FAKULTÄT FÜR
ELEKTROTECHNIK UND
INFORMATIONSTECHNIK
Faculty of Electrical Engineering and Information Technology

Combining Symmetric Projection Attractor Reconstruction with Machine Learning to Automatically Detect Atrial Fibrillation During Hemodialysis

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur/in

im Rahmen des Masterstudiums

Biomedical Engineering, 066 453

eingereicht von

Veronika Cap
Matrikelnummer 11802546

an der Fakultät für Elektrotechnik und Informationstechnik
der Technischen Universität Wien

Betreuung: Ao.Univ.Prof. Dipl.Ing. Dr.techn. Eugenijus Kaniusas

Mitwirkung: Univ.Lektor Dipl.-Ing. Dr.techn. Christopher Mayer

Wien, 02.11.2021

Veronika Cap

Veronika Cap

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit gemäß dem Code of Conduct – Regeln zur Sicherung guter wissenschaftlicher Praxis, insbesondere ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel, angefertigt wurde. Die aus anderen Quellen direkt oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder in ähnlicher Form in anderen Prüfungsverfahren vorgelegt.

Wien, 02.11.2021

Veronika Cap

Veronika Cap

Acknowledgements

First of all, I want to thank Christopher Mayer, my supervisor at the AIT Austrian Institute of Technology, for introducing me to the entire world of attractor reconstruction, fostering my interest in cardiological signal analysis and introducing me to his research contacts in this field. Thank you for providing me with this opportunity, keeping me on track with my thesis and always being available for my questions or concerns.

Thank you to Eugenijus Kaniusas not only for supervising this thesis on behalf of TU Wien, but also sparking my interest in biomedical signal analysis with his wonderful lectures. They were definitely a highlight of this master program and a weekly joy. Learning to view the human body with the eye of an engineer and understanding physiologic processes and concepts in a technical way, is something I will benefit from for the rest of my career.

My gratitude also goes out to Philip Aston, Jane Lyle and Manasi Nandi and Peter Charlton, from the University of Surrey and King's College London for inviting me to join their attractor meetings, to see Jane Lyle’s talk on ECG attractors at this year’s SIAM Conference on Applications of Dynamical Systems and providing me with their manuscript “Symmetric Projection Attractor Reconstruction: Sex Differences in the ECG” ahead of publication. Your work laid the foundations of this thesis.

Thank you to the Austrian Research Promotion Agency (FFG) for financing this thesis, via their FEMtech internship program. Supporting women to take up scientific and engineering positions is a wonderful cause and getting paid for your work is a pleasant change for a young scientist.

I also want to express my gratitude to my family, especially my parents, for supporting and encouraging me throughout my life and pushing me to do my very best instead of settling for okay or average. Thank you for believing in me and being there whenever I need you.

Last but not least, a massive thank you goes out to Marc Cudlik for his wonderful support, always, but especially during this time. Your tolerance for my incessant complaining and unnecessary phone calls is a miracle yet to be solved by science. You make everything better just by being there. I love you so much, thank you for being in my life!

Abstract

Atrial fibrillation (AF) is highly prevalent among patients suffering from renal failure and linked to a doubled one-year mortality rate in this group. Implementing a reliable algorithm that can detect AF on Electrocardiograms (ECGs) recorded during patients' regular dialysis treatments, would allow for routine monitoring, without further increasing patients' hospital time. Unfortunately, ECGs recorded during dialysis are harder to interpret via automated algorithms, because the ECG waveform is altered by fluid and electrolyte shifts. Symmetric projection attractor reconstruction (SPAR) is a new method of analyzing cardiovascular waveform data, that may be able to overcome these challenges. SPAR uses delay coordinates to convert a short time series signal into a so-called attractor, bound in three- or higher-dimensional phase space. Its rotationally symmetric, two-dimensional projections emphasize different aspects of the original signal. This can be used to better understand the underlying waveform.

This thesis combines SPAR with stacked *k*-nearest neighbor classifiers to automatically detect AF from a single lead ECG. The methodology was first established on ECGs from the open access database PhysioNet and then applied to ECGs recorded during dialysis. 30s ECG samples were taken from three different PhysioNet databases and extracted from 24h ECGs recorded during a study on end stage renal disease, at six different timepoints during and after dialysis. Models were trained on six different training and test compositions, each containing an approximately equal number of AF and control records. Controls for the dialysis records were matched based on age, gender and dialysis vintage in months. SPAR was used to generate 20 attractor projections from every ECG sample. These were quantified by calculating their angular density distribution, radial density distribution and attractor outline. By training one *k*-nearest neighbor classifier on each of the density curves, stacked models of 60 classifiers were trained for each training set. Test records were classified based on the mean posterior probability predictions of all classifiers in the model, that have a cross-validated accuracy of 70% or more based on 10-fold cross validation. All algorithms and data processing steps were implemented in MATLAB®.

The highest performing PhysioNet model achieved classification accuracies of 89.3% and 93.8% on the two PhysioNet test sets. The model trained on samples at the beginning, middle and end of dialysis classified samples from the start of dialysis test set with and average accuracy of 85.7%. Including both PhysioNet and dialysis samples in the training set showed no improvements for the classification of either category. The visual comparison of PhysioNet and dialysis three-point attractors and their densities also showed that the differences between the two groups exceed the differences between AF and no AF records of either group. This confirms that changes in the ECG caused by the dialysis are also visible in the ECG attractors.

These classification accuracies achieved by the models in this thesis compare well to other attempts at automated AF detection in the literature. Most of these approaches rely on convolutional neural networks or meticulously selected features. Compared to these attempts, the methodology presented in this thesis has the advantage of a simpler, less computationally expensive methodology, good stability towards outliers and noise and no necessity for feature selection. Its ability to classify very short ECG samples in a short time qualify it for real time monitoring applications. Experimenting with even less preprocessing and investigating methods of feature selection may be ways to further improve AF detection using SPAR.

Kurzfassung

Vorhofflimmern (VHF) ist eine häufige Nebenerkrankung bei Menschen mit chronischem Nierenversagen, die statistisch mit einer Verdopplung der Ein-Jahres-Mortalität einher geht. Routineüberwachung der kardialen Gesundheit, mit automatischer Detektion von VHF im Elektrokardiogramm (EKG) wäre ein wichtiger Schritt, um eine rechtzeitige Diagnose zu gewährleisten. Um zusätzliche Spitalszeiten zu vermeiden, sollten diese Messungen während der Dialyse durchgeführt werden. Diese EKGs sind allerdings schwerer auszuwerten, da die dabei auftretenden Elektrolyt- und Flüssigkeitsschwankungen, die EKG-Kurve verzerren.

„Symmetric projection attractor reconstruction" (SPAR) ist eine neuartige Methode der EKG-Auswertung, die diesen Herausforderungen gewachsen sein könnte. Mithilfe von Verzögerungskoordinaten werden dabei aus einem kurzen Zeitsignal Attraktoren im drei- oder höherdimensionalen Phasenraum errechnet. Die kreissymmetrischen, zweidimensionalen Projektionen dieser Attraktoren heben verschiedene Eigenschaften des Ursprungssignals hervor und können so zu einem besseren Verständnis der ursprünglichen Wellenform beitragen.

In dieser Arbeit wird SPAR mit Nächste-Nachbarn-Klassifikation (*k*-nearest neighbor) kombiniert, um automatisch VHF im Einkanal-EKG zu detektieren. Die Methodik wurde dabei zuerst an EKGs aus der frei zugänglichen Datenbank PhyioNet getestet und dann auf EKGs während der Dialyse angewandt. Zweitere wurden aus 24h EKGs entnommen, welche während einer Studie zu Nierenversagen und Dialyse aufgezeichnet wurden. Für die Analyse wurden 30s Ausschnitte zu sechs verschiedenen Zeitpunkten extrahiert. Zusammen mit 30s EKGs aus den PhysioNet Datenbanken wurden sechs verschiedene Trainings- und Testdatensätze zusammengestellt, wobei auf ein ausgeglichenes Verhältnis zwischen VHF und Kontrollgruppendaten geachtet wurde. Die Kontrollgruppe der Dialysedaten wurde dabei nach Alter, Geschlecht und Dialysemonaten gematcht. Mittels SPAR wurden aus jedem EKG Sample 20 rotationssymmetrische Attraktorprojektionen generiert, die anschließend über drei Dichteverteilungen quantifiziert wurden. Für jede dieser 60 Kurven wurde dann ein Nächste-Nachbarn-Klassifikator trainiert. Diese bilden zusammen ein Modell. Ob ein Test-Sample als VHF oder kein VHF klassifiziert wird, errechnet sich als Mittelwert der für das Sample ermittelten Zuordnungswahrscheinlichkeiten aller Klassifikatoren im Modell, welche mehr als 70% Klassifizierungsgenauigkeit bei einer fünffachen Kreuzvalidierung während des Trainingsprozesses erreichten. Die Implementierung dieser Schritte erfolgte in MATLAB®.

Das bessere der beiden Modelle, die nur an PhysioNet Daten trainiert wurden, erreichte eine mittlere Klassifizierungsgenauigkeit von 89.3% und 93.8% für die beiden PhysioNet Testdatensätze. Das beste Klassifikationsergebnis für am Anfang der Dialyse aufgezeichnete EKGs, zeigte ein Modell welches an EKGs von Anfang, Mitte und Ende der Dialyse trainiert wurde, mit einer mittleren Klassifikationsgenauigkeit von 85.7%. Das Kombinieren von PhysioNet und Dialyse Samples in den Trainingsdatensätzen verbesserte für keine der beiden Gruppen die Klassifizierung. Ein visueller Vergleich der Attraktorprojektionen und Dichtekurven für Beispiele der PhysioNet- und Dialysedaten, zeigte dass die Unterschiede zwischen den beiden Gruppen weitaus deutlicher sind als die Unterschiede zwischen VHF und kein VHF innerhalb der Dialysedaten. Dies bestätigt, dass die durch die Dialyse hervorgerufenen Veränderungen der EKG-Wellenform auch in den Attraktoren sichtbar sind.

Die Klassifikationsgenauigkeiten der in dieser Arbeit präsentierten Modelle sind mit den Ergebnissen weitaus komplexerer und rechenaufwändigerer Ansätze aus der Literatur, welche großteils auf Neuronalen Netzen basieren, vergleichbar. Ein weiterer Vorteil der SPAR Methode ist ihre Stabilität gegenüber Ausreißern oder Rauschen im Eingangssignal, wodurch weniger Vorverarbeitung des Rohsignals vonnöten ist. Da SPAR auch auf sehr kurze EKG Aufzeichnungen angewandt werden kann, eignet sich diese Methode auch für Echtzeitanwendungen. Mögliche weitere Schritte zur Verbesserung der hier präsentierten Methodik wären eine weitere Reduktion der Vorverarbeitung und Qualitätsprüfung des rohen EKG Signals und eine Vorauswahl der Inputfeatures für die einzelnen Klassifikatoren.

Abbreviations and Symbols

2D	Two dimensional
3D	Three dimensional
AF	Atrial Fibrillation
bpm	Beats per minute
ECG	Electrocardiogram
FN	False negative
FP	False positive
LA	Left atrium
LV	Left ventricle
RA	Right atrium
RV	Right ventricle
SD	Standard deviation
SPAR	Symmetric Projection Attractor Reconstruction
TN	True negative
TP	True positive
WFDB	Waveform database

Table of Contents

1. Introduction

1.1 Motivation

According to a 2018 report by the world health organization, an estimated 5 to 10 million annual deaths are attributable to kidney disease [1]. Even under regular hemodialysis, patients with renal failure suffer from increased cardiovascular morbidity and mortality [2]. Thus, frequent monitoring of dialysis patient's cardiac health is highly important to recognize risk factors and early signs of various cardiac conditions. One of the most important tools in cardiac diagnostics and risk stratification is electrocardiography (ECG), which uses skin electrodes to measure the heart's electrical excitation. Because dialysis patients already spend a significant amount of time in the hospital for their triweekly dialysis sessions, ECG measurements should be performed during these times, to avoid additional hospital time. Unfortunately, ECG data collected during the hemodialysis treatment are difficult to interpret, because amplitudes and intervals are altered by fluid and electrolyte shifts caused by the dialysis [2]. This is especially detrimental to automated ECG analysis algorithms, which traditionally rely on the automatic detection of specific points on the ECG and the subsequent calculation of amplitudes and time intervals based on these markers (e.g., detection of QRS-complex and T-wave, calculation of the QT-interval) [2]. Symmetric Projection Attractor Reconstruction (SPAR) is a novel approach of analyzing cardiovascular waveform data, which may be able to overcome these issues, since it does not require the detection of specific points in the ECG signal other than the easily identifiable R-peak. Additionally, attractor reconstruction has the advantage of analyzing the underlying waveform morphology which is discarded in conventional ECG analyses, even though it may hold more in-depth information about the patient's cardiovascular system [3–5].

There is a variety of cardiovascular morbidities that are prevalent in dialysis patients and diagnoseable through the ECG. This thesis focusses on atrial fibrillation (AF), the most common clinically significant cardiac arrythmia [6]. With a prevalence of approximately 10% [7], AF is a frequent diagnosis in patients suffering from renal failure. A 2011 study on the prevalence of AF among hemodialysis patients showed that one-year mortality rates were more than double in patients with AF compared with those without it [7]. Establishing an algorithm for automatic AF detection in ECGs recorded during dialysis would allow early diagnosis of this serious condition, without causing additional hospital time to the patient or straining hospital's limited resources of qualified diagnosticians.

1.2 Aim of the thesis

The aim of this thesis is to combine the SPAR methodology with machine learning techniques to develop an automated AF detection algorithm that is also successful in ECG data collected during hemodialysis. For this purpose, data from different sources are combined into six datasets that contain varying combinations of ECGs collected at 6 time points during or after a hemodialysis session, as well as short recordings from patients without renal failure, with or without AF. Each dataset is split into a training and test set with a 30% hold out approach and even distribution of AF characteristics. A MATLAB® algorithm is implemented, that automatically calculates and quantifies attractors from each ECG signal, following the SPAR methodology. The resulting attractor parameters are then used as feature inputs to multiple *k*-nearest neighbor classifiers that are combined into one model for each of the six training sets. Each model is then tested on multiple test sets to evaluate its performance in different applications. The highest performing models are compared to other AF detection models documented in the literature to see how the methodology performs in comparison to other approaches.

1.3 Thesis outline

This thesis is divided into 7 chapters, starting with an introduction that includes motivation, aim and thesis outline. This is followed by a background section that provides medical and technical information concerning the topics of this thesis. Section 3, methods and implementation, notes the workflow behind establishing the AF detection models, from data sourcing, sample preparation, feature extraction to model training and testing. The following section, statistics and visualization, contains the statistical tests and data visualization techniques used to generate the results. These are documented in section 5. The 6th section, the discussion, interprets, connects and contextualizes the results. The last chapter of this thesis, the conclusion, summarizes the most notable observations from the discussion and provides an outlook into further improvements and possible applications of the methodology for AF detection presented in this thesis.

2. Background

This chapter provides medical background information on the human heart, its anatomy, physiology and particularly its electrical excitation. It introduces the ECG as a biosignal, explains what atrial fibrillation (AF) is and how it is diagnosed. This section also provides an introduction to symmetric attractor reconstruction (SPAR), including an overview on origins and underlying concepts as well as the practical methodology necessary to apply SPAR to an ECG signal. The last subsection of this chapter introduces *k*-nearest neighbor classification, the classification algorithm that was used to build the AF detection models presented in this thesis.

2.1 Physiological Background

This section provides background information on the heart, the cardiac cycle and the ECG. If not otherwise specified, information presented in this section is based on textbooks of Faller and Schünke [8], Kaniusas [9], Betts et.al [10] and Rawshani [11].

2.1.1 Heart anatomy and blood flow

The human heart is a muscular hollow organ that uses rhythmic contractions to pump blood through the circulatory system. Located medially between the lungs in the thoracic cavity, a space also referred to as the mediastinum, it is surrounded by the pericardium, a fibroelastic sac, that separates the heart from other mediastinal structures. The great veins, superior and inferior vena cava, the pulmonary vein, and the great arteries, aorta and pulmonary artery, enter the heart at its superior surface, called the base, see figure 1. Through these vessels, blood enters and exits the heart. The cardiac septum separates the heart into its left and right side. Each side has one atrium that acts as a receiving chamber, and one ventricle, that propels the blood into the respective artery. Unidirectional flow in and around the heart is ensured by one-way valves opening and closing in response to pressure differences. The first set of valves, tricuspid and mitral valve, are located between the atrium and ventricle of each side and prevent backflow into the respective atrium as the ventricle contracts. The second set of valves, the semilunar valves, are located at the entrance of the pulmonary artery and aorta and prevent backflow while the ventricle refills.

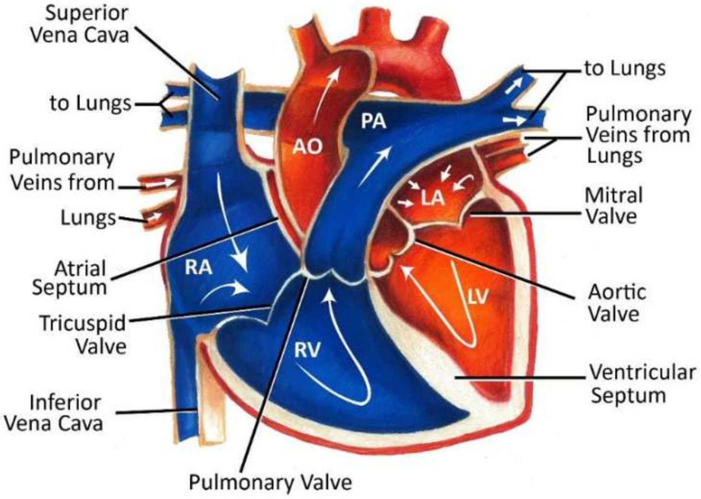


***Figure 1 – Heart anatomy and blood flow.** Blue color indicates deoxygenated blood, red oxygenated blood, arrows represent the directions of blood flow between the 4 ventricles, right atrium (RA), right ventricle (RV), left atrium (LA) and left ventricle (LV) and the two major arteries pulmonary artery (PA) and aorta (AO). [12]*

Functionally, the heart acts as a double pump, supplying both the systemic and pulmonary loop of the circulatory system. Deoxygenated blood enters the heart at the right atrium through superior and inferior vena cava (see figure 1). Contraction of the atrium pushes the blood into the right ventricle, from where it is ejected into the pulmonary artery. The pulmonary artery transports the deoxygenated blood to the lungs, for gas exchange. Oxygenated blood from the lungs is transported back to the left atrium through the pulmonary vein. Contraction of the left atrium pushes the blood into the left ventricle, from where it is ejected into the systemic circuit through the aorta. From there, large vessels branch off towards all areas of the human body, splitting into ever-smaller vessels. The smallest arteries branch off into tiny, thin-walled capillaries. These allow oxygen and nutrients to diffuse into the surrounding tissue, while carbon dioxide and other metabolic waste products are removed from the tissue. Vessels carrying this now deoxygenated blood combine to larger and larger veins, transporting the blood back to the heart.

2.1.2 The Cardiac Conduction System

Cardiac muscle is a highly specialized tissue found only in the heart. Like skeletal muscle, it is striated and organized in sarcomeres, but with shorter fibers, usually containing only one nucleus. What is especially unique about heart muscle, is its mechanism of excitation. While in other types of muscle each contraction is triggered by an impulse from a nerve fiber, the heart has its own pacemaking structures and conductive pathways made up of specialized cardiac muscle cells, see figure 2. These structures control and coordinate the contractions of the different chambers, to allow them to work together to efficiently pump blood through the two circulation loops. Located at the right atrium, the sinoatrial node is the heart's main pacemaker (see figure 2). It consists of self-excitable cells, that periodically depolarize at a frequency of approx. 60-70 beats per minute (bpm), although the exact frequency is modified by the autonomic nervous system via the vagus nerve, to adapt to varying demands. The atrioventricular node can act as a secondary pacemaker if needed, with a lower heart rate of 50 bpm. At an even lower frequency of 30 bmp, the Bundle of His and Purkinje fibers can act as tertiary pacemakers. In a healthy state, contractions are only initiated by the sinoatrial node.

The heart consists of two separate muscles, one for the two atria, one for the two ventricles. Atria and ventricles are electrically isolated from each other; the atrioventricular node is their only point of conduction. The conduction speed in the atrioventricular node is lower, which induces a necessary time delay in the conduction. This compensates for the faster electrical propagation compared to the mechanical pumping action of the atria, thus allowing the relaxed ventricles to fully fill before action potentials traveling down the Bundle of His and Purkinje Fibers cause the ventricles to contract.

