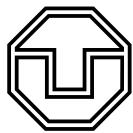


Entwicklung eines robusten Deep Learning Ansatzes zur Detektion von Vorhofflimmern in mobilen EKG-Aufnahmen

Antonia Gerdes

Masterarbeit



Entwicklung eines robusten Deep Learning Ansatzes zur Detektion von Vorhofflimmern in mobilen EKG-Aufnahmen

Antonia Gerdes

Geboren am: 21.02.1996 in Wilhelmshaven

Masterarbeit

zur Erlangung des akademischen Grades

Master of Science (M.Sc.)

Erstgutachter

Prof. Dr.-Ing. habil. H. Malberg

Zweitgutachter

Prof. Dr.-Ing. Raimund Dachselt

Betreuer

Dr.-Ing. Martin Schmidt

Dipl.-Wi.-Ing. Alexander Hammer

Betreuer Hochschullehrer

Prof. Dr.-Ing. habil. H. Malberg

Eingereicht am: 27. Februar 2025

Technische Universität Dresden
Fakultät Elektrotechnik und Informationstechnik
Institut für Biomedizinische Technik
Masterarbeit

Zielstellung der Arbeit

Vorhofflimmern (VHF) ist die verbreitetste anhaltende Herzrhythmusstörung in unserer Gesellschaft. Unbehandelt geht sie mit einer signifikant erhöhten Morbidität und Mortalität einher. VHF tritt anfänglich meist episodenweise (paroxysmales VHF) auf und chronifiziert sich mit der Zeit. Bei frühzeitiger Detektion von VHF ist eine effektive Intervention zur Prävention von Folgeerkrankungen möglich. Im Gegensatz zum Goldstandard, dem manuell ausgewerteten 12-Kanal-Elektrokardiogramm (EKG), ermöglichen mobile EKG-Patches ein kontinuierliches und umgebungsunabhängiges Monitoring der elektrischen Herzaktivität und erhöhen die Chancen der frühzeitigen Detektion von VHF besonders in Risikogruppen. Deep Learning-basierte Algorithmen erzielen hohe Genauigkeiten bei der Detektion von Vorhofflimmern anhand der Standard-EKG-Ableitungen. Allerdings ist ihre Übertragbarkeit auf Aufnahmen mobiler EKG-Patches begrenzt. Dies ist auf die im Vergleich zum Goldstandard reduzierte Kanalanzahl, die je nach Situation variierende Signalqualität und die bauformbedingte und platzierungsabhängige Signalmorphologie zurückzuführen. Besonders die Robustheit von Algorithmen gegen Letzteres ist für eine hohe Klassifikationsgenauigkeit in Aufnahmen mobiler EKG-Patches erforderlich. Das Ziel dieser Arbeit ist daher die Entwicklung eines Deep Learning-basierten Ansatzes zur Detektion von VHF auf Basis von Standard-12-Kanal-EKGs, der robust gegen signalmorphologische Veränderungen ist. Dieser Ansatz soll anschließend zur Validierung auf die Aufnahmen mobiler EKG-Patches, welche im Rahmen des TIMELY-Projekts entstanden sind, angewendet werden.

Im Rahmen der Arbeit sollen folgende Aufgaben bearbeitet werden:

- Einarbeitung in
 - o physiologische Grundlagen zu Vorhofflimmern und Möglichkeiten der signaltechnischen Erfassung
 - o Deep Learning Methoden zur Detektion von Vorhofflimmern im EKG
 - Identifikation von Ansätzen zur Realisierung der Übertragbarkeit von DL-basierten Ansätzen auf unbekannte und morphologisch veränderte Signale
 - Umsetzung und Evaluierung DL-basierter Ansätze zur Detektion von VHF
 - Bewertung und Diskussion der Übertragbarkeit auf Aufnahmen mobiler EKG-Patches des TIMELY-Projekts

