Heart Rate Variability and Irregularity in Atrial Fibrillation

Marna Devilde

June 2026

Master's Thesis in
Biomedical Engineering

Supervisor: Frida Sandberg
Examiner: Martin Stridh

LUND UNIVERSITY

Faculty of Engineering LTH
Department of Biomedical Engineering

Abstract

Atrial fibrillation (AF) is the most common cardiac arrhythmia worldwide, affecting an estimated 59.7 million people and imposing a growing burden on public healthcare systems. AF is characterized by a rapid and irregular heart rate. Specifically, understanding how rate-control therapies affect RR-interval variability and irregularity is essential for optimizing treatment. Accurate and reliable heart rate variability (HRV) analysis depends critically on robust QRS-detection and quality assurance of the resulting RR-interval series.

This thesis investigates HRV and beat-to-beat irregularity in patients with permanent AF using 24-hour Holter ECG recordings from the RATAF II study, a randomized trial comparing rate control with diltiazem and metoprolol. Three interconnected objectives were addressed. First, the Pan Tompkins QRS-detection algorithm was evaluated by comparing its output against RR intervals extracted by clinical hospital software, revealing generally close agreement with minor timing discrepancies attributable to the 125 Hz sampling resolution. Second, two quality assurance approaches were developed: an intersection-based method combining both detectors, and a Bidirectional Long Short-Term Memory (BiLSTM) deep learning classifier trained to distinguish clean beats from noise. The BiLSTM achieved 97.0% accuracy, 94.2% sensitivity, and 99.8% specificity on the validation set, though a tendency to reject beats affected by baseline wander was identified as a limitation.

Lastly, HRV parameters were computed across five distinct RR interval pipelines and compared systematically. Results were remarkably consistent across all pipelines, suggesting that for long 24-hour recordings in this patient population, the choice of QRS-detector and quality assurance method has limited impact on HRV metrics. Across all pipelines, rate control medication reduced heart rate and increased variance-based HRV parameters, while entropy-based irregularity measures remained stable, indicating that rate control does not alter the underlying irregular nature of AF.

Populärvetenskaplig Sammanfattning

Hjärtats variabilitet och irregularitet - spelar analysmetoden någon roll?

Förmaksflimmer är världens vanligaste arytmi och drabbar ungefär 59.7 miljoner människor globalt. Hjärtat slår oregelbundet och för snabbt, och det är just den oregelbundenheten som är kärnan i det här arbetet.

Hjärtfrekvensvariation (HRV) är ett mått på hur tidsskillnaden mellan olika hjärtslag varierar, och kan därmed ge information om hjärtats pumpförmåga. Men för att räkna ut HRV behöver man först identifiera varje enskilt hjärtslag i en EKG-inspelning. Det låter enkelt, men i verkligheten är 24-timmars inspelningar fyllda av brus, rörelsestörningar och signaler som ser ut som hjärtslag utan att vara det.

Det här masterarbetet undersöker om valet av metod för att hitta hjärtslagen, och filtrera bort brus eller artefakter, faktiskt påverkar slutresultatet av HRV-analysen. Data kom från RATAF II studien, där patienter med permanent förmaksflimmer behandlats med ett av två frekvensreglerande läkemedel för att sänka hjärtfrekvensen.

Fem olika analysmetoder jämfördes. Dels användes den klassiska Pan Tompkins algoritmen från 1985, dels en klinisk sjukhusprogramvara, och dessutom ett egenutvecklat neuralt nätverk, ett så kallat BiLSTM-nätverk som tränades att skilja riktiga hjärtslag från brus. Det neurala nätverket uppnådde 97% noggrannhet på testdata. Men nätverket visade sig ha en svaghet, den avvisade ibland tydliga hjärtslag som drabbats av baslinjevandring, vilket är ett vanligt fenomen vid långtidsinspelningar av EKG.

Det kanske mest intressanta fyndet är att alla fem metoderna gav väldigt lika HRV-värden, trots att metoderna ibland skiljde sig åt i vilka slag de räknade med. Det tyder på att HRV-analys vid förmaksflimmer är robust, åtminstone i den undersökta populationen.

Läkemedlen sänkte hjärtfrekvensen och ökade variationen, precis som förväntat. Men oregelbundenheten i sig, mätt med entropimått, förändrades inte. Förmaksflimret fortsätter att vara oregelbundet, även med frekvensreglerande läkemedel.

Acknowledgment

I want to extend my sincere thanks to my supervisor Frida Sandberg, whose guidance and support made this thesis possible. Your feedback, encouragement, and expertise throughout the process have been appreciated.

I am also grateful to my family and friends. Thank you for your patience, endless support and for always believing in me. Your encouragement throughout this process have been invaluable, and I am thankful for the strength and motivation you have given me.

Declaration of AI assistance

Generative AI tools were mainly used for language support, including grammar and spelling checks, and for improving the flow of the text. Besides writing aids, some Matlab code to generate nice and neat plots was also (partly) generated with AI.

No AI-generated content was used without review and revision by the author. All research, analysis, results and conclusions presented in this thesis are the author's own work.

Contents

1 Introduction

Atrial fibrillation is the most common cardiac arrhythmia in the world and its global burden on the public health care system continues to rise dramatically. [1] In 2019, an estimated 59.7 million people were living with AF around the globe. [1] Additionally, the prevalence of AF is projected to double over the coming decades, this driven by an aging population, increased clinical awareness, and the expanding use of diagnostic and screening technologies. [2] AF is characterized by a fast and irregularly contracting atria, due to misfiring of the sinus node. As a result, the AV node is flooded with signals which leads to the heart beating abnormally irregular and fast, creating symptoms like chest pain. [3] AF markedly impairs quality of life and elevates the risk of serious complications including ischaemic stroke and heart failure. [1][2] Improving the tools available for monitoring and characterizing AF is therefore highly relevant.

A large body of evidence links cardiac autonomic nervous system function directly to mortality risk. [4] Heart Rate Variability (HRV) has been recognized as a promising noninvasive marker for autonomic activity in sinus rhythm. HRV refers to the variation in the duration of successive cardiac cycles, typically quantified from series of RR-intervals extracted from electrocardiogram (ECG) recordings. [5] HRV and autonomic function is considerably more complex in AF than in sinus rhythm. Nevertheless, studying RR variability and irregularity in AF is clinically relevant, beat-to-beat variability effects the heart's pumping efficiency and can be linked to patient well-being. [6] In the RATAF II study, patients with permanent AF were randomized to rate control treatment with one of two medications. By computing HRV parameters before and after treatment, it is possible to examine whether and how the two drugs had an impact.

Any HRV analysis begins with accurate detection of R-peaks in the ECG signal and the construction of RR interval series. The Pan Tompkins algorithm, proposed in 1985, remains today the most widely used QRS-complex detector in cardiac monitoring applications. However, the Pan Tompkins algorithm struggles to reliably detect QRS-complexes in the presence of noise and poor signal quality. [7] Errors in QRS-detection will directly affect the HRV parameters, a missed beat produces an artificially long RR interval, while a false positive detection produces a spuriously short one, both distorting HRV parameters. The identification and removal of noisy or unreliable signal segments and inaccurate QRS-detections is, therefore, a critical step in a reliable HRV analysis.

1.1 Aims and Objectives

A key question is how much the choice of QRS-detector and noise removal technique actually matters for the resulting HRV parameters, particularly in the context of AF and long-term ECG recordings. If different pipelines yield substantially different HRV parameter values, then comparisons of HRV in AF patients across studies are extremely limited. On the contrary,

if HRV parameters are robust to reasonable pipeline choices, that would support confidence in existing literature. Addressing this question requires a systematic comparison of multiple RR interval series derived from the same recordings under different pipeline configurations.

This master thesis investigates HRV and beat-to-beat irregularity in patients with AF using 24-hour ECG recordings from the RATAF II study [8]. Three interconnected objectives are investigated: first, the Pan Tompkins QRS-detection algorithm is evaluated by comparing its output against RR intervals provided by clinical hospital software. Second, two quality assurance approaches are developed and applied, a simpler method based on the intersection of the two QRS-detection methods, and a more advanced noise classifier using a BiLSTM deep learning network. Third, a set of HRV parameters is computed across multiple RR interval series derived from the same ECG recordings but with different pipeline configurations, with the aim of assessing how detector and quality assurance choices affect the resulting HRV parameters.

1.2 Outline

The remainder of this thesis is structured as follows. Section 2 provides background on atrial fibrillation, the electrocardiogram, QRS-detection with a focus on the Pan Tompkins algorithm, deep learning with particular attention to LSTM and BiLSTM networks, and the HRV parameters used in this study. Section 3 describes the methods, covering the RATAF II dataset, ECG pre-processing, QRS-detection, the RR series comparison, the development and training of the BiLSTM beat classifier, and the HRV analysis pipeline. Section 4 presents the results of the RR series comparison between the Pan Tompkins algorithm and the clinical hospital software, the classifier evaluation on the validation set and on the RATAF II recordings, and the HRV parameter comparison across all five pipelines. Section 5 discusses the findings in relation to each objective and in the context of existing literature, and identifies limitations and directions for future work. Finally, Section 6 summarizes the main conclusions of the thesis.

2 Background

2.1 Atrial Fibrillation

Atrial fibrillation (AF) is one of the most common cardiac arrhythmias in the world. AF is characterized by irregular heartbeats and palpitation attacks caused by rapid and irregular contractions of the atria. Symptoms of AF include fatigue, chest pain, shortness of breath and dizziness. While 90% of patients report varying degrees of symptomatic severity, instances of asymptomatic AF are documented. Although, episodes of AF can be asymptomatic even in patients with symptoms. Diagnostic evaluation of AF includes a standard 12-lead ECG.

Furthermore, patients with AF have a higher risk of heart failure and stroke. [2] AF accounts for one in five of all strokes and is associated with significantly higher morbidity and mortality rates compared to non-AF-related strokes [9]. The absence or presence of symptoms is not linked to stroke or heart failure, but symptoms do impact the patient quality of life. [2]

There are different classifications of AF which are based on presentation, duration and spontaneous termination of episodes. If AF has not been diagnosed before, regardless of duration, pattern or symptoms, it is classified as *first-diagnosed AF*. The other main classes are *paroxysmal AF*, *persistent AF* and *permanent AF*. Paroxysmal AF is AF which terminates spontaneously within 7 days or with the help of intervention. Persistent AF is AF which does not self-terminate, but treatment options are available. However, persistent AF is when no further attempts of restoring sinus rhythm are planned. [2]

AF alone is not life threatening, but it is a serious condition that needs to be treated to prevent stroke, which can be fatal. One develops AF when changes to your heart's tissue or electrical system occur. The risk of AF increases with age, with high blood pressure and also with alcohol consumption. [10]

2.1.1 Physiology

The heart has four chambers, the two upper ones are called the atria and the two lower ones are referred to as ventricles. The sinoatrial node or sinus node is a group of cells in the upper right chamber, and is responsible for creating the signal that initiates each heartbeat. In a normal heartbeat, the signals move across the upper chambers and arrive at another group of cells, the atrioventricular (AV) node. In the AV node, the signals normally slow down and then enter the lower chambers. In a healthy heart, this process goes smoothly, and a normal resting heart rate is around 60 to 100 bpm. [10]

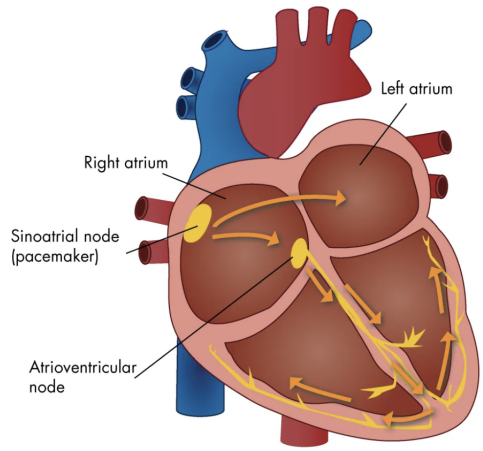

Figure 1: An illustrative image of the anatomy of the heart. [11]

However, in a heart with atrial fibrillation the cardiac conduction system does not work properly. In a healthy heart, the sinus node generates electrical impulses in a steady, coordinated pattern. However, during AF, this normal pacemaker activity is overridden by chaotic electrical impulses arising from multiple ectopic foci distributed across the tissue. [9] As a result, the atria contract quickly and irregularly. The AV node is flooded with atrial impulses making the ventricles also contract irregularly and fast. The typical heartbeat that a human feel is from the ventricles contracting. [3] The heart rate during atrial fibrillation may range from 100 to 175 bpm. [10]

2.1.2 Treatment

As mentioned before, most patients with AF need treatment or interventions to improve patient outcomes. The two most common concepts are rate control (slowing the heart rate) and rhythm control (restoring and maintaining sinus rhythm), but in reality most patients require a combination of both approaches. [2]

Rate control drugs are recommended for patients with AF, either as an initial treatment, as an adjunct to rhythm control approaches or as a sole treatment. This to control the heart rate and reduce symptoms. Beta-blockers or calcium channel-blockers are a common first-choice rate control drug for patients with AF. A lenient target of reducing the resting heart rate to under 110 bpm is recommended as the starting point for most patients. Stricter control (under 80 bpm) is only warranted if symptoms persist or if other arrhythmias are suspected. [2]

Rhythm control is never used on its own. It involves, for example, cardioversion (shocking or medicating the heart back into sinus rhythm), antiarrhythmic drugs (AADs), or catheter

ablation (burning away faulty electrical pathways). For stable patients, it is acceptable to wait up to 48 hours to see if the heart converts back to sinus rhythm on its own, rather than immediately intervening. However, if the patient is unstable, interventions are required right away. Rhythm control strategies have evolved a lot over the last few years, since the experience of using safe antiarrhythmic drugs has increased. Maintaining sinus rhythm through rhythm control strategies has been shown to improve patients' quality of life. [2]

2.2 Electrocardiogram (ECG)

ECG stands for electrocardiogram or electrocardiography and measures the electric activity of the heart. The heart is a pump, and is driven by intrinsic electrical impulses making it beat. [12] The heart contracts in response to potentials moving through the chambers during one cardiac cycle. When the cardiac cycle begins, the sinus node generates an electrical impulse, and cells in the cardiac tissue starts to depolarize creating an electrical wavefront moving through the heart with polarized cells at the front followed by depolarized cells at the back. This charge separation results in a dipole across the heart. The dipole is moving, therefore, it causes current flow in the surrounding body fluids, which results in a fluctuating electrical field around the heart. The electrical field can be captured by electrodes on the skin. The orientation of the electrodes with respects to the dipole impact the intensity of the detected voltage. The mass of the tissue involved in creating the dipole at any time is proportional to the amplitude of the voltage. This is what the ECG is showing. [13]

In other words, the ECG is a record of the electrical impulses flowing through the heart. However, these impulses can be viewed from many different angles to capture the electrical activity across the four chambers of the heart. The different views of the heart is referred to as different 'leads'. The leads do not describe the number of electrodes placed on the body. 12-lead ECG are most common, but there are instances where a 3-lead ECG may be sufficient (often seen in ambulance recordings). Single-lead ECG are becoming more and more popular, often implemented in smartphones and smartwatches. [12]

When the ECG is recording, signals of voltage vs time are gathered and displayed in millivolts (mV) vs seconds [13]. A normal rhythm is called a sinus rhythm and contains a P-wave, a QRS-complex and a T-wave, see Figure 2 for an example. The deflection from the baseline depends on the direction of the traveling electrical impulse. A positive deflection indicates an electrical impulse traveling towards the detecting electrode and a negative deflection means the electrical impulse is traveling away from the detecting electrode. [14]

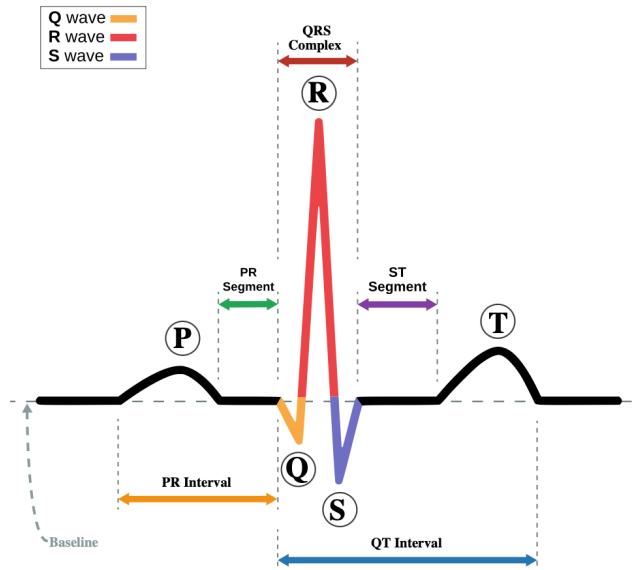

Figure 2: A single QRS-complex from an ECG recording. [15]

2.2.1 ECG and Atrial Fibrillation

AF presents in ECG recordings as irregular RR intervals (time between consecutive R-peaks), absence of distinct P-waves and presence of underlying fibrillatory waves (f-waves), see Figure 3. [9] In normal sinus rhythm, the sinus node fires a single, organized electrical impulse that spreads across the atria in a coordinated wave, this depolarizes the atria and shows up as a neat P wave on the ECG. [13] In AF, instead of one organized impulse, the atria fires chaotically with multiple disorganized electrical wavelets simultaneously. The atria never fully depolarizes or contracts properly, resulting in no P-wave showing in the ECG recording, instead f-waves are present representing the quivering of the atria. F-waves are low-amplitude and irregular waves at the baseline and they are the direct electrical signature of that chaotic atrial activity, see Figure 3. They reflect the hundreds of small, disorganized depolarization wavelets firing randomly across the atria.