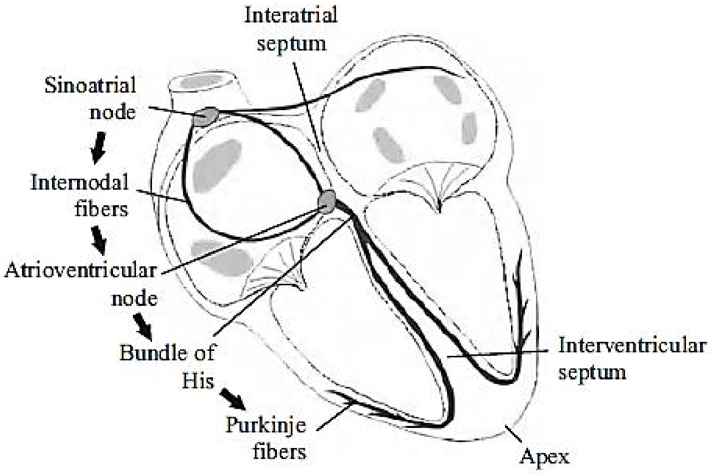


Figure 2 – Cardiac conduction system. *The coordinated contraction of cardiac muscle fibers is controlled by a system of specialized cardiac muscle cells that form conductive pathways through the heart. [9]*

2.1.3 The Cardiac Cycle

Because the left and the right side of the heart contract simultaneously, the cardiac cycle can be separated into two main phases, systole and diastole. Systole is the ventricular contraction phase, during which the atria are relaxed and passively refill. Diastole is the ventricular relaxation phase, during which the atria contract to assist with ventricular filling. Each heartbeat is triggered by an action potential from the sinoatrial node, that causes the two atria to depolarize. They contract, forcing the blood to fill the relaxed ventricle. The atrioventricular valves are open to allow the blood to enter the relaxed ventricle, the semilunar valves are still closed. As the excitation passes the AV node and propagates down the Bundle of His and Purkinje fibers, the ventricle starts to contract, the ventricular pressure exceeds the atrial pressure, causing the atrioventricular valves to close (time instance A in figure 3A). The ventricular pressure continues to increase, but as both sets of valves are closed, the volume remains constant. This first phase of systole is called isovolumetric contraction phase. As soon as the ventricular pressure exceeds the pressure inside the respective artery (approx. 10 mmHg for the pulmonary artery, 80 mmHg for the aorta) the semilunar valves open. The ventricular ejection phase starts, during which about two thirds of the blood contained in each ventricle is ejected. As the ventricle begins to relax, the intraventricular pressure decreases, falling below the arterial pressure. The semilunar valves close soon after, once the flow reaches zero as they are mechanically closed by a slight backflow in blood that fills their cusps. This causes a temporary drop in aortic pressure known as the dicrotic notch (see time instance C in figure 3A) and marks the end of systole. The ventricular pressure continues to drop until the ventricular pressure falls below the atrial pressure, the atrioventricular valves open and ventricular filling starts again (time instant D in figure 3A).

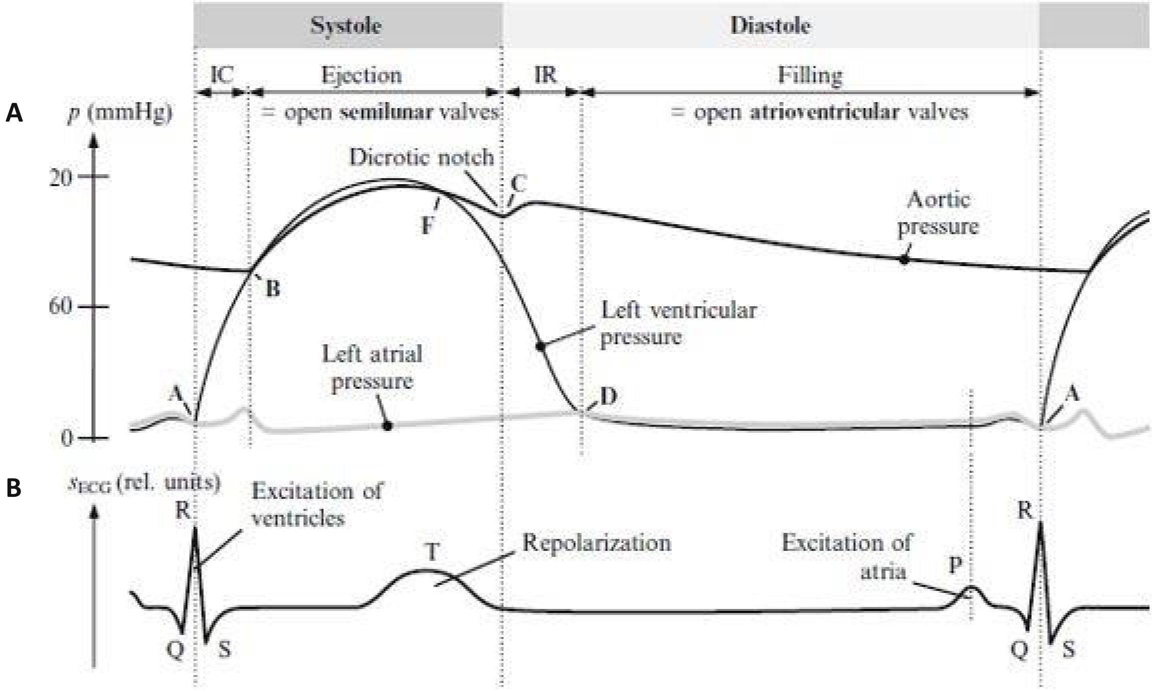


Figure 3 – Mechanical and electric activity of the heart during the cardiac cycle. *A) Left and right ventricular pressures and aortic pressure, results of the mechanical pumping activity of the heart. B) Einthoven I ECG signal, showing the depolarization and repolarization atria and ventricles that result in muscle contractions and relaxations causing the pressure changes shown in A. Adapted from [9]*

2.1.4 Generation of the ECG

Each mechanical pumping action performed by the heart during the cardiac cycle, corresponds to an electrical depolarization of the myocardial tissue in the respective area that triggers the necessary muscle contraction. This depolarization spreads through the muscle gradually, resulting an electrical potential difference of about 170 mV between the depolarized regions and regions that are still at their resting potential. This potential difference creates an electric field, that can be measured at the body's surface using skin electrodes. Because the mechanical and electric activity of the heart are so closely related, this electrocardiogram (ECG), contains a lot of information about the heart's function, making it an indispensable tool in clinical diagnostics.

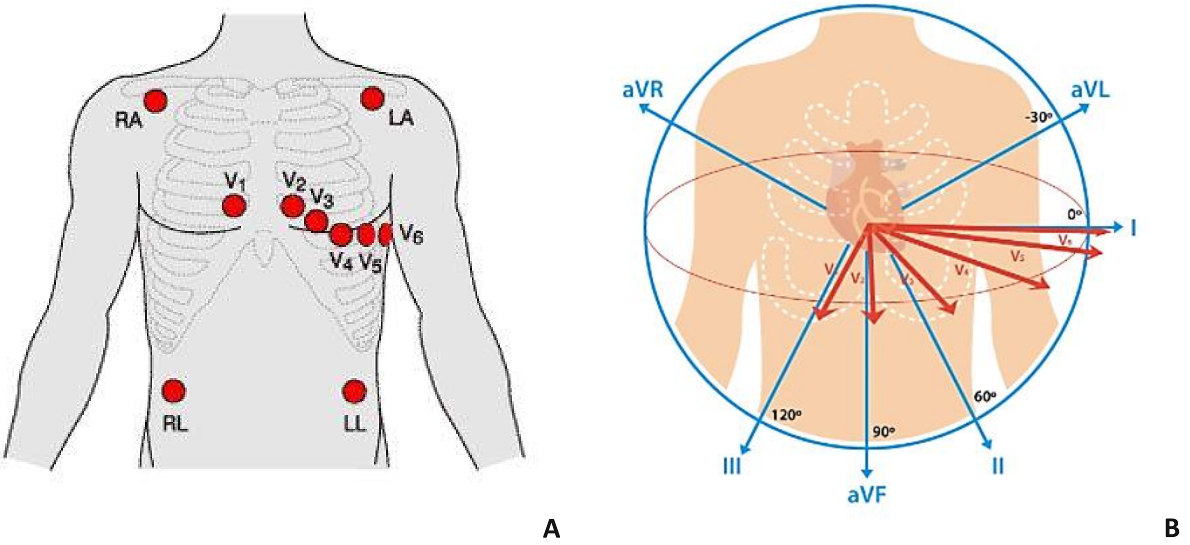


Figure 4 – ECG electrode placements and viewing angles of the 12 lead ECG. A) Standard positions for the 10 skin electrodes used in a clinical 12 lead ECG. Right arm (RA), left arm (LA), right leg (RL) and left leg (LL) electrodes may also be placed on the respective extremity [13]. B) ECG leads and their respective viewing angles in the frontal (blue) and horizontal plane (red) [14].

For a simple ECG tracing, three electrodes, one on each hand and a ground electrode on the right leg, are sufficient. The potential difference between left and right hand is measured over time, resulting in a characteristic waveform with an amplitude in the millivolt range, see figure 3B. This configuration is known as Einthoven I lead. The waves and peaks of the ECG signal correspond to specific phases in the cardiac cycle. After a heartbeat is triggered by the sinus node, the excitation spreads through the atria, which is visible in the ECG as the P wave. Once the atria are fully excited, the signal returns to its zero baseline until the excitation passes the atrioventricular node. Measuring this PQ interval can be used to determine the time delay induced by the atrioventricular node. Once the excitation passes the atrioventricular node, it quickly spreads through the ventricles resulting in the most prominent peak in the ECG, the QRS complex. Once the ventricles are fully excited, the signal returns to zero. Ventricular repolarization then causes another upward flexion in the ECG, the T wave. Atrial repolarization is not visible in the ECG as it occurs while the ventricular excitation spreads and is thus obscured by the more prominent QRS complex. The sign of waves and peaks and the exact shape of the waveform depends on the direction the excitation spreads in, in relation to the axis between the active electrodes.

Using several electrodes in different configurations makes it possible to view the electrical activity of the heart from multiple angles. These configurations are called ECG leads. A 12 lead ECG uses 10 skin electrodes in standardized positions on the human body to record cardiac excitation patterns from 12 different angles, see figure 4. The first six leads, or limb leads, use only three active electrodes, the right arm (RA), left arm (LA) and the left leg (LL) electrodes, which may either be placed on the respective extremity or in the locations shown in figure 4A. The right leg is used as a ground electrode in all 12 leads. Einthoven I, II and III are the potential differences between RA and LA, LL and RA and LL and LA electrodes, respectively. Goldberger's leads aVR, aVL and aVF measure the potential differences between one active electrode, RA, LA and LL, respectively, and the average of the other two. Einthoven and Goldberger's leads view the electric activity of the heart in the frontal plane. To also get information in the horizontal plane, a theoretical reference point, Wilson's central terminal, is used. Located in the center of the triangle formed by Einthoven's leads, which should approximately be in the center of the thorax, its potential is calculated by averaging the potentials of the three limb electrodes RA, LA and LL. The six chest leads V1 to V6 are the potential differences between each of the chest electrodes and Wilson's central terminal.

2.2 Atrial Fibrillation

Atrial fibrillation (AF) is cardiac arrythmia, during which the regular physiologic excitation patterns in the atria are replaced by diffuse and chaotic ones. This results in an irregularity of the heart rate, as well as symptoms such as chest pain, palpitations, shortness of breath and fatigue. AF can also be asymptomatic in some cases, or just accompanied by a general feeling of illness, which can be overlooked or attributed to other, known conditions [6]. Early diagnosis and treatment of AF is important since AF can be an indicator of other cardiac morbidities such as heart failure, as well as aggravate existing conditions of heart failure and increase the risk of thromboembolism such as stroke [6, 7].

AF is a common diagnosis in dialysis patients [7]. Although the exact reason is unknown, patients undergoing dialysis may be particularly prone to AF, because the periodical shifts in fluid and electrolyte levels, which build up in between dialysis sessions and are then rapidly removed during the treatment, strain the heart [7]. Recent evidence suggests that the treatment itself may also be a trigger for paroxysmal AF [2]. A 2011 study [7] showed age as the predominant risk factor for AF in dialysis patient. Compared with otherwise similar patients below 45 years of age, patients 85 or older had an almost 7-fold higher prevalence of AF. The same study showed that AF is more prevalent in men than women and dialysis vintage is related to a 2% increase in AF risk per year [7].

AF can be diagnosed from a single-lead ECG, by the absence of p-waves and the presence of an irregular ventricular rhythm without recurring pattern [6]. These changes can either be detected manually by a qualified physician, or using automated algorithms, that detect the peaks and waves of the ECG. Diagnosis by a physician's examination of ECG has the obvious disadvantage of straining the hospitals limited resources of qualified personnel. Especially in the case of paroxysmal (intermittent) AF, where longer recordings are needed, to include a period where AF is present, manual examination is not feasible. Automatically detecting and quantifying peaks and intervals is a good alternative, however the reliability of such algorithms is severely impaired when the signal quality is low, or the ECG is distorted by medications or other interventions such as hemodialysis [2]. Machine learning approaches to AF detection, that analyze the ECG waveform as a whole, have been investigated as well [15–20].

2.3 Symmetric Projection Attractor Reconstruction

Symmetric projection attractor reconstruction (SPAR) is a new method of analyzing an approximately periodic signal. It is well suited for analyzing noisy and strongly nonstationary data, qualifying it for ECG analysis [21]. Because SPAR does not require the detection of specific points on the ECG other than to determine the average heart rate, it may be the solution to overcoming the challenges of ECG monitoring during hemodialysis. Combined with machine learning techniques it may also be the key to learning more about ECG waveform morphology in general and especially when it comes to hemodialysis, since it has the advantage of analyzing the ECG waveform as a whole.

While the use of SPAR in biosignal analysis is in its early stages, attractor reconstruction itself is a well-established methodology in mathematics for dynamic systems, dating back to the works of Edward Lorenz in 1963 [22]. Studying atmospheric dynamics, Lorenz demonstrated, that recognizing patterns or behaviors in seemingly chaotic time series data, can be facilitated by visualizing the data in three-dimensional (3D) phase space. In his case, this meant plotting each of his three convection variables on one coordinate axis instead of visualizing them separately as time series. This allowed him to analyze changes in weather patterns. In 1981, Floris Takens [23] expanded the attractor to cases where a dynamic system is only described by one variable. His solution was the use of delay coordinates, N equally spaced points along every cycle of an approximately periodic signal, to form an attractor in N-dimensional phase space. SPAR uses this delay coordinated method to generate attractors and then adds an additional step of projecting the resulting attractor to a rotationally symmetric, two-dimensional (2D) image [24].

2.3.1 Generation and Projection of the Attractor

The simplest ECG attractor can be generated using $N = 3$ points. These are spaced at a distance of τ, which is calculated from the period T of the original signal, see formula 1. This period is determined as the mean r-interval (interval between adjacent r-peaks of the ECG) in each window. The time series of the three attractor coordinates, $x(t)$, $y(t)$ and $z(t)$ can be computed using equations 2 and 3 below. Plotting x over y over z results in a 3D image of multiple overlapping loops, one for each cycle in the signal, see figure 5A. In this attractor, the information contained in the (possibly lengthy) time series of the original signal is bound in 3D phase space and variation between the cycles and other aspects of the waveform are highlighted. The interpretability of this attractor can however be improved further by removing baseline variation in the signal, caused by e.g., respiration or motion. In the original signal, baseline variation can be understood as a vertical translation of all three points. In the bounded phase space of the attractor, this corresponds to variation in the direction of the vector (1,1,1). Projecting the attractor onto a plane orthogonal to this vector, removes baseline variation from the attractor and reduces the attractor's dimensionality to 2, which simplifies visualization, see figure 5B. Because the attractor is made up of multiple, partially overlapping lines, deriving a density and visualizing it using a color scale, can add additional detail to the attractor image by highlighting areas of the attractor that are frequented more often, see figure 5C. [21, 24, 25]

$$\tau = T/3 \tag{1}$$

$$y\,(t) = x(t-\tau) \tag{2}$$

$$z\,(t) = x(t-2\tau) \tag{3}$$

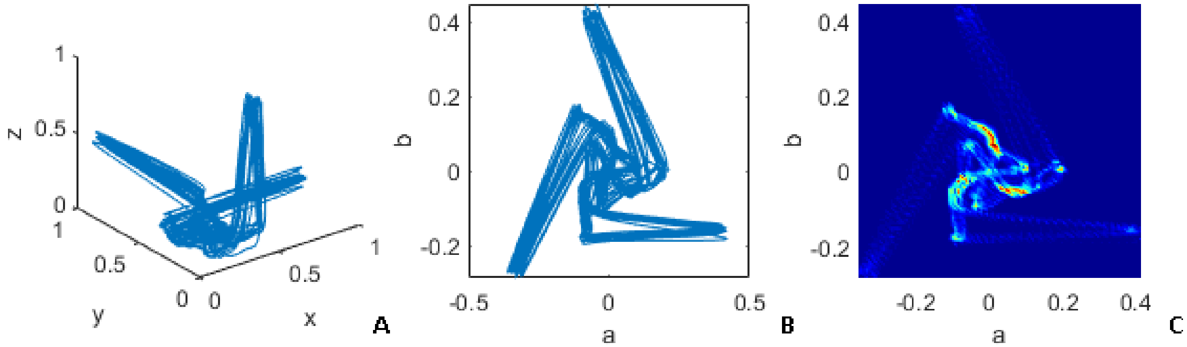


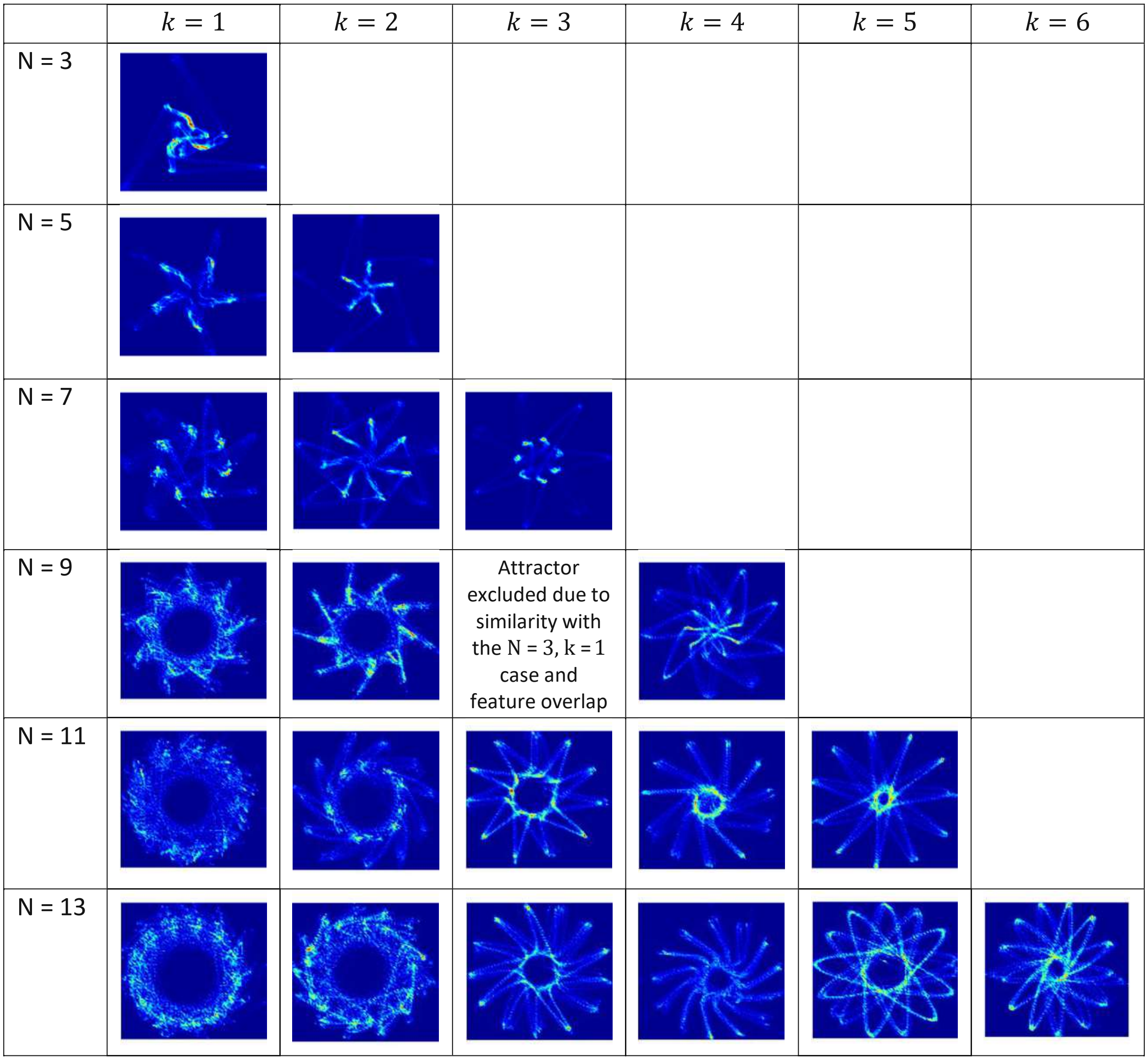
Figure 5 – Example of a three-point ECG attractor and its 2D projection with and without density. *A) 3D attractor generated using delay coordinates, evenly spaced along the signals period, B) 2D projection of the attractor onto a plane orthogonal to vector (1,1,1), C) 2D attractor with added density.*

2.3.2 Extension to Higher Dimensions

Attractor generation can be extended to more points and higher dimensions [24–26]. For higher dimensional attractors however, projection to a 2D image becomes a bit more complex. A continuous signal of period T can be embedded into $N \geq 3$ dimensions using a time delay $\tau = T/N$ using the coordinates $x_{N,j}$, see formula 4 [25]. It can be shown that this N-dimensional attractor has $(N-1)/2$ 2D projections that are rotationally symmetric about the origin. The new coordinates of these projections $a_{N,k}$ and $b_{N,k}$ can be computed using formula 5, $k = 1, \ldots, (N-1)/2$ [25]. Embeddings using odd numbers of points N have proven most useful, because they lead to less overlap of ECG features [24]. Figure 6 shows the 20 symmetric attractor projections of one ECG lead (Einthoven I) embedded with all odd numbers of points $N = 3, 5, 7, \ldots, 13$.

$$x_{\mathrm{N,j}}(t) = x(t - j\tau), \qquad j = 0, \ldots, N-1 \tag{4} [25]$$

$$a_{\mathrm{N,k}} = \frac{1}{\sqrt{N}} \sum_{j=0}^{N-1} \cos\left(\frac{2\pi jk}{N}\right) x_{\mathrm{N,j}}(t)\,, \qquad b_{\mathrm{N,k}} = -\frac{1}{\sqrt{N}} \sum_{j=0}^{N-1} sin\left(\frac{2\pi jk}{N}\right) x_{\mathrm{N,j}}(t) \qquad \text{(5) [25]}$$



Attractor excluded due to similarity with the N = 3, k = 1 case and feature overlap

Figure 6 – Higher dimensional ECG attractors. *20 symmetric attractor projections generated from a single ECG lead (Einthoven I), using all odd numbers of points* $N = 3, 5, \ldots\ 13$*, to embed the signal in* N *-dimensional phase space. Each* N*-dimensional attractor has* $(N-1)/2$ *rotationally symmetric 2D projections, numbered* $k = 1, \ldots, (N-1)/2$*. The attractor generated using* $N = 9, k = 3$ *is excluded due to its similarity with the* $N = 3, k = 1$ *case and overlap of features.*

2.4 *K*-Nearest Neighbor Classification

Classification is a common machine learning application, where an algorithm is trained on a set of labeled training data to automatically allocate unknown samples into one of two or more groups. Decisions are made based on a set of predictors also referred to as features. There is a variety of different algorithms available, that each have their own strengths and drawbacks and may be more or less successful depending on the dataset and application in question. *k*-nearest neighbor is one of the oldest and simplest classification algorithms yet continues to be widely used because of its easy implementation and powerful performance. It is a non-parametric classifier which means it is suitable for applications where there is little to no prior knowledge about the data distribution. An unknown sample is classified based on its *n* features, by mapping it into the *n*-dimensional feature space. Its distance to each of the labeled training samples is determined using a predetermined distance metric. The class of the unknown sample is predicted as the most frequent class among the *k* samples with the lowest distance from it, its so called *k* nearest neighbors [27]. The effect of choosing a different number of neighbors *k* is shown in figure 7 [27], using a simplified example with *n* = 2 features, two different classes and Euclidean distance measure. Selecting the optimal number of nearest neighbors and an appropriate distance function is the main challenge in this classification technique. [27]

Classifiers with a higher number of features require more training samples to maintain a sufficient density of samples in the feature space, because the distance between points increases exponentially with the dimensionality of the feature space, making it harder to find nearest neighbors. This problem is known as the “curse of dimensionality” or Hughes effect [28].