Betreuer: Dipl.-Wi.-Ing. Alexander Hammer
Dr.-Ing. Martin Schmidt

1. Prüfer: Prof. Dr.-Ing. habil. H. Malberg
2. Prüfer: Prof. Dr.-Ing. Raimund Dachselt

ausgehändigt am: 15.08.2024 **einzureichen bis:** 16.01.2025

Prof. Dr.-Ing. habil. H. Malberg
Verantwortlicher Hochschullehrer

Selbstständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit mit dem Titel *Entwicklung eines robusten Deep Learning Ansatzes zur Detektion von Vorhofflimmern in mobilen EKG-Aufnahmen* selbstständig und ohne unzulässige Hilfe Dritter verfasst habe. Es wurden keine anderen als die in der Arbeit angegebenen Hilfsmittel und Quellen benutzt. Die wörtlichen und sinngemäß übernommenen Zitate habe ich als solche kenntlich gemacht. Mir ist bekannt, dass die Nichteinhaltung dieser Erklärung zum nachträglichen Entzug des Hochschulabschlusses führen kann.

Antonia Gerdes
Dresden, 27. Februar 2025

Abstrakt

Vorhofflimmern ist eine der häufigsten Herzrhythmusstörungen und bringt ein erhöhtes Schlaganfallrisiko mit sich, was zu bleibenden Schäden bis hin zum vorzeitigen Tod führen kann. Die frühzeitige Detektion von Vorhofflimmern ist wünschenswert, da es andernfalls zur Chronifizierung kommen kann. Der Goldstandard zur Diagnose von Vorhofflimmern ist das 12-Kanal-EKG, welches durch medizinisches Fachpersonal interpretiert wird. Um das lediglich phasenweise auftretende paroxysmale Vorhofflimmern zu diagnostizieren, ist ein Langzeit-EKG nötig, welches mittels mobilen EKG-Patches aufgezeichnet werden kann.

Die manuelle Auswertung dieser EKGs ist jedoch ressourcenintensiv, weshalb eine automatisierte Auswertung mittels Deep Learning-Algorithmen nötig ist. Mobile EKG-Patches haben außerdem eine reduzierte Kanalanzahl und die Signalmorphologie unterscheidet sich von der der 12-Kanal-EKGs. Aufgrund dieser Abweichungen sind bereits vorhandene Klassifikationsansätze auf Aufnahmen mobiler EKG-Patches nur begrenzt übertragbar. Um dieses Problem zu lösen, wird in dieser Arbeit ein Deep Learning-basierter Ansatz zur Detektion von Vorhofflimmern auf Basis von 12-Kanal-EKGs entwickelt, welcher robust gegen signalmorphologische Veränderungen ist.

Das entwickelte gewichtete Ensemble aus 5 Domain Adversarial Neural Networks erreicht einen F1-Score von bis zu 0,962 auf einem Testdatensatz aus 12-Kanal-EKGs, sowie einen Recall von bis zu 0,955 und eine Specificity von bis zu 0,973. Auf Aufnahmen aus dem TIMELY-Projekt, welche mittels mobiler EKG-Patches aufgezeichnet wurden, erreicht das Ensemble einen F1-Score von bis zu 0,986, sowie einen Recall von bis zu 1,00 und eine Specificity von bis zu 0,979.

Abstract

Atrial fibrillation is one of the most common cardiac arrhythmias and leads to a higher risk of stroke, which can cause permanent damage and premature death. The early detection of atrial fibrillation is desirable, as untreated atrial fibrillation can become chronic. The gold standard for atrial fibrillation diagnosis is the 12-lead ECG, which is interpreted by medical personnel. Because paroxysmal atrial fibrillation only occurs sporadically, a longterm ECG, which can be recorded via mobile ECG patches, is required.

The manual interpretation of longterm ECGs is resource intensiv, which is why an automated interpretation using deep learning algorithms is necessary. Mobile ECG patches have a reduced lead count and the morphology of their signals is different compared to those of 12-lead ECGs. Due to these differences, available classifiers have a limited usefulness on records of mobile ECG patches. To solve this problem, in this work a deep learning algorithm is developed, which is trained on 12-lead ECGs and is robust against a change in signal morphology.

The developed weighted ensemble of 5 domain adversarial neural networks achieves an F1-score of up to 0,962, a recall of up to 0,955 and a specificity of up to 0,973 when used on as test dataset of 12-lead ECGs. On ECGs taken from the TIMELY project, which are recorded via mobile ECG patches, it achieves an F1-score of up to 0,986, a recall of up to 1,00 and a specificity of up to 0,979.