AF also presents as the heart beating irregularly, leading to irregular RR intervals. In AF, the AV node is bombarded with hundreds of chaotic impulses. Which impulses get through depends on the timing of each impulse relative to the AV node's refractory period. This produces an irregular ventricular response, which shows up on the ECG as RR intervals that vary unpredictably from beat to beat. [10]

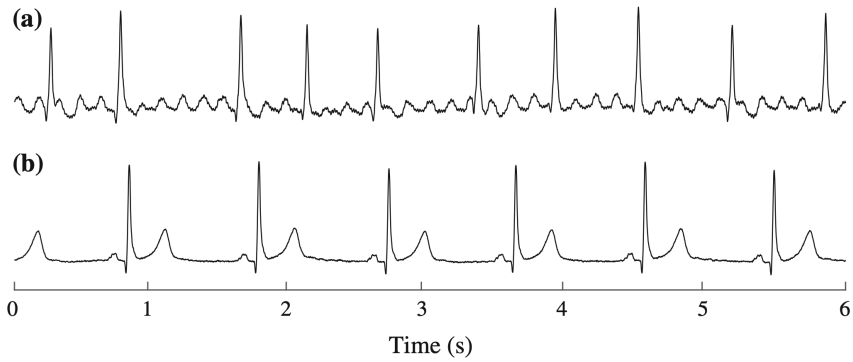

Figure 3: **(a)** An illustrative ECG recording with AF. **(b)** A representative ECG recording showing normal sinus rhythm. [9]

2.2.2 Ambulatory ECG Monitoring

Ambulatory ECG monitors are devices recording the heart's electrical activity while the patient is walking around (are ambulatory) and doing normal activities. These devices are given to patients suspected of having an arrhythmia to take home. These types of recordings are useful when a patient is suspected of having an intermittent arrhythmia, where an ECG test at the hospital might miss it. Most ambulatory ECG monitors use small, adhesive chest electrodes connected by wires to a compact recorder worn on a belt, sometimes they are referred to as Holter monitors. Patients are instructed to log the exact time of any symptoms in a diary. By correlating this diary with the recorded ECG data, physicians can determine if an arrhythmia is causing the patient's symptoms and maybe identify its type. [16]

However, compared to hospital ECG monitoring, ambulatory cardiac monitoring faces greater heartbeat detection difficulties due to more noise and artifact induced by daily patient activity. [17] The success behind computer-based ECG analysis comes from the ability to improve poor signal quality with different signal processing algorithms, which stems from knowledge of both signal properties and noise properties. It is therefore important be familiar with the most common types of noise and artifacts in an ECG recording, when dealing with ambulatory ECG recordings. [18]

Baseline wander is a type of low-frequency noise that appears on the ECG and can make accurate interpretation difficult, for an example see Figure 4. It is particularly problematic because it distorts the baseline, which is needed as a reference point for many ECG measurements. Common causes of baseline wander in ECG recordings include sweating, breathing, movement, and poor electrode contact, and it occurs most often during exercise. Despite usually being a low-frequency phenomenon, the amplitude of the wander can sometimes be large enough to drown out the QRS-complex entirely. [18] Breathing affects ECG recordings,

not just by influencing heart rate, but also by changing the shape of individual heartbeats. This happens because breathing causes the chest to move, shifts the position of the heart, and alters how well the lungs conduct electricity. Together, these changes affect the direction of the heart's electrical activity with each breath, which shows up as subtle variations in beat shape from one heartbeat to the next. One visible example of this is a rhythmic rise and fall in QRS amplitude that follows the breathing cycle, if this pattern repeats roughly every 5 seconds, it suggests the person is breathing at around 12 breaths per minute. While this kind of respiratory interference is generally an unwanted feature of the ECG signal, it can be put to use as a way of estimating the patient's breathing rate. [18]

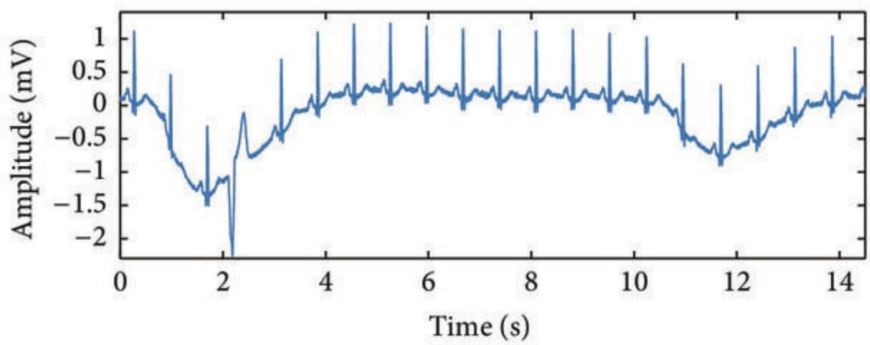

Figure 4: An ECG signal affected by baseline wander. [19]

When the skin stretches around an electrode it creates another common artifact in ECG recordings, see Figure 5. It changes the electrical properties of the skin and produces, what are called, electrode motion artifacts. These are harder to deal with than baseline wander because they occur at frequencies similar to the heart's own electrical signals. On the ECG they show up as large waveforms that can look like QRS-complexes, which is a particular problem during ambulatory monitoring, where they frequently cause heartbeats to be falsely detected. [18]

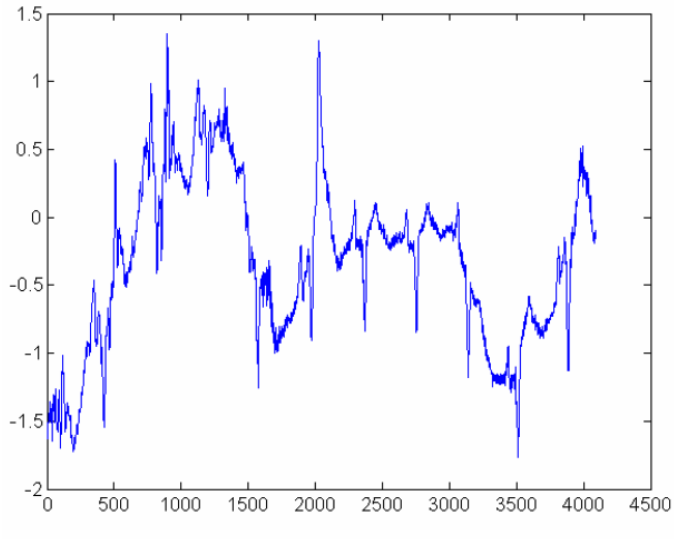

Figure 5: An ECG signal with electrode motion artifacts. [20]

When skeletal muscles contract, they produce electrical signals that can interfere with the ECG, this is called electromyographic noise (EMG noise) see Figure 6. It is commonly seen in ECGs recorded during movement or exercise, and can appear either as brief bursts caused by sudden movements or as more constant background interference. The problem with EMG noise is that it occurs at similar frequencies to the QRS-complex, making it very hard to remove without also distorting the heart signal itself. [18]

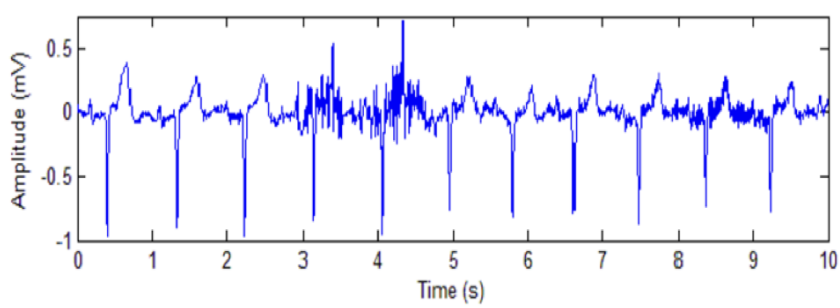

Figure 6: An ECG signal affected by electromyographic (EMG) noise. [21]

2.3 QRS-detection

2.3.1 The Pan Tompkins QRS-detection algorithm

The Pan Tompkins algorithm, published by Jiapu Pan and Willis Tompkins in 1985, is a real-time algorithm for detecting QRS-complexes in ECG-signals. It works by putting the

ECG-signal through a series of processing steps, each designed to make the QRS-complex easier to detect while suppressing noise. [22]

The first step of the Pan Tompkins QRS-detection algorithm is to apply a bandpass filter to the raw ECG-signal. This is done to remove unwanted ECG noise, like baseline wander at low frequencies and EMG/electrode motion artifacts at higher frequencies. The bandpass filter is designed to only keep the frequency range characteristic for QRS-complexes, this is approximately 5 to 15 Hz. [22]

After filtering, the signal is passed through a derivative filter to provide slope information. This highlights the steep part of the R-wave. The differentiated signal is then squared point by point. This makes all values positive and it amplifies larger values disproportionately, making higher amplitude QRS-features more prominent relative to smaller noise or other waveforms. [22]

Finally, a moving window integrator smooths the resulting signal into a waveform with distinct peaks corresponding to each QRS-complex. The window length is a critical parameter, it should be approximately the same width as the widest possible QRS-complex. Too wide a window risks merging the QRS-complex with T-waves, while to narrow a window may cause a single QRS-complex to produce several peaks. The width of the window is empirically decided to be around 150 ms. [22]

QRS-detection is then performed using an adaptive dual-threshold approach. Two sets of thresholds are maintained in parallel, one set applied to the integrated waveform and one applied to the bandpass filtered signal. The idea behind two sets of threshold is to increase reliability of detection, rather than just using one waveform. The algorithm keeps track of two running estimates at all times. One estimate is the average height of peaks it has already confirmed as QRS-complexes and the other estimate is the average height of peaks it has classified as noise. The threshold is then set somewhere between these two values, high enough to ignore noise, but low enough to still catch real QRS-complexes. Every time a new peak is detected and classified, both estimates update, so the threshold floats up or down to follow the signal. For a peak to be classified as a QRS-complex it must pass the threshold in both the integrated waveform and in the bandpass filtered signal. At the start of each ECG-signal, a learning phase of 2 seconds is needed to establish a signal estimate and a noise estimate. [22]

A search back strategy is used to find missed beats and therefore reduce the number of false negatives. The algorithm continuously keeps track of two RR interval averages as well. One average is of the eight most recent RR intervals and the other is the average of the eight most recent RR intervals which are physiologically plausible. This is done to be able to quickly

adapt to rapidly changing heart rates or irregular rhythms. If no QRS-complex is detected in 166% of the average RR interval, the algorithm searches back with a lower threshold. [22]

There is a 200 ms refractory period before the next QRS-complex can be detected once a valid QRS-complex is detected. The idea behind this is to eliminate multiple triggers on the same QRS-complex during this time since it is not physiologically possible to have two QRS-complexes within this time. However, if a QRS detection happens close to the end of the refractory period but within 360 ms of the previous QRS-complex, the question is whether this is a valid QRS-complex or a T-wave. Its slope is then compared to that of the preceding QRS-complex. If the slope is less than half of the previous QRS-slope, the peak is classified as a T-wave and discarded, and the noise level estimate is updated. [22]

2.4 Deep Learning

Deep learning is a subfield of machine learning in which models composed of multiple layers of interconnected nodes learn to extract patterns directly from data. Unlike traditional machine learning, deep learning does not require manual feature selection, instead representations are learned automatically during training. [23]

2.4.1 Neural Network Structure

The nodes in a neural network are loosely inspired by the human brain. Each node receives information via connections. It then computes a weighted sum of the inputs and passes the result through a transfer function to the next node. The different layers in a neural network are how the nodes are organized in linear arrays. A neural network usually consists of an input layer, zero or several hidden layers, and an output layer. Designing the network structure refers to determining the number of nodes in each layer, the number of layers and how the nodes are connected within each layer, for an example see Figure 7.

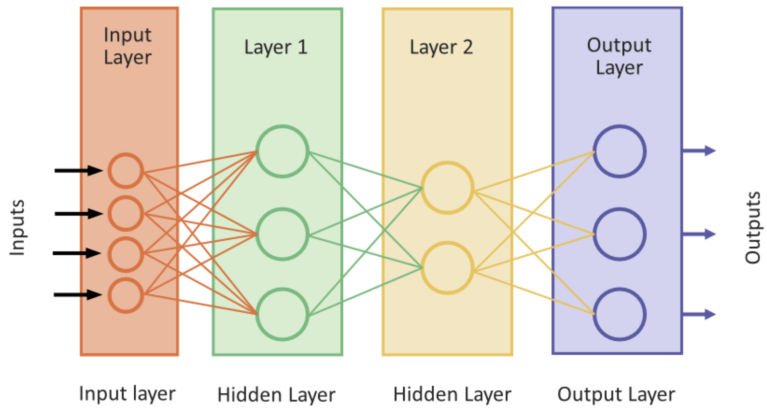

Figure 7: An illustrative example of a feed forward neural network with an input layer (red), two hidden layers (green and yellow) and an output layer (blue). [24]

There are typically two types of neural networks, feed forward networks and feed backward networks. In a feed forward network, the signal only travels in one direction (forward), while in a feed backward network the nodes can send signals back in cycles.

The learning process for a neural network involves adjusting weights to desired values. In supervised learning, the training data needs to be labeled, that is examples of input with the target output provided for the network to learn. The network learns by trying to minimize the error between the network output and the correct output by adjusting the weights. [25]

2.4.2 Recurrent Networks and LSTM

Standard neural networks cannot handle sequences of data, as they process each input independently. Recurrent neural networks (RNN) is a dynamic system which involves recurrence. With recurrence, the output data from a model is fed through again, but this time as input data in the next time-step. RNNs also contain a vector in their hidden units where they store information from past elements of the sequence. However, evidence has been shown that it is difficulty to store information for long. [23]

The Long Short-Term Memory (LSTM) network was developed to address this issue, specifically enabling the retention of information over longer periods of time. It uses a special unit called a memory cell, which has connections to itself in the next time-step with a weight of one. The LSTM network also contains another hidden unit which learns to decide when to clear the memory. [23]

A BiLSTM network is a neural network with two LSTM layers, one forward and one backward.

The forward layer propagates forward in the sequence, the input sequence is then reversed and put through a second LSTM layer. [26]

2.5 Heart Rate Variability (HRV) Analysis

The network of nerves that control unconscious processes throughout your body is called the autonomic nervous system. The autonomic nervous system is always active, and does things without you even thinking about them, like breathing and your heart beating. A well functioning autonomic nervous system is key for survival. [27] A significant relationship between the autonomic nervous system and cardiovascular mortality, including cardiac sudden death, has been observed and has spurred efforts for developing markers for autonomic activity. HRV is one of those promising markers and it has been popularized by its easy derivation and use. [5]

In normal sinus rhythm, HRV reflects the regulation of the sinus node by the autonomic nervous system [5]. However, in atrial fibrillation the interpretation of HRV is considerably more complex. AF rhythm is significantly more irregular than sinus rhythm, with patterns that are clearly visible on an ECG-recording. Despite this complexity, HRV analysis in AF has been shown to carry meaningful clinical information. HRV parameters have been associated with disease progression, treatment response, and long-term prognosis in AF patients. [28]

There are a number of different parameters and methods that can be used for HRV. Below is a brief description of the ones used in this thesis.

2.5.1 RR Variability

There are a number of methods that describe variations in heart rate. The simplest methods are in the time domain, either measuring heart rate or based on RR intervals. QRS-complexes are detected in a continuous ECG-signal and the RR intervals or the so-called normal-to-normal (NN) intervals are determined. NN intervals are all intervals between adjacent QRS-complexes which are deemed to be normal. Simple time domain variables can then be determined. [5]

Heart rate (HR) and the standard deviation of RR or NN intervals (SD) are perhaps the simplest variables to calculate. Standard deviation is a statistical measure of how dispersed or spread out numbers are in a data set relative to their mean, and are referred to as:

$$SD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(RR_i - \overline{RR})^2} \tag{1}$$

where, N denotes the total number of beats and $\overline{RR}$ refers to the mean RR interval. Another commonly used measure is rMSSD, which is derived from interval differences. rMSSD

denotes the square root of the mean squared differences of successive RR or NN intervals. rMSSD is given by:

$$rMSSD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N-1}(RR_{i+1} - RR_i)^2} \tag{2}$$

where N denotes the total number of beats and $(RR_{i+1} - RR_i)$ refers to successive RR interval differences. There is also pNN20, which is the number of interval differences of successive RR or NN intervals greater than 20 milliseconds ($NN20$) divided by the total number of intervals.

$$pNN20 = \frac{NN20}{N-1} \times 100\% \tag{3}$$

Similarly, there is pNN50 and pNN80 which are the same but greater than 50 and 80 milliseconds respectively. [5]

2.5.2 RR Irregularity

Approximate entropy (ApEn) is a statistical measure of complexity and regularity of time series data. It quantifies how repetitive or predictable a signal is. A signal that repeats patterns consecutively will generate a low ApEn value, while a more irregular and unpredictable signal will yield a high ApEn value. A key motivation for the development of ApEn was its applicability to the relatively short and noisy data sets commonly encountered in biomedical settings, such as heart rate, EEG recordings, and endocrine hormone recordings. [29]

The computation of ApEn requires two parameters to be specified in advance: m, the length of the pattern sequences being compared, and r, a tolerance threshold defining the maximum allowable difference between two data points for them to be considered similar. Given a time series of N data points, the algorithm then forms overlapping template vectors of length m, each representing m consecutive observations. For each template vector, the proportion of other template vectors in the series that fall within the tolerance r is computed, yielding a set of similarity counts $C_i^m(r)$. The natural logarithm of these counts is then averaged across all template vectors to produce the quantity:

$$\Phi^m(r) = (N-m+1)^{-1} \sum_{i=1}^{N-m+1} ln(C_i^m(r)). \tag{4}$$

ApEn is then defined as:

$$ApEn(m, r, N) = \Phi^m(r) - \Phi^{m+1}(r). \tag{5}$$

This difference reflects how frequently patterns that are similar over m consecutive observations remain similar when extended by one additional point. [29]