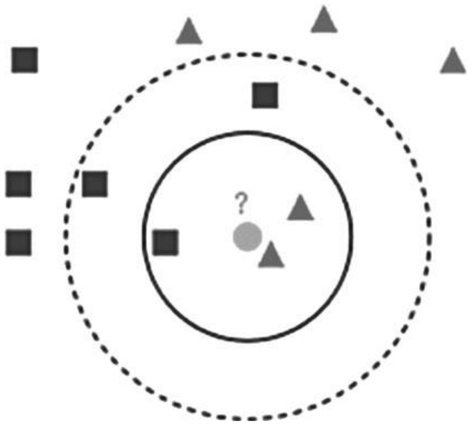


Figure 7 - Example of k-nearest neighbor classification. *Each sample is characterized by two features, its x and y coordinates in the two-dimensional feature space. The central circle represents the unknown sample. Depending on the number of neighbors k = 3 (solid line circle) or k = 5 (dashed line circle), the unknown sample will either be classified as a triangle or a square. A Euclidean distance measure is used in this example. [27]*

3. Methods and Implementation

To establish a reliable automated AF detection algorithm that is successful in ECG data collected during hemodialysis, ECGs with and without AF were sourced from a study on end-stage renal disease (ISAR study) [29] and PhysioNet, an open source archive of physiologic signals [30]. Appropriate records were selected, and 30s ECG samples extracted. Using SPAR, 20 attractors were generated from each sample. After experimentation with different quantification approaches, attractor projections were quantified by computing three density distributions for each of them. These attractor densities were then directly used as feature inputs into 60 *k*-nearest neighbor classifiers per model. Models were trained on six different training set compositions, containing ECGs recorded from six timepoints during or after dialysis and/or ECGs unrelated to dialysis. They were then tested on different test sets to analyze their accuracy when classifying unseen test records. The main steps of this workflow are summarized in Figure 8 and explained in more detail in the following chapters. Signal processing, feature extraction, model training and model testing was done in MATLAB® (R2018b, The MathWorks, Inc., Natick, Massachusetts, USA).

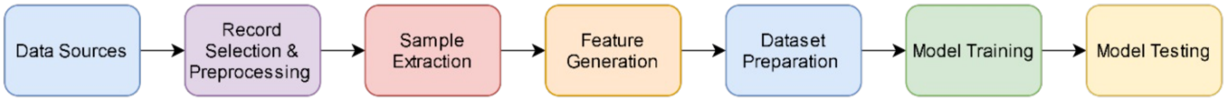


Figure 8 – Workflow summary. *Main steps of selecting and preparing the training and test data using ECGs from different sources and using them to train and test AF detection models.*

3.1 Data sources

ECGs from multiple sources were used in this thesis. ECG samples during or after dialysis were taken from 24h 12-channel ECG recordings, collected during an observational cohort study on risk stratification in end-stage renal disease (ISAR study) [29]. ECGs unrelated to hemodialysis or renal failure were taken from the open source archive PhysioNet [30]. Three different PhysioNet databases were used. For an overview of the data sources, see figure 9.

In cooperation with the computing in cardiology conference, PhysioNet issues annual challenges, where participants are encouraged to “tackle clinically interesting questions that are either unsolved or not well-solved” [31]. In these challenges, researchers are asked to establish working, open-source algorithms, using training sets, provided by PhysioNet. The submitted algorithms are scored based on their performance on a hidden test set. The 2017 PhysioNet/Computing in Cardiology Challenge [32] was titled “AF Classification from a Short Single Lead ECG Recording” and challenged the participants do develop algorithms that automatically sort 10-60 s long ECGs into one of four groups, normal sinus rhythm, AF, arrythmia other than AF or too noisy to be classified. The training set provided for this challenge contains 8,528 single lead ECG recordings from those four groups, sampled at 300 Hz and between 9 s to just over 60 s in length. For a more detailed data profile, see table 1. The “normal” and “AF” groups of this data set were used in this thesis.

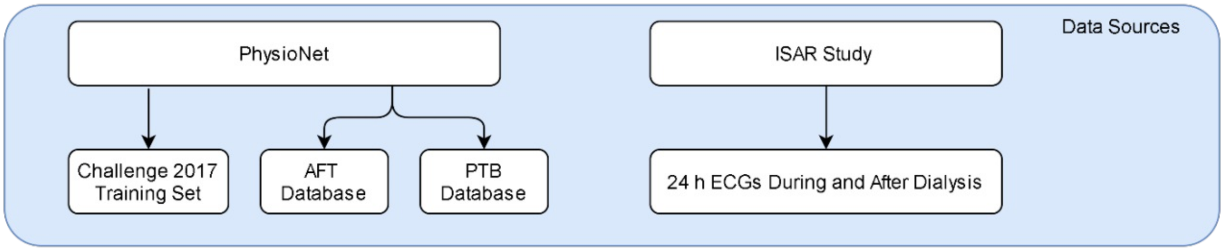

Data Sources
PhysioNet
ISAR Study
Challenge 2017 Training Set
AFT Database
PTB Database
24 h ECGs During and After Dialysis

Figure 9 – Data sources for training and testing the AF detection models. *Samples during or after dialysis were taken from 24h ECG recordings, collected during a study on risk stratification in end-stage renal disease (ISAR study). ECG samples unrelated to dialysis were taken from three different databases freely available via PhysioNet.*

Type	Number of Records	Time length (s)				
		Mean	SD	Max	Median	Min
Normal	5154	31.9	10.0	61.0	30	9.0
AF	771	31.6	12.5	60	30	10.0
Other rhythm	2557	34.1	11.8	60.9	30	9.1
Noisy	46	27.1	9.0	60	30	10.2
Total	**8528**	**32.5**	**10.9**	**61.0**	**30**	**9.0**

Table 1 – Data profile of the PhysioNet Challenge 2017 training set. *Number of available ECG signals with normal sinus rhythm, AF, arrhythmias other than AF and records that are too noisy to be classified in the training set provided for the 2017 PhysioNet/ Computing in Cardiology Challenge titled "AF Classification from a Short Single Lead ECG Recording" [33]*

The second PhysioNet database used was the Atrial Fibrillation Termination (AFT) database [34]. This database provides a total of 80 one-minute-long ECG samples with AF sampled at 128 Hz, that were also used in a challenge, the PhysioNet and Computers in Cardiology 2004 Challenge "Spontaneous Termination of Atrial Fibrillation". The goal of this challenge was to establish an algorithm that can predict if or when an episode of AF is going to terminate, by automatically sorting the samples into one of three groups, non-terminating AF, AF that will terminate within one minute after the end of the record and AF that will terminate within one second after the end of the record [34]. Because AF is present throughout each record, all 80 samples can be used as AF samples when training or testing an AF detection algorithm.

Additional control records, i.e. ECGs without AF, were taken from the PTB (Physikalisch-Technische Bundesanstalt) Diagnostic ECG Database [35], which is also available on PhysioNet. This database contains 549 records from 290 subjects. Most records include a detailed clinical summary about the subject's diagnosis and treatment plans, along with a 15 channel ECG (12 normal leads, see section 2.1.4, plus the 3 Frank lead ECGs, vx, vy, vz), sampled at 1 kHz. All records are multiple minutes in length. Included in this database are 80 ECGs from 52 healthy controls. These were used as healthy controls for the AF samples taken from the AFT database.

In total, 381 24h ECGs from subjects with end-stage renal disease were available from the ISAR study. This included ECG recordings of various quality, from subjects with or without pacemakers and different AF diagnoses. Additionally, a variety of other clinical information on each subject was collected during the study. The number of available records from subjects with and without a pacemaker and their AF diagnoses are summarized in Table 2. The ECGs were sampled at 128 Hz.

	Pacemaker	**No Pacemaker**	**Total**
Permanent AF	6	46	52
Paroxysmal/intermittent AF	4	35	39
No AF	17	268	285
Unknown AF Diagnosis	-	5	5
Total	27	354	381

Table 2 – Data profile ISAR study ECGs. *Number of available records from patients with or without a pacemaker and their AF diagnoses.*

3.2 Record Selection, Preprocessing and Sample Extraction

To generate training and test samples from the ECGs in the ISAR study and the three PhysioNet databases, they first had to be imported into MATLAB®. There, records with sufficient quality and length were selected from each data source. The raw signals were filtered if necessary and 30s samples extracted from them. The main steps of record selection, preprocessing and sample extraction for the different datasets is summarized in figure 10.

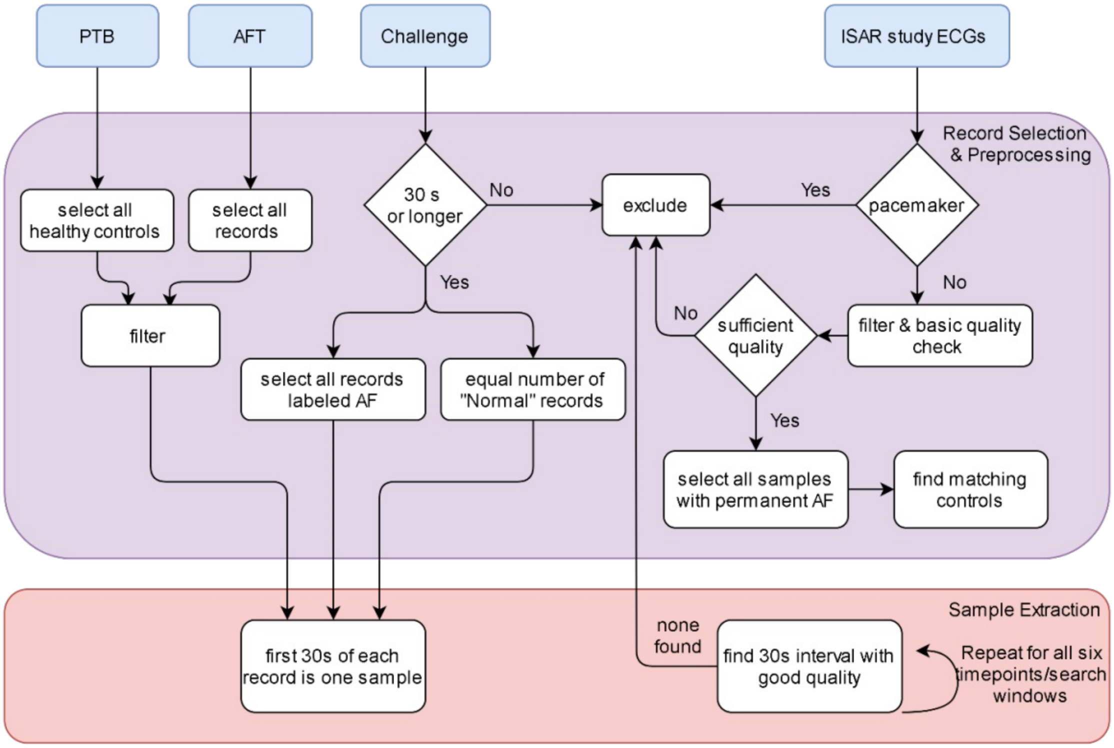

PTB
AFT
Challenge
ISAR study ECGs
Record Selection & Preprocessing
select all healthy controls
select all records
30 s or longer
No
Yes
exclude
Yes
pacemaker
No
filter
select all records labeled AF
equal number of "Normal" records
No
sufficient quality
filter & basic quality check
Yes
select all samples with permanent AF
find matching controls
Sample Extraction
first 30s of each record is one sample
none found
find 30s interval with good quality
Repeat for all six timepoints/search windows

Figure 10 – Preprocessing, record selection and sample extraction. *Overview of the different steps in preparing the ECGs from the different sources for feature generation and subsequent training and testing of different AF detection models.*

3.2.1 PhysioNet ECGs

The ECG data in the PhysioNet Computing in Cardiology Challenge 2017 training set are conveniently provided in *.mat* format, allowing them to be directly loaded into MATLAB®. As they were already bandpass filtered and of good quality, no preprocessing was necessary. Records shorter than 30s were excluded. With fewer available AF than normal records, all AF records and an equal number of records from the “Normal” category were selected. The first 30s of each record were used as one sample.

The ECGs in the AFT database are only available in PhysioNet’s *.dat* format. Using the WaveForm DataBase (WFDB) Toolbox for MATLAB® [36] the first 30s of each samples can however be automatically downloaded from PhysioNet’s servers and loaded into MATLAB® for further analysis. There the signals were filtered using three infinite impulse response, digital filters, designed via the MATLAB® filter designer add-in. High frequency noise was removed using a 6th order Butterworth lowpass filter, with a stopband frequency of 60 Hz and 80 dB stopband attenuation. Baseline wander was removed using a 5th order Butterworth highpass filter with a cutoff frequency of 1 Hz. 50 Hz powerline interference was removed using a 12th order Butterworth bandstop filter, with a lower stopband frequency of 49 Hz and an upper stopband frequency of 51 Hz. Magnitude response estimates for all three filters are depicted in figure 11.

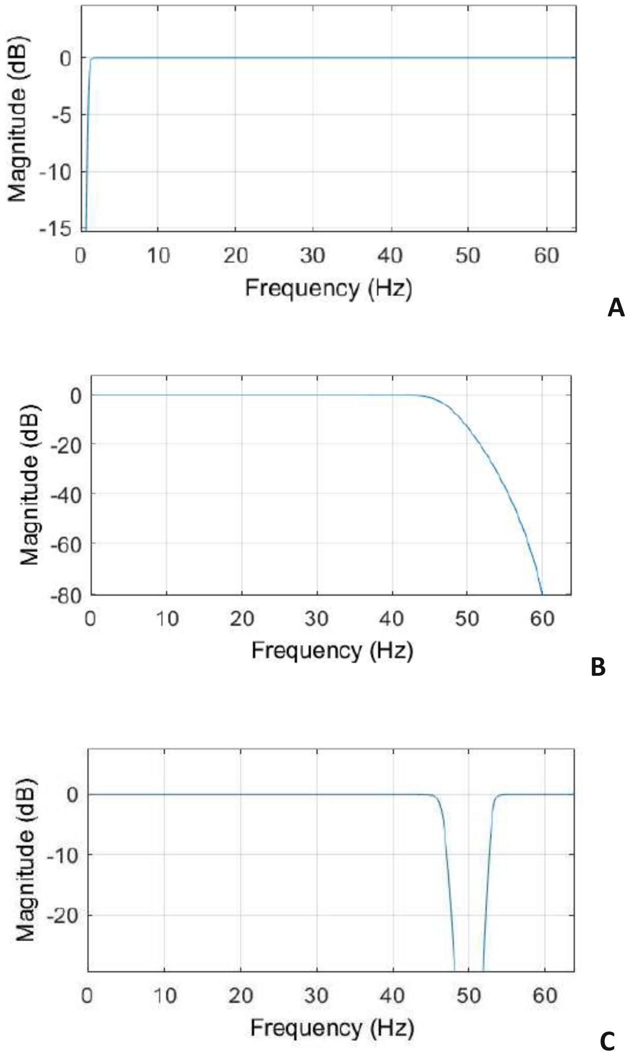


Figure 11 – Filter characteristics for f_s = 128 Hz. *Magnitude responses of the three infinite impulse response filters used to filter the ISAR study ECG data and PhysioNet AFT database ECGs A) 5th order Butterworth highpass B) 6th order Butterworth lowpass C) 12th order Butterworth bandstop*

The control ECGs from the PTB database were also downloaded using the WFDB Toolbox and filtered using three infinite impulse response filters, adapted to the much higher sampling frequency of 1 kHz. High frequency noise was removed using a 9th order Butterworth lowpass filter with a passband frequency of 40 Hz and a stopband frequency of 120 Hz. The stopband attenuation was set to 80 dB. Baseline wander was removed using a 5th order highpass with a cutoff frequency of 1 Hz. A 12th order bandstop with a lower stopband frequency of 49 Hz, upper stopband frequency of 51 Hz and 60 dB stopband attenuation was used to remove 50 Hz powerline interference. These filters' estimated magnitude responses are shown in figure 12. The first 30s of each filtered record was used as one sample.

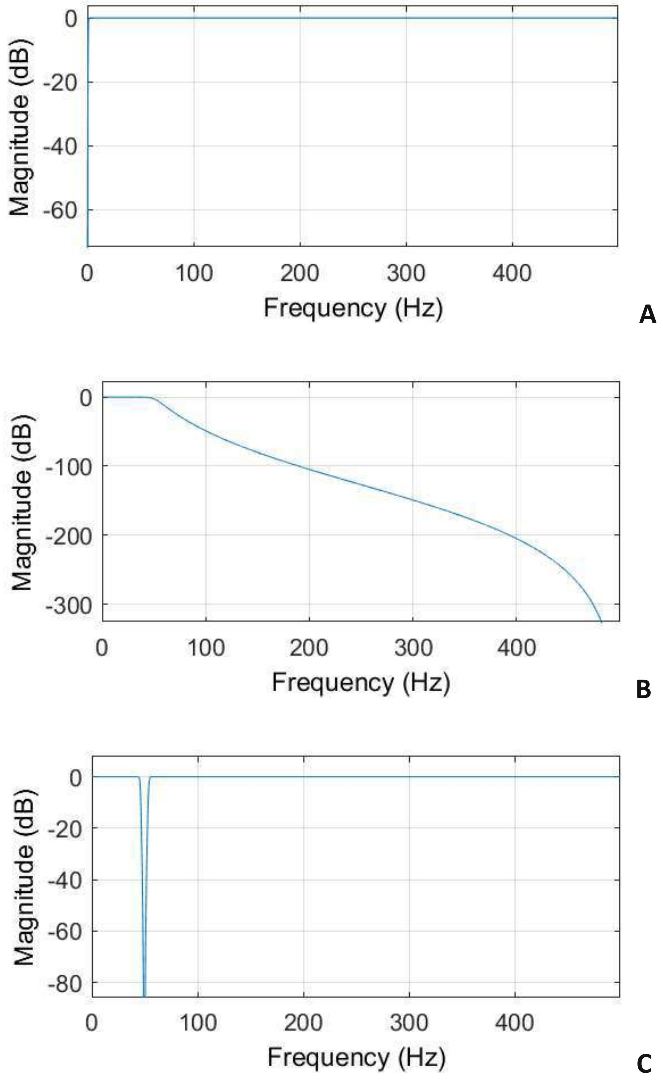


Figure 12 – Filter Characteristics for f_s = 1 kHz. *Magnitude responses of the three infinite impulse response filters used to filter the PhysioNet PTB database ECGs A) 5th order Butterworth highpass B) 9th order Butterworth lowpass C) 12th order Butterworth bandstop*

3.2.2 ISAR Study ECGs

The ISAR study ECGs included subjects with a pacemaker, which depending on the exact type and setting, can affect the ECG in various ways. Accounting for these effects would exceed the span of this thesis, so records from subjects with a pacemaker were excluded from further analysis. The remaining records were filtered using the same infinite impulse response filters as used on the PhysioNet AFT database ECGs, as both are sampled at 128 Hz. The magnitude response estimates of these filters are shown in figure 11. On the filtered ECGs, an initial quality check was performed. Because samples are taken from six different timepoints within each ECG, signal quality was analyzed within the search windows around these timepoints, see figure 13. The quality check was performed by detecting R-peaks in sliding 30s windows within the first 30min of each search window, using a validated r-peak detection algorithm [37]. If no 30s window with more than 20 detected and valid R-peaks is found at any of the six timepoints, the record was excluded.

