchi, E., Popescu, M. I., Rasmussen, L. H., Sinagra, G., Petrescu, L., Tavazzi, L., Maggioni, A. P., and Lip, G. Y. (May 2015). "Asymptomatic Atrial Fibrillation: Clinical Correlates, Management, and Outcomes in the EORP-AF Pilot General Registry". en. In: *The American Journal of Medicine* 128.5, 509–518.e2. DOI: 10.1016/j.amjmed.2014.11.026.

"Area under the Precision-Recall Curve" (2013). "Area under the Precision-Recall Curve: Point Estimates and Confidence Intervals". en. In: ed. by K. Boyd, K. Eng, and C. D. Page. Springer Berlin Heidelberg, pp. 451–466.

Breiman, L., Friedman, J., Stone, C., and Olshen, R. (1984). *Classification and Regression Trees*. Taylor & Francis.

Breiman, L. (2001). "Random Forest". In: *Machine Learning* 45.1, pp. 5–32. DOI: 10.1023/A:1010933404324.

Brennan, M., Palaniswami, M., and Kamen, P. (Nov. 2001). "Do existing measures of Poincare plot geometry reflect nonlinear features of heart rate variability?" In: *IEEE Transactions on Biomedical Engineering* 48.11, pp. 1342–1347. DOI: 10.1109/10.959330.

Burr, R. L. (July 2007). "Interpretation of normalized spectral heart rate variability indices in sleep research: a critical review". eng. In: *Sleep* 30.7, pp. 913–919. DOI: 10.1093/sleep/30.7.913.

Călburean, P.-A., Osório, T. G., Sieira, J., Ströker, E., Maj, R., Terasawa, M., Rizzo, A., Borio, G., Scala, O., Galli, A., Brugada, P., Chierchia, G.-B., and De Asmundis, C. (Jan. 2021). "High parasympathetic activity as reflected by deceleration capacity predicts atrial fibrillation recurrence after repeated catheter ablation procedure". eng. In: *Journal of Interventional Cardiac Electrophysiology: An International Journal of Arrhythmias and Pacing* 60.1, pp. 21–29. DOI: 10.1007/s10840-019-00687-9.

Castro, H., Garcia-Racines, J. D., and Bernal-Norena, A. (Nov. 2021). "Methodology for the prediction of paroxysmal atrial fibrillation based on heart rate

variability feature analysis". en. In: *Heliyon* 7.11, e08244. DOI: 10.1016/j.heliyon.2021.e08244.

Cerqueira, V., Torgo, L., and Mozetic, I. (Nov. 2020). "Evaluating time series forecasting models: An empirical study on performance estimation methods". In: *Machine Learning* 109.11. arXiv:1905.11744 [cs, stat], pp. 1997–2028. DOI: 10.1007/s10994-020-05910-7.

Chang, C.-C., Hsu, H.-Y., and Hsiao, T.-C. (Apr. 2014). "The interpretation of very high frequency band of instantaneous pulse rate variability during paced respiration". In: *BioMedical Engineering OnLine* 13.1, p. 46. DOI: 10.1186/1475-925X-13-46.

Chazal, P. de and Heneghan, C. (2001). "Automated assessment of atrial fibrillation". en. In: *Computers in Cardiology 2001. Vol.28 (Cat. No.01CH37287)*. Rotterdam, Netherlands: IEEE, pp. 117–120. DOI: 10.1109/CIC.2001.977605.

Chen, T. and Guestrin, C. (Aug. 2016). "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. arXiv:1603.02754 [cs], pp. 785–794. DOI: 10.1145/2939672.2939785.

Chen, Z., Yang, Y., Zou, C., Zhang, Y., Huang, X., Li, X., and Yang, X. (Apr. 2018). "Low heart deceleration capacity imply higher atrial fibrillation-free rate after ablation". en. In: *Scientific Reports* 8.1. Number: 1 Publisher: Nature Publishing Group, p. 5537. DOI: 10.1038/s41598-018-23970-7.

Chesnokov, Y. V. (June 2008). "Complexity and spectral analysis of the heart rate variability dynamics for distant prediction of paroxysmal atrial fibrillation with artificial intelligence methods". en. In: *Artificial Intelligence in Medicine* 43.2, pp. 151–165. DOI: 10.1016/j.artmed.2008.03.009.

Cho, J., Kim, Y., and Lee, M. (2018). "Prediction to Atrial Fibrillation Using Deep Convolutional Neural Networks". en. In: *PRedictive Intelligence in MEdicine*. Vol. 11121. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 164–171. DOI: 10.1007/978-3-030-00320-3_20.

Chollet, F. et al. (2015). *Keras*. https://keras.io.

Christensen, L. M., Krieger, D. W., Højberg, S., Pedersen, O. D., Karlsen, F. M., Jacobsen, M. D., Worck, R., Nielsen, H., Ægidius, K., Jeppesen, L. L., Rosenbaum, S., Marstrand, J., and Christensen, H. (2014). "Paroxysmal atrial fibrillation occurs often in cryptogenic ischaemic stroke. Final results from the SURPRISE study". en. In: *European Journal of Neurology* 21.6. _eprint:

https://onlinelibrary.wiley.com/doi/pdf/10.1111/ene.12400, pp. 884–889. DOI: 10.1111/ene.12400.

Christophersen, I. E., Yin, X., Larson, M. G., Lubitz, S. A., Magnani, J. W., McManus, D. D., Ellinor, P. T., and Benjamin, E. J. (Aug. 2016). "A comparison of the CHARGE–AF and the CHA2DS2-VASc risk scores for prediction of atrial fibrillation in the Framingham Heart Study". In: *American heart journal* 178, pp. 45–54. DOI: 10.1016/j.ahj.2016.05.004.

Chugh, S. S., Havmoeller, R., Narayanan, K., Singh, D., Rienstra, M., Benjamin, E. J., Gillum, R. F., Kim, Y.-H., McAnulty, J. H., Zheng, Z.-J., Forouzanfar, M. H., Naghavi, M., Mensah, G. A., Ezzati, M., and Murray, C. J. L. (Feb. 2014). "Worldwide epidemiology of atrial fibrillation: a Global Burden of Disease 2010 Study". eng. In: *Circulation* 129.8, pp. 837–847. DOI: 10.1161/CIRCULATIONAHA.113.005119.

Ciccone, A. B., Siedlik, J. A., Wecht, J. M., Deckert, J. A., Nguyen, N. D., and Weir, J. P. (Oct. 2017). "Reminder: RMSSD and SD1 are identical heart rate variability metrics". eng. In: *Muscle & Nerve* 56.4, pp. 674–678. DOI: 10.1002/mus.25573.

Citi, L., Brown, E., and Barbieri, R. (Aug. 2012). "A Real-Time Automated Point-Process Method for the Detection and Correction of Erroneous and Ectopic Heartbeats". In: *IEEE transactions on bio-medical engineering* 59, pp. 2828–37. DOI: 10.1109/TBME.2012.2211356.

Clifford, G. D., Liu, C., Moody, B., Lehman, L.-w. H., Silva, I., Li, Q., Johnson, A. E., and Mark, R. G. (Sept. 2017). "AF Classification from a Short Single Lead ECG Recording: the PhysioNet/Computing in Cardiology Challenge 2017". In: *Computing in cardiology* 44.

Cooper, J. M., Katcher, M. S., and Orlov, M. V. (June 2002). "Implantable Devices for the Treatment of Atrial Fibrillation". In: *New England Journal of Medicine* 346.26, pp. 2062–2068. DOI: 10.1056/NEJMra012144.

Costa, M. D., Redline, S., Davis, R. B., Heckbert, S. R., Soliman, E. Z., and Goldberger, A. L. (Sept. 2018). "Heart Rate Fragmentation as a Novel Biomarker of Adverse Cardiovascular Events: The Multi-Ethnic Study of Atherosclerosis". en. In: *Frontiers in Physiology* 9, p. 1117. DOI: 10.3389/fphys.2018.01117.

Costin, H., Rotariu, C., and Pasarica, A. (May 2013). "Atrial fibrillation onset prediction using variability of ECG signals". en. In: *2013 8TH INTERNATIONAL SYMPOSIUM ON ADVANCED TOPICS IN ELECTRICAL ENGI-*

NEERING (ATEE). Bucharest, Romania: IEEE, pp. 1–4. DOI: 10.1109/ATEE.2013.6563419.

Coumel, P. (Apr. 1994). "Paroxysmal Atrial Fibrillation: A Disorder of Autonomic Tone?" en. In: *European Heart Journal* 15.suppl A, pp. 9–16. DOI: 10.1093/eurheartj/15.suppl_A.9.

Coumel, P., Fayn, J., Maison-Blanche, P., and Rubel, P. (Jan. 1994). "Clinical relevance of assessing QT dynamicity in holter recordings". en. In: *Journal of Electrocardiology* 27, pp. 62–66. DOI: 10.1016/S0022-0736(94)80050-2.

Dabiré, H., Mestivier, D., Jarnet, J., Safar, M. E., and Chau, N. P. (Oct. 1998). "Quantification of sympathetic and parasympathetic tones by nonlinear indexes in normotensive rats". eng. In: *The American Journal of Physiology* 275.4, H1290–1297. DOI: 10.1152/ajpheart.1998.275.4.H1290.

De Giovanni, E., Aminifar, A., Luca, A., Yazdani, S., Vesin, J.-M., and Atienza, D. (Sept. 2017). "A Patient-Specific Methodology for Prediction of Paroxysmal Atrial Fibrillation Onset". en. In: DOI: 10.22489/CinC.2017.285-191.