Sample entropy (SampEn) is another statistical measure introduced by Richman and Moorman in 2000, as an improved ApEn. ApEn has two major limitations. The first concern is that ApEn has a self-counting bias. When counting how many template vectors match a given template, ApEn includes the template vector itself in the count. This will inflate the similarity count and introduce a systematic bias that undermines the consistency of ApEn as a complexity measure. The second limitation is dependence on data length, which makes comparison of ApEn values derived from recordings of different lengths unreliable. SampEn was developed to correct both of these shortcomings, yielding a measure that is more consistent, less biased, and more robust across varying data lengths. [30]

The computation of SampEn follows a similar structure to that of ApEn and requires the same two input parameters: m, the template length, and r, the tolerance threshold. Given a time series with N points, $u(1), u(2), ..., u(N)$, a window of length m slides over the signal creating overlapping template vectors. For each template vector, you compare it to every other template vector (not itself, this is the key difference from ApEn). Two vectors are considered a "match" if the maximum absolute difference between their corresponding elements is within the tolerance r:

$$d[x(i), x(j)] = max|u(i+k) - u(j+k)| \leq r \tag{6}$$

All such matching pairs across the entire time series are counted and the total count $B(m, r)$. Intuitively, $B(m, r)$ counts how many times a pattern of length m repeats somewhere else in the signal. The same thing is done with the template length of $m + 1$, creating $B(m + 1, r)$. The SampEn is then calculated as:

$$SampEn(m, r, N) = -ln(\frac{B(m+1, r)}{B(m, r)}). \tag{7}$$

Like ApEn, a low value of SampEn indicates a regular signal that repeats itself and a high value of SampEn suggests an irregular and complex signal. [30]

Regularity index (R) is yet another statistical measure, this is derived from the quantities called conditional entropy (CE) and corrected conditional entropy (CCE). The regularity index is a measure of the degree of regularity of a time series. Firstly, the signal needs to be normalized to have zero mean and unit variance for the rest of the calculations,

$$x(i) = \frac{X(i) - mean[X]}{std[X]} \tag{8}$$

where $mean[X]$ and $std[X]$ are the mean and standard deviation of the time series. From the normalized series, a L-dimensional phase space is constructed by forming vectors of L

consecutive samples:

$$x_L(i) = (x(i), x(i-1), ..., x(i-L+1)). \tag{9}$$

Each such vector represents a pattern of L consecutive samples. For a small L, the pattern is short and likely to repeat itself. As L increases, patterns become longer and more specific, and the likelihood of finding the same pattern twice decreases. Next, the Shannon entropy (SE) can be calculated of $x_L(i)$, [31]

$$E(L) = -\sum_{L} p_L log p_L \tag{10}$$

where p_L denotes the probability of a given length-L pattern. The Conditional Entropy (CE) is then defined as the difference in Shannon Entropy between successive dimensions:

$$E(L/L-1) = E(L) - E(L-1), \tag{11}$$

with $E(1/0) = E(1)$. The CE is known as how much new information is added by one extra sample, given that you already know the preceding $L-1$ samples. For a completely stochastic process such as white noise, CE remains constant and equal to $E(1)$, meaning each new sample carries the same amount of information regardless of the past. For a strictly periodic process, CE decreases to zero once the pattern length L is sufficient to detect the phase of the period. This reflects the fact that the next sample can be perfectly predicted from the previous ones. [31]

In practice, the true probabilities p_L are unknown and must be estimated from the available data. The estimate of CE can be obtained with:

$$\hat{E}(L/L-1) = \hat{E}(L) - \hat{E}(L-1), \tag{12}$$

where $\hat{E}(L)$ and $\hat{E}(L-1)$ are SE with estimates of p_L. However, the estimations introduces a problem when only short amounts of data are available. As L increases, the number of possible patterns grows, while the number of available samples remains the same. When L increases, the probability that a pattern will appear twice decreases. A pattern that appears only once is referred to as a single point. A single point creates a false certainty that the next value can be predicted, since the pattern only appeared once. In turn, this drives CE towards zero regardless of the regularity of the series. [31]

To overcome this limitation, a corrected conditional entropy (CCE) is proposed, see the following equation.

$$CCE(L) = \hat{E}(L/L-1) + E_c(L) \tag{13}$$

The corrective term $E_c(L)$ is defined as,

$$E_c(L) = perc(L) \times \hat{E}(1). \tag{14}$$

Here, $perc(L)$ represents the percentage of single points in the L-dimensional phase space and $\hat{E}(1)$ represents the estimated value of SE for $L = 1$. The corrective term thereby increases with L as single points increase. Since CE decreases and the corrective term increases with L, CCE exhibits a minimum. This minimum is taken as the regularity index, representing the best estimate of CE given the limited data available. A lower regularity index indicates a more regular process, while a higher value indicates more irregularity and randomness. [31]

3 Method

The overall workflow consists of four stages. First, 24-hour Holter ECG recordings from the RATAF II study were loaded and pre-processed. Second, QRS-detection was performed using the Pan Tompkins algorithm, and its output was compared to the RR intervals extracted by clinical hospital software. Third, two quality assurance approaches were developed: an intersection method combining both detectors, and a BiLSTM deep learning classifier trained to distinguish clean beats from noise. Together, these steps result in five different RR interval pipelines: the RR intervals from the clinical hospital software (RR), the normal-to-normal intervals from the clinical hospital software (NN), RR intervals extracted from the Pan Tompkins detections (PT), the intersection of the clinical hospital software and Pan Tompkins detections (RR $\cap$ PT), and the Pan Tompkins detections filtered by the BiLSTM classifier (PT + Classifier). Finally, a set of HRV parameters is computed from each pipeline and compared systematically. An overview of the different pipelines is illustrated in Figure 8.

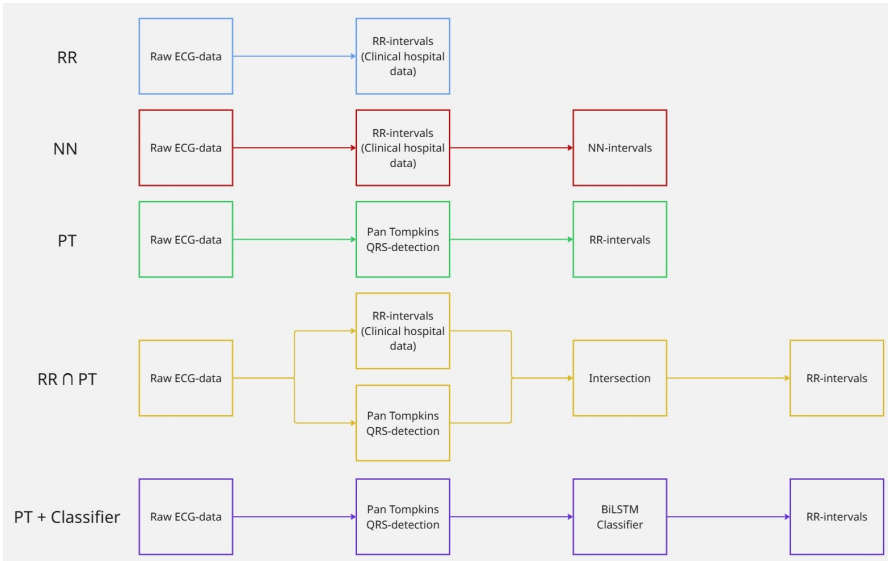

Figure 8: An simple overview of the five different pipelines.

3.1 RATAF II Data

The data used in this study were obtained from the RATAF II study. RATAF II was a randomized, parallel-group study designed to investigate the effects of two different drug regimens for rate control in permanent AF [8]. The study consisted of 93 patients, of which 28 were women, and the average age was 71 ± 7 years. To qualify for the study, patients needed to be diagnosed with symptomatic, permanent AF, have preserved left ventricular systolic function, have no coronary heart disease, and have a resting heart rate above 80 bpm. [32]

Following a two week long period free from drugs affecting heart rate, participants were randomized in a 1:1 ratio to receive either calcium channel blocker diltiazem (n = 49) or beta-blocker metoprolol (n = 44) once daily for six months. The medications are from now on referred to as D for diltiazem and M for metoprolol. The trial was approved by the Norwegian Medicines Agency and the Regional Ethics Committee, and all patients provided written informed consent in accordance with the Helsinki Declaration. [32]

As part of the study protocol, 24-hour Holter monitoring was performed on all the participants at three time points, at baseline, after one month of treatment, and after six months of treatment. The Holter recordings were acquired using three ECG leads with a sampling frequency of 125 Hz. In addition to the raw ECG data, RR interval series extracted from the same recordings by the hospital's clinical software were provided, along with annotations.

3.2 ECG Data Loading and Pre-processing

The raw ECG recordings from the RATAF II study were stored in ISHNE format, a standard format for long-term Holter recordings. Each recording was read into MATLAB using a custom ISHNE reader (read_ishne_RATAF.m, based on the technical specifications provided by the Telemetric and Holter ECG Warehouse [33]), extracting a fixed-length segment of 10 800 000 samples corresponding to a full 24-hour recording at 125 Hz sampling frequency. Each recording contained three leads, stored as a matrix with one column per lead.

Prior to QRS-detection, each lead was independently normalized to a zero mean and unit variance (z-score normalization) according to the formula:

$$x = \frac{X - mean(X)}{std(X)}.$$

where $mean(X)$ and $std(X)$ denote the mean and standard deviation of the respective lead computed over the full 24-hour recording. This normalization was applied consistently across all patients and time points, see Figure 9 for an example.

In parallel, the clinically extracted RR intervals provided by the hospital were loaded from text files. Each file contained beat timestamps, RR interval durations, and beat annotations, where beats labeled 'N' were identified as normal, 'S' means a supraventricular beat and 'A' indicates an abnormal beat. Start times were extracted from each file and used to align the RR interval series to the same timescale, expressed in seconds from midnight.

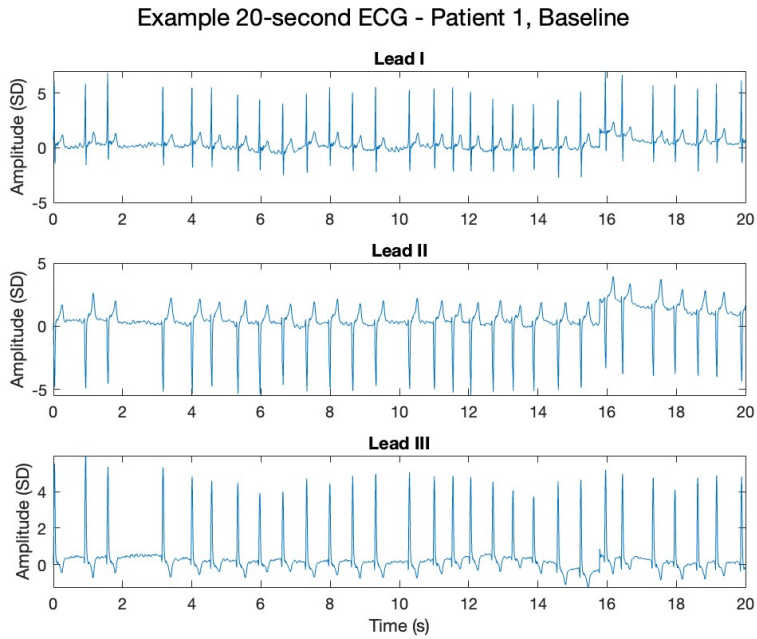

Figure 9: Representative 20-second example window of a normalized 24-hour ECG recording from one RATAF II patient at baseline, showing all three leads.

3.3 QRS-Detection

The first lead of the normalized raw ECG data was then fed through a Pan Tompkins QRS-detector (see Section 2.3.1). The implementation in MATLAB used was developed by Sedghamiz (2014) [34]. The output from this detector was the amplitude of R waves, the sample index of the detected R-peaks, and a sample delay from filtering [34].

3.4 RR Series Comparison

To assess the quality of the clinically extracted RR intervals, the R-peaks detected by the Pan Tompkins algorithm were compared against the RR intervals provided by the hospital software. The comparison was performed on every patient using a time-matching approach.

For each patient, R-peak times from the Pan Tompkins detector were calculated in seconds by dividing the sample indices by the sampling frequency, 125 Hz. The clinically detected R-peak timestamps were similarly expressed in seconds relative to the start of the recording. A clinically detected R-peak and a Pan Tompkins detected R-peak were considered a matched pair if their temporal difference was within a threshold of 50 ms from each other. The matching was done by iterating through each of the clinically detected R-peaks. The closest unmatched Pan Tompkins detected R-peak within the threshold was selected. Each detection could only

be matched once.

From this matching procedure, three quantities were computed for each patient: the number of matched beats, the number of missed beats (clinically detected beats with no matching Pan Tompkins detection within the threshold), and the number of extra beats (Pan Tompkins detections with no matching clinically detected beat). For all matched pairs, the timing difference (Pan Tompkins minus clinical software) was recorded.

3.5 Beat Classification Network

To improve the quality of the RR-interval series derived from the Pan Tompkins detector, a beat classification network was developed to distinguish clean QRS-complexes from noise or artifacts. The classifier was implemented as a Bidirectional Long Short-Term Memory (BiLSTM) network. The network was trained on labeled data from two publicly available databases: AF beats from the MIT-BIH Atrial Fibrillation Database and noise from the MIT-BIH Noise Stress Test Database. The following subsections describe the training data, network architecture, training procedure, and classifier evaluation.

3.5.1 MIT-BIH Atrial Fibrillation Database

The training data for the beat classifier was taken from the MIT-BIH Atrial Fibrillation Database, which is a publicly available annotated ECG database provided by PhysioNet. The database contains 23 10-hour ECG recordings from patients with atrial fibrillation (mostly paroxysmal), sampled at 250 Hz, with beat-by-beat annotations marking the location and type of each heart beat. [35]

3.5.2 MIT-BIH Noise Stress Test Database

Noisy training data was downloaded from the MIT-BIH Noise Stress Test Database on PhysioNet. Three half-hour recordings of realistic clinical noise in ECG recordings were used, electrode motion artifacts (EM), muscle artifacts (MA) and baseline wander (BW). The noise recordings were created by selecting intervals in recordings that contain predominantly baseline wander, muscle artifacts, and electrode motion, and assembling them. The sampling frequency for these recordings was 360 Hz. [36]

3.5.3 Training Data Preparation

For each of the 23 MIT-BIH Atrial Fibrillation recordings, the ECG signal was first normalized to zero mean and unit variance to ensure consistent scaling across recordings from different patients and equipment. A fixed-length window of 150 samples was then extracted around each annotated R-peak, consisting of 50 samples (200 ms) before the R-peak and 100 samples (400 ms) after the R-peak. This window length was chosen to capture the full QRS-complex

along with some context from the preceding and following signal morphology. Beats whose windows extended beyond the start and end of the recording were excluded.

To ensure that only beats which had a clearly visible QRS-complex was included in the training data, a peak-to-peak amplitude threshold was applied. Beats with a peak-to-peak amplitude lower than 1.0 normalized, were excluded. This thresholding retained 98.51 % of the annotated beats.

The three half-hour MIT-BIH Noise Stress Test Database were first resampled from 360 Hz to the sample frequency of the MIT-BIH Atrial Fibrillation recordings, 250 Hz, as well as normalized to zero mean and unit variance. To prevent data leakage between training and validation, the noise recordings were split into training and validation data. The first 80% of each recording was meant for training and the remaining 20% was assigned as validation data. No noise from the validation portion was used for training. Fixed-length noise windows were extracted from each of the noisy recordings, again with a length of 150 samples. Here, a sliding window with overlap approach were used, to be able to produce more noise windows while still avoiding identical windows. A step-size of 50 samples were used. To ensure that the noisy windows captured enough noise, windows with a standard deviation below 0.3 normalized units were discarded. This was done to avoid silent or near-silent periods that could be confused as low amplitude beats. This thresholding approach were applied to the EM and AM windows only, since BW by nature has such a low variance over short windows.

The noise training set was then composed of four equal quarters to expose the classifier to varied noise: (1) EM only, (2) MA only, (3) EM + BW and (4) MA + BW. This composition was designed to prevent the network from learning that noise always contains baseline wander, to avoid incorrectly rejecting clear QRS-complexes affected by mild baseline drift.

After filtering with thresholds, a number of 3437 noise windows were obtained for training and 970 for validation, see Figure 10 for some training data windows. An equal number of clean beat windows were randomly selected from the MIT-BIH Atrial Fibrillation Database to balance the two classes. The final training set consisted of 6874 samples and the final validation set became 1940 samples. Both sets were shuffled prior to training and a fixed random seed (42) was used throughout all random operations to ensure reproducibility of the results.

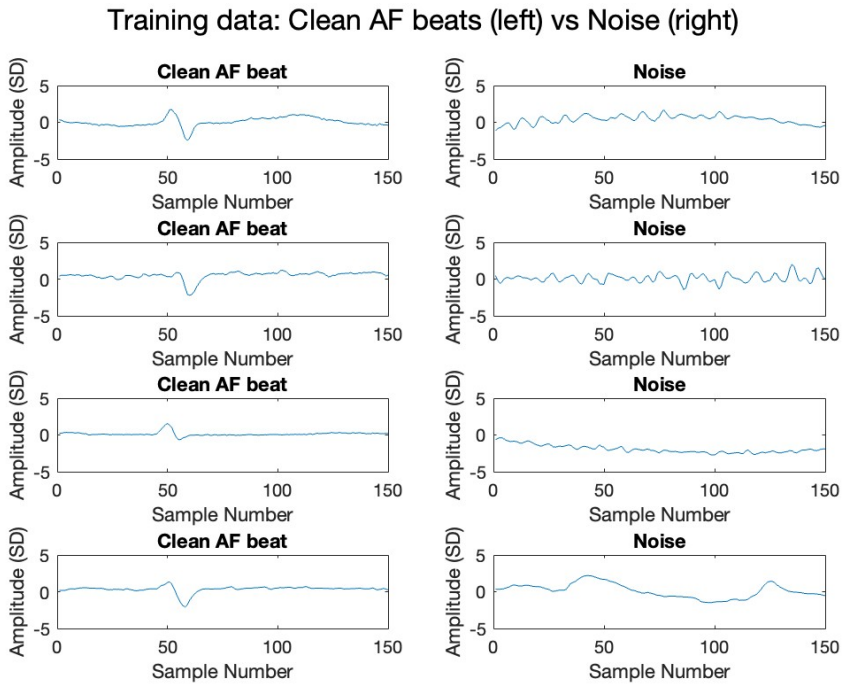

Figure 10: Eight examples of training data, training data labeled 'Beat' to the left and training data labeled 'Noise' to the right.