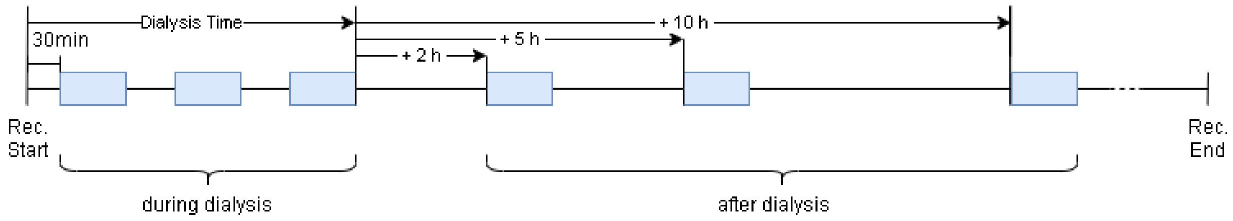


Figure 13 - Search windows for sample extraction from the ISAR study ECGs. *30s ECG samples were taken from the ISAR study ECGs at six different timepoints that are defined with respect to the beginning and end times of the dialysis session. Because the ECG quality may not be high enough in any arbitrarily set 30s window, a one-hour search window was placed around each timepoint (blue box) and a sample of sufficient quality was manually selected within that interval.*

To ensure that AF is present in every sample taken, only records from patients with permanent AF were used as AF samples, subjects with paroxysmal/intermittent AF were excluded. These were again fewer than the available controls (without AF). Because a variety of information is available on every subject in the ISAR study, it was possible to compare sample characteristics from both groups, that may be confounding factors in the analysis. The three characteristics considered in the match were gender, age and dialysis vintage, i.e., for how many months the patient has been undergoing regular dialysis for. Based on these factors, an equal number of matching controls were selected from the available records without AF. Controls were selected by going through one AF record after another, finding control subjects of the same gender and within ±10 years in age. From this group, the subject closest in dialysis vintage was selected as a matched control.

From the AF records and their matched controls, 30s samples had to be extracted at appropriate timepoints. Every recording is approx. 25h in length, with the start of the dialytic treatment occurring at some point during the first 30 min of the recording. As the ECG waveform is known to change during dialysis, samples were extracted at six different timepoints during and after dialysis. The duration of the dialysis treatment, i.e., the dialysis time, was noted for most subjects in the study, for the others it was estimated as the mean dialysis time of all other subjects. To ensure every sample is of sufficient quality to be classified, a one-hour search window was placed around each of the six timepoints, i.e. the beginning, middle and end of dialysis, as well as two, five and ten hours after the treatment, see Figure 13. Samples were then selected by manually sliding a 30s sampling window through the search window, until a sample of sufficient quality was found.

3.3 Attractor Reconstruction and Quantification

Attractor reconstruction was used to generate features from the prepared ECG samples. The workflow of transforming each 30s ECG sample into a 20-by-3-by-64 matrix of attractor features, is summarized in figure 14, and follows the methodology used by in [24]. The amplitude of an ECG signal depends on the electrical impedance between heart and electrode [9]. This impedance in turn is affected by parameters such as skin impedance and body fat. Because these factors are not relevant for this analysis, all samples were normalized to a range of zero to one before computing the attractors. The same R-peak detection algorithm as used previously [37], was used to determine the mean heart rate. Attractors were then generated using all odd numbers of points $N = 3,\ 5, \dots, 13$ following the SPAR methodology described in [24] and [25]. For more details, see section 2.3. For each attractor, the $k = 1, \dots, (N-1)/2$ two-dimensional projections were calculated using formula 6 in section 2.3.2. The $N = 9, k = 3$ case was excluded from further evaluation, due to its similarity to the $N = 3,\ k = 1$ projection and overlap of features. This resulted in a total of 20 attractor projections per sample.

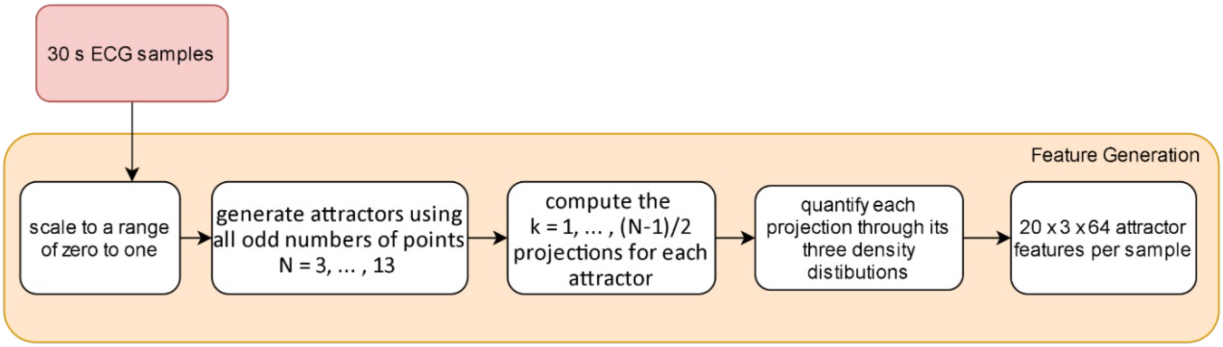

30 s ECG samples
Feature Generation
scale to a range of zero to one
generate attractors using all odd numbers of points N = 3, ... , 13
compute the k = 1, ... , (N-1)/2 projections for each attractor
quantify each projection through its three density distibutions
20 x 3 x 64 attractor features per sample

Figure 14 – Feature generation using attractor reconstruction and quantification. *20 attractor projections are generated from each 30s long ECG sample and quantified by computing its three density distributions. These attractor features can then be used as inputs to an AF detection model.*

3.3.1 Density Distributions

The projections were then quantified by transforming into polar coordinates and computing three summary measures of the attractor shape and density, namely angular density, radial density and attractor outline, see figure 15. These were calculated as follows:

1) To get the angular density, the angular range of 0 to 2π is split into 64 bins, see figure 15A. The number of data points in each bin is calculated and divided by the total number of datapoints in the sample to get a density measure.
2) The radial density is computed by splitting the range between the center of the attractor and its outline into 64 bins, counting the number of datapoints per bin and dividing by the total number of datapoints in the sample, see figure 15B.

3) To compute the attractor outline, the maximal radius in each angular bin is determined, see figure 15C.

Overall, this transforms every ECG sample into 60 (20x3) curves of 64 datapoints each, that describe its waveform morphology and characteristics.

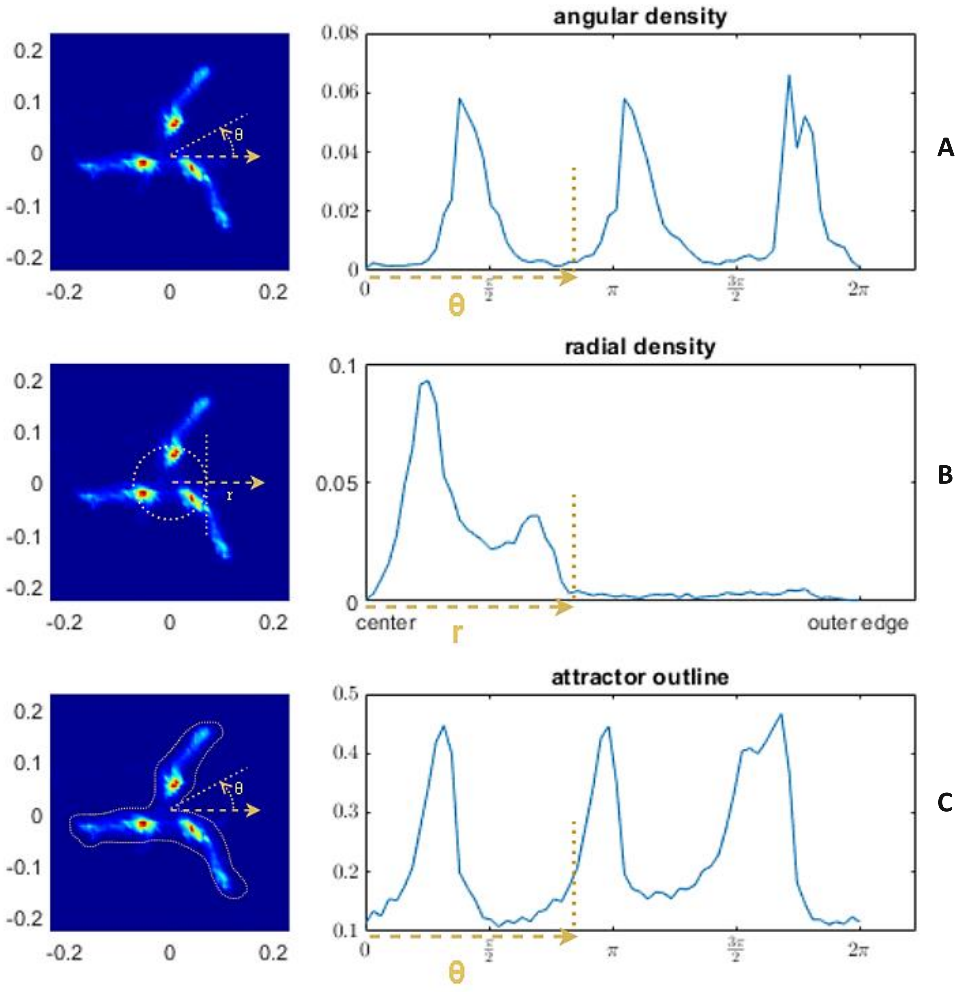


Figure 15 – Attractor quantification via density distributions. *The attractor shown is an N=3 k=1 attractor from a control sample in the PhysioNet dataset. The attractor is quantified by transforming its points into radial coordinates and deriving three density distributions from those. A) the angular density of the attractor is computed by dividing the angular range of the attractor from 0 to 2π into 64 bins and determining the relative number of datapoints per bin. B) the radial density of the attractor is computed by dividing the range between attractor center and outline into 64 bins and then counting the number of datapoints in each bin, divided by the total number of datapoints in the attractor. C) the attractor outline is defined as the maximal radius in each angular bin. Based on [24]*

3.3.2 Different Quantification Approaches

To be able to train an AF detection model, features had to be extracted from the generated attractor projections and/or their densities. Starting with the $N = 3, k = 1$ projections of healthy samples, several quantification approaches were tested. In a first attempt, the attractors were sorted into nine main types, based on a visual inspection of their shape, see figure 16. Using its three density distributions, the attractor type of an unknown attractor projection was to be determined automatically and features extracted accordingly. Unfortunately, this proved to be impossible, since assigning the attractors to types was often ambiguous, since many of them combined characteristics of multiple types. Once the attractor generation was extended to higher dimensions which increased the number of attractors from one to 20 per sample, this quantification approach was definitively out of the question.

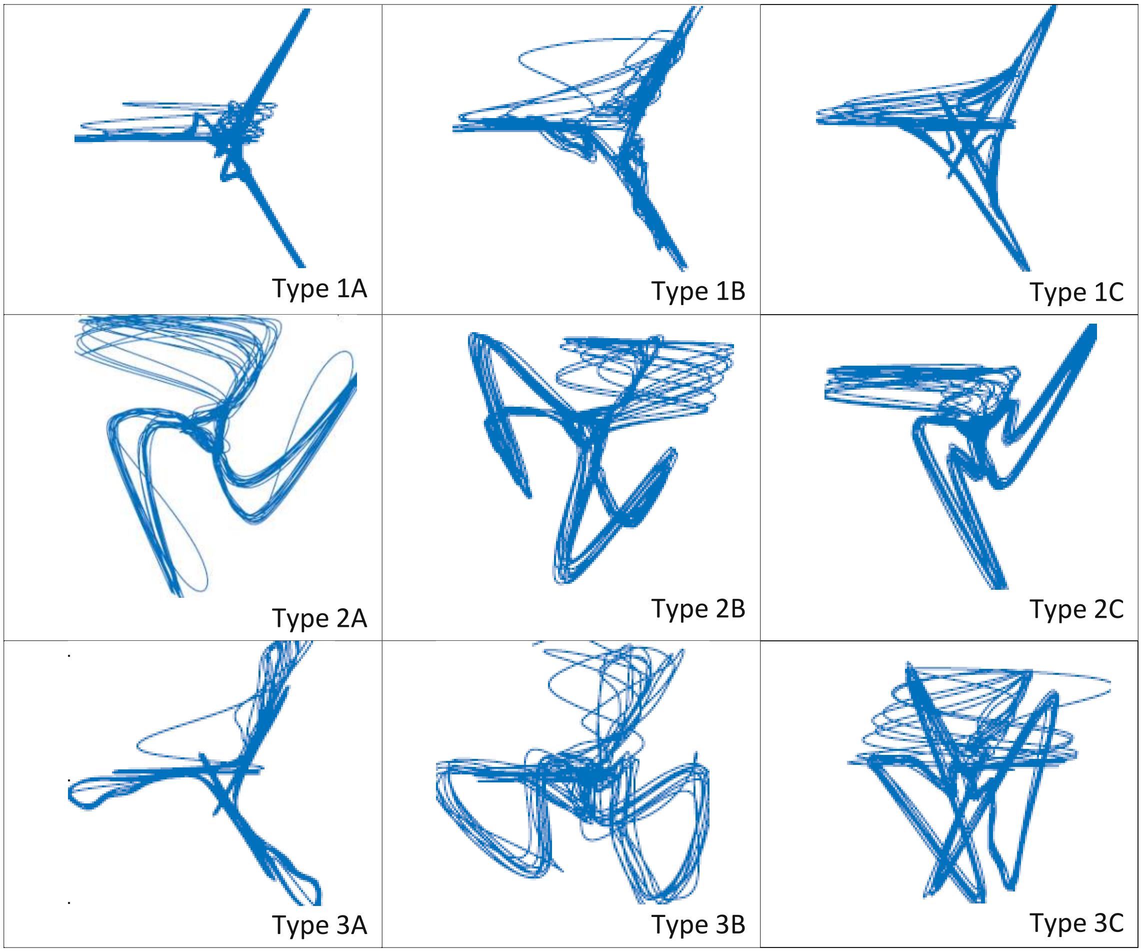


Figure 16 – Attractor types. *Examples of the nine typical shapes of N = 3, k = 1, attractor projections shapes generated from healthy ECG samples.*

A second quantification attempt was to train a shape model on the attractors. For this approach, a mean attractor shape was calculated from a set of healthy $N = 3, k = 1$ attractor samples. Principal component analysis was then used to analyze the variation in shape between them and describe it in terms of eigenvectors and eigenvalues of the covariance matrix. The shape of an unknown attractor could then be characterized by fitting the shape model, i.e., placing the mean attractor shape on top and varying it in the direction of the determined eigenvectors, until the distance between the fitted model and the attractor points reaches a minimum. The scaling values of the fitted shape for the first few principal components could then be used as features that describe the attractor's shape. Unfortunately, because the lines of the attractor overlap so frequently, the shape model was impossible to fit and the idea had to be given up.

A third attempt at quantification consisted of quantifying the attractor by calculating features that describe its three density distributions. A total of 18 features was selected that summarize the shape and symmetry of an $N = 3, k = 1$ attractor projections' density distributions, see table 3. These features showed some differences between samples with and without AF based on visual inspection using boxplots and when training a support vector machine for binary classification on them. Despite standardizing the input features, decision making in the classifier was however most heavily based on features eight and 15, the two features that count peaks in the angular density and attractor outline. Since these two depend on the peak prominence threshold set empirically, based on the visual inspection of a small number of examples, these features are rather unreliable and to a certain degree subjective.

Feature	Density distribution	Characteristic
1	radial density	maximum
2	radial density	location of the maximum
3	radial density	width of the highest peak at half prominence
4	angular density	maximum
5	angular density	location of the maximum
6	angular density	width of the highest peak
7	angular density	width of the widest peak
8	angular density	number of peaks with peak prominence > 0.02
9	angular density	Standard deviation (SD) of the three highest peaks' heights
10	angular density	SD of the three highest peaks' width at half prominence
11	attractor outline	maximum
12	attractor outline	location of the maximum
13	attractor outline	width of the highest peak
14	attractor outline	width of the widest peak
15	attractor outline	number of peaks with peak prominence > 0.2
16	attractor outline	SD of the three highest peaks' heights
17	attractor outline	SD of the three highest peaks' width at half prominence
18	mean absolute difference between angular density and outline	

Table 3 – Features selected for quantifying the attractor via its density distributions. *18 features calculated form the three density distributions, radial density, angular density and attractor outline of an attractor.*

Finally, the decision was made to follow the methodology used in [24] and directly use the density curves as feature inputs for the AF detection model. This approach outperformed the feature extraction from the density curves in classifying $N = 3, k = 1$ attractor projections as AF or control (no AF) using support a vector machine and *k*-nearest neighbor classification and was easier to extend to higher dimensions, making it a clear favorite.

3.4 Dataset Preparation and Training Set Compositions

To compile datasets with different characteristics for training and testing the models, the available samples were split into eight groups, the challenge group, AFT/PTB group and the six dialysis timepoints. Each contained an approximately equal number of AF and control records, with slight deviations caused by the dialysis timepoints where no sample of sufficient quality was found. Every group was then randomly split into a training and test portion using a 30% holdout, while maintaining the even ratio of AF and control samples. This was done using the built-in MATLAB® function *cvpartition*. The training and test subsets were then combined into six different training sets and six test sets, respectively. Which groups are included in which set is illustrated in figure 17. The sample sizes of each dataset and the size of its training and test portion were documented, along with the percentages of AF versus control records for each of them.

By using the same methodology on these six different training set compositions, six different AF detection models were trained:

1) The challenge set contained only the samples taken from the PhysioNet Challenge 2017 training set and could be used to train a first AF detection model that is unrelated to dialysis. This model served as a benchmark, to on the one hand see how well the SPAR method performs compared to previously published AF detection models working with the raw ECG samples, and on the other hand get an idea of how much more difficult ECGs collected during or after dialysis are to classify.
2) The second training set contained only the training subset of samples taken at the start of dialysis, so thirty minutes after the beginning of the measurement or up to 1h30min into the measurement, depending on the local signal quality. The model trained on this set was intended to investigate, if AF detection using ECGs recorded during dialysis is easier, if all samples are collected at approx. the same timepoint during the treatment. The start of dialysis was chosen as a timepoint because of its practical advantages in future applications, i.e., the ECG measurement can be started when starting the dialysis and deliver quick results. Because the exact timepoint each ECG sample was taken at, with respect to the dialysis, is however unknown, including a wider range of timepoints in the training set may produce a more successful model. This was tested using the third and fourth training sets.
3) The third set contains the training subset of all samples taken during dialysis, so start, middle and end of dialysis.
4) The fourth set contains the training subsets of all six dialysis timepoints.

5) The effect of including samples unrelated to dialysis in the training set was investigated using a fifth, combined training set, including the training subsets of all six timepoints plus the challenge data.
6) Finally, the AFT/PTB training set was used to see how a model performs if the AF and control samples it is trained on, originate from two different databases. Additionally, the AFT/PTB test set was used to test the challenge models accuracy on samples from a different database, to see if the model is overfitted to the challenge data or actually detecting AF.

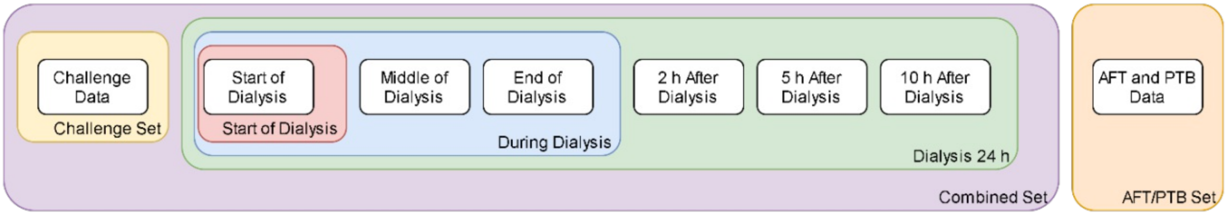


Figure 17 – Compositions of the six datasets. *Each data group (white boxes) is first split into a training and test portion with a 30% holdout. These are then combined into six training sets and six test sets, respectively, as indicated by the colored boxes.*

3.5 Model Training and Testing

Since the decision was made to directly use the density distributions as model inputs, each training set consisted of 64 datapoints times 3 densities times 20 attractor projections per sample. To limit the number of input features, a separate classifier had to be trained for each density distribution. Therefore, 60 individual classifiers were trained on the 64 density features of each attractor density. This was repeated for each of the training sets. The workflow of training and testing the AF detection models is summarized in figure 18.