Deschamps, E., Venier, S., Jacon, P., Carabelli, A., Peter, P., Desbiolles, A., Favre-Wiki, I., Cuisenier, P., Garambois, K., Detante, O., Jadidi, A., and Defaye, P. (Jan. 2023). "P-wave duration in cryptogenic stroke: A case control prospective study". In: *Archives of Cardiovascular Diseases Supplements*. JESFC 2023 15.1, p. 88. DOI: 10.1016/j.acvdsp.2022.10.169.

Dewland, T. A., Vittinghoff, E., Mandyam, M. C., Heckbert, S. R., Siscovick, D. S., Stein, P. K., Psaty, B. M., Sotoodehnia, N., Gottdiener, J. S., and Marcus, G. M. (Dec. 2013). "Atrial Ectopy as a Predictor of Incident Atrial Fibrillation: A Cohort Study". en. In: *Annals of Internal Medicine* 159.11, p. 721. DOI: 10.7326/0003-4819-159-11-201312030-00004.

Diao, C. and Cai, N. (Dec. 2022). "Temporal Variation Measure Analysis: An Improved Second-Order Difference Plot". en. In: *Complexity* 2022. Publisher: Hindawi, e8265275. DOI: 10.1155/2022/8265275.

Dupulthys, S., Dujardin, K., Anné, W., Pollet, P., Vanhaverbeke, M., McAuliffe, D., Lammertyn, P.-J., Berteloot, L., Mertens, N., and Jaeger, P. (Dec. 2023). "Single-lead ECG AI model with risk factors detects Atrial Fibrillation during Sinus Rhythm". In: *Europace*. DOI: 10.1093/europace/euad354.

Ebrahimzadeh, E., Kalantari, M., Joulani, M., Shahraki, R. S., Fayaz, F., and Ahmadi, F. (Oct. 2018). "Prediction of paroxysmal Atrial Fibrillation: A machine learning based approach using combined feature vector and mixture of expert classification on HRV signal". en. In: *Computer Methods and Programs in Biomedicine* 165, pp. 53–67. DOI: 10.1016/j.cmpb.2018.07.014.

Echt, D. S. and Ruskin, J. N. (Apr. 2020). "Use of Flecainide for the Treatment of Atrial Fibrillation". English. In: *American Journal of Cardiology* 125.7, pp. 1123–1133. DOI: 10.1016/j.amjcard.2019.12.041.

Eckmann, J.-P., Kamphorst, S., and Ruelle, D. (Nov. 1987). "Recurrence Plots of Dynamical Systems". In: *Europhysics Letters (epl)* 4, pp. 973–977. DOI: 10.1209/0295-5075/4/9/004.

Einthoven, W. (1906). "The telecardiogram". In: *American Heart Journal* 53.4, pp. 602–615. DOI: 10.1016/0002-8703(57)90367-8.

Erdenebayar, U., Kim, H., Park, J. U., Kang, D., and Lee, K. J. (Feb. 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. Publisher: Korean Academy of Medical Science. DOI: 10.3346/jkms.2019.34.e64.

Fagard, R. H., Pardaens, K., Staessen, J. A., and Thijs, L. (1998). "Power spectral analysis of heart rate variability by autoregressive modelling and fast Fourier transform: a comparative study". eng. In: *Acta Cardiologica* 53.4, pp. 211–218.

Faust, O., Shenfield, A., Kareem, M., San, T. R., Fujita, H., and Acharya, U. R. (Nov. 2018). "Automated detection of atrial fibrillation using long short-term memory network with RR interval signals". en. In: *Computers in Biology and Medicine* 102, pp. 327–335. DOI: 10.1016/j.compbiomed.2018.07.001.

Fawcett, T. (Jan. 2004). "ROC Graphs: Notes and Practical Considerations for Researchers". In: *Machine Learning* 31, pp. 1–38.

— (June 2006). "An introduction to ROC analysis". en. In: *Pattern Recognition Letters* 27.8, pp. 861–874. DOI: 10.1016/j.patrec.2005.10.010.

Feher, J. (2012). "The Electrocardiogram". en. In: *Quantitative Human Physiology*. Elsevier, pp. 467–476. DOI: 10.1016/B978-0-12-382163-8.00050-5.

Freyer, L., Stülpnagel, L. von, Spielbichler, P., Sappler, N., Wenner, F., Schreinlechner, M., Krasniqi, A., Behroz, A., Eiffener, E., Zens, M., Dolejsi, T., Massberg, S., Rizas, K. D., and Bauer, A. (Nov. 2021). "Rationale and design of a digital trial using smartphones to detect subclinical atrial fibrillation in a population at risk: The eHealth-based bavarian alternative detection of Atrial Fibrillation (eBRAVE-AF) trial". eng. In: *American Heart Journal* 241, pp. 26–34. DOI: 10.1016/j.ahj.2021.06.008.

Friberg, L. and Bergfeldt, L. (Nov. 2013). "Atrial fibrillation prevalence revisited". en. In: *Journal of Internal Medicine* 274.5, pp. 461–468. DOI: 10.1111/joim.12114.

Gabelica, M., Bojčić, R., and Puljak, L. (Oct. 2022). "Many researchers were not compliant with their published data sharing statement: a mixed-methods study". English. In: *Journal of Clinical Epidemiology* 150. Publisher: Elsevier, pp. 33–41. DOI: 10.1016/j.jclinepi.2022.05.019.

Gadaleta, M., Harrington, P., Barnhill, E., Hytopoulos, E., Turakhia, M. P., Steinhubl, S. R., and Quer, G. (Dec. 2023). "Prediction of atrial fibrillation from at-home single-lead ECG signals without arrhythmias". en. In: *npj Digital Medicine* 6.1. Number: 1 Publisher: Nature Publishing Group, pp. 1–9. DOI: 10.1038/s41746-023-00966-w.

Gavidia, M., Zhu, H., Montanari, A., Fuentes, J., Cheng, C., Dubner, S., Maison-Blanche, P., Rahman, M., Sassi, R., Badilini, F., Jiang, Y., Zhang, H.-T., Du, H., Teng, B., Yuan, Y., Wan, G., Tang, Z., He, X., and Goncalves, J. (2023). "Early Warning of Atrial Fibrillation". en. In: *Preprint*.

Gilon, C., Gregoire, J.-M., and Bersini, H. (July 2020). "Forecast of paroxysmal atrial fibrillation using a deep neural network". en. In: *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7. DOI: 10.1109/IJCNN48605.2020.9207227.

Gilon, C., Gregoire, J.-M., Hellinckx, J., Carlier, S., and Bersini, H. (Sept. 2022). "Reproducibility of machine learning models for paroxysmal atrial fibrillation onset prediction". en. In: *Computer in Cardiology*.

Gilon, C., Grégoire, J.-M., Mathieu, M., Carlier, S., and Bersini, H. (July 2023a). *IRIDIA-AF, a large paroxysmal atrial fibrillation long-term electrocardiogram monitoring database*. DOI: 10.5281/zenodo.8186846.

— (Oct. 2023b). "IRIDIA-AF, a large paroxysmal atrial fibrillation long-term electrocardiogram monitoring database". en. In: *Scientific Data* 10.1. Number: 1 Publisher: Nature Publishing Group, pp. 1–10. DOI: 10.1038/s41597-023-02621-1.

Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., and Stanley, H. E. (June 2000). "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals". en. In: *Circulation* 101.23. DOI: 10.1161/01.CIR.101.23.e215.

Goldenberg, I., Goldkorn, R., Shlomo, N., Einhorn, M., Levitan, J., Kuperstein, R., Klempfner, R., and Johnson, B. (Dec. 2019). "Heart Rate Variability for Risk Assessment of Myocardial Ischemia in Patients Without Known Coronary Artery Disease: The HRV-DETECT (Heart Rate Variability for the De-

tection of Myocardial Ischemia) Study†". In: *Journal of the American Heart Association* 8.24. Publisher: Wiley, e014540. DOI: 10.1161/JAHA.119.014540.

Goshvarpour, A. and Goshvarpour, A. (2020). "Diagnosis of epileptic EEG using a lagged Poincare plot in combination with the autocorrelation". en. In: *Image and Video Processing*, p. 10.

Grégoire, J.-M., Gilon, C., Carlier, S., and Bersini, H. (2022). "Role of the autonomic nervous system and premature atrial contractions in short-term paroxysmal atrial fibrillation forecasting: Insights from machine learning models". eng. In: *Archives of Cardiovascular Diseases* 115.6-7, pp. 377–387. DOI: 10.1016/j.acvd.2022.04.006.

Gruwez, H., Barthels, M., Haemers, P., Verbrugge, F. H., Dhont, S., Meekers, E., Wouters, F., Nuyens, D., Pison, L., Vandervoort, P., and Pierlet, N. (June 2023a). "Detecting Paroxysmal Atrial Fibrillation From an Electrocardiogram in Sinus Rhythm: External Validation of the AI Approach". en. In: *JACC: Clinical Electrophysiology*. DOI: 10.1016/j.jacep.2023.04.008.

Gruwez, H., Verbrugge, F. H., Proesmans, T., Evens, S., Vanacker, P., Rutgers, M. P., Vanhooren, G., Bertrand, P., Pison, L., Haemers, P., Vandervoort, P., and Nuyens, D. (Dec. 2023b). "Smartphone-based atrial fibrillation screening in the general population: feasibility and impact on medical treatment". In: *European Heart Journal - Digital Health* 4.6, pp. 464–472. DOI: 10.1093/ehjdh/ztad054.