3.5.4 Network Architecture and Training

The classifier was implemented as a Bidirectional Long Short-Term Memory (BiLSTM) network, following the architecture described in the MathWorks documentation for ECG signal classification. [37] Each beat window of 150 samples was fed to the network input layer, with one value per time step. The network consisted of four layers. First a sequence input layer accepting one feature per time step, then a BiLSTM layer with 50 hidden units and outputting only the final time step, a fully connected layer with two output nodes corresponding to the two classes (Noise and Beat), and a softmax layer converting the outputs to class probabilities. [37]

The network was trained using the Adam optimizer, a mini-batch size of 200, a gradient threshold of 1 to stabilize training, and a maximum of 50 epochs. The data was reshuffled at every epoch. [37] A fixed random seed (42) was set prior to training, again to ensure reproducibility. The model was immediately saved after training.

Training was performed in MATLAB using the trainnet function (Deep Learning Toolbox) with cross-entropy loss. [37] Training converged within 50 epochs, reaching a training accuracy of approximately 95 %, with no notable gap between the two indicating absence of

overfitting.

3.5.5 Classifier Evaluation

The trained BiLSTM network were than evaluated on the validation set. The validation set comprised of 1940 samples in total, 970 clean AF beats from the MIT-BIH Atrial Fibrillation Database and 970 noise windows from the MIT-BIH Noise Stress Test Database. Performance was assessed using three metrics: accuracy, sensitivity and specificity. Accuracy, sensitivity and specificity were calculated with the following equations.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{15}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{16}$$

$$Specificity = \frac{TN}{TN + FP} \tag{17}$$

A confusion matrix was then produced to visualize the distribution of correct and incorrect classifications. Lastly, the classifier was run on the patients from the RATAF II study, with the QRS-complexes detected by the Pan Tompkins algorithm. The raw ECG signals from the RATAF II study were first resampled to match the frequency of the training data and windows of 150 samples were extracted around each detected R-peak, mirroring the exact pre-processing used for the training data.

3.6 HRV Analysis

HRV analysis was performed on five different RR interval series to allow comparison across pipeline configurations: the RR intervals from the clinical hospital software (RR), the NN intervals from the clinical hospital software (NN), the RR intervals from the Pan Tompkins detector (PT), the intersection of beats detected by both the clinical software and Pan Tompkins (RR $\cap$ PT), and the Pan Tompkins beats filtered by the BiLSTM classifier (PT + Classifier).

For each pipeline, RR intervals were computed as the difference between successive beat timestamps. A physiological plausibility filter was applied, designed to retain only RR intervals in the range of 0.3-2.0 seconds (corresponding to a ranging heart rate between 200 and 30 bpm). This step is important as chunks of data could be lost for different pipelines, for example, the BiLSTM classifier may classify chunks of several seconds as noise. Those chunks are then removed, and the RR interval between the last and the first detected beats would be several seconds. That RR interval is not only physiologically impossible, it is not a true RR interval since data in between has been removed.

Before any HRV parameters were computed, the total beat count across all recordings was

calculated for each pipeline and compared pairwise. Two beats detected by different methods were considered a match if the absolute time difference between them was less than 50 ms, and each beat could only be matched once. This comparison allowed for a direct assessment of how many beats each pipeline retained and how much overlap existed between methods. However, a small number of recordings were of such poor signal quality that they were considered unusable for analysis, see Figure 11. A minimum threshold of 50 000 detected beats was therefore applied, corresponding to an average heart rate of approximately 35 BPM over 24 hours. Any recording falling below this threshold was excluded from any further analysis.

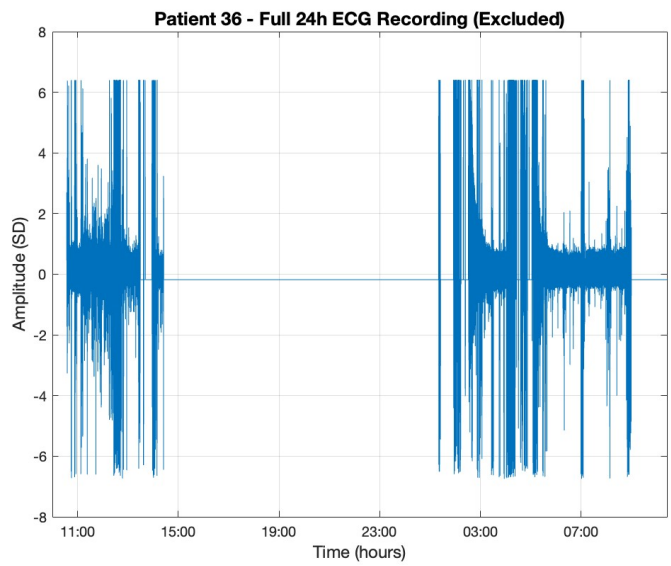

Figure 11: An example of a recording excluded from further analysis.

To replicate the analysis protocol for the RATAF I study [38], HRV parameters were computed over a 20-minute segment from 14.00 to 14.20 for each recording. Since recordings did not all begin at the same time of the day, the window selection was handled with some flexibility. If the recording started before 14.00, the segment 14.00-14.20 on the first day was used. If the recording started between 14.00 and 14.20, the window was shifted to 14.20-14.40 to ensure that a full 20 minute segment was available. If the recording started after 14.20, the 14.00-14.20 window on the second day was used instead.

All HRV parameters were then computed using MATLAB. Each 20 minute RR interval segment was expressed in milliseconds. The following parameters were computed:

- **HR:** mean heart rate in beats per minute
- **SD:** standard deviation of the RR intervals in milliseconds

- **pNN20, pNN50, pNN80:** the percentages of successive RR interval differences exceeding 20, 50, and 80 milliseconds, respectively
- **rMSSD:** the root mean square of successive RR interval differences in milliseconds
- **ApEn:** Approximate Entropy, computed with embedding dimension $m = 2$ and tolerance $r = 0.2$ (fraction of standard deviation)
- **SampEn:** Sample Entropy, computed with the same parameters ($m = 2$, $r = 0.2$)
- **R**: the Regularity Index, computed with embedding dimension $L = 3$ and 6 quantifications.

The RR irregularity measures ApEn, SampEn and R were calculated using a toolbox in MATLAB: A set of Entropy measures for temporal series (1D signals) by Jesús Monge-Álvarez (2026) [39].

To evaluate whether the choice of pipeline had a significant effect on the computed HRV parameters, pairwise differences between all five pipelines were computed for each metric (HR, SD, pNN20, pNN50, pNN80, rMSSD, ApEn, SampEn, and R). For each pairwise comparison, the difference was computed on a recording-by-recording basis, i.e. each recording was processed by all five pipelines, and the difference in a given HRV metric between two pipelines was taken for that same recording. Data from all time points (Baseline, 1 month, and 6 months) and both drug conditions were pooled, yielding a single difference vector per pipeline pair per metric.

Prior to significance testing, the normality of each pairwise difference distribution was tested using the Lilliefors test. None of the difference vectors passed the normality criteria, parametric testing was then ruled out for all comparisons. Pairwise statistical significance was therefore assessed using a two-sided Wilcoxon signed-rank test, which tests whether the median of the differences deviates significantly from zero without assuming normality. Results were reported as medians with the 25th and 75th percentiles (interquartile range, IQR) and marked as significant or not significant based on the signed-rank test. In total, 10 pairwise comparisons were performed across all 9 HRV metrics, yielding 90 hypothesis tests in total. However, the more tests carried out, the higher the odds are of getting a false positive result just by chance. This is called the multiple comparisons problem. Therefore, a Bonferroni correction was applied, adjusting the significance level from $\alpha = 0.05$ to $\alpha = \frac{0.05}{90} \approx 0.00056$. A pipeline comparison was considered statistically significant only if the corrected threshold was met.

4 Results

4.1 RR Series Comparison

The Pan Tompkins QRS-detection algorithm was applied to all 24-hour recordings. Figure 12 shows a typical 20-second ECG window from one patient, with R-peak detections from both the Pan Tompkins algorithm and the clinical software overlaid. In this example, the two detectors show close agreement, with all detections placed at the same heart beats. However, minor timing offsets between the two methods are visible in some beats.

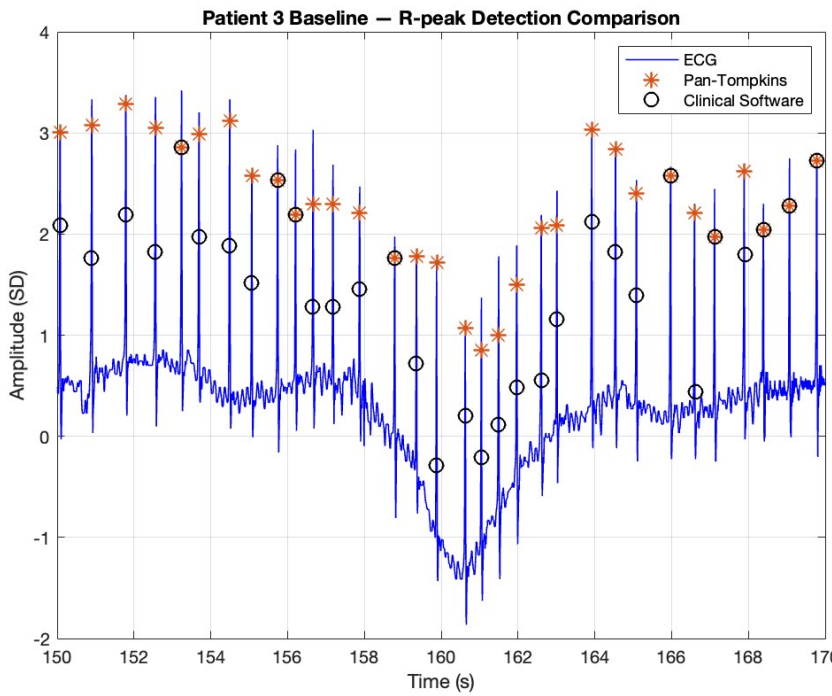

Figure 12: Representative 20-second ECG example window from a RATAF II patient with R-peak detections overlaid. Stars denote Pan Tompkins detections, and circles denote beats detected by the clinical software.

There are instances where the two QRS-detecting algorithms does not agree. Figure 13, shows one example where the clinical hospital software failed and detected two QRS-complexes where there is only one, see between timestamps 146 s and 147 s. Figure 14 shows one example where the Pan Tompkins algorithm mistakes fibrillation for a QRS-complex. Both algorithms struggles when they encounter noisy parts, see example in Figure 15.

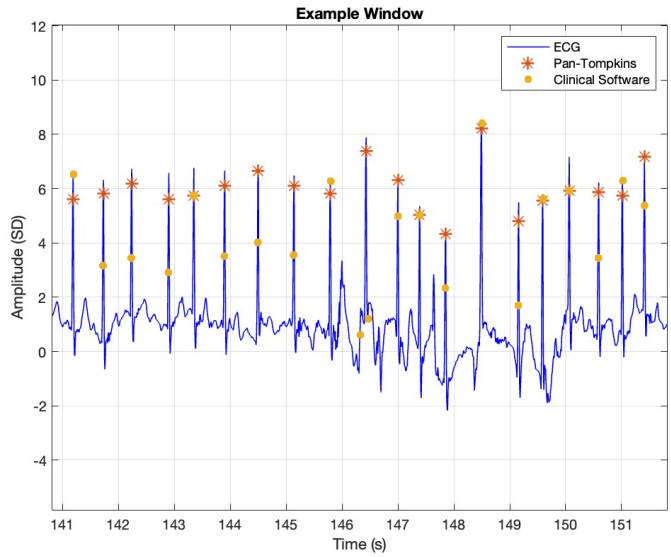

Figure 13: An example window where the clinical hospital software failed.

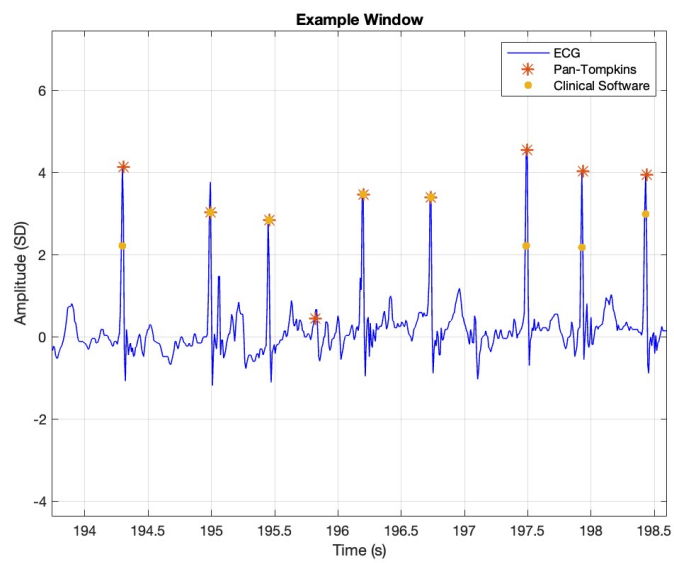

Figure 14: An example window that shows one instance where the Pan Tompkins algorithm failed.

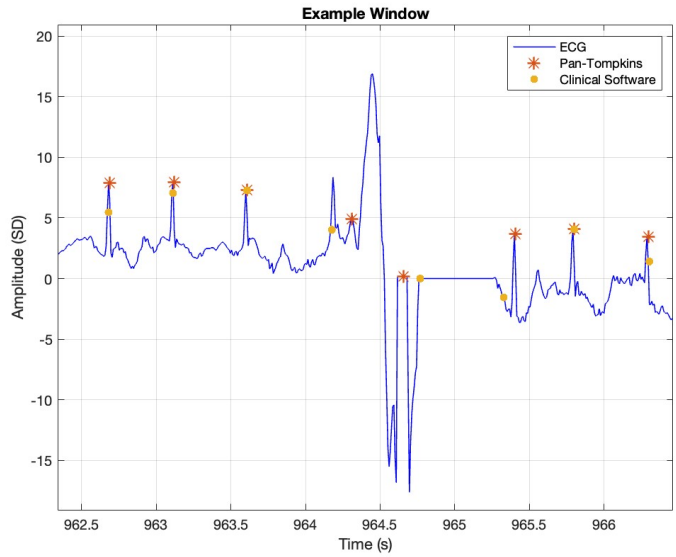

Figure 15: An example of where both of the QRS-detection algorithms failed.

Figure 16 shows the distribution of R-peak timing differences between the Pan Tompkins algorithm and the clinical software, combining all matched beat pairs across all recordings. The distribution is centered near zero and shows a discrete structure, with bars spaced approximately 8 ms apart. This spacing corresponds exactly to one sample at the recording sampling frequency of 125 Hz, $\frac{1}{125\,Hz} = 0.008$ s. The Pan Tompkins detector can only place detections at sample indices.

The concentration of beats within ±16 ms of zero indicates on average ±2 samples difference between the detection methods. Larger differences approaching the ±50 ms matching threshold are present but only represent a small minority of beats.

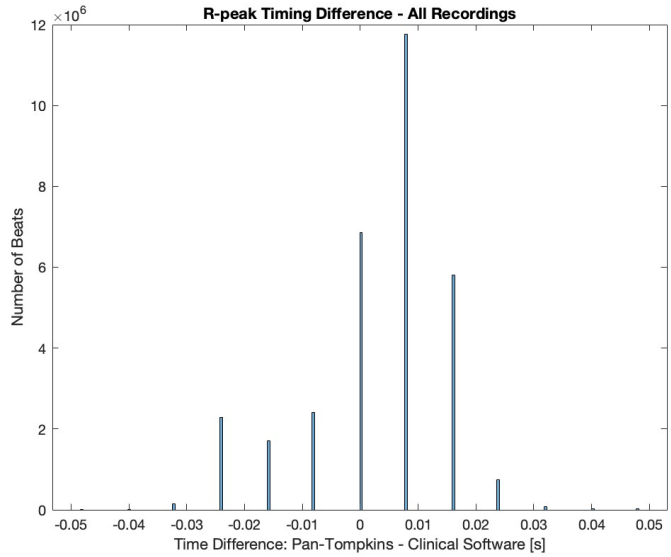

Figure 16: Distribution of R-peak timing differences between the PanTompkins algorithm and the clinical software, across all recordings.

Figures 17 and 18 show the distribution of missed and extra beats by the Pan Tompkins detector across all recordings. The majority of recordings had very few missed beats, with the distribution heavily concentrated near zero. However, the distribution has a long right tail, with some recordings showing up to 18 000 missed beats. This indicating that a small number of recordings had large disagreement between the two detectors. A similar pattern is seen for extra beats (Figure 18), where most recordings are concentrated near zero but a handful of recordings show up to 25 000 extra detections by Pan Tompkins. For the majority of recordings the two detectors are in close agreement.