K-nearest neighbor was used as the classification algorithm. For more information on this classifier, see section 2.4. The classifiers were trained using the built-in MATLAB® function *fitcknn [38]*, with the 64 datapoints of one density distribution from all samples in the dataset as feature inputs. The sample labels (AF or control) were used as label inputs. Using the automatic hyperparameter optimization included in this function, optimal distance measure and number of neighbors were determined for each classifier by minimizing five-fold cross validation loss [38]. The classification error of each classifier was determined via ten-fold cross validation and stored along with the trained classifiers in another 20x3x2 cell array for easy access.

The trained models were tested using the unknown samples in the prepared test sets and comparing the labels predicted by the model to the actual labels. Unknown samples were labeled based on the combined outputs of the highest performing classifiers in the model. Using the function *predict*, each trained classifier in the model was used to compute a posterior probability score for every test sample, based on its corresponding density features. The sample is labeled based on the mean posterior probability of all classifiers in the model that have a cross-validated accuracy of 70% or more. Model performance was assessed using two different metrics, namely classification accuracy and F1 statistic, see section 4.4. To get an overview of the performance of the different models for different datasets, the challenge model and dialysis models, were each tested on all five test set compositions, resulting in a five-by-five table of classification accuracies. The AFT/PTB and challenge model were also tested on both their test sets, results were summarized in a two-by-two table.

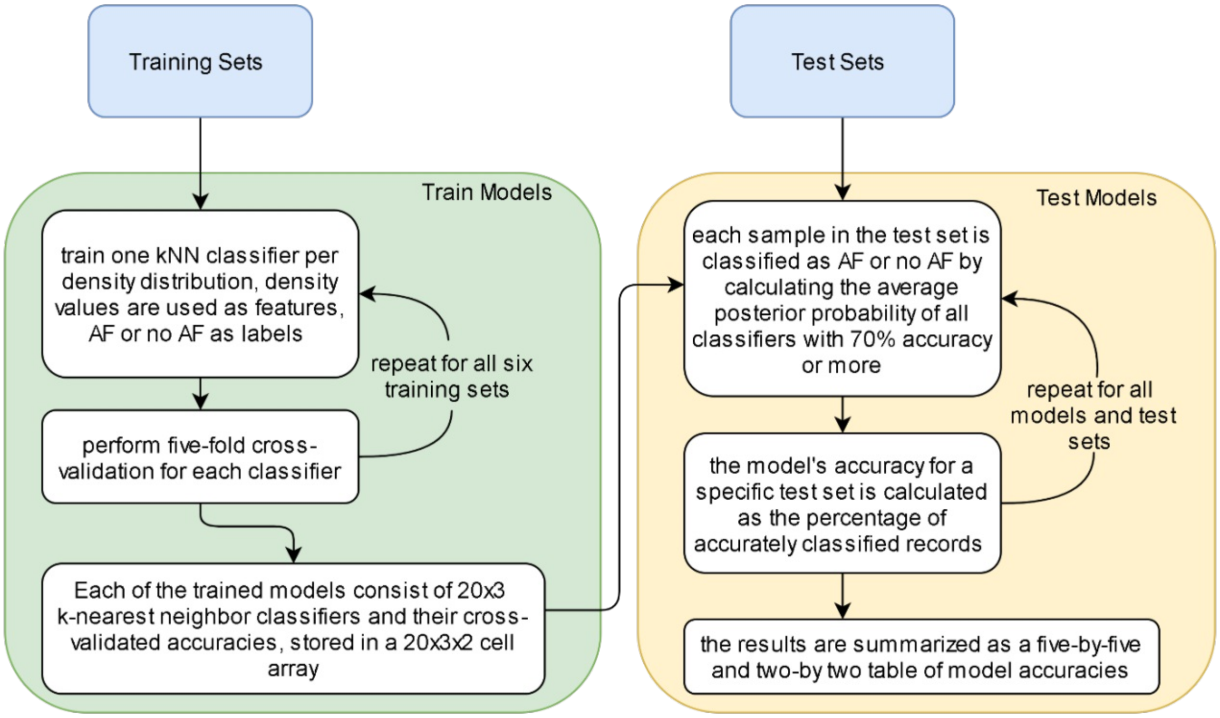

Training Sets
Test Sets
Train Models
train one kNN classifier per density distribution, density values are used as features, AF or no AF as labels
repeat for all six training sets
perform five-fold cross-validation for each classifier
Each of the trained models consist of 20x3 k-nearest neighbor classifiers and their cross-validated accuracies, stored in a 20x3x2 cell array
Test Models
each sample in the test set is classified as AF or no AF by calculating the average posterior probability of all classifiers with 70% accuracy or more
repeat for all models and test sets
the model's accuracy for a specific test set is calculated as the percentage of accurately classified records
the results are summarized as a five-by-five and two-by two table of model accuracies

Figure 18 – Workflow for training and testing the AF detection models. *Using six different training set compositions, six different models were trained. Each model was tested on its corresponding test set as well as other test sets, to understand more about each model's performance on different datasets.*

3.6 Practical Implementation

The practical implementation of the steps described in the previous subsections, was repeated three times, due to an error when setting the filter characteristics and some uncertainties during sample selection, i.e., manual quality check of the dialysis samples. In a first implementation, the dialysis ECGs were mistakenly filtered using a band stop filter set at 60 Hz, instead of the 50 Hz power line frequency used in Germany, where the ISAR study data was collected. After the filter was corrected and set to what is documented in section 3.2.1, the ECG quality in a large subset of records was outstanding. A small portion of records clearly had to be excluded due to low quality. The rest of the ECGs however fit into neither of those categories. Because ECG sample selection was done manually, the decision of which of these ECGs where good enough to keep, and which would have to be discarded, was difficult and subjective.

To understand how sensitive the method is to lower quality samples in the training or test sets, two rounds of sample selection were performed on these correctly filtered data. In a first round, samples were only excluded if artefacts made up more than a third of the 30s sampling window, or noise completely obscured the ECG waveform. In a second round, an effort was made to find the best possible quality in every search window and samples with visible noise and artefacts were excluded. Examples of which records were kept or discarded in each case are documented in the results, see section 5.1.

Because the effects of sample selection and ECG filtering on the success of the resulting AF detection model may be of interest to the reader, the process of training and testing the models was performed on all three iterations, i.e., filtering error, moderate quality check and strict quality check. The results of all three are documented and discussed in the following chapters. The challenge dataset remained unchained but was newly split into a training and test portion for each round. The variance in the challenge model's classification accuracy between the iterations, was used to get an estimate of how much results vary simply based on the division into training and test portions, despite working with the same dataset. The AFT/PTB model was trained only once.

4. Statistics and Visualization

Various statistical tests and visualization techniques were used to document, emphasize or confirm the results, including boxplots, a bar chart and two-sample student's t-tests. Model performance was assessed using two different performance metrics, namely classification accuracy and F1 score.

4.1 Control Matching

Control matching results were documented by summarizing the number of male and female records as well as mean and standard deviation of age and dialysis vintage before and after matching in a table. To determine if control matching was necessary, two sample Student's t-tests were used to test if age and dialysis vintage of the AF records significantly differed from all available controls [39]. Characteristics were considered significantly different if the null hypothesis, that both samples have equal means was rejected at a level of significance α = 0.05. Because variances between the groups were unequal, Satterthwaite's approximation was used to estimate the effective degrees of freedom [40]. Two sample t-tests were used to confirm that matching was successful, by testing the equality of mean age and dialysis vintage between the AF records and matched controls. If these tests fail to reject the null hypothesis of both distributions having equal means, matching was successful.

Similarity in age and dialysis vintage between the AF group and the matched controls was documented graphically using boxplots. Boxplots can be used to visualize a distribution through its minimum, maximum, median, and 75th and 25th percentile. These values are determined by sorting all values from smallest to largest. Minimum and maximum are the values at either end of the list, the median is the value in the middle of the sorted list, or arithmetic mean of the two central values in case of an even number of samples. 75th and 25th percentile, are also referred to as upper and lower quartile. They are calculated as the median of the upper and lower half of your sorted list, i.e., the values that are larger than 75% and 25% of the values in your list, respectively [39]. In a boxplot, the upper and lower quartiles plotted as horizontal lines that form the upper and lower edges of a box. The median is indicated with a red horizontal line inside the box. Minimum and maximum values are also plotted as vertical lines, and connected to the box through a dashed vertical line, forming the so-called whiskers [39].

4.2 Visual Comparison

To visualize the difference in attractor shape and density between the PhysioNet Challenge samples with and without AF and ECGs collected during dialysis with and without AF, the $N = 3, k = 1$ attractor projection (three-point attractor) and its three density distributions were plotted for each group. To reduce the influence of individual differences, 40 samples from each group were selected. Their attractors were overlayed and plotted with density. The resulting image shows how frequently a certain area is visited across the 40 attractor examples, making it possible to compare the attractor characteristics of the different groups. A logarithmic color scale was used to increase the visibility of the attractor's outer areas, which have a much lower density than the center. The three density distributions' mean across each group were plotted as well and color coded for the four groups.

4.3 Classifier Characteristics

To compare the different models in terms of the numbers of neighbors used in their individual classifiers, boxplots were used to summarize these values. Three individual boxes grouped together were used for the models that were trained multiple times. To understand which density distributions of which attractor projection achieved the best classification results, the cross-validated accuracies were visualized in three sets of bar charts, one per density distribution, with the 20 attractor projections on the x-axis and the classifiers cross-validated accuracy on the y-axis. The models were color coded and visualized as one bar per attractor. For models that were trained multiple times, the mean across the three iterations is plotted.

4.4 Performance Metrics

The quality of a model can be assessed using different performance metrics. The parameters they are calculated from can be summarized in a so-called confusion matrix. The confusion matrix for binary classification as AF or control is shown in table 4.

	Predicted Classification		
	AF	**Control**	**Total**
AF	TP	FN	∑AF
Control	FP	TN	∑CT
Total	∑P	∑N	

Table 4 – Confusion matrix for binary classification of AF or control. *Samples can be categorized into one of four categories, true positive (TP), samples that are correctly labeled as AF, false negatives (FN), samples that are labeled as control by the model despite showing AF, false positive (FP), control samples labeled as AF, and true negatives (TN), correctly identified by the model. The total number of AF and control samples and the number of samples labeled as positive or negative are needed to calculate accuracy and F1 performance metrics for a model based on the number of samples in particular categories.*

The simplest metric to assess a model's classification performance, is calculating the percentage of correctly labeled samples, i.e., its accuracy, see formula 7. This metric works well as long as the classes are balanced, i.e., the number of samples in both classes is equal or approximately equal [41]. Since equal numbers of AF and control records were used to train the AF detection models, accuracy was used as the main quality score in this thesis.

To be able to compare the results to previous studies, another performance statistic was computed for the challenge model and the highest performing dialysis model, namely the F1 score. This score was used in the original PhysioNet challenge [32], and is used in most of the publication since then, that worked with the same dataset. The score can be computed using formula 7 and 8 [32]. For the models trained in this thesis, equal or approximately equal numbers of AF and control records were used, resulting in very similar accuracy and F1 scores (compare formulas 7 and 9 for $\sum \mathrm{AF} \approx \sum \mathrm{CT} \approx \sum \mathrm{P} \approx \sum N$).

$$Accuracy = \frac{TP + TN}{\sum AF + \sum CT} \times 100\% \qquad \text{(6) [41]}$$

$$F1_{CT} = \frac{2 \cdot TN}{\sum \mathrm{CT} + \sum \mathrm{N}}, \qquad F1_{AF} = \frac{2 \cdot TP}{\sum \mathrm{AF} + \sum \mathrm{P}} \qquad \text{(7) [32]}$$

$$F1 = \frac{F1_{CT} + F1_{AF}}{2} = \frac{TN}{\sum \mathrm{CT} + \sum \mathrm{N}} + \frac{TP}{\sum \mathrm{AF} + \sum \mathrm{P}} \qquad \text{(8) [32]}$$

5. Results

This chapter starts with the results of sample selection and control matching, as well as record selection during the manual quality check. Three-point attractors with and without AF, from the start of dialysis and the challenge dataset, i.e., unrelated to dialysis, are visualized for comparison (i.e., sensitivity analysis), along with their three density distributions. The results of the three iterations of training the AF detection models and testing them on unseen test samples are also included in this section. Furthermore, this chapter explores the effects of using unmatched training data, by comparing the two models trained on PhysioNet data. Classification accuracies and number of neighbors for the individual classifiers of each model are also included in this section.

5.1 Sample Size and Dataset Characteristics

The PhysioNet Challenge 2017 training set included 648 records with AF that were over 30s in length. 648 samples from the normal rhythm group that were also 30s or longer, were selected as controls. The 80 records available from PhysioNet's AFT database were used as AF records for the second PhysioNet dataset. The 80 healthy controls available from the PTB database were used as controls. The resulting sample sizes for these two datasets are shown in table 5, along with the sample sizes and AF percentages after splitting them into a training and test portion using a 30% hold out.

	Available		**Training**		**Test**	
	total	AF	total	AF	total	AF
Challenge dataset	1 296	648 (50%)	908	454 (50%)	388	194 (50%)
AFT/PTB dataset	160	80 (50%)	112	56 (50%)	48	24 (50%)

Table 5 – Sample Sizes for the two PhysioNet datasets. *The challenge dataset uses AF and control samples from the PhysioNet Challenge 2017 training set. AF samples for the AFT/PTB set were taken from the AFT database, controls from the PTB diagnostic database. Datasets were randomly split into a training and test portion using a 30% hold out and even distribution of AF and control records.*

The available ECGs from the ISAR study that passed the basic quality check included 46 records from subjects with permanent AF and no pacemaker. 222 records with sufficient quality, no pacemaker and no AF were available as controls. From these, 46 matching controls were selected based on their age, gender and dialysis vintage. The subject characteristics of these three groups are summarized in table 6. Paired t-tests showed significant differences in age and dialysis vintage between the AF samples and all available controls but not to the matched controls (α = 0.05). Graphically, the similarity in age and dialysis vintage after matching is shown in the boxplots in figure 19.

	AF samples	Control samples	Matched controls
Number of records	46	222	46
Age, years	76.9 (8.3)	58.0 (14.7)	75.9 (8.8)
Gender, m/f	33/13	149/73	33/13
Dialysis vintage, months	32.1 (26.7)	71.3 (63.9)	34.8 (24.1)

***Table 6 - Subject characteristics of the ISAR study ECGs before and after matching.** Number of records, age, gender and dialysis vintage in months of the remaining AF and control samples after the initial quality check compared to the matched the controls. Age and dialysis months are given as mean (standard deviation), gender as number of male subjects/number of female subjects in the sample.*

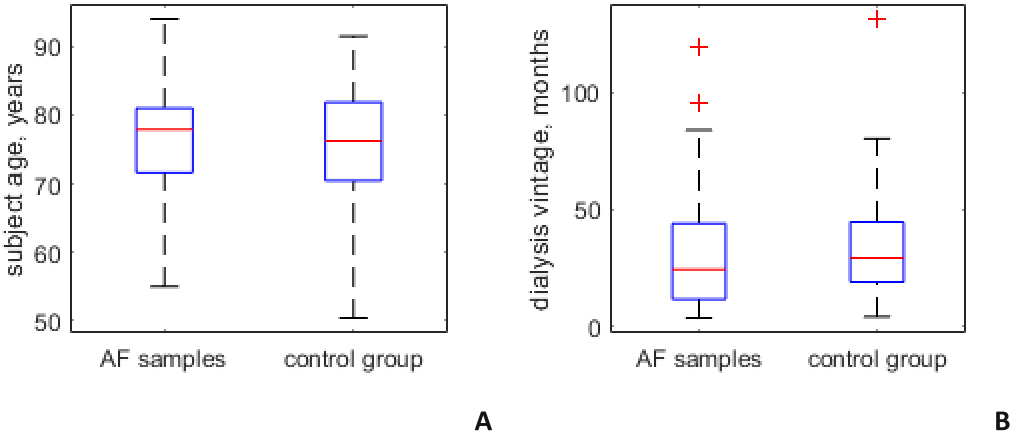


***Figure 19 – Control matching results.** Comparison of A) age in years and B) dialysis vintage in months between the subjects with AF and their matched controls. The red line indicates the median, upper and lower edge of the blue box mark 75th and 25th percentile, respectively. The most extreme points not considered outliers (inside of ±2.7 standard deviations assuming normally distributed data) are marked as the black horizontal bars, outliers are plotted individually in red.*

Three iterations of the manual quality check were performed on the ECG samples during and after dialysis, see section 3.6. Examples of the different ECG quality levels are shown in figure 20. Perfect quality ECG samples such as the one in figure 20A were included in every iteration. Samples with some noise or artefacts in part of the sampling window, see figure 20B, were included as is during the moderate quality check. During the strict quality check, the sampling window was moved within the search window to find a better sample, i.e., one comparable to figure 20A. Figure 20C shows an example of the ECG quality that would be included during moderate quality check but excluded during the strict check. Samples with quality levels comparable to figure 20D were excluded during every iteration. Since a sample of sufficient ECG quality could not be found in every search window of every record, the number of samples per search window per phase differed slightly from the number of selected records in table 6. Because the exclusion criteria also differed between the iterations, the training and test set sample sizes differed slightly between the iterations and the ratio between AF and control samples is only approximately 50%. The sample sizes for the four dialysis datasets, start of dialysis, during dialysis, dialysis 24h and combined set, their partition into training and test portion for each of the three iterations and the corresponding AF percentages are summarized in table 7.

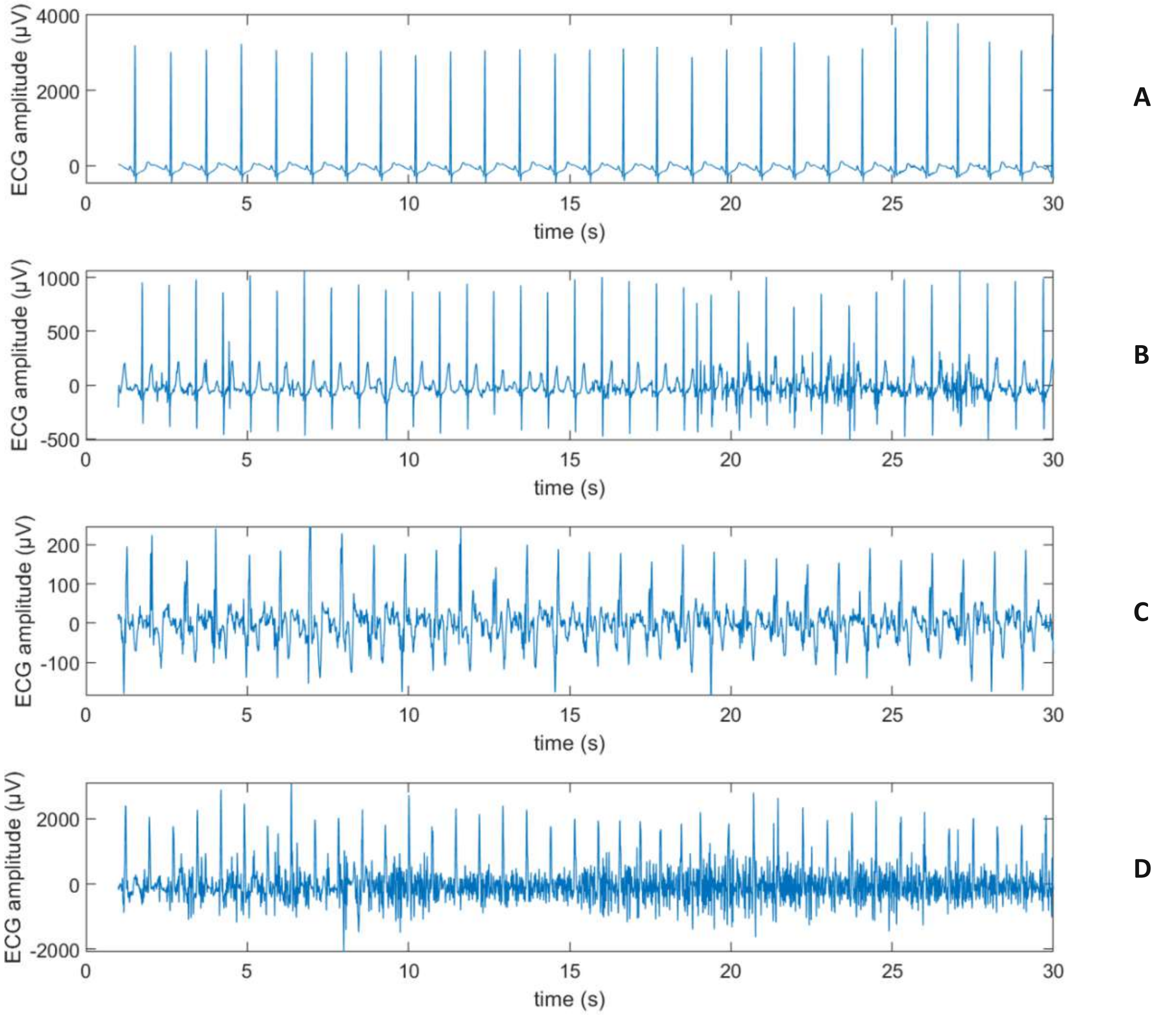


Figure 20 – ECG samples of different signal quality. *All samples were taken from ISAR records without AF. A) Example of an ECG with excellent quality that is included in all iterations of the manual quality check. B) Example of an ECG with some irregularity in part of the signal. During the strict quality check, the sampling window was moved to find a better-quality sample within the search window. In the moderate quality check dataset, samples like this were included as-is. C) Example of a signal quality that was included in the moderate quality check dataset but removed during the strict quality check. D) Example of the signal quality that was considered too low to include in either iteration of the quality check.*

During the visual inspection of samples for the manual quality check, several records with rhythm disorders other than AF were found, see figure 21 C for an example. To increase generalization in the model, these samples were not excluded, but treated as regular control samples, because they did not show AF.