Inhaltsverzeichnis

Abbildungsverzeichnis	viii
Tabellenverzeichnis	ix
Abkürzungsverzeichnis	x
1 Einleitung	1
2 Physiologische Grundlagen	4
2.1 Erregungsleitsystem des Herzens	4
2.2 Funktionsweise der Kardiomyozyten	6
2.3 Pathophysiologie von Vorhofflimmern	8
2.3.1 Auslöser von Vorhofflimmern	8
2.3.2 Aufrechterhaltung von Vorhofflimmern	10
2.4 Diagnostik von Vorhofflimmern	11
2.4.1 Charakteristika des gesunden Elektrokardiogramms	11
2.4.2 Vorhofflimmern im Elektrokardiogramm	12
3 Machine Learning	14
3.1 Grundlagen des Machine Learnings	14
3.2 Deep Learning	15
3.2.1 Artificial Neural Network	15
3.2.2 Convolutional Neural Network	17
3.2.3 Training von Lernalgorithmen	17

3.2.4	Optimierung der Hyperparameter	18
3.3	Bewertung von Klassifikatoren	19
3.3.1	Metriken zur Bewertung der Klassifikationsgüte	19
3.4	Domain Shift	21
3.4.1	Domain Adaptation und Transfer Learning	21
3.4.2	Domain Generalization	22
3.4.3	Domain Adversarial Learning	22
4	Stand der Technik	24
4.1	Ansätze zur Zeitreihenklassifikation	24
4.2	Ansätze zur Deep Learning-basierten Vorhofflimmern-Detektion in EKGs	26
4.3	Verwendung von Domain Generalization in der EKG-Klassifikation	27
5	Methodik	31
5.1	Genutztes Datenmaterial	31
5.2	Vorverarbeitung der Datensätze	33
5.3	Modellarchitektur	35
5.4	Trainings- und Optimierungsprozess	37
6	Ergebnisse	40
6.1	Ergebnisse der Hyperparameteroptimierung	40
6.2	Evaluierung auf Daten der Quelldomäne	42
6.3	Evaluierung auf Daten der Zieldomänen	45
6.4	Untersuchung des Einflusses der Normalisierung der Daten	47
7	Diskussion	49
7.1	Allgemeine Einordnung der Klassifikationsgüte der Modelle	49
7.2	Ergebnisse der Hyperparameteroptimierung	50
7.3	Diskussion der Ergebnisse auf unterschiedlichen Testdatensätzen	51
7.4	Einfluss von Normalisierung	57
8	Zusammenfassung und Ausblick	58
Quellenverzeichnis		x
Anhang		xvi

A Anhang	xvii
A.1 Suchstrings der systematischen Literaturrecherche	xvii
A.2 Grid Search Ergebnisse für InceptionTime	xix

Abbildungsverzeichnis

2.1	Darstellung des Erregungsleitsystems des Herzens	5
2.2	Darstellung von Aktionspotentialen verschiedener Zellen	7
2.3	Diagramm zur Pathophysiologie von VHF	9
2.4	Musterschlag im EKG	11
2.5	VHF im EKG	13
3.1	Feed Forward Neural Network	16
3.2	Domain Adversarial Nerual Network Prinzip	23
4.1	Critical Difference Diagramm aktueller Deep Neural Networks	25
4.2	Critical Difference Diagramm aktueller Deep Neural Networks und InceptionTime	25
4.3	Darstellung eines InceptionTime Klassifikators	26
4.4	Darstellung eines einzelnen Inception Moduls	27
5.1	DANN Architektur	36
5.2	DANNdirect Architektur	37
6.1	DANN Training Plots	42
6.2	Label Predictor ROC-Kurve	44
6.3	Domain Classifier ROC-Kurven	44
7.1	Plot einer Icentia11k-Aufnahme mit geringer Auflösung	52
7.2	Plot einer verrauschten Icentia11k-Aufnahme	52
7.3	SHDB-AF F1-Score vs. durchschnittliche Sicherheit für VHF	54
7.4	SHDB-AF Aufnahme von Patient 003 mit AFIB Annotation	54
7.5	SHDB-AF Aufnahme von Patient 003 mit AFL Annotation	55
7.6	SHDB-AF Aufnahme von Patient 111 mit AT Annotation	55
7.7	SHDB-AF F1-Score vs. durchschnittliche Sicherheit für nicht-VHF	56