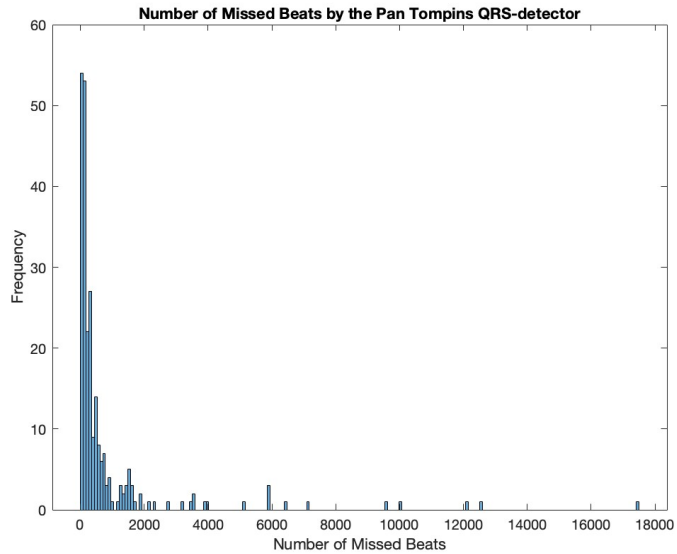

Figure 17: Distribution of the number of missed beats by the Pan Tompkins QRS-detector across all recordings. This shows the number of beats which were detected by the clinical hospital software and not the Pan Tompkins QRS-detection algorithm.

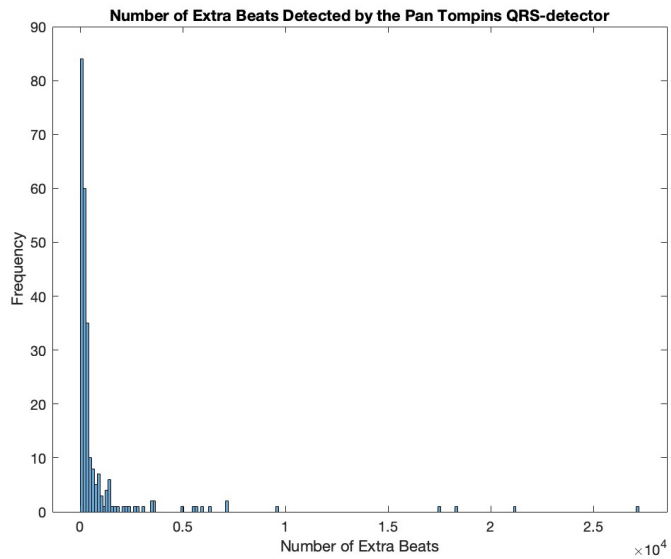

Figure 18: Distribution of the number of extra beats detected by the Pan Tompkins QRS-detector across all recordings. This shows the number of beats which were detected by the QRS-detector and not by the clinical hospital software.

4.2 Beat Classifier Evaluation

The BiLSTM classifier was trained for 50 epochs on a dataset of 6874 samples consisting of equal numbers of clean AF beat windows and real noise windows. Training converged smoothly with no notable overfitting, as seen by the close tracking of training and validation accuracy throughout all epochs. The training progress plot is shown in Figure 19.

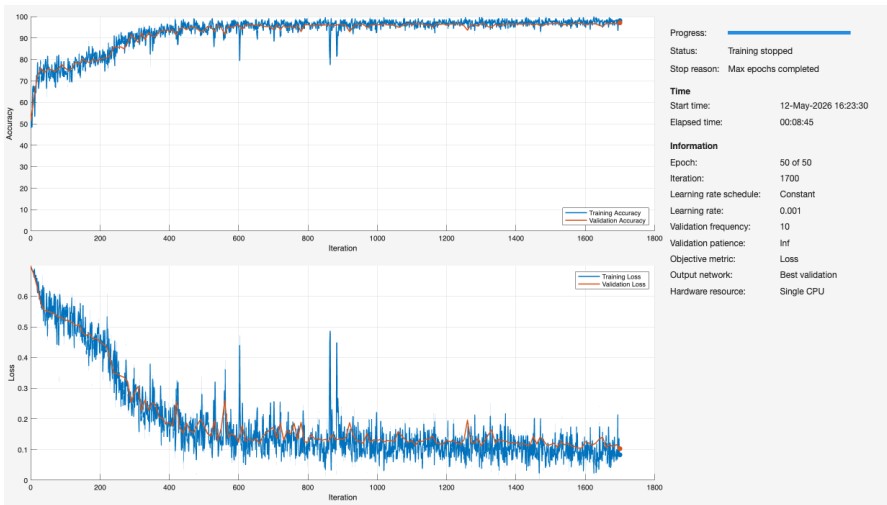

Figure 19: The training progress plot, showing both accuracy and loss over iterations for both training and validation.

Following training, the classifier was evaluated on the held-out validation set of 1940 samples (970 clean beats and 970 noise windows). The classifier achieved an overall accuracy of 97.0%, a sensitivity of 94.2% and a specificity of 99.8%. The confusion matrix is shown in Figure 20. Of the 970 noise windows, 968 were correctly rejected and only 2 were incorrectly passed through as clean beats. Of the 970 clean beat windows, 914 were correctly accepted and 56 were incorrectly rejected as noise. The high specificity indicates that the classifier is highly effective at rejecting genuine noise, while the slightly lower sensitivity reflects a careful classification behavior that preferentially discards borderline beats rather than risking inclusion of noisy data.

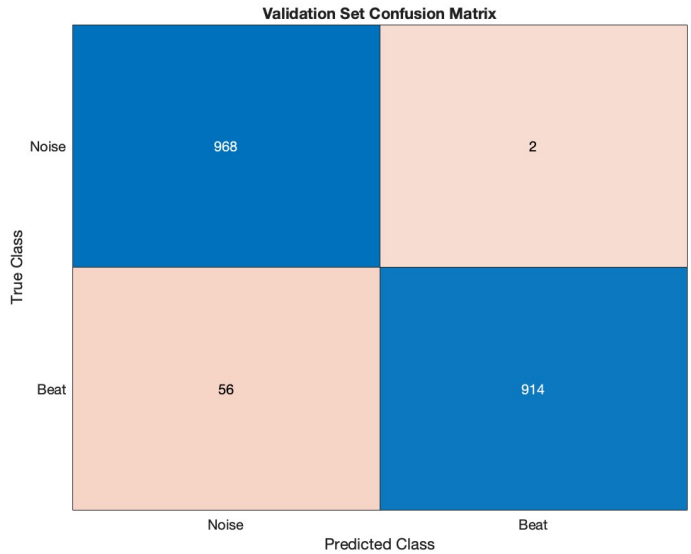

Figure 20: The confusion matrix for the trained network on the validation set.

The trained classifier was applied to the 24-hour ECG recordings from the RATAF II study, in windows and resampled from 125 Hz to 250 Hz to match the training data. Figure 21 illustrates a successful noisy beat classification, while Figure 22 and 23 show less successful examples. When the classifier encountered extremely noisy segments, as shown in Figure 23, it was far from perfect. The majority of beats in these parts were correctly labeled as noise, but the classifier missed quite a few. A significant problem was identified, with how the classifier behaved when encountering baseline wander. Despite clearly visible QRS-complexes, the classifier discarded beats affected by baseline wander, see Figure 22. This is not an ideal outcome since beats with clear QRS-complexes are wanted.

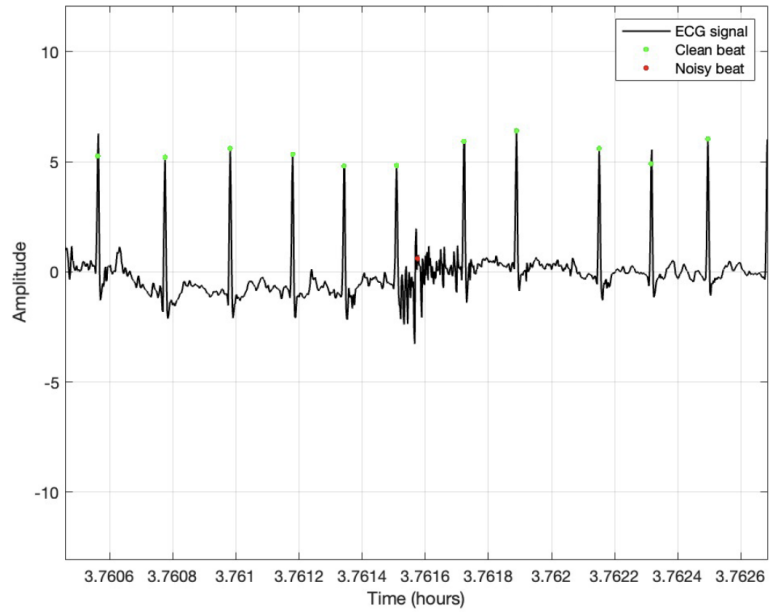

Figure 21: An illustrative example of when the BiLSTM network correctly classified noise. Green dots indicates QRS-complexes which were labeled clean (free from noise), and red dots indicates noisy QRS-complexes.

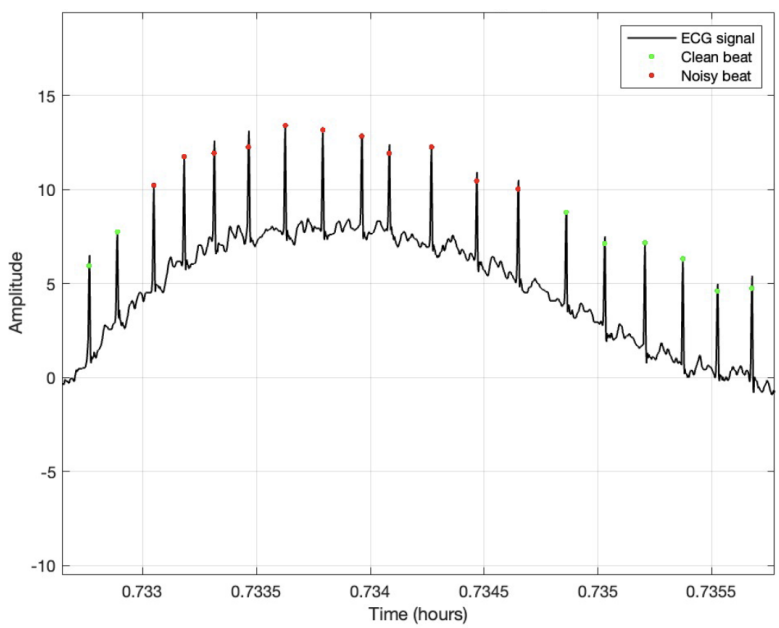

Figure 22: An example of a problem with the BiLSTM network. The network falsely classified some QRS-complexes as noise even though it is a clean QRS-complex.

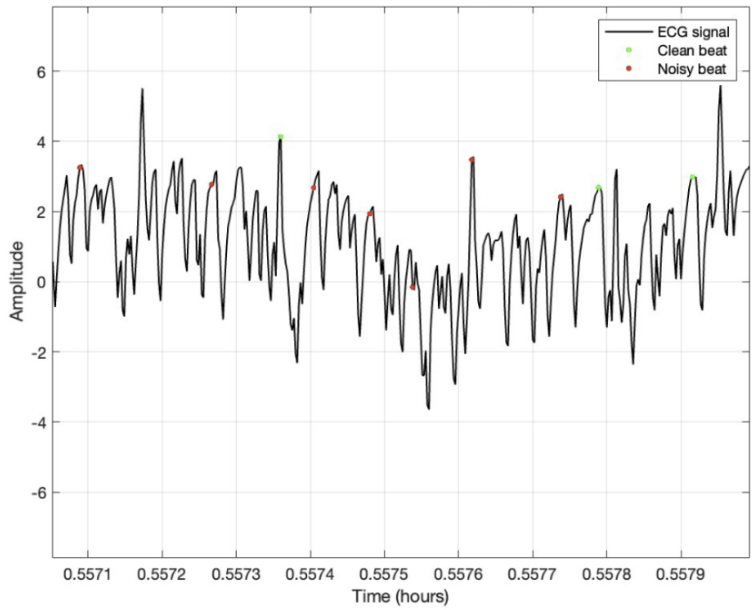

Figure 23: An example of how the BiLSTM network behaves when encountering a very noisy segment.

4.3 Pipeline Comparison

Table 1 summarizes the total number of beats detected by each pipeline across all non-excluded recordings. The hospital software RR series (RR) contained the most beats in total (32 168 700), followed by the Pan Tompkins series (PT, 32 103 789). Their intersection (RR ∩ PT) had 31 650 037 beats, and the hospital NN interval series had the fewest (28 148 088), reflecting the removal of non-normal beats by the clinical software. The PT + Classifier pipeline retained a comparable number of beats to the raw PT series, indicating that the BiLSTM classifier rejected a relatively small proportion overall. The non-diagonal entries show the number of beats shared between pairs of pipelines.

Table 1: Total number of beats across all recordings for the intersection of each pair of RR interval pipelines. Diagonal entries (bold) indicate the total number of beats in that pipeline.

	RR	NN	PT	RR ∩ PT	PT + Classifier
RR	**32168700**	28148088	31650037	31650037	27835781
NN	28148088	**28148088**	27756727	27745727	24410507
PT	31650037	27756727	**32103789**	31650037	28168021
RR ∩ PT	31650037	27756727	31650037	**31650037**	27831661
PT + Classifier	27835781	24410507	28168021	27831661	**31852921**

4.4 HRV Analysis Comparison

Tables 11 to 15, in Appendix A, present HRV parameters, replicated from the RATAF I study, across the five different approaches. Figures 24 to 32 show the same results, plotted for easier comparison. Several patterns can be seen across all pipelines. Heart rate was highest at baseline, around 105 bpm, and decreased substantially under both treatments. The entropy measures (ApEn, SampEn) and the regularity index (R) remained relatively stable across time points and treatments, suggesting that the irregularity of AF was not substantially altered by the rate control medication.

Reassuringly, the HRV values are strikingly consistent across all five pipelines at every time point, as seen in Figures 24 to 32, where the error bars of all pipelines overlap almost entirely in most cases.

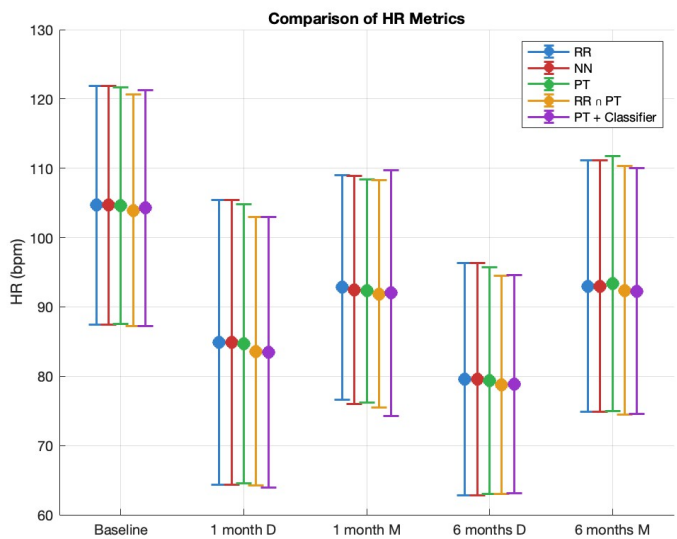

Figure 24: Comparison of heart rate (HR) measurements across all 5 pipelines at all timestamps. Mean heart rate (represented by dots) and standard deviation (error bars) are shown for Baseline, 1-month, and 6 month across all the different pipelines.

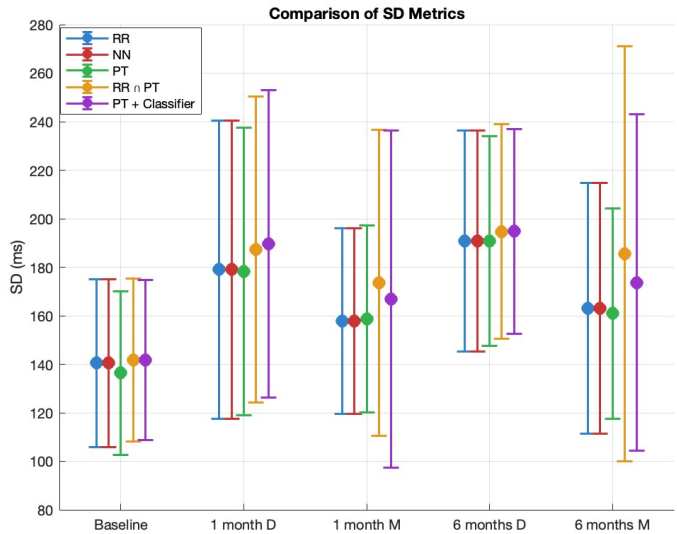

Figure 25: Comparison of Standard Deviation (SD) measurements across all 5 pipelines at all timestamps. Mean SD (represented by dots) and standard deviation (error bars) are shown for Baseline, 1-month, and 6 month across all the different pipelines.

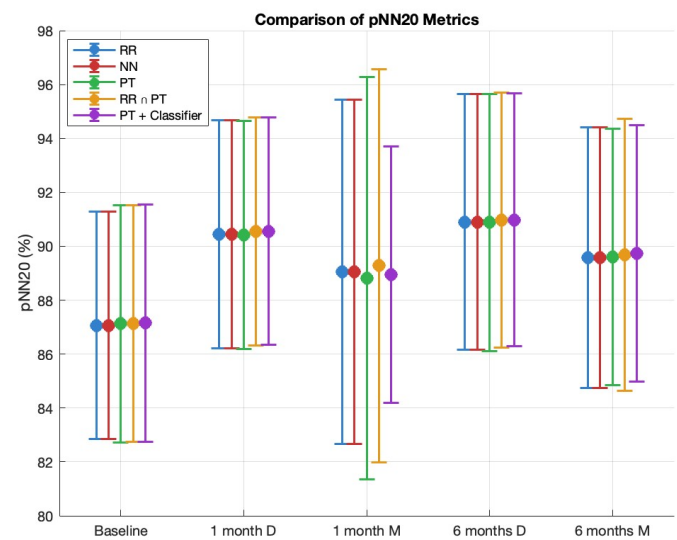

Figure 26: Comparison of pNN20 measurements across all 5 pipelines at all timestamps. Mean pNN20 (represented by dots) and standard deviation (error bars) are shown for Baseline, 1-month, and 6 month across all the different pipelines.