Guichard, J. B., Pichot, V., Hupin, D., Celle, S., Da Costa, A., Barthelemy, J. C., and Roche, F. (Oct. 2022). "Heart rate fragmentation as a marker of altered global autonomic nervous system activity: a novel predictor of atrial fibrillation occurrence in the general population". In: *European Heart Journal* 43.Supplement_2, ehac544.511. DOI: 10.1093/eurheartj/ehac544.511.

Guillaudeux, M., Rousseau, O., Petot, J., Bennis, Z., Dein, C.-A., Goronflot, T., Vince, N., Limou, S., Karakachoff, M., Wargny, M., and Gourraud, P.-A. (Mar. 2023). "Patient-centric synthetic data generation, no reason to risk re-identification in biomedical data analysis". en. In: *npj Digital Medicine* 6.1. Number: 1 Publisher: Nature Publishing Group, pp. 1–10. DOI: 10.1038/s41746-023-00771-5.

Guo, Y., Wang, H., Zhang, H., Liu, T., Li, L., Liu, L., Chen, M., Chen, Y., and Lip, G. Y. (Dec. 2021). "Photoplethysmography-Based Machine Learning Approaches for Atrial Fibrillation Prediction". en. In: *JACC: Asia* 1.3, pp. 399–408. DOI: 10.1016/j.jacasi.2021.09.004.

Haïssaguerre, M., Jaïs, P., Shah, D. C., Takahashi, A., Hocini, M., Quiniou, G., Garrigue, S., Le Mouroux, A., Le Métayer, P., and Clémenty, J. (Sept. 1998). "Spontaneous initiation of atrial fibrillation by ectopic beats originating in the pulmonary veins". eng. In: *The New England Journal of Medicine* 339.10, pp. 659–666. DOI: 10.1056/NEJM199809033391003.

Hammer, " Malberg, H., and Schmidt", M. (Dec. 2022). "Towards the Prediction of Atrial Fibrillation Based on Interpretable ECG Features". en. In: DOI: 10.22489/CinC.2022.236.

Hämmerle, P. et al. (Aug. 2020). "Heart Rate Variability Triangular Index as a Predictor of Cardiovascular Mortality in Patients With Atrial Fibrillation". In: *Journal of the American Heart Association* 9.15. Publisher: American Heart Association, e016075. DOI: 10.1161/JAHA.120.016075.

Han, L., Zhang, Q., Chen, X., Zhan, Q., Yang, T., and Zhao, Z. (Sept. 2017). "Detecting work-related stress with a wearable device". In: *Computers in Industry* 90, pp. 42–49. DOI: 10.1016/j.compind.2017.05.004.

Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., and Ng, A. Y. (2019). "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network". In: *Nature Medicine* 25, pp. 65–69. DOI: 10.1038/s41591-018-0268-3.

Hari, K. J., Nguyen, T. P., and Soliman, E. Z. (Nov. 2018). "Relationship between P-wave duration and the risk of atrial fibrillation". eng. In: *Expert Review of Cardiovascular Therapy* 16.11, pp. 837–843. DOI: 10.1080/14779072.2018.1533814.

Hartikainen, S., Lipponen, J. A., Hiltunen, P., Rissanen, T. T., Kolk, I., Tarvainen, M. P., Martikainen, T. J., Castren, M., Väliaho, E.-S., and Jäntti, H. (May 2019). "Effectiveness of the Chest Strap Electrocardiogram to Detect Atrial Fibrillation". In: *The American Journal of Cardiology* 123.10, pp. 1643–1648. DOI: 10.1016/j.amjcard.2019.02.028.

Hayano, J. and Yuda, E. (Mar. 2019). "Pitfalls of assessment of autonomic function by heart rate variability". In: *Journal of Physiological Anthropology* 38.1, p. 3. DOI: 10.1186/s40101-019-0193-2.

He, K., Zhang, X., Ren, S., and Sun, J. (Dec. 2015). *Deep Residual Learning for Image Recognition*. en. arXiv:1512.03385 [cs].

Heathers, J. A. J. (May 2014). "Everything Hertz: methodological issues in short-term frequency-domain HRV". en. In: *Frontiers in Physiology* 5. DOI: 10.3389/fphys.2014.00177.

Heeringa, J., Kuip, D. A. M. van der, Hofman, A., Kors, J. A., Herpen, G. van, Stricker, B. H. C., Stijnen, T., Lip, G. Y. H., and Witteman, J. C. M. (Apr. 2006). "Prevalence, incidence and lifetime risk of atrial fibrillation: the Rotterdam study". eng. In: *European Heart Journal* 27.8, pp. 949–953. DOI: 10.1093/eurheartj/ehi825.

Hickey, B. and Heneghan, C. (2002). "Screening for paroxysmal atrial fibrillation using atrial premature contractions and spectral measures". en. In: *Computers in Cardiology*. Memphis, TN, USA: IEEE, pp. 217–220. DOI: 10.1109/CIC.2002.1166746.

Hindricks, G. et al. (Feb. 2021). "2020 ESC Guidelines for the diagnosis and management of atrial fibrillation developed in collaboration with the European Association for Cardio-Thoracic Surgery (EACTS)". en. In: *European Heart Journal* 42.5, pp. 373–498. DOI: 10.1093/eurheartj/ehaa612.

Hochreiter, S. and Schmidhuber, J. (Dec. 1997). "Long Short-term Memory". In: *Neural computation* 9, pp. 1735–80. DOI: 10.1162/neco.1997.9.8.1735.

Holst, H., Ohlsson, M., Peterson, C., and Edenbrandt, L. (Oct. 1999). "A confident decision support system for interpreting electrocardiograms: A confident decision support system". en. In: *Clinical Physiology* 19.5, pp. 410–418. DOI: 10.1046/j.1365-2281.1999.00195.x.

Hovsepian, K., al'Absi, M., Ertin, E., Kamarck, T., Nakajima, M., and Kumar, S. (Sept. 2015). "cStress: Towards a Gold Standard for Continuous Stress Assessment in the Mobile Environment". In: *Proceedings of the ... ACM International Conference on Ubiquitous Computing . UbiComp (Conference)* 2015, pp. 493–504. DOI: 10.1145/2750858.2807526.

Hu, R., Chen, J., and Zhou, L. (May 2022). "A transformer-based deep neural network for arrhythmia detection using continuous ECG signals". en. In: *Computers in Biology and Medicine* 144, p. 105325. DOI: 10.1016/j.compbiomed.2022.105325.

Hygrell, T., Viberg, F., Dahlberg, E., Charlton, P. H., Kemp Gudmundsdottir, K., Mant, J., Hörnlund, J. L., and Svennberg, E. (Apr. 2023). "An artificial intelligence–based model for prediction of atrial fibrillation from single-lead sinus rhythm electrocardiograms facilitating screening". In: *EP Europace* 25.4, pp. 1332–1338. DOI: 10.1093/europace/euad036.

Jalali, A. and Lee, M. (Feb. 2020). "Atrial Fibrillation Prediction With Residual Network Using Sensitivity and Orthogonality Constraints". en. In: *IEEE Journal of Biomedical and Health Informatics* 24.2, pp. 407–413. DOI: 10.1109/JBHI.2019.2957809.

Jeppesen, J., Beniczky, S., Johansen, P., Sidenius, P., and Fuglsang-Frederiksen, A. (Aug. 2014). "Using Lorenz plot and Cardiac Sympathetic Index of heart rate variability for detecting seizures for patients with epilepsy". en. In: *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Chicago, IL: IEEE, pp. 4563–4566. DOI: 10.1109/EMBC.2014.6944639.

Kharbanda, R. K., Does, W. F. B. van der, Staveren, L. N. van, Taverne, Y. J. H. J., Bogers, A. J. J. C., and Groot, N. M. S. de (Apr. 2022). "Vagus Nerve Stimulation and Atrial Fibrillation: Revealing the Paradox". eng. In: *Neuromodulation: Journal of the International Neuromodulation Society* 25.3, pp. 356–365. DOI: 10.1016/j.neurom.2022.01.008.

Al-Khatib, S. M., Wilkinson, W. E., Sanders, L. L., McCarthy, E. A., and Pritchett, E. L. (July 2000). "Observations on the transition from intermittent to permanent atrial fibrillation". en. In: *American Heart Journal* 140.1, pp. 142–145. DOI: 10.1067/mhj.2000.107547.

Kher, R. (Dec. 2019). "Signal Processing Techniques for Removing Noise from ECG Signals". en. In: *jber*. DOI: 10.17303/jber.2019.3.101.

Khurshid, S., Friedman, S., Reeder, C., Di Achille, P., Diamant, N., Singh, P., Harrington, L. X., Wang, X., Al-Alusi, M. A., Sarma, G., Foulkes, A. S., Ellinor, P. T., Anderson, C. D., Ho, J. E., Philippakis, A. A., Batra, P., and Lubitz, S. A. (Jan. 2022). "ECG-Based Deep Learning and Clinical Risk Factors to Predict Atrial Fibrillation". en. In: *Circulation* 145.2, pp. 122–133. DOI: 10.1161/CIRCULATIONAHA.121.057480.

Kikillus, N., Hammer, G., Wieland, S., and Bolz, A. (2007). "Algorithm for identifying patients with paroxysmal atrial fibrillation without appearance on the ECG". eng. In: *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference* 2007, pp. 275–278. DOI: 10.1109/IEMBS.2007.4352277.