Tabellenverzeichnis

4.1 Deep Learning Ansätze zur VHF-Klassifikation	27
4.2 Deep Learning Ansätze zur DG in der EKG-Klassifikation	30
5.1 Klassenverteilung der xECGArch-Datenbank	32
5.2 Klassenverteilung im lcentia11k-Datensatz	34
5.3 Klassenverteilung in der SHDB-AF-Datenbank	34
5.4 Klassenverteilung im TIMELY-Datensatz	35
5.5 Klassenverteilung Cross Validation	39
6.1 Hyperparameteroptimierung des DANNs	41
6.2 Ergebnisse der Evaluation auf der Quelldomäne xECGArch	42
6.3 Ergebnisse der Evaluation auf der Quelldomäne xECGArch pro Ableitung	43
6.4 Ergebnisse der Evaluation auf der Zieldomäne lcentia11k	45
6.5 Recall und Specificity lcentia11k	45
6.6 Ergebnisse der Evaluation auf der Zieldomäne TIMELY	46
6.7 Recall und Specificity TIMELY	46
6.8 Ergebnisse der Evaluation auf der Zieldomäne SHDB-AF	47
6.9 Recall und Specificity SHDB-AF	47
6.10 Ergebnisse der Evaluation auf der Quelldomäne mit nicht-normalisierten Daten	48
6.11 Ergebnisse der Evaluation auf nicht-normalisierten TIMELY-Daten	48
A.1 Ergebnisse der InceptionTime Grid Search	xix

Abkürzungsverzeichnis

ANN	Artificial Neural Network
AUC	Area Under Curve
AV-Knoten	Atrioventrikulknoten
CNN	Convolutional Neural Network
CPSG2018	China Physiological Signal Challenge 2018
DANN	Domain Adversarial Neural Network
DAp	Domain Adaptation
DG	Domain Generalization
DL	Deep Learning
DNN	Deep Neural Network
EKG	Elektrokardiogramm
FC	Fully Connected
FFNN	Feed Forward Neural Network
FN	False Negatives
FP	False Positives
GAP	Global Average Pooling
GRL	Gradient Reversal Layer
HPC	High Performance Computer
HRV	Herzratenvariabilität
LSTM	Long Short-Term Memory
ML	Machine Learning
PPG	Photoplethysmographie
ReLU	Rectified Linear Unit
ResNet	Residual Network
ROC	Receiver-Operating-Characteristic
SHDB-AF	Saitama Heart Database Atrial Fibrillation
TN	True Negatives
TP	True Positives
UCR	University of California, Riverside

UEA	University of East Anglia
VHF	Vorhofflimmern

1 Einleitung

Vorhofflimmern (VHF) ist nach ventrikulären und supraventrikulären Extrasystolen die häufigste Herzrhythmusstörung [1]. Die Global Burden of Disease 2010 Study [2] gibt an, dass die geschätzte Zahl von an VHF leidenden Menschen weltweit im Jahr 2010 bei 33,5 Millionen lag, von denen die Zahl der Männer 20,9 Millionen und die Zahl der Frauen 12,6 Millionen beträgt. Weiterhin prognostiziert die Studie in den USA mehr als eine Verdopplung der VHF-Fälle bis zum Jahr 2050. VHF erhöht unbehandelt aufgrund von Thrombenbildung das Schlaganfallrisiko signifikant, was zu bleibenden Beeinträchtigungen bis hin zum vorzeitigen Tod führen kann [1]. Die altersstandardisierte Mortalitätsrate weltweit hat im Zeitraum von 1990 bis 2021 von 1,46 auf 1,50 pro 100 000 Menschen stetig zugenommen [3].