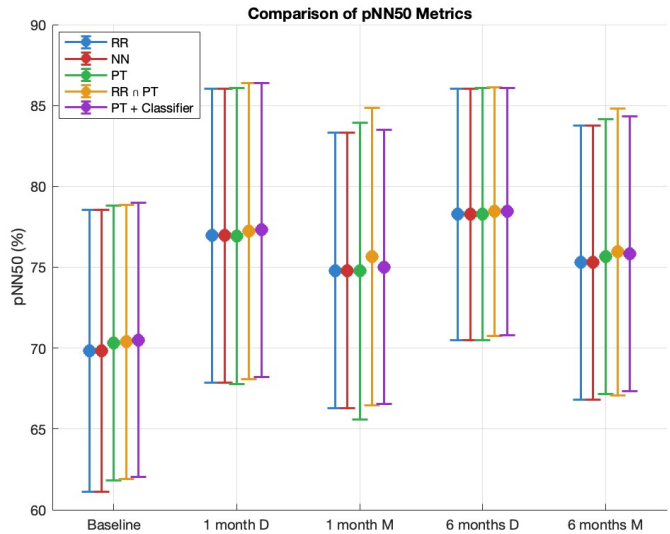

Figure 27: Comparison of pNN50 measurements across all 5 pipelines at all timestamps. Mean pNN50 (represented by dots) and standard deviation (error bars) are shown for Baseline, 1-month, and 6 month across all the different pipelines.

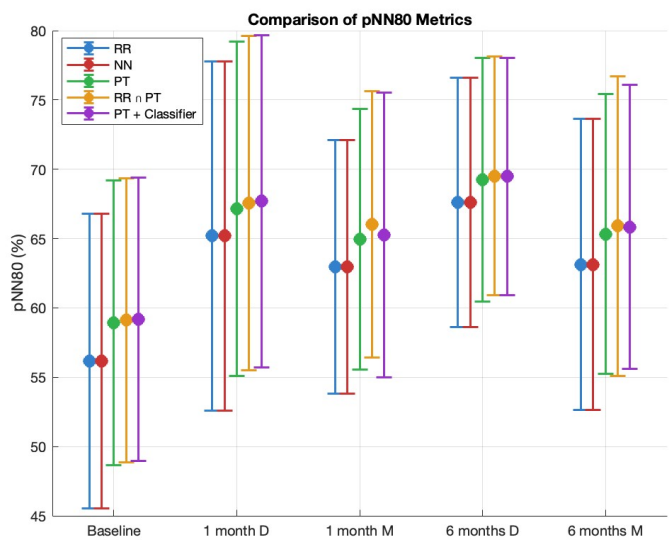

Figure 28: Comparison of pNN80 measurements across all 5 pipelines at all timestamps. Mean pNN80 (represented by dots) and standard deviation (error bars) are shown for Baseline, 1-month, and 6 month across all the different pipelines.

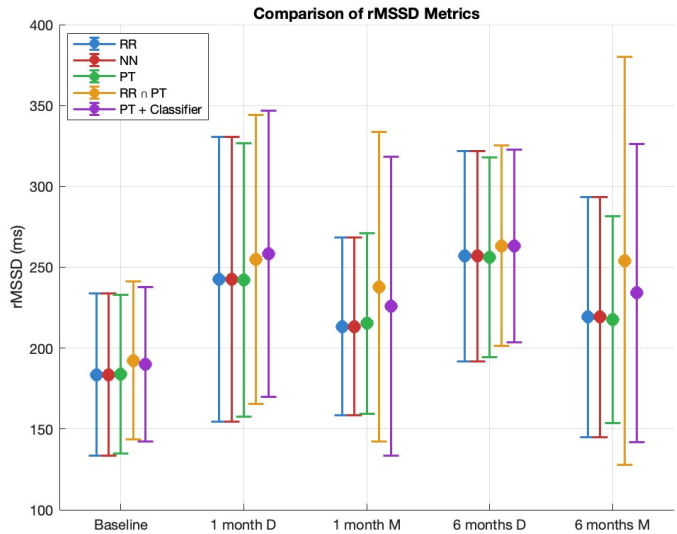

Figure 29: Comparison of rMSSD measurements across all 5 pipelines at all timestamps. Mean rMSSD (represented by dots) and standard deviation (error bars) are shown for Baseline, 1-month, and 6 month across all the different pipelines.

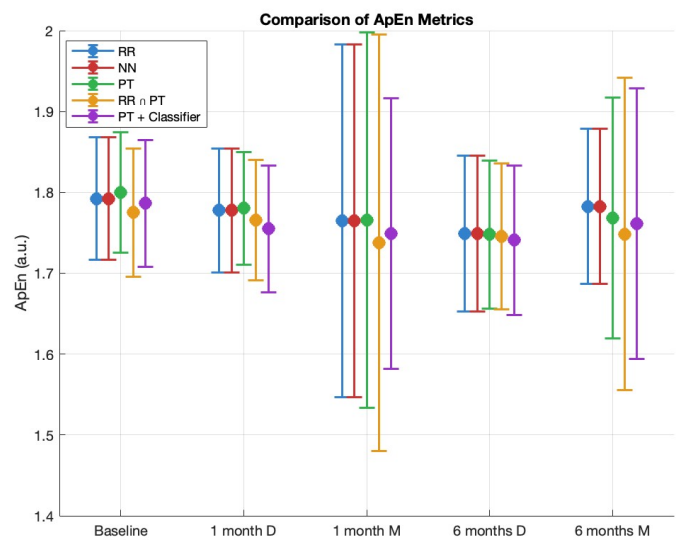

Figure 30: Comparison of ApEn measurements across all 5 pipelines at all timestamps. Mean ApEn (represented by dots) and standard deviation (error bars) are shown for Baseline, 1-month, and 6 month across all the different pipelines.

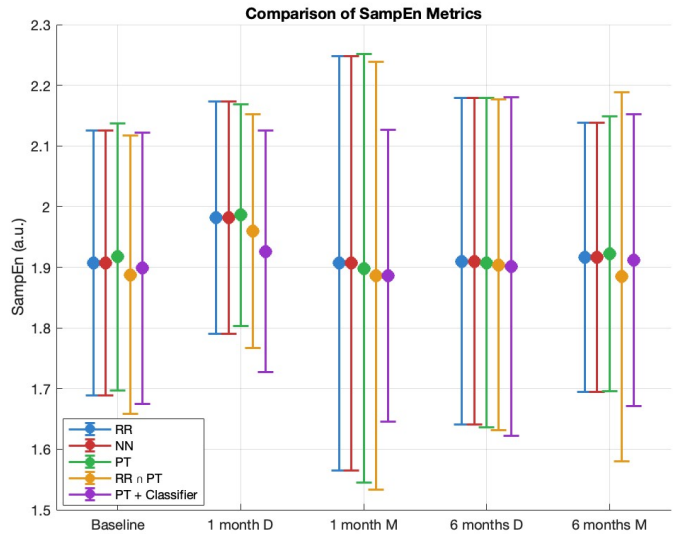

Figure 31: Comparison of SampEn measurements across all 5 pipelines at all timestamps. Mean SampEn (represented by dots) and standard deviation (error bars) are shown for Baseline, 1-month, and 6 month across all the different pipelines.

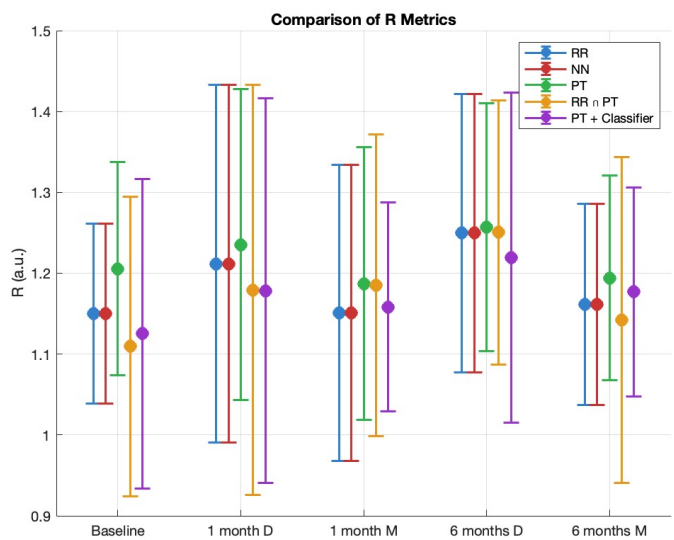

Figure 32: Comparison of R measurements across all 5 pipelines at all timestamps. Mean R (represented by dots) and standard deviation (error bars) are shown for Baseline, 1-month, and 6 month across all the different pipelines.

Tables 2 to 10 present the pairwise statistical comparisons between pipelines. Values marked

with * indicate statistically significant differences after Bonferroni correction $\alpha = \frac{0.05}{90} \approx 0.00056$.

The pairwise statistical comparisons largely confirm the visual impression of Figures 24 to 32. The RR and NN series produced identical results across all metrics at all time points, with a median difference of exactly zero for every parameter, indicating that the removal of non-normal beats by the clinical software had no detectable impact on HRV metrics in this population.

The comparisons between the clinical software pipelines (RR, NN) and the Pan Tompkins based pipelines (PT, RR ∩ PT, PT + Classifier) revealed a small number of statistically significant differences, though the magnitudes were very small.

Table 2: Pairwise differences in Heart Rate (HR) metrics. Values represent the Median [25th, 75th percentiles] for each comparison. Differences are calculated as [Row] − [Column].

	RR	NN	PT	RR ∩ PT	PT + Classifier
RR	–	0.0000 [0.0000, 0.0000]	0.0015 [-0.0693, 0.1067]	0.0804* [-0.0020, 0.4999]	0.0660* [-0.0693, 0.1067]
NN	–	–	0.0008 [-0.0512, 0.0773]	0.0804* [0.0012, 0.4077]	0.0601* [-0.0047, 0.4892]
PT	–	–	–	0.0726* [0.0000, 0.2994]	0.0500* [0.0000, 0.2716]
RR ∩ PT	–	–	–	–	0.0000 [-0.0966, 0.0546]
PT + Classifier	–	–	–	–	–

Table 3: Pairwise differences in Standard Deviation (SD) metrics. Values represent the Median [25th, 75th percentiles] for each comparison. Differences are calculated as [Row] − [Column].

	RR	NN	PT	RR ∩ PT	PT + Classifier
RR	–	0.0000 [0.0000, 0.0000]	-0.1039 [-0.8362, 0.7773]	-0.7655* [-3.2205, 0.03860]	-0.0008* [-4.2135, 0.0929]
NN	–	–	-0.1039 [-0.8362, 0.7773]	-0.7655* [-3.2204, 0.0386]	-0.7882* [-4.2135, 0.0928]
PT	–	–	–	-0.4608* [-3.0380, 0.0000]	-0.4964* [–2.8813, 0.0000]
RR ∩ PT	–	–	–	–	0.0000 [-1.3029, 0.7487]
PT + Classifier	–	–	–	–	–

Table 4: Pairwise differences in pNN20 metrics. Values represent the Median [25th, 75th percentiles] for each comparison. Differences are calculated as [Row] − [Column].

	RR	NN	PT	RR ∩ PT	PT + Classifier
RR	–	0.0000 [0.0000, 0.0000]	-0.0082 [-0.4022, 0.4472]	-0.08548 [-0.5475, 0.3675]	-0.0414 [-0.4455, 0.3669]
NN	–	–	-0.0082 [-0.4022, 0.4472]	-0.08162 [-0.5019, 0.3675]	-0.0414 [-0.4455, 0.3669]
PT	–	–	–	0.0000* [-0.0473, 0.0043]	0.0000* [-0.0425, 0.0046]
RR ∩ PT	–	–	–	–	0.0000 [-0.0173, 0.0214]
PT + Classifier	–	–	–	–	–

Table 5: Pairwise differences in pNN50 metrics. Values represent the Median [25th, 75th percentiles] for each comparison. Differences are calculated as [Row] − [Column].

	RR	NN	PT	RR ∩ PT	PT + Classifier
RR	–	0.0000 [0.0000, 0.0000]	-0.1078 [-0.5568, 0.2833]	-0.1681* [-0.6282, 0.1810]	-0.2058* [-0.6282, 0.1810]
NN	–	–	-0.1078 [-0.5568, 0.2833]	-0.1681* [-0.1282, 0.1810]	-0.2058* [-0.5820, 0.1814]
PT	–	–	–	0.0000* [-0.1107, 0.0000]	0.0000* [-0.0953, 0.0000]
RR ∩ PT	–	–	–	–	0.0000 [-0.0532, 0.0248]
PT + Classifier	–	–	–	–	–

Table 6: Pairwise differences in pNN80 metrics. Values represent the Median [25th, 75th percentiles] for each comparison. Differences are calculated as [Row] − [Column].

	RR	NN	PT	RR ∩ PT	PT + Classifier
RR	–	0.0000 [0.0000, 0.0000]	-1.9036* [-2.5889, -1.3634]	-1.9531* [-2.8306, -1.4654]	-1.9633* [-2.7825, -1.4747]
NN	–	–	-1.9036* [-2.5889, -1.3634]	-1.9531* [-1.8306, -1.4654]	-1.9633* [-2.7825, -1.4747]
PT	–	–	–	0.0000* [-0.1618, 0.0000]	0.0000* [-0.1195, 0.0000]
RR ∩ PT	–	–	–	–	0.0000 [-0.0485, 0.0596]
PT + Classifier	–	–	–	–	–

Table 7: Pairwise differences in rMSSD metrics. Values represent the Median [25th, 75th percentiles] for each comparison. Differences are calculated as [Row] − [Column].

	RR	NN	PT	RR ∩ PT	PT + Classifier
RR	–	0.0000 [0.0000, 0.0000]	-0.3537 [-2.1030, 0.8476]	-1.9804* [-6.0981, -0.0378]	-0.5936* [-4.0516, 0.0000]
NN	–	–	-0.3537 [-2.1030, 0.8476]	-1.9804*[-6.0908, -0.0379]	-1.8760* [-6.8448, 0.0176]
PT	–	–	–	-0.6777* [-4.7814, 0.0000]	-0.5936* [-4.0516, 0.0000]
RR ∩ PT	–	–	–	–	0.0000 [-1.2802, 0.9973]
PT + Classifier	–	–	–	–	–

Table 8: Pairwise differences in ApEn metrics. Values represent the Median [25th, 75th percentiles] for each comparison. Differences are calculated as [Row] − [Column].

	RR	NN	PT	RR ∩ PT	PT + Classifier
RR	–	0.0000 [0.0000, 0.0000]	-0.0004 [-0.0114, 0.0065]	0.0015 [-0.0090, 0.0110]	0.0018 [-0.0087, 0.0121]
NN	–	–	-0.0004 [-0.0114, 0.0065]	0.0015 [-0.0090, 0.0110]	0.0018 [-0.0087, 0.0121]
PT	–	–	–	0.0000* [0.0000, 0.0034]	0.0000* [0.0000, 0.0033]
RR ∩ PT	–	–	–	–	0.0000 [-0.0013, 0.0024]
PT + Classifier	–	–	–	–	–

Table 9: Pairwise differences in SampEn metrics. Values represent the Median [25th, 75th percentiles] for each comparison. Differences are calculated as [Row] − [Column].

	RR	NN	PT	RR ∩ PT	PT + Classifier
RR	–	0.0000 [0.0000, 0.0000]	-0.0014 [-0.0164, 0.0093]	-0.0014 [-0.0150, 0.0127]	-0.0014 [-0.0149, 0.0118]
NN	–	–	-0.0014 [-0.0164, 0.0093]	-0.0014 [-0.0150, 0.0127]	-0.0014 [-0.0149, 0.0118]
PT	–	–	–	0.0000 [-0.0021, 0.0002]	0.0000 [-0.0021, 0.0000]
RR ∩ PT	–	–	–	–	0.0000 [-0.0023, 0.0014]
PT + Classifier	–	–	–	–	–

Table 10: Pairwise differences in R metrics. Values represent the Median [25th, 75th percentiles] for each comparison. Differences are calculated as [Row] − [Column].

	RR	NN	PT	RR ∩ PT	PT + Classifier
RR	–	0.0000 [0.0000, 0.0000]	-0.0019 [-0.0388, 0.0107]	0.0009 [-0.0164, 0.0311]	0.0018 [-0.0192, 0.0591]
NN	–	–	-0.0019 [-0.0388, 0.0107]	0.0009 [-0.0164, 0.0311]	0.0018 [-0.0192, 0.0591]
PT	–	–	–	0.0000 [-0.0024, 0.0288]	0.0000 [-0.0016, 0.0411]
RR ∩ PT	–	–	–	–	0.0000 [-0.0072, 0.0150]
PT + Classifier	–	–	–	–	–

5 Discussion

5.1 RR Series Comparison

The Pan Tompkins algorithm was originally published by Pan and Tompkins in 1985 and the method was validated on the MIT-BIH Arrhythmia Database [22]. Since then, it remains a commonly used QRS-detection algorithm in cardiac monitoring applications. However, its performance was established on recordings containing primarily sinus rhythm and common arrhythmias, not specifically on permanent AF. I applied the Pan Tompkins to 24-hour ECG recordings from the RATAF II study and compared its detections to detections by standard practice clinical hospital software.

Visual inspection of individual recordings, see Figure 12, showed close overall agreement between the detectors. Both detectors identify the same beats in most cases. However, minor timing differences were visible on individual beats. Figure 10 quantifies this across all recordings. The distribution of R-peak timing differences is centered around zero but it shows a discrete structure, with bars spread approximately ± 8 ms apart. This space corresponds exactly to one sample with the sampling frequency of 125 Hz (1/125 Hz = 8 ms). The Pan Tompkins algorithm used can only place R-peak detections at integer samples indices. Therefore, this is not a sign of any wrong detection by the algorithm, it is rather an unavoidable consequence of the sampling rate. The practical implication is that individual RR intervals detected by the Pan Tompkins algorithm carry an uncertainty of ± 1 sample. This uncertainty is unlikely to create major differences in HRV parameters, but it is important context when comparing RR intervals at high precision.