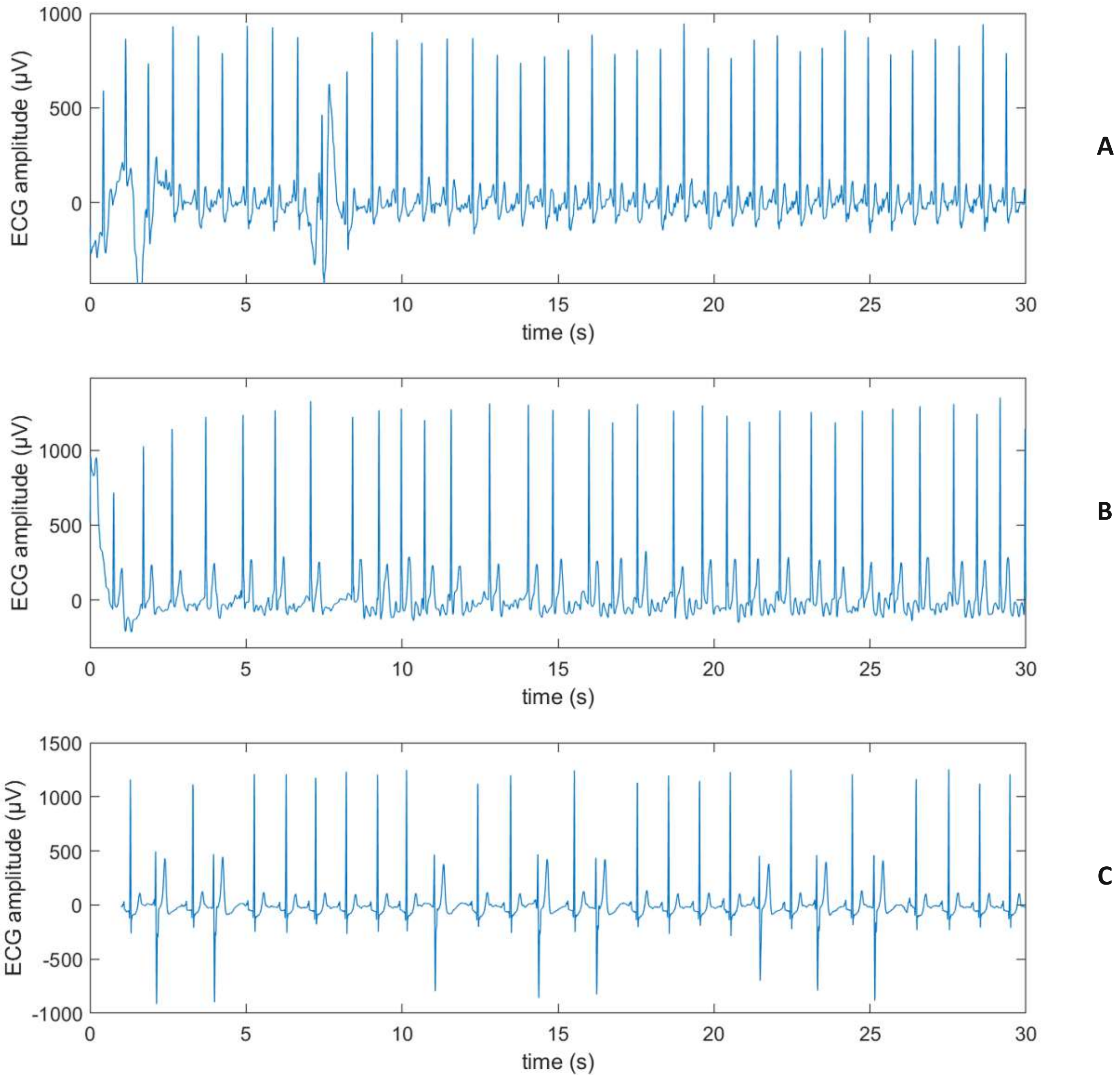


Figure 21 - ECG examples with different rhythms. *A) Control record from the challenge dataset showing normal sinus rhythm B) ECG with AF from the challenge dataset C) ISAR study ECG that shows an arrythmia other than AF. This record was included in the controls of this dataset, since AF is not present, even though it did not show sinus rhythm upon visual inspection.*

Iteration I – Filtering Error						
	Available		Training		Test	
	total	AF	total	AF	total	AF
Start of dialysis	90	44 (48.9%)	63	31 (49.2%)	27	13 (48.1%)
During dialysis	260	129 (49.6%)	183	91 (49.7%)	77	38 (49.4%)
Dialysis 24h	513	254 (49.5%)	362	178 (49.2%)	151	76 (50.3%)
Combined	1 809	902 (49.9%)	1 270	632 (49.8%)	539	270 (50.1%)
Iteration II – Moderate Quality Check						
	Available		Training		Test	
	total	AF	total	AF	total	AF
Start of dialysis	87	43 (49.4%)	61	31 (50.8%)	26	12 (46.2%)
During dialysis	262	131 (50.0%)	184	92 (50.0%)	78	39 (50.0%)
Dialysis 24h	520	259 (49.8%)	366	183 (50.0%)	154	76 (49.4%)
Combined	1 816	907 (49.9%)	1 274	637 (50.0%)	542	270 (49.8%)
Iteration III – Strict Quality Check						
	Available		Training		Test	
	total	AF	total	AF	total	AF
Start of dialysis	83	40 (48.2%)	59	28 (47.5%)	24	12 (50.0%)
During dialysis	253	125 (49.4%)	179	89 (49.7%)	74	36 (48.6%)
Dialysis 24h	497	245 (49.3%)	351	174 (49.6%)	146	71 (48.6%)
Combined	1 793	893 (49.8%)	1 259	628 (49.9%)	534	265 (49.6%)

Table 7 – Sample sizes of the four datasets containing dialysis samples and their partition into training and test sets for each iteration. *Number of AF and N (control, i.e., no AF) samples in the four datasets for each iteration. Sets were randomly split into training and test portions using a 30% hold out and even distribution of AF and control records.*

5.2 Visual Comparison

Figure 22 shows examples of three-point attractors ($N = 3,\ k = 1$ attractor projections) from PhysioNet (challenge set) and start of dialysis samples with and without AF. Each figure contains the overlayed attractors of 40 samples from the respective group to reduce the influence of individual differences and highlight the differences between the groups. The attractors are plotted with density, to show which areas are visited most frequently. Angular density, radial density and attractor outline for the three-point attractors in figure 22 are shown in figure 23.

As seen in figure 22, there is an obvious difference in attractor shape between the PhysioNet attractors with and without AF. While the attractor without AF is more of a rounded star shape, the AF attractor is almost triangular. This is also visible in the mean attractor outlines in figure 23C. There, the attractor outline of the PhysioNet samples with AF has three clearly detectable peaks, while the one without AF has six smaller peaks and is generally flatter.

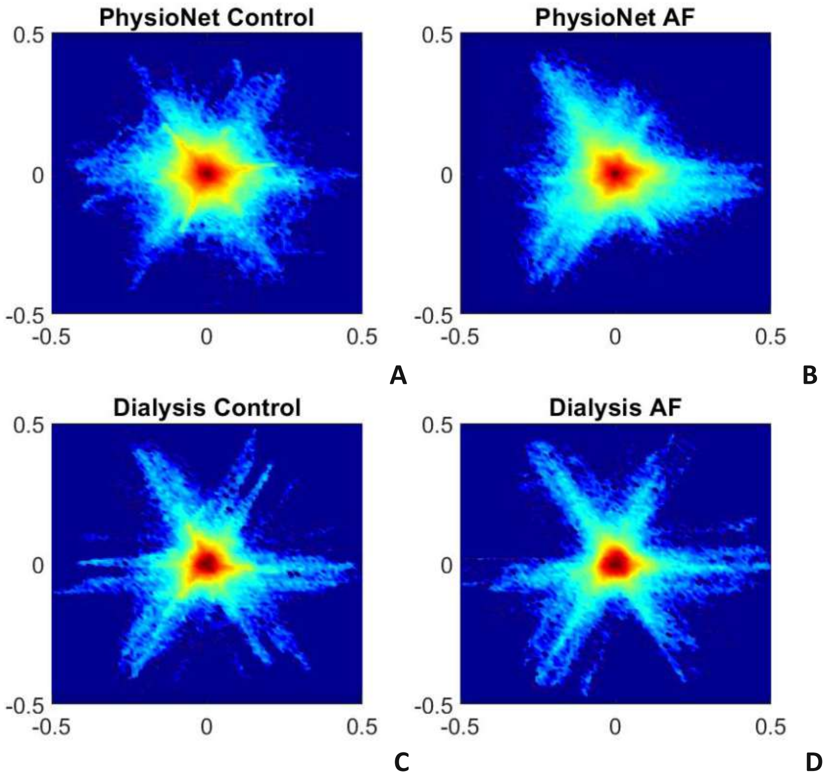


Figure 22 – Three-point attractors of PhysioNet and dialysis ECGs with and without AF. *A) PhysioNet samples without AF, taken from the "normal" group of the challenge 2017 training set, B) PhysioNet samples with AF, from the AF group of the challenge 2017 training set C) start of dialysis samples without AF and D) start of dialysis samples with AF. The attractors of 40 samples from each group are overlayed and plotted with density to create the equivalent of a mean attractor that highlights the differences between the groups by reducing individual differences.*

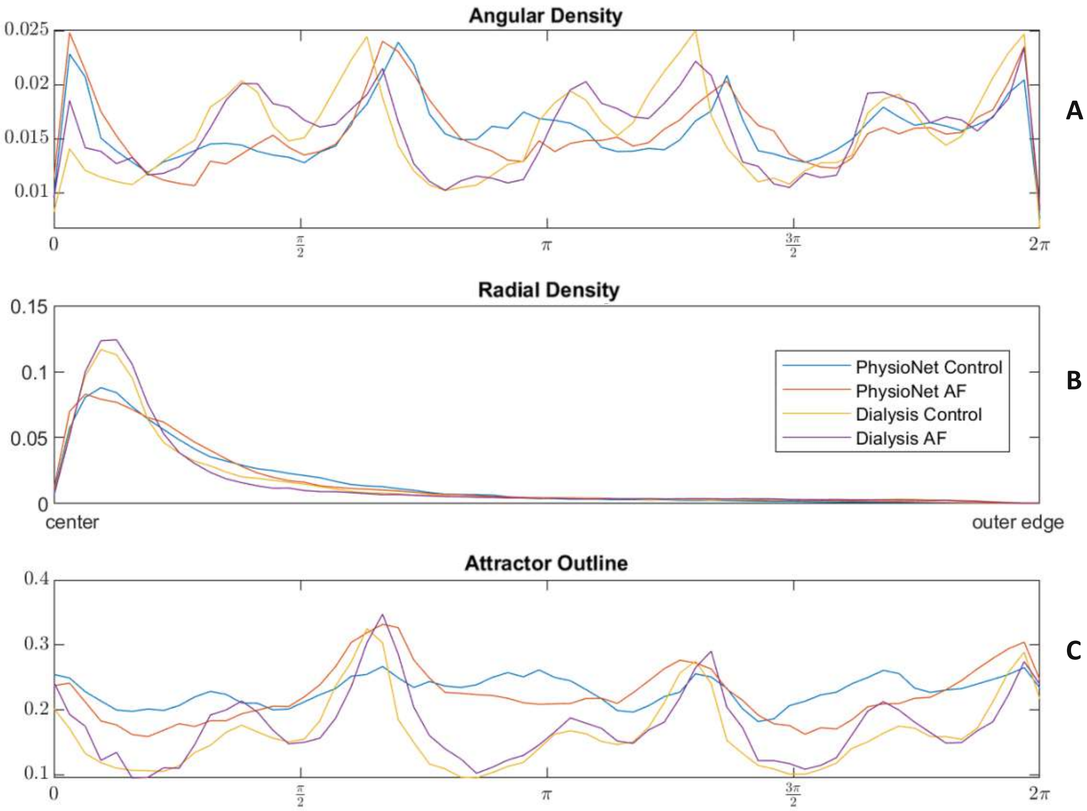


Figure 23 – Comparison of the three-point attractor density distributions of the PhysioNet and dialysis data with and without AF. *Mean A) angular density, B) radial density and C) attractor outline of the three-pint attractors of 40 samples from each of four groups, the PhysioNet control group, i.e., normal records from the PhysioNet Challenge 2017 training set, PhysioNet samples with AF, i.e., AF samples from the 2017 challenge dataset, samples from the start of dialysis without AF and start of the dialysis with AF. Dialysis samples were extracted from 24h ECGs collected during the ISAR study. The corresponding attractors are shown in figure 22. The legend in subplot B is valid for all three plots.*

The attractors of ECGs taken at the start of dialysis are very sharp and defined star shapes, both for the samples with and without AF, see figure 22C and D. This is also visible in their attractor outlines, which have six peaks like the PhysioNet control group, but deeper valleys between them, see figure 23C. The radial density of the start of dialysis examples is also consistent with this sharper and more defined attractor than the PhysioNet examples, which are more spread out across the radial range. The mean angular density distributions of all four groups are very similar, although the differences between PhysioNet and dialysis still exceed those between AF and control. Differences between the AF and no AF attractor examples from the start of dialysis are hard to find in both the attractor images as well as the three density distributions. Because the ECG signals were scaled to a range of zero to one, all four attractors are the same size.

5.3 Model Results

This section contains the results of testing the three iterations of the challenge model, the three dialysis models and the combined model on unseen test samples from their corresponding test set, as well as the other four sets. The AFT/PTB and challenge models are compared in terms of classification accuracy on their own test sets, and each other's test set. More details on the individual *k*-nearest neighbor classifiers the models are comprised of, namely the number of neighbors used and their cross-validated accuracies, is summarized and visualized in the last part of this section. The approximate training time per model, i.e., total time of fitting each set of 60 classifiers, was around 45 to 50 minutes.

5.3.1 Classification Performance

The classification accuracies of the challenge model, the three dialysis models and the combined model for the five test sets are noted in table 8. These data show that the challenge model performed very well when classifying samples in the challenge test set, with an average classification accuracy of 89.3% across the three iterations. That is higher than any of the dialysis models or the combined model, when tested on their corresponding test sets. Its performance on the three dialysis test sets is however lower than that of the during dialysis, dialysis 24h and combined models. The challenge model's performance on the combined test set is good, with almost 85% classification accuracy for all three iterations.

The during dialysis model was the highest performing model for the start of dialysis and during dialysis test sets, with cross validated accuracies higher or equal to those of all other models when tested on these sets. Its mean classification accuracy for the start of dialysis test set was 85.7%. This model was trained on all samples extracted from the three timepoints during the dialysis but no PhysioNet or after dialysis samples. The dialysis 24h model classified start of dialysis test samples with an average accuracy of 81.6% across the three iterations. The combined model's performance on the challenge test set never significantly exceeds that of the challenge model, which was only trained on the challenge training samples. Its average classification accuracy for the start of dialysis samples was 81.7 %. The model trained on just the samples from the beginning of dialysis (30min into the measurement) showed the lowest classification accuracies across all iterations and test sets.

The AFT/PTB model achieved a perfect accuracy of 100% on the training portion of the same set, but only 60% accuracy when tested on the challenge test set, see table 9. The challenge model on the other hand, also performed excellently when tested on PhysioNet data from a different database. All three iterations achieved over 90% classification accuracies (mean classification accuracy of 93.8%) when tested on the AFT/PTB test data, an even better result than the accuracy on the test portion of the same dataset.

Computing the F1 score for the three iterations of the challenge model, resulted in an average score of 0.892 for the challenge test set and 0.937 for the AFT/PTB test set. Each individual score was within the range of ± 0.001 of the corresponding accuracy score when expressed as a ratio. The during dialysis model scored an average F1 of 0.855 on the start of dialysis test set.

Iteration I – Filtering error

Training Test set	Challenge model	Start of dialysis	During dialysis	Dialysis 24h	Combined model
Challenge set	91.0 %	61.3 %	67.0 %	72.2 %	88.1 %
Start of dialysis	66.7 %	63.0 %	85.2 %	85.2 %	85.2 %
During dialysis	71.4 %	75.3 %	84.4 %	83.1 %	84.4 %
Dialysis 24h	71.5 %	76.8 %	84.8 %	83.4 %	82.1 %
Combined set	85.5 %	65.7 %	72.0 %	75.3 %	86.5 %

Iteration II – Moderate quality check

Training Test set	Challenge model	Start of dialysis	During dialysis	Dialysis 24h	Combined model
Challenge set	89.4 %	60.1 %	72.4 %	69.1 %	86.6 %
Start of dialysis	76.9 %	73.1 %	88.5 %	84.6 %	80.8 %
During dialysis	71.8 %	69.2 %	78.2 %	76.9 %	73.1 %
Dialysis 24h	70.8 %	70.1 %	77.3 %	78.6 %	76.6 %
Combined set	84.1 %	62.9 %	73.8 %	71.8 %	83.8 %

Iteration III – Strict quality check

Training Test set	Challenge model	Start of dialysis	During dialysis	Dialysis 24h	Combined model
Challenge set	87.4 %	58.5 %	69.3 %	72.2 %	88.9 %
Start of dialysis	75.0 %	75.0 %	83.3 %	75.0 %	79.2 %
During dialysis	74.3 %	74.3 %	81.1 %	79.7 %	79.7 %
Dialysis 24h	76.7 %	76.0 %	82.9 %	82.9 %	80.1 %
Combined set	84.5 %	63.3 %	73.0 %	75.1 %	86.5 %

Table 8 – Classification accuracies of the challenge model and the four dialysis models when tested on unseen test samples. *Each row corresponds to a test set, each cells contains the percentage of samples from this set that was correctly labeled by the model of the respective column. Models with 80% classification accuracy or more are highlighted in green. Each iteration corresponds to a slightly different sample size and quality level for the dialysis related ECGs and a new random split into training and test portions for the challenge samples in the challenge and combined datasets (see explanation section 3.6).*

	AFT/PTB Model	Challenge Model
AFT/PTB test set	100.0 %	91.7 % / 93.8 % / 95.8 %
Challenge test set	60.3 %	91.0 % / 89.4 % / 87.4 %

Table 9 – Classification accuracies of the two PhysioNet models when tested on unseen test samples. *The AF and control samples used to train the AFT/PTB model come from two different databases. The challenge model was trained on samples from one database specifically intended for training an AF detection algorithm. Three different accuracies were noted for the challenge model because it was trained and tested three times with the same underlying dataset but different random split into training and test portion.*

5.3.2 Classifier Characteristics and Details

Each model is made up of 60 k-nearest neighbor classifiers, one per density distribution per attractor projection. The number of neighbors k is determined for each of them via an automated optimization process that minimizes five-fold cross-validation loss [38]. The resulting numbers of neighbors for the different models are visualized as boxplots, see figure 24. This figure shows, that the AFT/PTB model is comprised of classifiers with a very low number of neighbors. Since the median number of neighbors is one, the lowest possible number of neighbors, more than half of the classifiers in this model use only one nearest neighbor. The challenge model's classifiers use a similar number of neighbors for all three iterations, with a median of 11.5 for the first iteration and 11 for the latter two. Results for the combined model are in the same range, the first iteration being slightly higher, with a median of 15. The classifiers in the during dialysis and dialysis 24h models use lower numbers of neighbors than the challenge and combined model, with medians ranging between 3.5 and seven. Numbers of neighbors in the start of dialysis model differ among the iterations, with medians between four and fourteen.

The performance of each individual classifier is assessed by calculating its accuracy based on ten-fold cross-validation. These values are visualized in figure 25, with the full values included in tables 12-15 in the appendix. For the models that were trained multiple times, the mean across the three iterations is plotted. The challenge model's classifiers perform well, with particular success on the angular densities. 26, 28 and 24 out of the 60 classifiers exceed the 70% threshold for being used in the classification process of an unseen test sample, for the three iterations respectively.

The classifiers in the start of dialysis model show low accuracies for most of the density distributions and attractor projections. Only few of them achieve the 70% accuracy required to be used in the classification of a test sample. The classifier trained on the $N = 3,\ k = 1$ attractor projection's radial density performs particularly poorly, with an accuracy that barely exceeds the 50% statistically anticipated value for a binary classification. The classifiers in the during dialysis and dialysis 24h models achieve accuracies around 65% to slightly more than 70%, with no obvious preferences for a certain density or attractor. The combined models' classifier accuracies are in a similar range for the radial density and attractor outline, but slightly higher for the angular densities.