Kim, J. Y., Kim, K. G., Tae, Y., Chang, M., Park, S.-J., Park, K.-M., On, Y. K., Kim, J. S., Lee, Y., and Jang, S.-W. (July 2022). "An Artificial Intelligence Algorithm With 24-h Holter Monitoring for the Identification of Occult Atrial Fibrillation During Sinus Rhythm". In: *Frontiers in Cardiovascular Medicine* 9, p. 906780. DOI: 10.3389/fcvm.2022.906780.

Kingma, D. P. and Ba, J. (Jan. 2017). "Adam: A Method for Stochastic Optimization". In: *arXiv:1412.6980 [cs]*. arXiv: 1412.6980 version: 8.

Kirchhof, P., Toennis, T., Goette, A., Camm, A. J., Diener, H. C., Becher, N., Bertaglia, E., Lundqvist, C. B., Borlich, M., Brandes, A., Cabanelas, N., Calvert, M., Chlouverakis, G., and Dan, G.-A. (2023). "Anticoagulation with Edoxaban in Patients with Atrial High-Rate Episodes". en. In: *n engl j med*.

Kisohara, M., Masuda, Y., Yuda, E., Ueda, N., and Hayano, J. (Mar. 2020). *Optimal Length of R-R Interval Segment Window for Lorenz Plot Detection of Paroxysmal Atrial Fibrillation by Machine Learning*. en. preprint. In Review. DOI: 10.21203/rs.3.rs-19433/v1.

Kleiman, R., Darpo, B., Brown, R., Rudo, T., Chamoun, S., Albert, D. E., Bos, J. M., and Ackerman, M. J. (Nov. 2021). "Comparison of electrocardiograms (ECG) waveforms and centralized ECG measurements between a simple 6-lead mobile ECG device and a standard 12-lead ECG". en. In: *Annals of Noninvasive Electrocardiology* 26.6. DOI: 10.1111/anec.12872.

Koch, B., Denton, E., Hanna, A., and Foster, J. G. (Aug. 2021). "Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research". en. In.

Krijthe, B. P., Kunst, A., Benjamin, E. J., Lip, G. Y., Franco, O. H., Hofman, A., Witteman, J. C., Stricker, B. H., and Heeringa, J. (Sept. 2013). "Projections on the number of individuals with atrial fibrillation in the European Union, from 2000 to 2060". In: *European Heart Journal* 34.35, pp. 2746–2751. DOI: 10.1093/eurheartj/eht280.

Kubios (2023). *User's Guide: Kubios HRV Scientific (version 4.1)*.

La Rovere, M. T., Pinna, G. D., Maestri, R., Mortara, A., Capomolla, S., Febo, O., Ferrari, R., Franchini, M., Gnemmi, M., Opasich, C., Riccardi, P. G., Traversi, E., and Cobelli, F. (Feb. 2003). "Short-Term Heart Rate Variability Strongly Predicts Sudden Cardiac Death in Chronic Heart Failure Patients". In: *Circulation* 107.4. Publisher: American Heart Association, pp. 565–570. DOI: 10.1161/01.CIR.0000047275.25795.17.

Lakkireddy, D., Pillarisetti, J., Patel, A., Boc, K., Bommana, S., Sawers, Y., Vanga, S., Sayana, H., Chen, W., Nath, J., and Vacek, J. (June 2009). "Evolution of Paroxysmal Atrial Fibrillation to Persistent or Permanent Atrial Fibrillation: Predictors of Progression". en. In: *Journal of Atrial Fibrillation* 1.7, p. 536. DOI: 10.4022/jafib.v1i7.536.

Lee, H., Shin, S.-Y., Seo, M., Nam, G.-B., and Joo, S. (Aug. 2016). "Prediction of Ventricular Tachycardia One Hour before Occurrence Using Artificial Neural Networks". en. In: *Scientific Reports* 6.1. Number: 1 Publisher: Nature Publishing Group, p. 32390. DOI: 10.1038/srep32390.

Lee, S.-H., Kang, D.-W., and Lee, K.-J. (Apr. 2018). "Prediction of Paroxysmal Atrial Fibrillation using Time-domain Analysis and Random Forest". en. In: *Journal of Biomedical Engineering Research* 39.2, pp. 69–79. DOI: 10.9718/JBER.2018.39.2.69.

Levasseur, G. and Bersini, H. (July 2022). "Time Series Representation for Real-World Applications of Deep Neural Networks". In: *2022 International Joint Conference on Neural Networks (IJCNN)*. ISSN: 2161-4407, pp. 1–8. DOI: 10.1109/IJCNN55064.2022.9892244.

Lewis, T. (Nov. 1909). "Report Cxix. Auricular Fibrillation: A Common Clinical Condition". en. In: *Br Med J* 2.2552. Publisher: British Medical Journal Publishing Group Section: The Science Committee of the British Medical Association, pp. 1528–1528. DOI: 10.1136/bmj.2.2552.1528.

Li, Y.-G., Pastori, D., Farcomeni, A., Yang, P.-S., Jang, E., Joung, B., Wang, Y.-T., Guo, Y.-T., and Lip, G. Y. H. (Mar. 2019). "A Simple Clinical Risk Score (C2HEST) for Predicting Incident Atrial Fibrillation in Asian Subjects: Derivation in 471,446 Chinese Subjects, With Internal Validation and External Application in 451,199 Korean Subjects". eng. In: *Chest* 155.3, pp. 510–518. DOI: 10.1016/j.chest.2018.09.011.

Lip, G. Y. H., Fauchier, L., Freedman, S. B., Van Gelder, I., Natale, A., Gianni, C., Nattel, S., Potpara, T., Rienstra, M., Tse, H.-F., and Lane, D. A. (Mar. 2016). "Atrial fibrillation". en. In: *Nature Reviews Disease Primers* 2.1, p. 16016. DOI: 10.1038/nrdp.2016.16.

Lip, G. Y. H., Nieuwlaat, R., Pisters, R., Lane, D. A., and Crijns, H. J. G. M. (Feb. 2010). "Refining Clinical Risk Stratification for Predicting Stroke and Thromboembolism in Atrial Fibrillation Using a Novel Risk Factor-Based Approach: The Euro Heart Survey on Atrial Fibrillation". en. In: *Chest* 137.2, pp. 263–272. DOI: 10.1378/chest.09-1584.

Liu, C. and Li, J., eds. (2020). *Feature Engineering and Computational Intelligence in ECG Monitoring*. en. Singapore: Springer Singapore. DOI: 10.1007/978-981-15-3824-7.

Liu, H., Chen, D., Chen, D., Zhang, X., Li, H., Bian, L., Shu, M., and Wang, Y. (Dec. 2022). "A large-scale multi-label 12-lead electrocardiogram database with standardized diagnostic statements". en. In: *Scientific Data* 9.1, p. 272. DOI: 10.1038/s41597-022-01403-5.

Liu, N., Guo, D., Koh, Z. X., Ho, A. F. W., and Ong, M. E. H. (Feb. 2019). *Heart Rate n-Variability (HRnV): A Novel Representation of Beat-to-Beat Variation in*

Electrocardiogram. en. Pages: 449504 Section: New Results. DOI: 10.1101/449504.

Lombardi, F. (Sept. 2002). "Clinical implications of present physiological understanding of HRV components". eng. In: *Cardiac Electrophysiology Review* 6.3, pp. 245–249. DOI: 10.1023/a:1016329008921.

Lombardi, F., Tarricone, D., Tundo, F., Colombo, F., Belletti, S., and Fiorentini, C. (July 2004). "Autonomic nervous system and paroxysmal atrial fibrillation: a study based on the analysis of RR interval changes before, during and after paroxysmal atrial fibrillation". eng. In: *European Heart Journal* 25.14, pp. 1242–1248. DOI: 10.1016/j.ehj.2004.05.016.

Lopez Perales, C. R., Van Spall, H. G. C., Maeda, S., Jimenez, A., Laţcu, D. G., Milman, A., Kirakoya-Samadoulougou, F., Mamas, M. A., Muser, D., and Casado Arroyo, R. (Jan. 2021). "Mobile health applications for the detection of atrial fibrillation: a systematic review". eng. In: *Europace: European Pacing, Arrhythmias, and Cardiac Electrophysiology: Journal of the Working Groups on Cardiac Pacing, Arrhythmias, and Cardiac Cellular Electrophysiology of the European Society of Cardiology* 23.1, pp. 11–28. DOI: 10.1093/europace/euaa139.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. (Jan. 2020). "From local explanations to global understanding with explainable AI for trees". en. In: *Nature Machine Intelligence* 2.1. Number: 1 Publisher: Nature Publishing Group, pp. 56–67. DOI: 10.1038/s42256-019-0138-9.

Lynn, K. and Chiang, H. (2001). "A two-stage solution algorithm for paroxysmal atrial fibrillation prediction". en. In: *Computers in Cardiology 2001. Vol.28 (Cat. No.01CH37287)*. Rotterdam, Netherlands: IEEE, pp. 405–407. DOI: 10.1109/CIC.2001.977678.

Maier, C., Bauch, M., and Dickhaus, H. (2001). "Screening and prediction of paroxysmal atrial fibrillation by analysis of heart rate variability parameters". en. In: *Computers in Cardiology 2001. Vol.28 (Cat. No.01CH37287)*. Rotterdam, Netherlands: IEEE, pp. 129–132. DOI: 10.1109/CIC.2001.977608.

Maji, U., Mitra, M., and Pal, S. (Jan. 2014). "Differentiating normal sinus rhythm and atrial fibrillation in ECG signal: A phase rectified signal averaging based approach". In: *Proceedings of The 2014 International Conference on Control, Instrumentation, Energy and Communication (CIEC)*. Calcutta, India: IEEE, pp. 176–180. DOI: 10.1109/CIEC.2014.6959073.