The two detectors did not always agree, see Figures 13, 14 and 15. Figure 13 depicts an instance where the clinical hospital software produced a false positive detection, splitting a single QRS-detection into two whereas the Pan Tompkins detector correctly identified one. In contrast, Figure 14 shows a case where Pan Tompkins mistook a fibrillatory wave for a QRS-complex, adding a false detection. The performance of Pan Tompkins decreases in low-quality and noisy signals. AF presents a real challenge with no organized atrial activity, it

produces smaller fibrillatory waves that can be misclassified as QRS-complexes by threshold based methods. Another challenge for these detectors is the fact that these are 24-hour Holter recordings, meaning that the presence of noise is unavoidable. Figure 15 illustrates one example where both detection methods failed, mistaking noise for a QRS-complex.

The distributions of missed beats (Figure 17) and extra beats (Figure 18) across all recordings show that the majority of recordings had very few such errors, with both distributions concentrated near zero. This is reassuring and consistent with the generally high agreement seen in individual windows. However, instances with higher disagreement between the two detection methods are present. There is no way of knowing which of the two detection methods is right without manually going through all the data. However, if a beat is missed by a detector, it inserts an artificially long RR interval into the series and an extra beat will insert an artificially short one. Both cases will affect HRV parameters. Long and short artificial intervals affect variance based parameters, such as pNN and rMSSD, and introduces false complexity which will affect the entropy measures. This motivates a quality assurance step, discussed in the next section.

The Pan Tompkins algorithm has been the dominant QRS-detector in cardiac monitoring for four decades now, but when used for ambulatory ECG recordings, some questions are raised. As mentioned before, the performance of the algorithm decreases with poorer signal quality and noisier segments. ECG data acquired during daily activities present a challenge for QRS-detection, for example, by increasing noise and artifacts due to patient movement. These limitations have been acknowledged and addressed in literature. Imtiaz and Khan (2022) introduced Pan Tompkins++ [7], an improved version of the original algorithm. This improved algorithm applies a bandpass filter of 5-18 Hz followed by an N-point moving average filter to suppress noise without discarding significant signal components, and uses three adaptive thresholds. When evaluated, the Pan Tompkins++ achieved 2.8% less false positives, a 1.8% reduction in false negatives and a 33% improvement in computational time compared to the original Pan Tompkins algorithm. The failures of Pan Tompkins in the present study, occasionally misclassified fibrillatory waves as QRS-complexes or missed beats in noisy segments, are exactly the types of errors Pan Tompkins++ were designed to reduce.

Liu et al. (2018) [40] systematically evaluated ten common QRS-detectors across six ECG databases and found that all algorithms achieved high detection accuracy above 99% on clean, high-quality ECG signals. Performance dropped substantially to below 80% for poor signal quality data. [40] This highlights that the challenges encountered in this thesis, occasional false detections in noisy segments, are not specific to the Pan Tompkins algorithm, but reflect a general limitation of common QRS-detectors when applied to real-world noisy recordings.

5.2 Quality Assurance

The simpler quality assurance approach consisted of retaining only beats detected by both the clinical hospital software and the Pan Tompkins algorithm (RR∩PT). As seen in Table 1, this pipeline retained 31 650 037 beats across all recordings, slightly fewer than either detector alone, which is expected since any beat missed by either method was excluded. This approach provides a simple quality filter. If a beat is detected by only one method, it is more likely to be a false detection or a missed beat caused by noise, and excluding it reduces the risk of incorrect RR intervals entering the HRV analysis. Despite retaining fewer beats, the HRV parameters computed from this pipeline were consistent with those from the individual detector series, see Figures 24 to 32 where the means lies within each others standard deviation. However, the median difference between this pipeline and the individual detectors was significant in most of the HRV parameters, although the magnitudes was minor. The main flaw in this approach is that it doesn't actually detect noise. It just flags a problem when the two detectors disagree. Therefore, as shown in Figure 15, it won't catch errors that both detectors make at the same time.

The BiLSTM achieved an overall accuracy of 97% on the validation set, with a sensitivity of 94.2% and a specificity of 99.8%. These results indicate that the classifier is highly effective at rejecting real noise. Of the 970 noisy windows in the validation set, 968 were correctly rejected, only 2 incorrectly passed through as clean beats. This specificity means that beats accepted by the classifier are very likely to be genuinely clean.

However, the sensitivity score is lower. This means that the classifier preferably discards borderline windows rather than risk including noisy data. For HRV-calculations, this is preferable, a false negative (passing noise as clean) is more damaging than a false positive (discarding a clean beat). This is because artificial RR intervals directly corrupt metrics such as rMSSD, SampEn, and ApEn, whereas discarding a small number of clean beats only marginally reduces the available data from a 24-hour recording.

The BiLSTM architecture is well suited for this classification task. The context and chronological order are important in ECG data analytics. A sudden spike in amplitude or morphological change only makes sense if the context and surrounding samples are known. A BiLSTM network extracts information both before and after the beat, making it suitable for detecting noise. The training data was created from the MIT-BIH Atrial Fibrillation Database and the MIT-BIH Noise Stress Test Database. Individual beats were extracted from the MIT-BIH Atrial Fibrillation Database and real noise in the form of baseline wander, electrode motion artifacts or muscle artifacts from the MIT-BIH Noise Stress Test Database were extracted. This was done to mimic the types of noise that appears in a 24-hour Holter recording. The balanced training set of 6874 samples and smooth convergence over 50 epochs with no notable gap between training and validation accuracy indicates that the model generalized well and

did not overfit to the training data.

Despite the strong overall performance, a notable limitation emerged when the classifier was applied to the RATAF II recordings. The network frequently classified beats affected by baseline wander as noisy, even when the QRS-complex itself was clearly visible, see Figure 22. This is a problem because baseline wander is a common phenomenon in long-term ECG monitoring, it arises for example from patient movement and respiration. In the majority of cases, baseline wander does not obscure the QRS-complexes or invalidate RR intervals, making it unnecessary to discard such beats, since it limits the amount of available data. It could also introduce unnecessary bias to the data if baseline wander occurs non-randomly across the recording. More of beats during physical activity will be discarded compared to those while the patient is at rest, because baseline wander is more frequent during physical activity.

The root cause likely lies in the training data composition. The training data was carefully constructed to include a combination of noise types (EM, MA, EM + BW and AM + BW). Baseline wander, by nature has low frequency and standard deviation over a 150 sample short window. It was noticed that the baseline wander did not distort any QRS-complexes, therefore the decision was made to not include baseline wander by it self in the training data, only in combination with another noise type. However, when the classifier encountered beats with a drifting baseline it triggered a noise classification anyway, despite the careful consideration.

Normalizing each window relative to its local baseline before classification, or adding the peak-to-peak amplitude relative to the local mean as an explicit feature, could perhaps help the network distinguish baseline wander from true noise that corrupts QRS-complexes.

The classifier was trained on MIT-BIH Atrial Fibrillation recordings and MIT-BIH Noise Stress Test, then applied to RATAF II data resampled to match the training sampling frequency (250 Hz). While the resampling aligns the temporal resolution, the RATAF II recordings were acquired in a different clinical setting and potentially with different equipment, meaning the noise characteristics may differ from those in the training set. The strong performance seen in the validation results may therefore not fully transfer to all RATAF II recordings, particularly those with unusual noise profiles. Future work should consider tuning the classifier on a labeled subset of RATAF II data.

Moreover, the limitations of classical QRS-detectors have motivated a growing amount of research into deep learning for ECG analysis. HeartNetEC is one such example, a deep learning architecture for ECG beat classification by Deevi et al (2021) [41]. The data for training and testing of HeartNetEC, was a combination of 12-lead ECG recordings and the MIT-BIH Noise Stress Test Database, which is similar to the data used for this thesis. The BiLSTM

classifier developed in this thesis takes a step in this direction, but work like HeartNetEC illustrates that the field is moving towards fully integrated deep learning approaches, handling everything from denoising, detection and classification. This is something that would improve robustness in challenging recordings such as 24-hour AF data used here.

5.3 HRV Analysis

The most remarkable finding from the HRV comparison is how consistent the results are across all five pipelines. Despite meaningful differences in the total number of beats retained (Table 1), ranging from 28 148 088 in the NN series to 32 168 700 in the raw RR series, the HRV parameters computed from each pipeline are very similar at every time point. Heart rate, SD, rMSSD, pNN measures, entropy values, and the regularity index all fall within each other's standard deviation ranges no matter the pipeline. This suggests that for 24-hour recordings in this patient population, the choice of QRS-detector and quality screening method has limited impact on these HRV parameters, a reassuring result for consistency in HRV in AF research.

Tables 11 and 12 (in Appendix A) and Figures 24 to 32, computed from the hospital RR and NN intervals, produced identical HRV results across all metrics and time points, despite the NN series retaining approximately 4 million fewer beats. In sinus rhythm, the distinction between RR and NN intervals is clinically meaningful because extra or skipped heartbeats produce shorter or longer intervals that affect variability estimates. In persistent AF, however, there are no truly normal beats and the rhythm is irregular. The removal of non-normal beats by the clinical software appears to have had negligible impact on the resulting HRV analysis. This finding suggests that in persistent AF, the RR/NN differentiation may be unnecessary for HRV analysis.

The PT + Classifier pipelines showed elevated SD, rMSSD and pNN80 at some time points compared to the other pipelines. As discussed in the quality assurance section, the BiLSTM classifier has a tendency to reject beats affected by baseline wander even when the QRS-complex is clearly visible. If these rejected beats are not randomly distributed across the RR series, then their removal could leave a bias, affecting HRV metrics. This is a subtle but worth noting effect of the classifier. It also brings up the fact that aggressive quality assurance could not always be neutral with respect to HRV outcomes.

Although the different pipelines mostly yielded consistent HRV parameters, some differences between pipelines were statistically significant. Statistical significance here does not imply clinical relevance. Even very small differences can reach significance if they are consistent across many recordings. The actual median differences were consistently very small, for example, the significant HR difference between RR and RR $\cap$ PT had a median of only 0.08 bpm, and the significant SD difference between PT and RR $\cap$ PT was less than 0.5 ms. These

magnitudes are very small and is possibly not meaningful for clinical interpretation. However, that is up to a physician to conclude.

The variance-based parameters SD, pNN and rMSSD showed the most statistically significant pairwise differences across pipelines, particularly for comparisons involving the intersection pipeline (RR ∩ PT) and the PT + Classifier pipeline. SD is the standard deviation of all RR intervals in the segment. It measures how spread out the intervals are around their mean. If a pipeline includes a few extra long or short RR intervals, caused by a missed beat or a false detection, those extreme values pull the mean slightly and increase the spread, inflating SD. On the contrary, if a pipeline removes beats that produce unusually long or short intervals, SD will decrease slightly. SD treats every interval equally, meaning that even a small number of extreme values at the edges of the distribution can shift the result noticeably.

rMSSD is even more sensitive. It is computed by taking the difference between each pair of successive RR intervals, squaring those differences, averaging them, and then taking the square root. The squaring step is what makes it sensitive, it means that large differences contribute disproportionately more to the result than small ones. In a 20-minute segment containing perhaps 2000 beats, even one or two extreme intervals introduced or removed by a pipeline can shift the rMSSD metric.

The pNN metrics (pNN20, pNN50, pNN80) are somewhat less sensitive. These metrics count the proportion of successive RR interval differences that exceed a fixed threshold, 20, 50, or 80 milliseconds respectively. Like rMSSD, they are focused on successive differences rather than the intervals themselves, so they are also affected by the presence or absence of beats that produce large interval jumps. However, pNN metrics express their result as a percentage of all intervals, and they apply a hard threshold rather than squaring the differences. This means that an extreme interval cannot contribute more than once to the percentage, no matter how extreme it is. This threshold-based counting is therefore less affected by outliers than the squaring in rMSSD, making the pNN metrics somewhat more stable across pipelines.

However, an important limitation to mention is the 20-minute segment extracted from all recordings to replicate the analysis protocol of the RATAF I study. Table 1 clearly shows that the five pipelines differ in their total beat counts, for example, the NN series retains approximately 4 million fewer beats than the RR series when considering all recordings. When looking at a short 20-minute window instead of a full 24-hours, there are far fewer heartbeats to analyze. This means any small error, like a missed beat, a false detection, or a noisy patch, has a much bigger impact. While a single missed beat doesn't matter over 24 hours, it can change the outcome in a short window. The question also becomes, what is the probability that a faulty detection or a noisy segment appears in that specific 20-minute

segment. Since the resulting HRV tables from the RR series and the NN series are identical over a relatively short window, one can argue that the beats in the window were mostly labeled normal. These findings from short windows should therefore be interpreted with some caution.

Across all the pipelines, the fact remained that with medication the heart rate decreases substantially from baseline to both one month and six month follow-up. This is expected since this is the goal with rate control treatments. In the same way, the variability metrics all increased from baseline to follow-ups with medication. There is a well-known inverse relationship between mean heart rate and beat-to-beat variability [42]. This is because the higher the heart rate, the shorter the RR intervals. Shorter RR intervals are less likely to present variability. Therefore, when comparing HRV metrics, it is important to compare patients with similar heart rates. [42]

On the contrary, the entropy measures (ApEn, SampEn) and the regularity index (R) remained notably stable across all time points and both treatments. This is an important finding, unlike variability metrics, entropy captures the irregularity and complexity of the RR series. The stability of these measures suggests that rate control medication altered the speed of the ventricular response without changing the underlying irregularity problem of AF. However, that was to be expected since both drugs are part of a rate control regime, therefore not altering the rhythm or atrial activity.

The HRV parameters computed in this thesis are roughly consistent with those reported in the RATAF I study by Corino et al.[38], which investigated rate control in a similar permanent AF patient population. Baseline values across all variance-based parameters agree between the two studies: baseline HR was 110 ± 18 bpm in RATAF I versus approximately 104 ± 17 bpm here, baseline SD was 133 ± 37 ms versus 140 ± 35 ms, baseline rMSSD was 171 ± 47 ms versus 183 ± 47 ms, and the pNN metrics were nearly identical. The results of the treatments were also similar, both studies showed reductions in heart rate and an increase in variance-based parameters under treatment.

The numerical values of ApEn, SampEn and R differ between the two studies, which was expected since the RATAF I study did not specify the exact parameters used for their computation, such as embedding dimension L or tolerance threshold r. Since these choices directly affect the values of ApEn, SampEn and R, comparison of these metrics between studies are not meaningful without identical parameters. What can be compared is the behavior. The entropy measures and the regularity index remained stable across treatment conditions in both studies, supporting the conclusion that rate control medication does not alter the irregularity of AF.

5.4 Limitations and Future Work

There are several aspects that can be improved in this thesis. First, the BiLSTM classifier could be improved to better handle baseline wander, for example by normalizing each beat window relative to its local baseline before classification, or by fine-tuning the model on a labeled subset of RATAF II data to get a better performance on this population. Second, replacing the Pan Tompkins QRS-detector algorithm with a more robust detector, such as Pan Tompkins++ mentioned before or a fully deep learning-based approach. Finally, the 20-minute segment analysis approach could be expanded. Either trying shorter segments, longer segments or both to see how the HRV metrics behave across pipelines.

6 Conclusion

This thesis developed and evaluated five full pipelines for HRV analysis in patients with permanent AF, using 24-hour ECG recordings from the RATAF II study. The Pan Tompkins algorithm showed generally close agreement with the clinical hospital software, with most disagreements attributable to the 125 Hz sampling resolution or isolated failure cases in noisy and irregular signal segments. A BiLSTM noise classifier was developed and achieved high accuracy on the validation data. However, when applied to RATAF II patient's ECG-recordings, a tendency to reject clear QRS-complexes affected by baseline wander was noticed and noted as a limitation. HRV parameters were computed across five distinct RR interval pipelines, and the results were astoundingly consistent across all of them, suggesting that for full 24-hour recordings, the choice of QRS-detector and quality assurance method has limited impact on HRV estimates in this patient population. Across all the different pipelines, rate control medication reduced heart rate and increased the variance-based HRV parameters, while entropy measures remained quite stable. All of these findings suggest that HRV analysis in AF is robust to reasonable pipeline choices over long recordings, while highlighting that methodological decisions remain important to consider, especially when analyzing shorter segments.

References

[1] L. Deng, Q. Li, H. Li, B. Li, Z. Cheng, and Y. Xiao, "The burden and trend prediction of atrial fibrillation and flutter associated with lead exposure: insights from the global burden of disease study 2021," *Frontiers in Cardiovascular Medicine*, vol. 12, p. 1638747, 2025.

[2] I. C. Van Gelder, M. Rienstra, K. V. Bunting, R. Casado-Arroyo, V. Caso, H. J. Crijns, T. J. De Potter, J. Dwight, L. Guasti, T. Hanke *et al.*, "2024 esc guidelines for the management of atrial fibrillation developed in collaboration with the european association

for cardio-thoracic surgery (eacts) developed by the task force for the management of atrial fibrillation of the european society of cardiology (esc), with the special contribution of the european heart rhythm association (ehra) of the esc. endorsed by the european stroke organisation (eso)," *European heart journal*, p. ehae176, 2024.

[3] Cleveland Clinic. (2025, Apr.) Atrial fibrillation (AFib). https://my.clevelandclinic.org/health/diseases/16765-atrial-fibrillation-afib. Accessed: 2026-05-02.