The classifiers of the AFT/PTB model all show very high cross-validated accuracies, with every single one exceeding the 70% accuracy requirement for their predictions to be considered during the classification process. Most of them exceeded 80 or even 90% cross validated accuracy, with the radial density distributions performing particularly well.

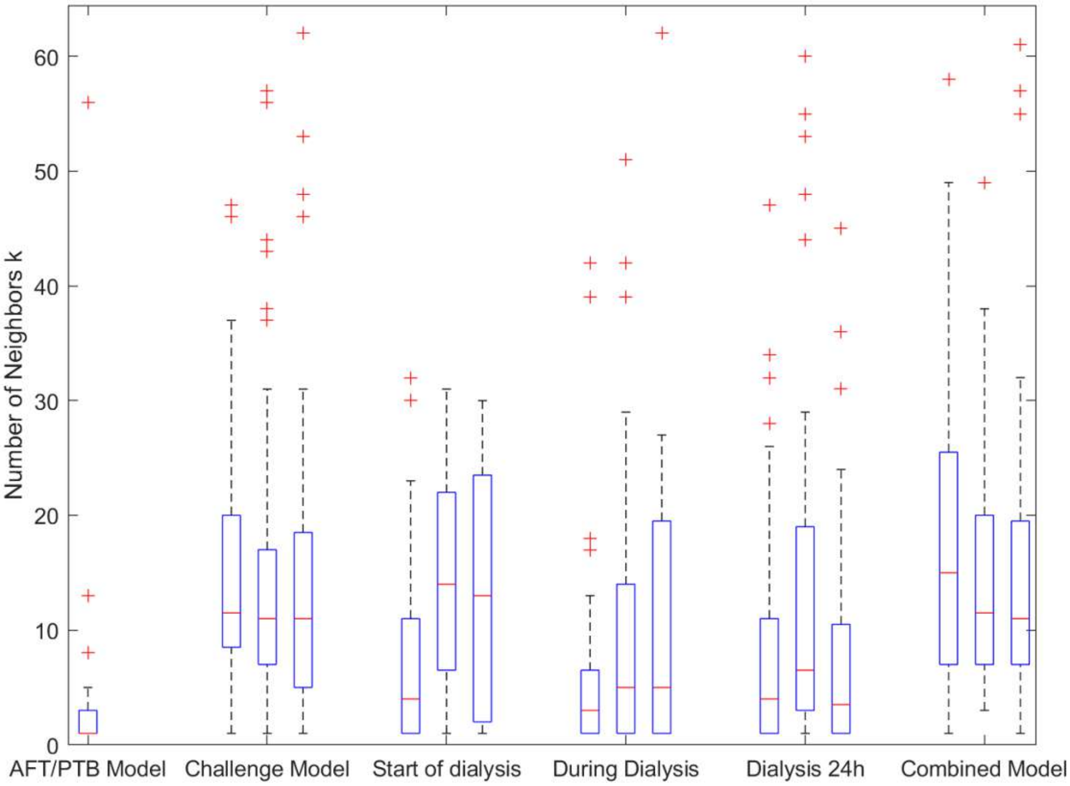


Figure 24 – Number of nearest neighbors used in the individual k-nearest neighbor classifiers of each model. *Models other than the AFT/PTB model were trained three times on slightly different training data due to uncertainties in the manual quality check and an error when filtering the raw data. The red line indicates the median, upper and lower edge of the blue box mark 75th and 25th percentile, respectively. The most extreme points not considered outliers (inside of ±2.7 standard deviations assuming normally distributed data) are marked as the black horizontal bars, outliers are plotted individually in red. Some extreme points marked as outliers are out of range, and thus not visible in the figure.*

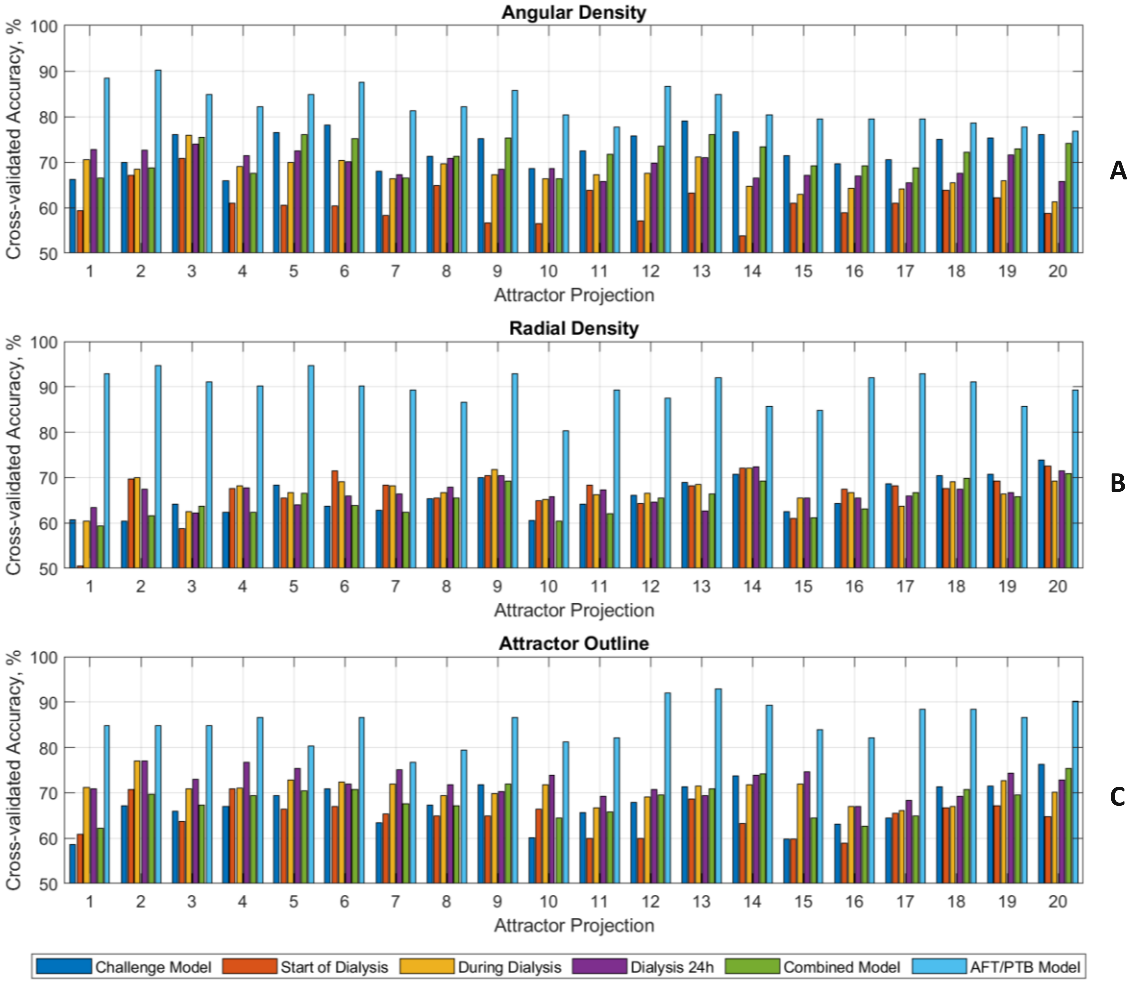


Figure 25 – Cross-validated accuracies of the individual classifiers the six AF detection models are comprised of. *One k-nearest neighbor classifier is trained per density curve (angular density, radial density and attractor outline) for each of the twenty attractor projections. Accuracies are based on 10-fold cross validation. For the challenge model, the combined model and the three dialysis models, the mean across the three iterations is shown. Classifiers of more than 70% accuracy are used in the classification of an unseen test record.*

6. Discussion

This section interprets, connects and contextualizes the results from the previous section. It discusses the results of control matching the ISAR study records, and the effects of not matching the AF and control groups of a training set. The effectiveness of combining SPAR with *k*-nearest neighbor classification for AF detection is analyzed by comparing the challenge model to entries from the original challenge and other AF detection algorithms in terms of performance, stability and complexity of the algorithm. The final subsection of this chapter uses the three iterations of the dialysis models and the combined model to analyze how well the proposed methodology performs on ECGs during or after dialysis. Furthermore, the effects of different levels of pre-processing and training set compositions are discussed.

6.1 Control Matching and Overfitting

The paired t-tests comparing mean age and dialysis vintage between the dialysis subjects with and without AF, show that without control matching, there are significant differences in both parameters between the two groups. This confirms that control matching should be performed to avoid confounding factors in the classification [42]. The paired t-tests performed after matching and the boxplots shown in figure 19, show that the control matching approach of selecting subjects of the same gender, that are ±10 years in age and picking the closest match in dialysis vintage, was successful in matching those parameters. Influences of other differences between the two groups that were not considered in the match are however still possible.

The AFT/PTB model demonstrates the effects of not matching the AF and control group data. Because the AF and control samples in this training set are from two different databases, there is a variety of differences between them, that may influence the model. Aside from the differences in cardiac rhythm, factors such as subject age and cardiac health may cause visible changes in the attractors and their densities. Additionally, the control samples taken from the PTB database were sampled at a much higher frequency of 1 kHz than the samples in the AFT database, which were sampled at 128 Hz. This affected the filters used to remove noise from the ECG signals, particularly with regards to the lowpass filter. Despite increasing the filter order from six to nine, the lowpass applied to the PTB data only achieved an attenuation of - 23 dB at a frequency of 60 Hz instead of -80 dB. These different frequency characteristics in the filtered ECG samples may also be visible in the attractors.

The AFT/PTB model performs extremely well when tested on unseen samples from the same source (perfect accuracy of 100%), but poorly when tested on a different dataset (60% accuracy on the challenge test set). This is a problem known as overfitting, where a model focusses too strongly on the particular training data at hand and then lacks the generalization needed to be successful in classifying records from other datasets [28]. In this case, the different subject and frequency characteristics of the two groups are more notable in the attractor densities than the difference between AF and sinus rhythm that the model is intended to detect. The model is perfect at identifying if a sample is from the AFT or the PTB database, but not if AF is present. This is also reflected in the high number of classifiers in the model that use only one nearest neighbor, which can often lead to overfitting [28].

While the AFT/PTB set proved unsuitable for training a functioning AF detection model, it can still be used to test the challenge model. Unlike the AF/PTB model the challenge model showed excellent results when tested on samples from a different source, i.e., the AFT/PTB test set, which supports that this model is not overfitted to training set. The AF and sinus rhythm data in the PhysioNet 2017 Challenge training set were all recorded on the same device and underwent the same preprocessing steps [32].

6.2 AF Detection During Dialysis

The visual comparison of the three-point attractors from the start of dialysis showed little visual distinction between the AF and control (no AF) samples. This is consistent with the low cross validated accuracies of the classifiers trained on the $N = 3,\ k = 1$ attractor projections during the three iterations of the start of dialysis model. While the classifiers trained on some of the higher dimensional attractors have slightly higher cross validated accuracies, the overall performance of the model is still low. This may also be due to the small sample size of this training set, which is only around 60 samples (63, 61 and 59 for iterations I, II and III respectively), which is very small for a classifier with 64 features. Higher dimensional classifiers require a larger training sample to maintain the necessary density of datapoints in the feature space to find nearest neighbors [28], see section 2.4.

Dialysis is known to distort the ECG waveform due to fluid and electrolyte shifts [2]. The visible differences in attractor shape and density between the PhysioNet and start of dialysis samples, show that these changes are also reflected in the three-point ECG attractors and its densities. Presuming this is also true for the higher dimensional attractors, this explains why the challenge model has a low cross validated accuracy for the start of dialysis test set, as the differences between the two datasets exceeds the difference between AF and control that is to be detected by the model. The challenge model's good performance on the combined test set is due to PhysioNet samples making up 70% of this test set, which compensates for the worse results of the model in the dialysis portion of the combined test set.

The fact that the dialysis 24h and combined models never markedly outperform the during dialysis model when classifying dialysis test samples shows that adding post-dialytic ECG samples or samples unrelated to dialysis to the training set, brings no benefit for AF detection during dialysis. The combined model's performance on the challenge test set is also never better than the challenge model' performance. These results show that adding dialysis samples to the training set when aiming to classify records unrelated to dialysis, also does not improve the model. Since the attractors and densities of PhysioNet and dialysis samples appeared so different upon visual inspection of the $\mathrm{N} = 3,\ \mathrm{k} = 1$ case, it is possible that the features of PhysioNet and dialysis samples are so different that their training samples are located in separate regions of the feature space. This would mean that for a dialysis test sample, few if any PhysioNet sample will be amongst its nearest neighbors and thus affecting its classification, even if they are included in the training set, or vice versa.

Overall, the dialysis models show slightly lower classification accuracies than the challenge model. This is consistent with the expectation that samples collected during dialysis are more difficult to classify because of the aforementioned distortion of the ECG waveform. Additionally, samples with other arrhythmias than AF are included as controls in the dialysis data sets but excluded from the challenge set. Another advantage the challenge model has over the dialysis models, is the fact that its training set is considerably larger than any of the dialysis training sets. Since the performance of the dialysis 24h model is however approximately equal to the during dialysis model, which is trained on only around half the number of samples, this effect may be negligible, as 180 training samples may simply be sufficient for this application. Despite all these challenges, the during dialysis model still performed well, with an 85.7% accuracy for the start of dialysis test samples.

The small variance between the iterations of the dialysis and combined models suggests that the differences in filtering and sample selection have little to no effect on the models' accuracy. This is consistent with previous studies working with SPAR that have highlighted the method's robustness towards outliers and noise [21, 25, 26]. Experimenting with even less preprocessing and/or filtering and may be a topic for further research, as it would further reduce the effort of generating training samples and extracting features.

6.3 Classification Performance

Since the results from the original challenge are published [32], the results of the challenge model can be used to compare the proposed methodology to other entries in the challenge. In the challenge, models were scored using mean F1 scores. For comparability, this metric was also computed. In this case F1 scores are almost identical to the accuracy scores, as the models were trained on an equal number of AF and control records resulting in very balanced classes, see section 4.4 for more information on the two metrics.

The F1 scores of the challenge winners are summarized in table 10 [32]. Winners were chosen based on the algorithm's performance on a hidden test set. The model's scores on the training set and a 300-sample subset of the training set referred to as validation set, are also reported, to indicate if a model has been over-trained on the training data [32]. Three of the winning algorithms were based on complex machine learning techniques, namely combining features from deep neural networks with extreme gradient boosting [43], a multi-layer cascaded binary classification approach [44] or recurrent neural networks [45], while the fourth used a random forest with manually selected features [46].

A direct comparison of those algorithms to the challenge model is unfortunately impossible, since the training set continues to be unavailable to the public. Additionally, the participants of the challenge were also scored on their algorithms ability to distinguish between three categories (AF, normal and arrythmia other than AF) [32]. With F1 scores of 0.892 on the challenge test set and 0.937 on the AFT/PTB set, the challenge model's performance is however in a similar range as the challenge winners, which is an excellent result considering how computationally inexpensive the combination of SPAR and k-nearest neighbor is.

Additionally, in the years following the challenge, several articles were published that propose automated AF detection algorithms, that work with the same publicly available training portion of the challenge dataset, as was used in this thesis. Their F1 scores (mean of F1n and F1a, see section 4.4) and if available classification accuracies are summarized in table 11. These results show that the combination of SPAR and multilayer *k*-nearest neighbor classification used in this thesis, can absolutely match the results of state-of-the-art approaches to AF detection. Compared to these approaches, the algorithm presented in this thesis has the advantage of requiring less preprocessing, due to SPAR being very robust to noise and outliers [21, 25, 26] and requiring no feature selection. Furthermore, the computational cost and complexity of the model is far below than that of a neural network or the approach of [15], where both traditional ECG features are detected and deep learning features extracted by two different networks, before using discriminant canonical correlation analysis feature fusion.

Despite the additional challenges involved in processing ECGs recorded during dialysis, the during dialysis model's performance on the samples from the start of dialysis is in a similar performance range as other AF detection models published in the last years, see table 11. This is a good result, that highlights what a powerful feature extraction tool SPAR is, when it comes to AF detection.

Entry	**Test**	**Validation**	**Training**
Teijeiro et al. [45]	**0.831**	0.912	0.893
Datta et al. [44]	**0.829**	0.990	0.970
Zabihi et al. [46]	**0.826**	0.968	0.951
Hong et al. [43]	**0.826**	0.968	0.951

Table 10 – F1 scores of the 2017 challenge winners. *The scores in the test column mark the model's performance on the hidden test set. Validation is the model's performance on a 300-sample subset of the training set. [32]*

Method	F1 score	Accuracy
Convolutional recurrent neural network [16]	0.869	87.5%
Decision tree ensemble [19]	0.84	–
16-layer 1D residual convolutional network [17]	0.86	80.2%
Convolutional neural network containing residual blocks and recurrent layers [20]	0.889	–
Discriminant canonical correlation analysis feature fusion [15]	0.907	91.7%
Challenge Model	**0.892**	**89.3%**
During Dialysis Model	**0.855**	**85.7%**

Table 11 – Performance of other AF detection algorithms working with the same data set. *F1 and accuracy scores (if available) of AF detection algorithms working with the publicly available 2017 PhysioNet/Computing in Cardiology Challenge training set. F1 scores were calculated as the mean between F1n and F1a to summarize the model's performance for the binary classification task of AF or control. For the challenge model the mean scores across the three iterations, when tested on the challenge test set are given here.*

6.4 Limitations

The main limitations of the dialysis related models presented in this thesis are the unknown dialysis time for some records and the uncertain start time of the measurement with regards to the dialysis treatment. Also, the time of day each recording was started at was unknown. Changes in the ECG during dialysis also depend on the patient's pre-dialytic electrolyte levels and the composition of the dialysate [2]. Both factors were not considered in the analysis.

The comparability between the challenge model, and the during dialysis model is limited by the tenfold difference in training sample size. The challenge model and the classification algorithms published in the literature were also trained on different sample sizes with regards of the control portion of the training set. The challenge model was trained on an equal number of AF and control samples, while the others used the full training set available. Whether this affects the model's performance was not investigated in this thesis.

7. Conclusion

The high classification accuracies of the challenge model for the two PhysioNet test sets, show that SPAR is a valid and useful tool for extracting features from an ECG time series for the purpose of detecting AF. While none of the models trained in this thesis significantly outperform previously published AF detection algorithms [15–18, 20], the combination of SPAR and *k*-nearest neighbor is far less complex and computationally expensive than most if not all of these algorithms. Additionally, little effort was made to fine tune or optimize the models in this thesis. Despite the high number of features for each classifier, no attempts at feature selection were made. Both distance measure and number of neighbors for each classifier were chosen by MATLAB®'s built-in optimization algorithm. The fact that the models still performed so well speaks to the immense potential this method has and warrants future research into further improvements to the algorithm. Since the SPAR methodology has again proven to be very stable towards outliers and signal noise, experimenting with even less preprocessing and quality selection may also be a topic for future research.

Even though the differences in ECG waveform between samples recorded during dialysis and other ECG recordings are also notable in the attractor shape and densities, the results of this thesis show that automatic AF detection with reasonable accuracy is also possible during dialysis, when using the methodology presented in this thesis. Since SPAR can be used to extract features from very short recordings, 30s ECGs in this case but also 10s samples in [24], it qualifies for real time monitoring approaches.

References

[1] V. A. Luyckx, M. Tonelli, and J. W. Stanifer, "The global burden of kidney disease and the sustainable development goals," *Bulletin of the World Health Organization*, vol. 96, no. 6, 414-422D, 2018, doi: 10.2471/BLT.17.206441.

[2] D. Poulikakos and M. Malik, "Challenges of ECG monitoring and ECG interpretation in dialysis units," *Journal of electrocardiology*, vol. 49, no. 6, pp. 855–859, 2016, doi: 10.1016/j.jelectrocard.2016.07.019.

[3] P. J. Aston, M. I. Christie, Y. H. Huang, and M. Nandi, "Beyond HRV: attractor reconstruction using the entire cardiovascular waveform data for novel feature extraction," *Physiological measurement*, vol. 39, no. 2, p. 24001, 2018, doi: 10.1088/1361-6579/AAA93D.

[4] M. Nandi and P. J. Aston, "Extracting new information from old waveforms: Symmetric projection attractor reconstruction: Where maths meets medicine," *Experimental physiology*, vol. 105, no. 9, pp. 1444–1451, 2020, doi: 10.1113/EP087873.

[5] Jane V Lyle, Peter H Charlton, Esther Bonet-Luz, Gary Chaffey, Mark Christie, Manasi Nandi, Ed., *Beyond HRV: Analysis of ECG Signals Using Attractor Reconstruction*: Computing in Cardiology, 2017.

[6] P. Zimetbaum, "Atrial Fibrillation," *Annals of internal medicine*, vol. 166, no. 5, ITC33-ITC48, 2017, doi: 10.7326/AITC201703070.

[7] W. C. Winkelmayer, A. R. Patrick, J. Liu, M. A. Brookhart, and S. Setoguchi, "The increasing prevalence of atrial fibrillation among hemodialysis patients," *Journal of the American Society of Nephrology : JASN*, vol. 22, no. 2, pp. 349–357, 2011, doi: 10.1681/ASN.2010050459.

[8] M. Schünke, *Der Körper des Menschen: Einführung in Bau und Funktion,* 16th ed. Stuttgart: Thieme, 2012.

[9] E. Kaniusas, *Biomedical Signals and Sensors I*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.