Die Detektion von VHF stellt eine Herausforderung dar, denn gerade zu Beginn tritt VHF lediglich phasenweise auf und endet von allein (paroxysmales VHF). Eine frühzeitige Diagnose von VHF ist wünschenswert, da sich aus unbehandeltem paroxysmalem VHF ein persistentes und aus diesem ein permanentes, also ein ununterbrochen andauerndes und nicht mehr konventionell behandelbares, VHF entwickeln kann. Der Goldstandard zur VHF-Diagnose ist das 12-Kanal-Elektrokardiogramm (EKG), welches durch medizinisches Fachpersonal interpretiert wird. Aufgrund des episodischen Auftretens von paroxysmalem VHF gekoppelt mit der Tatsache, dass 50% der Patienten¹ keinerlei Symptome verspüren [1], bleibt paroxysmales VHF mit konventionellen Elektrokardiographen häufig undagnostiziert. [4]

Aufgrund dieser Probleme erfordert die Detektion von paroxysmalem VHF die Analyse von Langzeit-EKGs, welche bspw. mit mobilen EKG-Patches aufgenommen werden können. Im Vergleich zu herkömmlichen 12-Kanal-Langzeit-EKG-Rekordern sind mobile EKG-Patches weniger invasiv und im Alltag tragbar. Die manuelle Auswertung eines Langzeit-EKGs durch medizinisches Fachpersonal ist zeitaufwändig, kostenintensiv und aufgrund der monotonen Arbeit fehleranfällig. Deshalb ist eine automatisierte Auswertung notwendig, für welche bspw. Machine Learning (ML)- oder Deep Learning (DL)-Algorithmen genutzt werden können.

Mobile EKG-Patches haben meist eine reduzierte Kanalanzahl und aufgrund der Positionierung der Elektroden unterscheidet sich die Signalmorphologie der Ableitungen von der der 12-Kanal-EKGs. Aufgrund dieser Abweichungen zu klassischen EKG-Aufnahmen sind bereits

¹In dieser Arbeit wird aufgrund der besseren Lesbarkeit bewusst auf eine geschlechtsneutrale Formulierung verzichtet. Sämtliche männliche Schreibweisen beziehen sich dabei gleichermaßen auf alle Geschlechter.

vorhandene Klassifikationsansätze auf Aufnahmen mobiler EKG-Patches nur begrenzt übertragbar. Die Robustheit von Algorithmen gegen signalmorphologische Veränderungen ist wichtig für eine hohe Klassifikationsgenauigkeit in Aufnahmen mobiler EKG-Patches.

Ziel dieser Arbeit ist die Entwicklung eines DL-basierten Ansatzes zur Detektion von VHF auf Basis von Standard-12-Kanal-EKGs, welcher robust gegen signalmorphologische Veränderungen ist. Der Ansatz wird mit Aufnahmen aus dem TIMELY Projekt² [5] evaluiert, in dessen Rahmen Aufnahmen von Patienten mit koronarer Herzkrankheit mit mobilen EKG-Patches aufgezeichnet wurden. Der Ansatz wird außerdem auf zwei weiteren öffentlich zugänglichen Datenbanken, bestehend aus Aufnahmen mobiler EKG-Rekordern, mit annotierten Vorhofflimmerepisoden getestet.

Als Ansatz für ein solches DL-Modell wird Domain Adversarial Learning auf Basis von InceptionTime [6] genutzt. Ein Domain Adversarial Neural Network (DANN) wird darauf optimiert, domäneninvariante Merkmale zu erlernen, wobei die Domäne in diesem Fall die EKG-Ableitung ist. Somit ist das Modell robuster gegen Veränderungen der Domäne. [7]

Da als Domänen in dieser Arbeit die verschiedenen Ableitungen eines EKGs genutzt werden und die Signalmorphologie von Ableitung zu Ableitung unterschiedlich ist, wird die Vermutung aufgestellt, dass solch ein Modell ebenfalls robuster gegenüber der veränderten Signalmorphologie eines mobilen EKG-Patches ist. Das so entwickelte DANN wird im Ensemble betrieben. Als Vergleichsmodell wird ein klassisches InceptionTime Ensemble verwendet. Im Rahmen dieser Arbeit werden folgende Hypothesen aufgestellt:

1. Haupthypothese: Ein DANN erzielt auf EKG-Aufnahmen mit veränderter Morphologie eine bessere Leistung als das InceptionTime Vergleichsmodell.
2. Teilhypotesen
 - a) Es ist möglich, ein DANN zu erstellen und mit 12-Kanal-EKGs zu trainieren.
 - b) Das DANN ist in der Lage, VHF in Standard-12-Kanal-EKGs zu detektieren.
 - c) Das DANN ist in der Lage, VHF in morphologisch veränderten EKGs zu detektieren.
 - d) Das DANN erzielt eine bessere Leistung, wenn es im Ensemble genutzt wird.
 - e) Das DANN erzielt auf dem Testdatensatz der 12-Kanal-EKGs eine schlechtere Leistung als das Vergleichsmodell, da es aufgrund des Domain Adversarial Learnings weniger stark overfittet.