Makowski, D., Pham, T., Lau, Z. J., Brammer, J. C., Lespinasse, F., Pham, H., Schölzel, C., and Chen, S. H. A. (Aug. 2021). "NeuroKit2: A Python toolbox

for neurophysiological signal processing". en. In: *Behavior Research Methods* 53.4, pp. 1689–1696. DOI: 10.3758/s13428-020-01516-y.

Malik, M., Hnatkova, K., Huikuri, H. V., Lombardi, F., Schmidt, G., and Zabel, M. (May 2019). "Rebuttal from Marek Malik, Katerina Hnatkova, Heikki V. Huikuri, Federico Lombardi, Georg Schmidt and Markus Zabel". en. In: *The Journal of Physiology* 597.10, pp. 2603–2604. DOI: 10.1113/JP277962.

Marrouche, N. F., Brachmann, J., Andresen, D., Siebels, J., Boersma, L., Jordaens, L., Merkely, B., Pokushalov, E., Sanders, P., Proff, J., Schunkert, H., Christ, H., Vogt, J., and Bänsch, D. (Feb. 2018). "Catheter Ablation for Atrial Fibrillation with Heart Failure". In: *New England Journal of Medicine* 378.5, pp. 417–427. DOI: 10.1056/NEJMoa1707855.

Martinez, A., Abásolo, D., Alcaraz, R., and Rieta, J. J. (July 2015). "Alteration of the P-wave non-linear dynamics near the onset of paroxysmal atrial fibrillation". en. In: *Medical Engineering & Physics* 37.7, pp. 692–697. DOI: 10.1016/j.medengphy.2015.03.021.

Marwan, N., Carmen Romano, M., Thiel, M., and Kurths, J. (Jan. 2007). "Recurrence plots for the analysis of complex systems". In: *Physics Reports* 438.5, pp. 237–329. DOI: 10.1016/j.physrep.2006.11.001.

Mathur, S., Patel, J., Goldstein, S., Hendrix, J. M., and Jain, A. (2024). "Bispectral Index". eng. In: *StatPearls*. Treasure Island (FL): StatPearls Publishing.

McCulloch, W. S. and Pitts, W. (Dec. 1943). "A logical calculus of the ideas immanent in nervous activity". en. In: *The Bulletin of Mathematical Biophysics* 5.4, pp. 115–133. DOI: 10.1007/BF02478259.

Meeus, P., Dalcq, V., Beauport, D., Declercq, K., and Swine, B. (Aug. 2023). "Medical practice variations: Holter monitor ECG 2022". en. In.

Mendez, M. M., Hsu, M.-C., Yuan, J.-T., and Lynn, K.-S. (Feb. 2022). "A Heart Rate Variability-Based Paroxysmal Atrial Fibrillation Prediction System". en. In: *Applied Sciences* 12.5, p. 2387. DOI: 10.3390/app12052387.

Mitchell, T. M. (1997). *Machine Learning*. en. McGraw-Hill series in computer science. New York: McGraw-Hill.

Mitrega, K., Lip, G. Y. H., Sredniawa, B., Sokal, A., Streb, W., Przyludzki, K., Zdrojewski, T., Wierucki, L., Rutkowski, M., Bandosz, P., Kazmierczak, J., Grodzicki, T., Opolski, G., and Kalarus, Z. (May 2021). "Predicting Silent Atrial Fibrillation in the Elderly: A Report from the NOMED-AF Cross-Sectional Study". eng. In: *Journal of Clinical Medicine* 10.11, p. 2321. DOI: 10.3390/jcm10112321.

Miyakawa, T. (Feb. 2020). "No raw data, no science: another possible source of the reproducibility crisis". In: *Molecular Brain* 13.1, p. 24. DOI: 10.1186/s13041-020-0552-2.

Mohebbi, M. and Ghassemian, H. (Jan. 2012). "Prediction of paroxysmal atrial fibrillation based on non-linear analysis and spectrum and bispectrum features of the heart rate variability signal". en. In: *Computer Methods and Programs in Biomedicine* 105.1, pp. 40–49. DOI: 10.1016/j.cmpb.2010.07.011.

Moody, G., Goldberger, A., McClennen, S., and Swiryn, S. (2001a). "Predicting the onset of paroxysmal atrial fibrillation: the Computers in Cardiology Challenge 2001". en. In: *Computers in Cardiology 2001. Vol.28*. Rotterdam, Netherlands: IEEE, pp. 113–116. DOI: 10.1109/CIC.2001.977604.

Moody, G. and Mark, R. (1983). "A new method for detecting atrial fibrillation using R-R intervals". In: *Computers in Cardiology*.

Moody, G. and Mark, R. (June 2001b). "The impact of the MIT-BIH Arrhythmia Database". en. In: *IEEE Engineering in Medicine and Biology Magazine* 20.3, pp. 45–50. DOI: 10.1109/51.932724.

Mota, S., Ros, E., Toro, F. de, and Ortega, J. (2003). "Genetic Algorithm applied to Paroxysmal Atrial Fibrillation Prediction". en. In: *Artificial Neural Nets Problem Solving Methods*. Ed. by J. Mira and J. R. Álvarez. Vol. 2687. Series Title: Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 345–352. DOI: 10.1007/3-540-44869-1_44.

Myrovali, E., Hristu-Varsakelis, D., Tachmatzidis, D., Antoniadis, A., and Vassilikos, V. (Mar. 2023). "Identifying patients with paroxysmal atrial fibrillation from sinus rhythm ECG using random forests". en. In: *Expert Systems with Applications* 213, p. 118948. DOI: 10.1016/j.eswa.2022.118948.

Narin, A., Isler, Y., Ozer, M., and Perc, M. (Nov. 2018). "Early prediction of paroxysmal atrial fibrillation based on short-term heart rate variability". en. In: *Physica A: Statistical Mechanics and its Applications* 509, pp. 56–65. DOI: 10.1016/j.physa.2018.06.022.

Nasario-Junior, O., Benchimol-Barbosa, P. R., and Nadal, J. (May 2014). "Refining the deceleration capacity index in phase-rectified signal averaging to assess physical conditioning level". In: *Journal of Electrocardiology* 47.3. Publisher: Churchill Livingstone, pp. 306–310. DOI: 10.1016/J.JELECTROCARD.2013.12.006.

Nguyen, D. M. D., Miah, M., Bilodeau, G.-A., and Bouachir, W. (Aug. 2022). "Transformers for 1D signals in Parkinson's disease detection

from gait". English. In: IEEE Computer Society, pp. 5089–5095. DOI: 10.1109/ICPR56361.2022.9956330.

Noseworthy, P. A., Attia, Z. I., Behnken, E. M., Giblon, R. E., Bews, K. A., Liu, S., Gosse, T. A., Linn, Z. D., Deng, Y., Yin, J., Gersh, B. J., Graff-Radford, J., Rabinstein, A. A., Siontis, K. C., Friedman, P. A., and Yao, X. (Oct. 2022). "Artificial intelligence-guided screening for atrial fibrillation using electrocardiogram during sinus rhythm: a prospective non-randomised interventional trial". en. In: *The Lancet* 400.10359, pp. 1206–1212. DOI: 10.1016/S0140-6736(22)01637-3.

Olesen, J. B., Lip, G. Y. H., Hansen, M. L., Hansen, P. R., Tolstrup, J. S., Lindhardsen, J., Selmer, C., Ahlehoff, O., Olsen, A.-M. S., Gislason, G. H., and Torp-Pedersen, C. (Jan. 2011). "Validation of risk stratification schemes for predicting stroke and thromboembolism in patients with atrial fibrillation: nationwide cohort study". en. In: *BMJ* 342.jan31 1, pp. d124–d124. DOI: 10.1136/bmj.d124.

Ozcan, N. and Kuntalp, M. (2017). "Determining Best HRV Indices for PAF Screening using Genetic Algorithm". en. In: p. 4.

Padsalgikar, A. D. (2017). "Cardiovascular System: Structure, Assessment, and Diseases". en. In: *Plastics in Medical Devices for Cardiovascular Applications*. Elsevier, pp. 103–132. DOI: 10.1016/B978-0-323-35885-9.00005-9.

Pan, J. and Tompkins, W. J. (Mar. 1985). "A Real-Time QRS Detection Algorithm". en. In: *IEEE Transactions on Biomedical Engineering* BME-32.3, pp. 230–236. DOI: 10.1109/TBME.1985.325532.

Pan, Q., Zhou, G., Wang, R., Cai, G., Yan, J., Fang, L., and Ning, G. (Dec. 2016). "Do the deceleration/acceleration capacities of heart rate reflect cardiac sympathetic or vagal activity? A model study". eng. In: *Medical & Biological Engineering & Computing* 54.12, pp. 1921–1933. DOI: 10.1007/s11517-016-1486-9.

Panusittikorn, M., Uchaipichat, N., Tantibundhit, C., and Buakamsri, A. (May 2010). "Prediction of paroxysmal atrial fibrillation occurrence with wavelet-based markers". In: *ECTI-CON2010: The 2010 ECTI International Confernce on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*, pp. 342–345.

Parsi, A., Glavin, M., Jones, E., and Byrne, D. (June 2021). "Prediction of paroxysmal atrial fibrillation using new heart rate variability features". en. In: *Computers in Biology and Medicine* 133, p. 104367. DOI: 10.1016/j.compbiomed.2021.104367.

Paszke, A. et al. (Dec. 2019). "PyTorch: an imperative style, high-performance deep learning library". In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. 721, pp. 8026–8037.