[10] J. G. Betts *et al., Anatomy and physiology*. Houston, Texas: OpenStax College, Rice University, 2017.

[11] Dr. Araz Rawshani, *The ECG leads: electrodes, limb leads, chest (precordial) leads, 12-Lead ECG (EKG) – ECG & ECHO.* [Online]. Available: https://ecgwaves.com/topic/ekg-ecg-leads-electrodes-systems-limb-chest-precordial/ (accessed: May 17 2021).

[12] Pediatric Heart Specialists, *Normal Heart Anatomy and Blood Flow.* [Online]. Available: https://pediatricheartspecialists.com/heart-education/14-normal/152-normal-heart-anatomy-and-blood-flow (accessed: May 14 2021).

[13] John Furst, *12 Lead ECG Diagram.* [Online]. Available: https://www.firstaidforfree.com/recording-a-12-lead-ecgekg/12-lead-ecg-diagram/ (accessed: May 17 2021).

[14] L. L. Cables and Sensors, *12-Lead ECG Placement Guide with Illustrations.* [Online]. Available: https://www.cablesandsensors.com/pages/12-lead-ecg-placement-guide-with-illustrations (accessed: May 17 2021).

[15] J. Shi, C. Chen, H. Liu, Y. Wang, M. Shu, and Q. Zhu, "Automated Atrial Fibrillation Detection Based on Feature Fusion Using Discriminant Canonical Correlation Analysis," *Computational and mathematical methods in medicine*, vol. 2021, p. 6691177, 2021, doi: 10.1155/2021/6691177.

[16] Jérôme Van Zaen, Olivier Chételat, Mathieu Lemay, Enric Calvo, and Ricard Delgado-Gonzalo, "Classification of Cardiac Arrhythmias from Single Lead ECG with a Convolutional Recurrent Neural Network," in 2021, pp. 33–41. Accessed: Mar. 14 2019. [Online]. Available: https://www.scitepress.org/Link.aspx?doi=10.5220/0007347900330041

[17] *Robust ECG signal classification for detection of atrial fibrillation using a novel neural network*, 2017.

[18] D. Hutchison *et al., Eds., Pervasive Computing and the Networked World*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013.

[19] M. Rizwan, B. M. Whitaker, and D. V. Anderson, "AF detection from ECG recordings using feature selection, sparse coding, and ensemble learning," *Physiological measurement*, vol. 39, no. 12, p. 124007, 2018, doi: 10.1088/1361-6579/aaf35b.

[20] Z. Xiong, M. P. Nash, E. Cheng, V. V. Fedorov, M. K. Stiles, and J. Zhao, "ECG signal classification for the detection of cardiac arrhythmias using a convolutional recurrent neural network," *Physiological measurement*, vol. 39, no. 9, p. 94006, 2018, doi: 10.1088/1361-6579/aad9ed.

[21] J. Lyle *et al.,* "Beyond HRV: Analysis of ECG Signals using Attractor Reconstruction," in *2017 Computing in Cardiology Conference (CinC)*, 2017.

[22] E. N. Lorenz, "Deterministic Nonperiodic Flow," *J. Atmos. Sci.*, vol. 20, no. 2, pp. 130–141, 1963, doi: 10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.

[23] F. Takens, "Detecting strange attractors in turbulence," in *Lecture Notes in Mathematics, Dynamical Systems and Turbulence, Warwick 1980*, D. Rand and L.-S. Young, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 1981, pp. 366–381.

[24] Jane Lyle, Manasi Nandi, and Philip Aston, "Symmetric Projection Attractor Reconstruction: Sex Differences in the ECG," *Frontiers in Cardiovascular Medicine*, 2021. [Online]. Available: https://kclpure.kcl.ac.uk/portal/en/publications/symmetric-projection-attractor-reconstruction-sex-differences-in-the-ecg(43c17170-0ddc-47c1-a50f-3740e5b6d113).html

[25] Jane V Lyle, "Symmetric Projection Attractor Reconstruction: Embedding of Physiological Time Series in Higher Dimensions," Virtual Conference, 24 May, 2021. Accessed: 29 July, 2021. [Online]. Available: https://www.siam.org/conferences/cm/conference/ds21

[26] J. Lyle, P. Aston, and M. Nandi, "Investigating the Response to Dofetilide with Symmetric Projection Attractor Reconstruction of the Electrocardiogram," in *2019 Computing in Cardiology Conference (CinC)*, 2019.

[27] H. A. Abu Alfeilat *et al.,* "Effects of Distance Measure Choice on K-Nearest Neighbor Classifier Performance: A Review," *Big Data*, vol. 7, no. 4, pp. 221–248, 2019, doi: 10.1089/big.2018.0175.

[28] O. Kramer, "K-Nearest Neighbors," in *Intelligent Systems Reference Library, Dimensionality Reduction with Unsupervised Nearest Neighbors*, O. Kramer, Ed., Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 13–23.

[29] C. Schmaderer *et al.,* "Rationale and study design of the prospective, longitudinal, observational cohort study "rISk strAtification in end-stage renal disease" (ISAR) study," *BMC Nephrol*, vol. 17, no. 1, p. 161, 2016, doi: 10.1186/s12882-016-0374-8.

[30] A. L. Goldberger *et al.,* "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, E215-20, 2000, doi: 10.1161/01.CIR.101.23.e215.

[31] PhysioNet, *The PhysioNet/Computing in Cardiology Challenges.* [Online]. Available: https://physionetchallenges.org/ (accessed: Aug. 20 2021).

[32] G. D. Clifford *et al.,* "AF Classification from a Short Single Lead ECG Recording: the PhysioNet/Computing in Cardiology Challenge 2017," *Computing in cardiology*, vol. 44, 2017, doi: 10.22489/CinC.2017.065-469.

[33] G. D. Clifford *et al., The PhysioNet Computing in Cardiology Challenge 2017 Webpage.* [Online]. Available: https://www.physionet.org/content/challenge-2017/1.0.0/ (accessed: Jul. 20 2021).

[34] G. B. Moody, "AF Termination Challenge Database," 2003.

[35] R. Bousseljot, D. Kreiseler, and A. Schnabel, "Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet," *Biomedizinische Technik / Biomedical Engineering*, vol. 40, s1, pp. 317–318, 2009, doi: 10.1515/bmte.1995.40.s1.317.

[36] I. Silva and G. B. Moody, "An Open-source Toolbox for Analysing and Processing PhysioNet Databases in MATLAB and Octave," *Journal of Open Research Software*, vol. 2, no. 1, 2014, doi: 10.5334/jors.bi.

[37] M. Bachler, C. Mayer, B. Hametner, S. Wassertheurer, and A. Holzinger, "Online and Offline Determination of QT and PR Interval and QRS Duration in Electrocardiography," in *Lecture Notes in Computer Science, Pervasive Computing and the Networked World*, D. Hutchison et al., Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 1–15.

[38] The MathWorks, Inc., *Fit k-nearest neighbor classifier - MATLAB fitcknn.* [Online]. Available: https://de.mathworks.com/help/stats/fitcknn.html (accessed: Oct. 15 2021).

[39] M. J. Campbell, S. J. Walters, and D. Machin, *Medical statistics: A textbook for the health sciences / Michael J. Campbell, David Machin, Stephen J. Walters,* 4th ed. Chichester: Wiley, 2007.

[40] F. E. Satterthwaite, "An Approximate Distribution of Estimates of Variance Components," *Biometrics Bulletin*, vol. 2, no. 6, p. 110, 1946, doi: 10.2307/3002019.

[41] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation," in *Lecture Notes in Computer Science, AI 2006: Advances in Artificial Intelligence*, D. Hutchison et al., Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 1015–1021.

[42] M. A. Mansournia, N. P. Jewell, and S. Greenland, "Case-control matching: effects, misconceptions, and recommendations," *European journal of epidemiology*, vol. 33, no. 1, pp. 5–14, 2018, doi: 10.1007/s10654-017-0325-0.

[43] S. Hong *et al.,* "ENCASE: an ENsemble ClASsifiEr for ECG Classification Using Expert Features and Deep Neural Networks," in *2017 Computing in Cardiology Conference (CinC)*, 2017. Accessed: Oct. 15 2021. [Online]. Available: http://www.cinc.org/archives/2017/pdf/178-245.pdf

[44] S. Datta *et al.,* "Identifying Normal, AF and other Abnormal ECG Rhythms using a Cascaded Binary Classifier," in *2017 Computing in Cardiology Conference (CinC)*, 2017. Accessed: Oct. 15 2021. [Online]. Available: http://www.cinc.org/archives/2017/pdf/173-154.pdf

[45] T. Teijeiro, C. A. Garcia, D. Castro, and P. Flix, "Arrhythmia Classification from the Abductive Interpretation of Short Single-Lead ECG Records," in *2017 Computing in Cardiology Conference (CinC)*, 2017. Accessed: Oct. 15 2021. [Online]. Available: http://www.cinc.org/archives/2017/pdf/166-054.pdf

[46] M. Zabihi, A. Bahrami Rad, A. K. Katsaggelos, S. Kiranyaz, S. Narkilahti, and M. Gabbouj, “Detection of Atrial Fibrillation in ECG Hand-held Devices Using a Random Forest Classifier,” in *2017 Computing in Cardiology Conference (CinC)*, 2017. Accessed: Oct. 15 2021. [Online]. Available: http://www.cinc.org/archives/2017/pdf/069-336.pdf

Appendix

Challenge Model			Start of Dialysis			During Dialysis			Dialysis 24h			Combined Model		
AD	RD	AO	AD	RD	AO	AD	RD	AO	AD	RD	AO	AD	RD	AO
65.0	61.0	59.3	61.3	61.3	67.7	72.8	61.1	70.6	76.0	62.7	71.9	66.7	57.5	61.4
70.6	58.3	64.8	58.1	79.0	87.1	70.6	68.3	80.0	66.9	64.3	77.4	68.2	59.8	68.7
75.7	64.0	66.0	69.4	58.1	58.1	76.7	66.7	72.2	71.0	60.4	70.5	73.7	62.3	65.6
64.5	60.2	67.3	66.1	64.5	74.2	70.0	66.1	68.9	73.5	63.2	74.4	66.4	59.7	68.4
76.0	69.7	70.4	59.7	67.7	69.4	64.4	66.7	70.6	73.0	63.8	72.7	76.4	67.2	69.4
79.0	65.9	70.5	61.3	74.2	77.4	66.7	67.2	68.3	68.8	66.0	72.1	74.5	64.1	72.1
68.9	60.2	64.0	58.1	59.7	71.0	71.1	66.1	76.7	68.2	59.9	74.9	66.4	61.0	65.9
71.5	65.9	65.8	66.1	59.7	67.7	73.9	60.6	68.9	71.3	68.2	71.0	71.8	64.8	65.1
75.0	68.4	69.9	54.8	71.0	69.4	67.2	68.9	68.3	64.6	74.4	69.9	74.0	67.9	70.5
68.6	59.1	59.2	64.5	53.2	71.0	70.0	64.4	67.8	67.4	65.2	73.3	65.0	59.6	63.7
71.8	63.7	64.9	54.8	59.7	53.2	63.9	62.2	61.1	64.9	68.2	62.4	71.9	61.6	63.0
75.1	68.1	66.5	61.3	66.1	64.5	71.7	63.0	71.7	69.4	64.1	72.1	74.0	67.4	68.2
77.8	70.0	71.0	64.5	69.4	72.6	71.1	71.1	70.0	70.2	67.1	70.2	76.1	65.8	70.3
77.4	72.0	74.1	51.6	71.0	54.8	66.7	69.4	70.6	66.0	72.4	74.1	73.2	69.7	73.6
71.1	57.9	59.6	58.1	62.9	67.7	64.4	65.6	68.3	67.7	67.1	71.9	69.0	60.6	63.2
68.8	63.4	61.8	54.8	71.0	50.0	64.4	70.6	66.1	68.0	64.3	64.1	69.8	61.6	61.5
70.8	69.8	63.3	62.9	71.0	64.5	63.3	68.9	68.3	66.0	68.0	69.4	68.2	66.3	62.9
74.4	70.4	68.5	64.5	66.1	58.1	72.2	66.7	70.0	69.6	66.0	71.3	70.4	69.8	68.5
74.8	73.1	70.4	62.9	69.4	61.3	66.1	65.6	73.3	71.9	64.1	71.9	72.8	65.2	65.9
75.8	74.0	75.8	54.8	74.2	74.2	63.3	73.9	71.7	66.0	72.1	74.4	74.9	66.4	74.7

Table 12 – Cross-validated accuracies of the individual k-nearest neighbors classifiers trained during the first iteration (filtering error), %. *Each model is made up of one k-nearest neighbor classifier per density distribution per attractor projection. Each row corresponds to one attractor projection, with its angular density (AD) in column one, radial density (RD) in column two and attractor outline (AO) in column three. Classification accuracies were determined via 10-fold cross validation. Classifiers with 70% or more accuracy are used in the prediction of an unknown sample and highlighted in green.*

Challenge Model			Start of Dialysis			During Dialysis			Dialysis 24h			Combined Model		
AD	RD	AO	AD	RD	AO	AD	RD	AO	AD	RD	AO	AD	RD	AO
68.2	62.5	58.9	55.7	34.4	59.0	74.5	57.6	66.3	69.9	60.7	70.2	68.2	61.4	63.0
70.5	63.8	69.0	73.8	68.9	65.6	70.1	71.2	72.8	73.8	67.2	73.8	69.2	62.8	70.8
76.5	63.6	64.8	72.1	59.0	70.5	76.1	58.7	69.0	74.9	62.0	69.7	76.0	64.1	67.1
66.6	64.3	68.7	52.5	67.2	62.3	66.8	65.8	74.5	69.7	64.8	76.0	68.0	63.0	70.0
77.0	66.8	69.0	50.8	54.1	65.6	75.0	62.5	76.6	71.3	62.0	73.8	76.0	63.8	71.2
79.0	62.7	71.9	65.6	68.9	60.7	70.1	65.2	71.2	73.5	61.7	68.3	75.3	62.9	70.4
68.8	63.2	63.7	57.4	67.2	65.6	66.3	65.8	72.8	68.6	65.8	75.7	67.5	63.8	68.7
71.1	65.0	69.4	63.9	67.2	59.0	69.0	69.0	69.6	73.0	67.8	69.1	71.1	65.6	67.5
75.5	71.4	73.6	50.8	65.6	52.5	67.9	75.0	71.7	70.5	66.4	67.2	74.9	70.0	72.7
67.7	60.7	62.0	50.8	73.8	65.6	65.8	65.2	76.1	69.1	63.1	73.2	66.4	60.3	63.6
72.1	65.4	67.6	68.9	67.2	59.0	67.9	65.2	66.8	66.9	63.9	68.6	72.7	61.5	66.6
75.7	65.7	70.4	55.7	57.4	52.5	67.4	64.7	69.6	68.6	59.6	68.3	73.5	63.4	70.3
79.8	68.1	72.2	65.6	63.9	67.2	71.2	64.1	70.7	69.1	54.9	65.6	76.4	66.0	71.9
75.3	71.7	73.9	57.4	70.5	67.2	65.2	72.3	70.7	65.6	71.6	72.1	74.6	68.3	74.4
73.0	67.2	62.6	65.6	59.0	52.5	64.7	66.8	73.9	68.6	62.6	74.0	69.2	61.5	64.9
70.0	65.2	64.9	62.3	68.9	62.3	64.1	64.7	66.3	68.6	63.4	67.5	69.3	64.1	63.2
70.0	70.1	65.7	62.3	63.9	60.7	66.8	57.6	63.6	65.0	63.9	65.6	68.9	66.2	64.2
76.7	70.4	72.8	62.3	68.9	67.2	63.6	69.0	64.1	66.7	64.5	67.5	73.4	69.4	71.5
75.4	70.4	72.2	57.4	70.5	68.9	67.9	62.0	72.8	71.0	67.8	73.0	71.4	66.4	70.8
76.5	72.6	77.4	63.9	70.5	60.7	59.2	65.8	66.3	65.6	68.6	73.2	74.9	71.2	76.0

Table 13 – Cross-validated accuracies of the individual k-nearest neighbors classifiers trained during the second iteration (moderate quality check), %. *Each model is made up of one k-nearest neighbor classifier per density distribution per attractor projection. Each row corresponds to one attractor projection, with its angular density (AD) in column one, radial density (RD) in column two and attractor outline (AO) in column three. Classification accuracies were determined via 10-fold cross validation. Classifiers with 70% or more accuracy are used in the prediction of an unknown sample and are highlighted in green.*

Challenge Model			Start of Dialysis			During Dialysis			Dialysis 24h			Combined Model		
AD	RD	AO	AD	RD	AO	AD	RD	AO	AD	RD	AO	AD	RD	AO
65.5	58.6	57.6	61.0	55.9	55.9	64.2	62.6	76.5	72.4	66.7	70.7	64.7	59.0	62.2
68.6	59.3	67.4	69.5	61.0	59.3	64.8	70.4	78.2	77.2	70.9	80.1	69.0	62.3	69.6
75.9	64.6	67.0	71.2	59.3	62.7	74.9	62.0	71.5	76.1	64.1	78.6	76.5	64.5	69.3
66.5	62.4	65.2	64.4	71.2	76.3	70.4	72.6	69.8	70.9	75.2	80.1	68.2	64.1	69.9
76.6	68.5	68.6	71.2	74.6	64.4	70.4	70.9	71.5	73.2	66.1	79.5	76.0	68.8	70.7
76.7	62.4	70.4	54.2	71.2	62.7	74.3	74.9	77.7	67.8	70.1	75.2	75.7	64.6	69.9
66.4	64.9	62.4	59.3	78.0	59.3	61.5	72.6	66.5	65.0	73.2	74.6	65.4	62.4	68.0
71.2	65.3	66.7	64.4	69.5	67.8	65.9	70.4	69.8	68.4	67.5	75.2	71.1	66.1	68.7
75.1	70.3	71.7	64.4	74.6	72.9	66.5	71.5	69.3	70.4	70.7	73.8	77.0	69.9	72.5
69.6	62.0	59.2	54.2	67.8	62.7	63.1	65.9	71.5	65.0	69.2	75.2	67.5	61.2	65.9
73.5	63.2	64.4	67.8	78.0	67.8	69.8	71.5	72.1	65.2	69.5	76.9	70.5	63.1	67.6
76.5	64.3	66.7	54.2	69.5	62.7	63.7	72.1	65.9	71.5	70.1	71.8	73.2	65.7	70.1
79.6	68.6	70.8	59.3	71.2	66.1	70.9	70.4	73.7	73.5	65.8	72.4	75.7	67.2	70.3
77.2	68.6	73.3	52.5	74.6	67.8	62.0	74.3	74.3	67.8	73.2	75.2	72.1	69.7	74.6
69.9	62.3	57.1	59.3	61.0	59.3	59.8	64.2	73.7	65.0	66.7	78.1	69.4	61.1	65.1
70.2	64.3	62.4	59.3	62.7	64.4	64.2	64.8	68.7	64.4	68.9	69.5	68.7	63.5	63.1
70.6	66.0	64.3	57.6	69.5	71.2	62.0	64.8	66.5	60.4	66.1	70.1	69.3	67.5	67.7
73.8	70.5	72.9	64.4	67.8	74.6	60.3	71.5	67.0	66.4	71.8	68.9	72.6	70.2	72.2
75.9	68.8	71.8	66.1	67.8	71.2	63.7	71.5	72.1	71.8	68.4	78.1	74.3	66.1	71.8
75.7	75.1	75.8	57.6	72.9	59.3	61.5	68.2	72.6	65.5	73.8	70.9	72.7	75.1	75.5

Table 14 – Cross-validated accuracies of the individual k-nearest neighbors classifiers trained during the third iteration (strict quality check), %. *Each model is made up of one k-nearest neighbor classifier per density distribution per attractor projection. Each row corresponds to one attractor projection, with its angular density (AD) in column one, radial density (RD) in column two and attractor outline (AO) in column three. Classification accuracies were determined via 10-fold cross validation. Classifiers with 70% or more accuracy are used in the prediction of an unknown sample and are highlighted in green.*

Angular Density	Radial Density	Attractor Outline
88.4 %	92.9 %	84.8 %
90.2 %	94.6 %	84.8 %
84.8 %	91.1 %	84.8 %
82.1 %	90.2 %	86.6 %
84.8 %	94.6 %	80.4 %
87.5 %	90.2 %	86.6 %
81.3 %	89.3 %	76.8 %
82.1 %	86.6 %	79.5 %
85.7 %	92.9 %	86.6 %
80.4 %	80.4 %	81.3 %
77.7 %	89.3 %	82.1 %
86.6 %	87.5 %	92.0 %
84.8 %	92.0 %	92.9 %
80.4 %	85.7 %	89.3 %
79.5 %	84.8 %	83.9 %
79.5 %	92.0 %	82.1 %
79.5 %	92.9 %	88.4 %
78.6 %	91.1 %	88.4 %
77.7 %	85.7 %	86.6 %
76.8 %	89.3 %	90.2 %

Table 15 - Cross-validated accuracies of the individual k-nearest neighbors classifiers trained for the AFT/PTB model. *Each row corresponds to one attractor projection, with its angular density in column one, radial density in column two and attractor outline in column three. Classification accuracies were determined via 10-fold cross validation. Classifiers with 70% or more accuracy are used in the prediction of an unknown sample and are highlighted in green.*