[4] P. Hämmerle, C. Eick, S. Blum, V. Schlageter, A. Bauer, K. D. Rizas, C. Eken, M. Coslovsky, S. Aeschbacher, P. Krisai *et al.*, "Heart rate variability triangular index as a predictor of cardiovascular mortality in patients with atrial fibrillation," *Journal of the American Heart Association*, vol. 9, no. 15, p. e016075, 2020.

[5] T. F. o. t. E. S. o. C. t. N. A. S. o. P. Electrophysiology, "Heart rate variability: standards of measurement, physiological interpretation, and clinical use," *Circulation*, vol. 93, no. 5, pp. 1043–1065, 1996.

[6] F. Shaffer and J. P. Ginsberg, "An overview of heart rate variability metrics and norms," *Frontiers in public health*, vol. 5, p. 290215, 2017.

[7] M. N. Imtiaz and N. Khan, "Pan-tompkins++: A robust approach to detect r-peaks in ecg signals," in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2022, pp. 2905–2912.

[8] ClinicalTrials.gov, "Rate control in atrial fibrillation ii (ratafii)," https://clinicaltrials.gov/study/NCT02695992, U.S. National Library of Medicine, 2016, Last updated 2022, identifier: NCT02695992.

[9] L. Sörnmo, *Atrial fibrillation from an engineering perspective*. Springer, 2018, vol. 10.

[10] Mayo Clinic Staff. (2026, Jan.) Atrial fibrillation – symptoms and causes. https://www.mayoclinic.org/diseases-conditions/atrial-fibrillation/symptoms-causes/syc-20350624. Accessed: 2026-05-02.

[11] Queensland Cardiovascular Group, "The conduction system," https://qcg.com.au/healthy-heart/conduction-system/, n.d.

[12] J. Hampton, J. Hampton, and D. Adlam, *The ecg made easy e-book: The ecg made easy e-book*. Elsevier Health Sciences, 2024.

[13] S. Vieau and P. A. Iaizzo, "Basic ecg theory, 12-lead recordings, and their interpretation," in *Handbook of Cardiac Anatomy, Physiology, and Devices*. Springer, 2015, pp. 321–334.

[14] S. Meek and F. Morris, "Introduction. i—leads, rate, rhythm, and cardiac axis," *Bmj*, vol. 324, no. 7334, pp. 415–418, 2002.

[15] Wikipedia Commons, "SinusRhythmLabels," https://commons.wikimedia.org/wiki/File:SinusRhythmLabels.svg, 2007, accessed: 2026-05-18.

[16] H. Q. Ontario *et al.*, "Long-term continuous ambulatory ecg monitors and external cardiac loop recorders for cardiac arrhythmia: a health technology assessment," *Ontario health technology assessment series*, vol. 17, no. 1, p. 1, 2017.

[17] Z. F. Mohd Apandi, R. Ikeura, S. Hayakawa, and S. Tsutsumi, "An analysis of the effects of noisy electrocardiogram signal on heartbeat detection performance," *Bioengineering*, vol. 7, no. 2, p. 53, 2020.

[18] L. Sörnmo and P. Laguna, *Bioelectrical signal processing in cardiac and neurological applications*. Academic press, 2005.

[19] G. Lenis, N. Pilia, A. Loewe, W. H. Schulze, and O. Dössel, "Comparison of baseline wander removal techniques considering the preservation of st changes in the ischemic ecg: a simulation study," *Computational and mathematical methods in medicine*, vol. 2017, no. 1, p. 9295029, 2017.

[20] C. Stolojescu, "Estimation of noise in ecg signals using wavelets," in *2008 11th International Conference on Optimization of Electrical and Electronic Equipment*. IEEE, 2008, pp. 113–118.

[21] R. Kher *et al.*, "Signal processing techniques for removing noise from ecg signals," *J. Biomed. Eng. Res*, vol. 3, no. 101, pp. 1–9, 2019.

[22] J. Pan and W. J. Tompkins, "A real-time qrs detection algorithm," *IEEE transactions on biomedical engineering*, no. 3, pp. 230–236, 1985.

[23] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[24] Practicum AI, "What are neural networks?" https://practicumai.org/deep_learning/01.1_what_are_neural_networks/, n.d.

[25] J. Zou, Y. Han, and S.-S. So, "Overview of artificial neural networks," *Artificial neural networks: methods and applications*, pp. 14–22, 2009.

[26] Y. Sun, J. Zhang, Z. Yu, Y. Zhang, and Z. Liu, "Bidirectional long short-term neural network based on the attention mechanism of the residual neural network (resnet–bilstm–attention) predicts porosity through well logging parameters," *ACS omega*, vol. 8, no. 26, pp. 24 083–24 092, 2023.

[27] Cleveland Clinic. (2022) Autonomic nervous system. Cleveland Clinic. Accessed: 2026-05-11. [Online]. Available: https://my.clevelandclinic.org/health/body/23273-autonomic-nervous-system

[28] L. Zhang, B. Li, and L. Wu, "Heart rate variability in patients with atrial fibrillation of sinus rhythm or atrial fibrillation: chaos or merit?" *Annals of medicine*, vol. 57, no. 1, p. 2478474, 2025.

[29] S. Pincus, "Approximate entropy (apen) as a complexity measure," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 5, no. 1, pp. 110–117, 1995.

[30] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," *American journal of physiology-heart and circulatory physiology*, vol. 278, no. 6, pp. H2039–H2049, 2000.

[31] A. Porta, G. Baselli, D. Liberati, N. Montano, C. Cogliati, T. Gnecchi-Ruscone, A. Malliani, and S. Cerutti, "Measuring regularity by means of a corrected conditional entropy in sympathetic outflow," *Biological cybernetics*, vol. 78, no. 1, pp. 71–78, 1998.

[32] K. Enge, S. R. Ulimoen, S. Enger, S. Onarheim, M. Olufsen, A. H. Pripp, T. Steinsvik, C. Hall, M. Hetland, and A. Tveit, "Effects of diltiazem and metoprolol on levels of high-sensitivity troponin i in patients with permanent atrial fibrillation: a randomized trial," *BMC Cardiovascular Disorders*, vol. 25, no. 1, p. 181, 2025.

[33] Telemetric and Holter ECG Warehouse (THEW), "The ISHNE Holter standard file format," accessed: May 31, 2026. [Online]. Available: http://thew-project.org/thewfileformat.htm

[34] H. Sedghamiz, "Matlab implementation of pan tompkins ecg qrs detector," *Code Available at the File Exchange Site of MathWorks*, pp. 1–3, 2014.

[35] G. B. Moody and R. G. Mark, "A new method for detecting atrial fibrillation using R-R intervals," *Computers in Cardiology*, vol. 10, pp. 227–230, 1983.

[36] G. B. Moody, W. E. Muldrow, and R. G. Mark, "A noise stress test for arrhythmia detectors," *Computers in Cardiology*, vol. 11, pp. 381–384, 1984.

[37] MathWorks, "Classify ecg signals using long short-term memory networks," MathWorks Help Center, 2024, accessed: May 12, 2026. [Online]. Available: https://se.mathworks.com/help/signal/ug/classify-ecg-signals-using-long-short-term-memory-networks.html

[38] V. D. Corino, S. R. Ulimoen, S. Enger, L. T. Mainardi, A. Tveit, and P. G. Platonov, "Rate-control drugs affect variability and irregularity measures of rr intervals in patients with permanent atrial fibrillation," *Journal of Cardiovascular Electrophysiology*, vol. 26, no. 2, pp. 137–141, 2015.

[39] Jesús Monge-Álvarez, "A set of entropy measures for temporal series (1d signals)," MATLAB Central File Exchange, 2026, retrieved May 12, 2026. [Online]. Available: (https://se.mathworks.com/matlabcentral/fileexchange/50289-a-set-of-entropy-measures-for-temporal-series-1d-signals)

[40] F. Liu, C. Liu, X. Jiang, Z. Zhang, Y. Zhang, J. Li, and S. Wei, "Performance analysis of ten common qrs detectors on different ecg application cases," *Journal of healthcare engineering*, vol. 2018, no. 1, p. 9050812, 2018.

[41] S. A. Deevi, C. P. Kaniraja, V. D. Mani, D. Mishra, S. Ummar, and C. Satheesh, "Heartnetec: a deep representation learning approach for ecg beat classification," *Biomedical Engineering Letters*, vol. 11, no. 1, pp. 69–84, 2021.

[42] T. Nieminen, M. Kähönen, T. Kööbi, K. Nikus, and J. Viik, "Heart rate variability is dependent on the level of heart rate," *American heart journal*, vol. 154, no. 1, p. e13, 2007.

7 Appendix A

This appendix features the HRV tables replicated from the RATAF I study, across all pipelines.

Table 11: HRV metrics (mean ± standard deviation across patients) computed from the RR intervals extracted by the clinical hospital software, at baseline, one and six months of medication.

		1 Month		6 Months	
Metric	**Baseline**	**D**	**M**	**D**	**M**
HR (bpm)	104.6585 ± 17.1720	84.8471 ± 20.5525	91.3857 ± 15.5242	80.3473 ± 16.9107	92.5906 ± 18.4229
SD (ms)	140.5212 ± 34.6226	179.0718 ± 61.4890	157.7961 ± 38.0134	190.6896 ± 45.4277	162.8068 ± 51.6176
pNN20 (%)	87.0573 ± 4.2180	90.4399 ± 4.2330	89.0415 ± 6.3869	90.8989 ± 4.7382	89.5757 ± 4.8378
pNN50 (%)	69.8272 ± 8.7064	76.9521 ± 9.0743	74.7909 ± 8.5122	78.2695 ± 7.7751	75.2912 ± 8.4773
pNN80 (%)	56.1736 ± 10.6241	65.1888 ± 12.6069	62.9525 ± 9.1564	67.6234 ± 9.0030	63.1277 ± 10.4930
rMSSD (ms)	183.5460 ± 50.2073	242.7283 ± 88.0324	213.2890 ± 55.0332	256.8840 ± 64.9670	219.2200 ± 74.1614
ApEn (a.u.)	1.7922 ± 0.0757	1.7778 ± 0.0765	1.7680 ± 0.2201	1.7491 ± 0.0980	1.7782 ± 0.0975
SampEn (a.u.)	1.9114 ± 0.2181	1.9815 ± 0.1918	1.9067 ± 0.3385	1.9092 ± 0.2776	1.9167 ± 0.2278
R (a.u.)	1.1526 ± 0.1135	1.2118 ± 0.2214	1.1456 ± 0.1847	1.2455 ± 0.1757	1.1673 ± 0.1263

Table 12: HRV metrics (mean ± standard deviation across patients) computed from the NN intervals extracted by the clinical hospital software, at baseline, one and six months of medication.

		1 Month		6 Months	
Metric	**Baseline**	**D**	**M**	**D**	**M**
HR (bpm)	104.6585 ± 17.1720	84.8471 ± 20.5525	91.3857 ± 15.5242	80.3473 ± 16.9107	92.5906 ± 18.4229
SD (ms)	140.5212 ± 34.6226	179.0718 ± 61.4890	157.7961 ± 38.0134	190.6896 ± 45.4277	162.8068 ± 51.6176
pNN20 (%)	87.0573 ± 4.2180	90.4399 ± 4.2330	89.0415 ± 6.3869	90.8989 ± 4.7382	89.5757 ± 4.8378
pNN50 (%)	69.8272 ± 8.7064	76.9521 ± 9.0743	74.7909 ± 8.5122	78.2695 ± 7.7751	75.2912 ± 8.4773
pNN80 (%)	56.1736 ± 10.6241	65.1888 ± 12.6069	62.9525 ± 9.1564	67.6234 ± 9.0030	63.1277 ± 10.4930
rMSSD (ms)	183.5460 ± 50.2073	242.7283 ± 88.0324	213.2890 ± 55.0332	256.8840 ± 64.9670	219.2200 ± 74.1614
ApEn (a.u.)	1.7922 ± 0.0757	1.7778 ± 0.0765	1.7680 ± 0.2201	1.7491 ± 0.0980	1.7782 ± 0.0975
SampEn (a.u.)	1.9114 ± 0.2181	1.9815 ± 0.1918	1.9067 ± 0.3385	1.9092 ± 0.2776	1.9167 ± 0.2278
R (a.u.)	1.1526 ± 0.1135	1.2118 ± 0.2214	1.1456 ± 0.1847	1.2455 ± 0.1757	1.1673 ± 0.1263

Table 13: HRV metrics (mean ± standard deviation across patients) computed from the RR intervals extracted by the Pan Tompkins QRS-detector, at baseline, one and six months of medication

		1 Month		6 Months	
Metric	**Baseline**	**D**	**M**	**D**	**M**
HR (bpm)	104.5865 ± 17.0719	84.6977 ± 20.1013	92.7702 ± 17.2586	80.1891 ± 16.4683	92.6064 ± 18.6070
SD (ms)	136.3791 ± 33.8022	178.2715 ± 59.2738	158.7381 ± 38.6127	190.5723 ± 43.1363	160.4070 ± 43.4017
pNN20 (%)	87.1223 ± 4.4038	90.4134 ± 4.2357	88.8152 ± 7.4733	90.8863 ± 4.7714	89.6076 ± 4.7596
pNN50 (%)	70.3270 ± 8.5041	76.9362 ± 9.1514	74.7666 ± 9.1722	78.2889 ± 7.7967	75.6474 ± 8.4860
pNN80 (%)	58.9348 ± 10.2559	67.1369 ± 12.0478	64.9560 ± 9.4128	69.2410 ± 8.7999	65.3306 ± 10.1100
rMSSD (ms)	183.8893 ± 48.9687	242.0600 ± 84.5901	215.3495 ± 55.8364	256.2677 ± 61.6878	217.5152 ± 64.0828
ApEn (a.u.)	1.7999 ± 0.0743	1.7802 ± 0.0700	1.7697 ± 0.2338	1.7486 ± 0.0930	1.7625 ± 0.1541
SampEn (a.u.)	1.9291 ± 0.2203	1.9863 ± 0.1827	1.8982 ± 0.3460	1.9078 ± 0.2798	1.9223 ± 0.2335
R (a.u.)	1.2072 ± 0.1325	1.2354 ± 0.1924	1.1927 ± 0.1684	1.2569 ± 0.1560	1.1888 ± 0.1292

Table 14: HRV metrics (mean ± standard deviation across patients) computed from the intersection of RR intervals from the clinical hospital software and the Pan Tompkins detector, at baseline, one and six months of medication.

		1 Month		6 Months	
Metric	**Baseline**	**D**	**M**	**D**	**M**
HR (bpm)	103.9209 ± 16.7335	83.5956 ± 19.3843	89.7010 ± 16.6479	79.5223 ± 15.8788	90.1832 ± 19.5459
SD (ms)	141.8313 ± 33.5748	187.3657 ± 63.1077	173.5457 ± 63.1667	194.6804 ± 44.2411	185.5493 ± 85.4894
pNN20 (%)	87.1433 ± 4.3902	90.5587 ± 4.2341	89.2762 ± 7.2893	90.9780 ± 4.7295	89.6862 ± 5.0363
pNN50 (%)	70.3814 ± 8.4903	77.2392 ± 9.1524	75.6497 ± 9.1915	78.4557 ± 7.6828	75.9502 ± 8.8640
pNN80 (%)	59.1161 ± 10.2396	67.5412 ± 12.0657	66.0489 ± 9.6041	69.5153 ± 8.6085	65.8981 ± 10.7907
rMSSD (ms)	192.3490 ± 48.7185	254.7722 ± 89.3293	237.8900 ± 95.8035	263.2910 ± 62.0963	253.0826 ± 126.1560
ApEn (a.u.)	1.7750 ± 0.0789	1.7656 ± 0.0746	1.7377 ± 0.2599	1.7461 ± 0.0922	1.7402 ± 0.3383
SampEn (a.u.)	1.8917 ± 0.2297	1.9594 ± 0.1929	1.8864 ± 0.3455	1.9039 ± 0.2813	1.8843 ± 0.3144
R (a.u.)	1.1121 ± 0.1861	1.1792 ± 0.2535	1.1745 ± 0.1761	1.2507 ± 0.1673	1.1205 ± 0.1817

Table 15: HRV metrics (mean ± standard deviation across patients) computed from the Pan Tompkins RR intervals filtered by the BiLSTM noise classifier, at baseline, one and six months of medication.

		1 Month		6 Months	
Metric	**Baseline**	**D**	**M**	**D**	**M**
HR (bpm)	104.2851 ± 17.0287	83.4721 ± 19.4986	91.8262 ± 17.7944	79.6001 ± 15.8914	90.9089± 18.6212
SD (ms)	141.7744 ± 32.9515	189.6134 ± 63.3099	166.8811 ± 69.4318	194.7470 ± 42.2374	173.7094± 69.4318
pNN20 (%)	87.1573 ± 4.4020	90.5540 ± 4.2151	88.9545 ± 4.7553	90.9729 ± 4.6921	89.7294 ± 4.7553
pNN50 (%)	70.5008 ± 8.4840	77.3012 ± 9.0982	75.0183 ± 8.4783	78.4469 ± 7.6321	75.8333 ± 8.4783
pNN80 (%)	59.1962 ± 10.2244	67.6958 ± 11.9723	65.2605 ± 10.2491	69.4871 ± 8.5605	65.8274 ± 10.2491
rMSSD (ms)	189.9930 ± 47.9488	258.2703 ± 88.5103	225.8317 ± 92.2756	263.0708 ± 59.4849	234.0991 ± 92.2756
ApEn (a.u.)	1.7863 ± 0.1826	1.7549 ± 0.0785	1.7490 ± 0.2424	1.7412 ± 0.0940	1.7471 ± 0.3198
SampEn (a.u.)	1.9084 ± 0.2235	1.9263 ± 0.1991	1.8857 ± 0.3462	1.9005 ± 0.2879	1.9114 ± 0.2477
R (a.u.)	1.1278 ± 0.1923	1.1783 ± 0.2378	1.1632 ± 0.1295	1.2197 ± 0.2074	1.1704 ± 0.1296