Pereira, T., Tran, N., Gadhoumi, K., Pelter, M. M., Do, D. H., Lee, R. J., Colorado, R., Meisel, K., and Hu, X. (2020). "Photoplethysmography based atrial fibrillation detection: a review". In: *npj Digital Medicine* 3.1. Publisher: Springer US. DOI: 10.1038/s41746-019-0207-9.

Perez, M. V. et al. (Nov. 2019). "Large-Scale Assessment of a Smartwatch to Identify Atrial Fibrillation". en. In: *New England Journal of Medicine* 381.20, pp. 1909–1917. DOI: 10.1056/NEJMoa1901183.

Petrutiu, S., Sahakian, A. V., and Swiryn, S. (July 2007). "Abrupt changes in fibrillatory wave characteristics at the termination of paroxysmal atrial fibrillation in humans". In: *EP Europace* 9.7, pp. 466–470. DOI: 10.1093/europace/eum096.

Pichon, A., Roulaud, M., Antoine-Jonville, S., Bisschop, C. de, and Denjean, A. (Jan. 2006). "Spectral analysis of heart rate variability: interchangeability between autoregressive analysis and fast Fourier transform". eng. In: *Journal of Electrocardiology* 39.1, pp. 31–37. DOI: 10.1016/j.jelectrocard.2005.08.001.

Pinhas, I., Toledo, E., Aravot, D., and Akselrod, S. (Oct. 2004). "Bicoherence analysis of new cardiovascular spectral components observed in heart transplant patients: statistical approach for bicoherence thresholding". In: *IEEE Transactions on Biomedical Engineering* 51.10. Conference Name: IEEE Transactions on Biomedical Engineering, pp. 1774–1783. DOI: 10.1109/TBME.2004.831519.

Podd, S. J., Sugihara, C., Furniss, S. S., and Sulke, N. (July 2016). "Are implantable cardiac monitors the 'gold standard' for atrial fibrillation detection? A prospective randomized trial comparing atrial fibrillation monitoring using implantable cardiac monitors and DDDRP permanent pacemakers in post atrial fibrillation ablation patients". In: *EP Europace* 18.7, pp. 1000–1005. DOI: 10.1093/europace/euv367.

Pomeranz, B., Macaulay, R. J., Caudill, M. A., Kutz, I., Adam, D., Gordon, D., Kilborn, K. M., Barger, A. C., Shannon, D. C., and Cohen, R. J. (Jan. 1985). "Assessment of autonomic function in humans by heart rate spectral analysis". eng. In: *The American Journal of Physiology* 248.1 Pt 2, H151–153. DOI: 10.1152/ajpheart.1985.248.1.H151.

Pourbabaee, B. and Lucas, C. (Dec. 2008). "Automatic Detection and Prediction of Paroxysmal Atrial Fibrillation based on Analyzing ECG Signal Feature Classification Methods". en. In: *2008 Cairo International Biomedical Engineering Conference*. Cairo, Egypt: IEEE, pp. 1–4. DOI: 10.1109/CIBEC.2008.4786068.

Pourbabaee, B., Roshtkhari, M. J., and Khorasani, K. (Dec. 2018). "Deep Convolutional Neural Networks and Learning ECG Features for Screening Paroxysmal Atrial Fibrillation Patients". en. In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 48.12, pp. 2095–2104. DOI: 10.1109/TSMC.2017.2705582.

Proesmans, T., Mortelmans, C., Van Haelst, R., Verbrugge, F., Vandervoort, P., and Vaes, B. (Mar. 2019). "Mobile Phone–Based Use of the Photoplethysmography Technique to Detect Atrial Fibrillation in Primary Care: Diagnostic Accuracy Study of the FibriCheck App". en. In: *JMIR mHealth and uHealth* 7.3, e12284. DOI: 10.2196/12284.

Proietti, M., Mairesse, G. H., Goethals, P., Scavee, C., Vijgen, J., Blankoff, I., Vandekerckhove, Y., and Lip, G. Y. (May 2016). "A population screening programme for atrial fibrillation: a report from the Belgian Heart Rhythm Week screening programme". en. In: *Europace*, euw069. DOI: 10.1093/europace/euw069.

Raghunath, S. et al. (Mar. 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". en. In: *Circulation* 143.13, pp. 1287–1298. DOI: 10.1161/CIRCULATIONAHA.120.047829.

Rebecchi, M. et al. (May 2023). "The Autonomic Coumel Triangle: A New Way to Define the Fascinating Relationship between Atrial Fibrillation and the Autonomic Nervous System". en. In: *Life* 13.5, p. 1139. DOI: 10.3390/life13051139.

Reiffel, J. A., Blomström-Lundqvist, C., Boriani, G., Goette, A., Kowey, P. R., Merino, J. L., Piccini, J. P., Saksena, S., and Camm, A. J. (June 2023). "Real-world utilization of the pill-in-the-pocket method for terminating episodes of atrial fibrillation: data from the multinational Antiarrhythmic Interventions for Managing Atrial Fibrillation (AIM-AF) survey". In: *EP Europace* 25.6, euad162. DOI: 10.1093/europace/euad162.

Reyna, M. A., Nsoesie, E. O., and Clifford, G. D. (July 2022). "Rethinking Algorithm Performance Metrics for Artificial Intelligence in Diagnostic Medicine". en. In: *JAMA* 328.4, p. 329. DOI: 10.1001/jama.2022.10561.

Ribeiro, A. H., Ribeiro, M. H., Paixão, G. M., Oliveira, D. M., Gomes, P. R., Canazart, J. A., Ferreira, M. P., Andersson, C. R., Macfarlane, P. W., Wagner, M., Schön, T. B., and Ribeiro, A. L. P. (2020). "Automatic diagnosis of the 12-lead ECG using a deep neural network". In: *Nature Communications* 11.1. arXiv: 1904.01949. DOI: 10.1038/s41467-020-15432-4.

Rivero-Ayerza, M., Scholte op Reimer, W., Lenzen, M., Theuns, D. A., Jordaens, L., Komajda, M., Follath, F., Swedberg, K., and Cleland, J. G. (July 2008). "New-onset atrial fibrillation is an independent predictor of in-hospital mortality in hospitalized heart failure patients: results of the EuroHeart Failure Survey". In: *European Heart Journal* 29.13, pp. 1618–1624. DOI: 10.1093/eurheartj/ehn217.

Rizas, K. D., Freyer, L., Sappler, N., Stülpnagel, L. von, Spielbichler, P., Krasniqi, A., Schreinlechner, M., Wenner, F. N., Theurl, F., Behroz, A., Eiffener, E., Klemm, M. P., Schneidewind, A., Zens, M., Dolejsi, T., Mansmann, U., Massberg, S., and Bauer, A. (Sept. 2022). "Smartphone-based screening for atrial fibrillation: a pragmatic randomized clinical trial". en. In: *Nature Medicine* 28.9. Number: 9 Publisher: Nature Publishing Group, pp. 1823–1830. DOI: 10.1038/s41591-022-01979-w.

Rooney, S. R., Kaufman, R., Murugan, R., Kashani, K. B., Pinsky, M. R., Al-Zaiti, S., Dubrawski, A., Clermont, G., and Miller, J. K. (Nov. 2023). "Forecasting imminent atrial fibrillation in long-term electrocardiogram recordings". In: *Journal of Electrocardiology* 81, pp. 111–116. DOI: 10.1016/j.jelectrocard.2023.08.011.

Ros, E., Mota, S., Fernández, F., Toro, F., and Bernier, J. (Dec. 2004). "ECG Characterization of paroxysmal atrial fibrillation: parameter extraction and automatic diagnosis algorithm". en. In: *Computers in Biology and Medicine* 34.8, pp. 679–696. DOI: 10.1016/j.compbiomed.2003.10.002.

Rosenblatt, F. (1958). "The perceptron: A probabilistic model for information storage and organization in the brain." en. In: *Psychological Review* 65.6, pp. 386–408. DOI: 10.1037/h0042519.

Roth, G. A. et al. (Dec. 2020). "Global Burden of Cardiovascular Diseases and Risk Factors, 1990-2019: Update From the GBD 2019 Study". eng. In: *Journal of the American College of Cardiology* 76.25, pp. 2982–3021. DOI: 10.1016/j.jacc.2020.11.010.

S. Pandya, R., Mao, L., Zhou, H., Zhou, S., Zeng, J., John Popp, A., and Wang, X. (June 2011). "Central Nervous System Agents for Ischemic Stroke: Neuro-

protection Mechanisms". en. In: *Central Nervous System Agents in Medicinal Chemistry* 11.2, pp. 81–97. DOI: 10.2174/187152411796011321.

Saito, T. and Rehmsmeier, M. (Mar. 2015). "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets". en. In: *PLOS ONE* 10.3. Ed. by G. Brock, e0118432. DOI: 10.1371/journal.pone.0118432.

Salahuddin, L., Cho, J., Jeong, M. G., and Kim, D. (2007). "Ultra short term analysis of heart rate variability for monitoring mental stress in mobile settings". eng. In: *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference* 2007, pp. 4656–4659. DOI: 10.1109/IEMBS.2007.4353378.

Saliu, S., Birand, A., and Kudaiberdieva, G. (2002). "Bispectral Analysis of Heart Rate Variability Signal". en. In: p. 4.

Samuel, A. L. (July 1959). "Some Studies in Machine Learning Using the Game of Checkers". In: *IBM Journal of Research and Development* 3.3, pp. 210–229. DOI: 10.1147/rd.33.0210.

Samuel Lévy, Philippe Ricard, Chu-Pak Lau, Ngai-Sang Lok, A. John Camm, Murgatroyd, F. D., Jordaens, L. J., Kappenberger, L. J., Brugada, P., and Ripley, K. L. (Mar. 1997). "Multicenter Low Energy Transvenous Atrial Defibrillation (XAD) Trial Results in Different Subsets of Atrial Fibrillation". In: *Journal of the American College of Cardiology* 29.4. Publisher: American College of Cardiology Foundation, pp. 750–755. DOI: 10.1016/S0735-1097(96)00583-9.

Sanna, T., Diener, H.-C., Passman, R. S., Di Lazzaro, V., Bernstein, R. A., Morillo, C. A., Rymer, M. M., Thijs, V., Rogers, T., Beckers, F., Lindborg, K., and Brachmann, J. (June 2014). "Cryptogenic Stroke and Underlying Atrial Fibrillation". In: *New England Journal of Medicine* 370.26. Publisher: Massachusetts Medical Society, pp. 2478–2486. DOI: 10.1056/NEJMoa1313600.

Sarkar, S., Ritscher, D., and Mehra, R. (Mar. 2008). "A Detector for a Chronic Implantable Atrial Tachyarrhythmia Monitor". en. In: *IEEE Transactions on Biomedical Engineering* 55.3, pp. 1219–1224. DOI: 10.1109/TBME.2007.903707.

Schreier, G., Kastner, P., and Marko, W. (2001). "An automatic ECG processing algorithm to identify patients prone to paroxysmal atrial fibrillation". en. In: *Computers in Cardiology 2001. Vol.28 (Cat. No.01CH37287)*. Rotterdam, Netherlands: IEEE, pp. 133–135. DOI: 10.1109/CIC.2001.977609.

Sebestyén, V., Szabó, Z., Sebestyén, V., and Szabó, Z. (Oct. 2016). "The Role of Electrocardiographic Markers in the Prevention of Atrial and Ventricular

Arrhythmias". en. In: *Echocardiography in Heart Failure and Cardiac Electrophysiology*. IntechOpen. DOI: 10.5772/64460.

Segan, L., Canovas, R., Nanayakkara, S., Chieng, D., Prabhu, S., Voskoboinik, A., Sugumar, H., Ling, L.-H., Lee, G., Morton, J., LaGerche, A., Kaye, D. M., Sanders, P., Kalman, J. M., and Kistler, P. M. (Sept. 2023). "New-onset atrial fibrillation prediction: the HARMS2-AF risk score". In: *European Heart Journal* 44.36, pp. 3443–3452. DOI: 10.1093/eurheartj/ehad375.

Sekelj, S., Sandler, B., Johnston, E., Pollock, K. G., Hill, N. R., Gordon, J., Tsang, C., Khan, S., Ng, F. S., and Farooqui, U. (May 2021). "Detecting undiagnosed atrial fibrillation in UK primary care: Validation of a machine learning prediction algorithm in a retrospective cohort study". en. In: *European Journal of Preventive Cardiology* 28.6, pp. 598–605. DOI: 10.1177/2047487320942338.

Sessa, F., Anna, V., Messina, G., Cibelli, G., Monda, V., Marsala, G., Ruberto, M., Biondi, A., Cascio, O., Bertozzi, G., Pisanelli, D., Maglietta, F., Messina, A., Mollica, M. P., and Salerno, M. (Feb. 2018). "Heart rate variability as predictive factor for sudden cardiac death". In: *Aging (Albany NY)* 10.2, pp. 166–177. DOI: 10.18632/aging.101386.

Shaffer, F. and Ginsberg, J. P. (Sept. 2017). "An Overview of Heart Rate Variability Metrics and Norms". en. In: *Frontiers in Public Health* 5, p. 258. DOI: 10.3389/fpubh.2017.00258.

Shaffer, F., McCraty, R., and Zerr, C. L. (Sept. 2014). "A healthy heart is not a metronome: an integrative review of the heart's anatomy and heart rate variability". en. In: *Frontiers in Psychology* 5. DOI: 10.3389/fpsyg.2014.01040.

Shah, R. U., Bress, A. P., and Vickers, A. J. (Apr. 2022). "Do Prediction Models Do More Harm Than Good?" en. In: *Circulation: Cardiovascular Quality and Outcomes* 15.4. DOI: 10.1161/CIRCOUTCOMES.122.008667.

Shim, M., Lee, S.-H., and Hwang, H.-J. (Dec. 2021). "Inflated prediction accuracy of neuropsychiatric biomarkers caused by data leakage in feature selection". en. In: *Scientific Reports* 11.1, p. 7980. DOI: 10.1038/s41598-021-87157-3.

Singh, J. P., Fontanarava, J., Massé, G. de, Carbonati, T., Li, J., Henry, C., and Fiorina, L. (July 2022). "Short-term prediction of atrial fibrillation from ambulatory monitoring ECG using a deep neural network". en. In: *European Heart Journal - Digital Health* 3.2, pp. 208–217. DOI: 10.1093/ehjdh/ztac014.

Singh, S., Pandey, S. K., Pawar, U., and Janghel, R. R. (2018). "Classification of ECG Arrhythmia using Recurrent Neural Networks". In: *Procedia Computer*

Science 132.Iccids. Publisher: Elsevier B.V. ISBN: 0000000000, pp. 1290–1297. DOI: 10.1016/j.procs.2018.05.045.

Song, H.-S. and Lehrer, P. M. (Mar. 2003). "The Effects of Specific Respiratory Rates on Heart Rate and Heart Rate Variability". en. In: *Applied Psychophysiology and Biofeedback* 28.1, pp. 13–23. DOI: 10.1023/A:1022312815649.

Sun, R. and Wang, Y. (Nov. 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. DOI: 10.1016/j.medengphy.2008.01.008.

Surucu, M., Isler, Y., Perc, M., and Kara, R. (Nov. 2021). "Convolutional neural networks predict the onset of paroxysmal atrial fibrillation: Theory and applications". en. In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 31.11, p. 113119. DOI: 10.1063/5.0069272.

Suzuki, S., Motogi, J., Nakai, H., Matsuzawa, W., Takayanagi, T., Umemoto, T., Hirota, N., Hyodo, A., Satoh, K., Otsuka, T., Arita, T., Yagi, N., and Yamashita, T. (Feb. 2022). "Identifying patients with atrial fibrillation during sinus rhythm on ECG: Significance of the labeling in the artificial intelligence algorithm". en. In: *IJC Heart & Vasculature* 38, p. 100954. DOI: 10.1016/j.ijcha.2022.100954.

Svennberg, E., Friberg, L., Frykman, V., Al-Khalili, F., Engdahl, J., and Rosenqvist, M. (Oct. 2021). "Clinical outcomes in systematic screening for atrial fibrillation (STROKESTOP): a multicentre, parallel group, unmasked, randomised controlled trial". English. In: *The Lancet* 398.10310. Publisher: Elsevier, pp. 1498–1506. DOI: 10.1016/S0140-6736(21)01637-8.

Tan, M. and Le, Q. V. (June 2021). *EfficientNetV2: Smaller Models and Faster Training*. en. arXiv:2104.00298 [cs].

Taran, L. M. and Szilagyi, N. (Jan. 1947). "The duration of the electrical systole, Q-T, in acute rheumatic carditis in children". eng. In: *American Heart Journal* 33.1, pp. 14–26. DOI: 10.1016/0002-8703(47)90421-3.

Tarvainen, M. P., Niskanen, J. P., Lipponen, J. A., Ranta-aho, P. O., and Karjalainen, P. A. (Jan. 2014). "Kubios HRV – Heart rate variability analysis software". In: *Computer Methods and Programs in Biomedicine* 113.1. Publisher: Elsevier, pp. 210–220. DOI: 10.1016/J.CMPB.2013.07.024.

Task Force of The European Society of Cardiology and The North American (Mar. 1996). "Heart Rate Variability: Standards of Measurement, Physiological Interpretation, and Clinical Use". en. In: *Circulation* 93.5, pp. 1043–1065. DOI: 10.1161/01.CIR.93.5.1043.

Tateno, K. and Glass, L. (Nov. 2001). "Automatic detection of atrial fibrillation using the coefficient of variation and density histograms of RR and deltaRR intervals". eng. In: *Medical & Biological Engineering & Computing* 39.6, pp. 664–671. DOI: 10.1007/BF02345439.

Tharwat, A. (Jan. 2020). "Classification assessment methods". In: *Applied Computing and Informatics* 17.1. Publisher: Emerald Publishing Limited, pp. 168–192. DOI: 10.1016/j.aci.2018.08.003.

Thong, T., McNames, J., Aboy, M., and Goldstein, B. (Apr. 2004). "Prediction of Paroxysmal Atrial Fibrillation by Analysis of Atrial Premature Complexes". en. In: *IEEE Transactions on Biomedical Engineering* 51.4, pp. 561–569. DOI: 10.1109/TBME.2003.821030.

Thuraisingham, R. A. (2009). "A Classification System to Detect Congestive Heart Failure Using Second-Order Difference Plot of RR Intervals". en. In: *Cardiology Research and Practice* 2009, pp. 1–7. DOI: 10.4061/2009/807379.

Tison, G. H., Sanchez, J. M., Ballinger, B., Singh, A., Olgin, J. E., Pletcher, M. J., Vittinghoff, E., Lee, E. S., Fan, S. M., Gladstone, R. A., Mikell, C., Sohoni, N., Hsieh, J., and Marcus, G. M. (May 2018). "Passive detection of atrial fibrillation using a commercially available smartwatch". In: *JAMA Cardiology* 3.5. Publisher: American Medical Association, pp. 409–416. DOI: 10.1001/jamacardio.2018.0136.

Toichi, M., Sugiura, T., Murai, T., and Sengoku, A. (Jan. 1997). "A new method of assessing cardiac autonomic function and its comparison with spectral analysis and coefficient of variation of R–R interval". In: *Journal of the Autonomic Nervous System* 62.1, pp. 79–84. DOI: 10.1016/S0165-1838(96)00112-9.

Turakhia, M. P., Hoang, D. D., Zimetbaum, P., Miller, J. D., Froelicher, V. F., Kumar, U. N., Xu, X., Yang, F., and Heidenreich, P. A. (Aug. 2013). "Diagnostic utility of a novel leadless arrhythmia monitoring device". eng. In: *The American Journal of Cardiology* 112.4, pp. 520–524. DOI: 10.1016/j.amjcard.2013.04.017.

Tzou, H.-A., Lin, S.-F., and Chen, P.-S. (Nov. 2021). "Paroxysmal atrial fibrillation prediction based on morphological variant P-wave analysis with wideband ECG and deep learning". en. In: *Computer Methods and Programs in Biomedicine* 211, p. 106396. DOI: 10.1016/j.cmpb.2021.106396.

Vandewiele, G., Dehaene, I., Kovács, G., Sterckx, L., Janssens, O., Ongenae, F., De Backere, F., De Turck, F., Roelens, K., Decruyenaere, J., Van Hoecke, S., and Demeester, T. (Jan. 2021). "Overly Optimistic Prediction Results on

Imbalanced Data: a Case Study of Flaws and Benefits when Applying Oversampling". en. In: *Artificial Intelligence in Medicine* 111. arXiv:2001.06296 [cs, eess, stat], p. 101987. DOI: 10.1016/j.artmed.2020.101987.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (Dec. 2017). *Attention Is All You Need*. en. arXiv:1706.03762 [cs].

Vijayan, V., Nys, J., Joosen, P., and Vandendriessche, B. (Nov. 2023). "A digital health solution for long-term cardiac monitoring supporting a value-based stroke care path". In: *European Heart Journal* 44.Supplement_2, ehad655.2961. DOI: 10.1093/eurheartj/ehad655.2961.

Vikman, S., Mäkikallio, T. H., Yli-Mäyry, S., Pikkujämsä, S., Koivisto, A. M., Reinikainen, P., Airaksinen, K. E., and Huikuri, H. V. (1999). "Altered complexity and correlation properties of R-R interval dynamics before the spontaneous onset of paroxysmal atrial fibrillation". In: *Circulation* 100.20. DOI: 10.1161/01.CIR.100.20.2079.

Vries, A. d. (Oct. 2023). "The growing energy footprint of artificial intelligence". English. In: *Joule* 0.0. DOI: 10.1016/j.joule.2023.09.004.

Wafa, H. A., Wolfe, C. D., Emmett, E., Roth, G. A., Johnson, C. O., and Wang, Y. (Aug. 2020). "Burden of Stroke in Europe: Thirty-Year Projections of Incidence, Prevalence, Deaths, and Disability-Adjusted Life Years". en. In: *Stroke* 51.8, pp. 2418–2427. DOI: 10.1161/STROKEAHA.120.029606.

Wagner, P., Strodthoff, N., Bousseljot, R.-D., Kreiseler, D., Lunze, F. I., Samek, W., and Schaeffter, T. (Dec. 2020). "PTB-XL, a large publicly available electrocardiography dataset". en. In: *Scientific Data* 7.1, p. 154. DOI: 10.1038/s41597-020-0495-6.

Waktare, J. E., Hnatkova, K., Sopher, S. M., Murgatroyd, F. D., Guo, X., Camm, A. J., and Malik, M. (Feb. 2001). "The role of atrial ectopics in initiating paroxysmal atrial fibrillation". eng. In: *European Heart Journal* 22.4, pp. 333–339. DOI: 10.1053/euhj.2000.2517.

Waller, A. D. (1887). "A Demonstration on Man of Electromotive Changes accompanying the Heart's Beat". en. In: *The Journal of Physiology* 8.5, pp. 229–234. DOI: 10.1113/jphysiol.1887.sp000257.

Walsh, I. et al. (Oct. 2021). "DOME: recommendations for supervised machine learning validation in biology". In: *Nature Methods* 18.10. arXiv: 2006.16189 Publisher: Nature Research, pp. 1122–1127. DOI: 10.1038/s41592-021-01205-4.

Wang, X., Ma, C., Zhang, X., Gao, H., Clifford, G., and Liu, C. (2021). *Paroxysmal Atrial Fibrillation Events Detection from Dynamic ECG Recordings: The 4th China Physiological Signal Challenge 2021*. DOI: 10.13026/KSYA-QW89.

Wasserlauf, J., You, C., Patel, R., Valys, A., Albert, D., and Passman, R. (June 2019). "Smartwatch Performance for the Detection and Quantification of Atrial Fibrillation". en. In: *Circulation: Arrhythmia and Electrophysiology* 12.6, e006834. DOI: 10.1161/CIRCEP.118.006834.

Welch, P. (June 1967). "The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms". en. In: *IEEE Transactions on Audio and Electroacoustics* 15.2, pp. 70–73. DOI: 10.1109/TAU.1967.1161901.

Wellens, H. J. J., Lau, C.-P., Lüderitz, B., Akhtar, M., Waldo, A. L., Camm, A. J., Timmermans, C., Tse, H.-F., Jung, W., Jordaens, L., and Ayers, G. (Oct. 1998). "Atrioverter: An Implantable Device for the Treatment of Atrial Fibrillation". In: *Circulation* 98.16. Publisher: American Heart Association, pp. 1651–1656. DOI: 10.1161/01.CIR.98.16.1651.

WHO (Nov. 2022). *Cardiovascular diseases (CVDs)*. en.

Wolf, P. A., Abbott, R. D., and Kannel, W. B. (Aug. 1991). "Atrial fibrillation as an independent risk factor for stroke: the Framingham Study." en. In: *Stroke* 22.8, pp. 983–988. DOI: 10.1161/01.STR.22.8.983.

Xia, Y., Wulan, N., Wang, K., and Zhang, H. (Feb. 2018). "Detecting atrial fibrillation by deep convolutional neural networks". en. In: *Computers in Biology and Medicine* 93, pp. 84–92. DOI: 10.1016/j.compbiomed.2017.12.007.

Yan, B. P., Lai, W. H. S., Chan, C. K. Y., Au, A. C. K., Freedman, B., Poh, Y. C., and Poh, M.-Z. (Jan. 2020). "High-Throughput, Contact-Free Detection of Atrial Fibrillation From Video With Deep Learning". In: *JAMA Cardiology* 5.1, pp. 105–107. DOI: 10.1001/jamacardio.2019.4004.

Yang, A. and Yin, H. (2001). "Prediction of paroxysmal atrial fibrillation by footprint analysis". en. In: *Computers in Cardiology 2001. Vol.28*. Rotterdam, Netherlands: IEEE, pp. 401–404. DOI: 10.1109/CIC.2001.977677.

Yu, S.-N. and Lee, M.-Y. (Aug. 2012). "Bispectral analysis and genetic algorithm for congestive heart failure recognition based on heart rate variability". In: *Computers in Biology and Medicine* 42.8, pp. 816–825. DOI: 10.1016/j.compbiomed.2012.06.005.

Yuan, N., Duffy, G., Dhruva, S. S., Oesterle, A., Pellegrini, C. N., Theurer, J., Vali, M., Heidenreich, P. A., Keyhani, S., and Ouyang, D. (Oct. 2023). "Deep Learning of Electrocardiograms in Sinus Rhythm From US Veterans to Pre-

dict Atrial Fibrillation". In: *JAMA Cardiology*. DOI: 10.1001/jamacardio.2023.3701.

Zheng, J., Zhang, J., Danioko, S., Yao, H., Guo, H., and Rakovski, C. (Feb. 2020). "A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients". en. In: *Scientific Data* 7.1, p. 48. DOI: 10.1038/s41597-020-0386-x.

Zhu, C., Hanna, P., Rajendran, P. S., and Shivkumar, K. (Aug. 2019). "Neuromodulation for Ventricular Tachycardia and Atrial Fibrillation". In: *JACC: Clinical Electrophysiology* 5.8. Publisher: American College of Cardiology Foundation, pp. 881–896. DOI: 10.1016/j.jacep.2019.06.009.

Zhu, H., Cheng, C., Yin, H., Li, X., Zuo, P., Ding, J., Lin, F., Wang, J., Zhou, B., Li, Y., Hu, S., Xiong, Y., Wang, B., Wan, G., Yang, X., and Yuan, Y. (July 2020). "Automatic multilabel electrocardiogram diagnosis of heart rhythm or conduction abnormalities with deep learning: a cohort study". English. In: *The Lancet Digital Health* 2.7. Publisher: Elsevier, e348–e357. DOI: 10.1016/S2589-7500(20)30107-2.

Zimetbaum, P. and Goldman, A. (Oct. 2010). "Ambulatory Arrhythmia Monitoring". In: *Circulation* 122.16. Publisher: American Heart Association. DOI: 10.1161/CIRCULATIONAHA.109.925610.

Zong, W., Mukkamala, R., and Mark, R. (Sept. 2001). "A methodology for predicting paroxysmal atrial fibrillation based on ECG arrhythmia feature analysis". In: *Computers in Cardiology 2001. Vol.28 (Cat. No.01CH37287)*. ISSN: 0276-6547, pp. 125–128. DOI: 10.1109/CIC.2001.977607.