Die Arbeit ist wie folgt aufgebaut: Zunächst werden in Kapitel 2 physiologische Grundlagen des Erregungsleitsystems des Herzens sowie die Pathophysiologie und Diagnostik von VHF erläutert. Anschließend werden in Kapitel 3 für das Verständnis der Arbeit notwendige Grundlagen im Bereich ML und DL geschaffen, sowie das Domain Shift-Problem und Lösungsansätze mit Hilfe von Domain Generalization (DG), insbesondere Domain Adversarial Learning, erklärt. In Kapitel 4 wird ein Überblick über den aktuellen Stand der Technik im Bereich Zeitreihenklassifikation, EKG-Klassifikation mittels DL und DG in der EKG-Klassifikation gegeben. Genutztes Datenmaterial und wie es vorverarbeitet wurde, sowie Modellarchitektur und Trainingsprozess

²<https://www.timely-project.com/>

werden in Kapitel 5 vorgestellt. Die Ergebnisse des Trainings, der Evaluation mit 12-Kanal-EKGs, sowie die der Evaluation mit Aufnahmen von mobilen EKG-Patches werden in Kapitel 6 präsentiert. Die Bewertung und Diskussion der Ergebnisse findet in Kapitel 7 statt. Zusammenfassung und Ausblick sind in Kapitel 8 zu finden.

2 Physiologische Grundlagen

Um einen Algorithmus zur automatisierten Detektion von VHF im EKG zu designen, sowie die durch den Algorithmus erstellten Klassifikationsergebnisse zu interpretieren, wird das Verständnis von physiologischen Grundlagen der Erregungsabläufe im Herzen sowie der Pathophysiologie bei VHF benötigt. Aus diesem Grund wird in diesem Kapitel zuerst in Abschnitt 2.1 das Erregungsleitsystem des gesunden Herzens erläutert, sowie in dem darauffolgenden Abschnitt 2.2 die Funktionsweise von Kardiomyozyten. Die Entstehung von VHF wird in Abschnitt 2.3 erklärt. Anschließend wird in Abschnitt 2.4 auf die signaltechnische Erfassung, sowie die Diagnostik von VHF mittels EKG eingegangen.

2.1 Erregungsleitsystem des Herzens

Das Herz dient der Blutzirkulation im Körper und stößt im Ruhezustand durchschnittlich mit einer Frequenz von 70 Schlägen/min sauerstoffreiches Blut in den Körper- und sauerstoffarmes Blut in den Lungenkreislauf aus [8]. Es teilt sich durch die Herzscheidewand (Septum) in zwei Hälften, welche jeweils einen Vorhof (Atrium) und eine Kammer (Ventrikel) besitzen. Zwischen dem rechten Vorhof und der rechten Kammer befindet sich die Trikuspidalklappe, zwischen dem linken Vorhof und der linken Kammer die Bikuspidalklappe (Mitralklappe). In den vom Herzen wegführenden Gefäßen (der Aorta und der Lungenschlagader Truncus pulmonalis) befinden sich Taschenklappen. [9]

Die Herzmuskulatur (Myokard) besitzt die Fähigkeit, ohne externe Nervenimpulse Erregungen auszubilden und zu kontrahieren. Die Herzmuskelzellen (Kardiomyozyten) in verschiedenen Bereichen des Herzens besitzen diese Fähigkeit je nach Zelltyp in unterschiedlich starken Ausprägungen. Spezialisierte Myozyten im Herzen bilden das Erregungsleitungssystem, welches für eine koordinierte Erregungsbildung und -weiterleitung zuständig ist. Dies ist in Abb. 2.1 schematisch dargestellt. Als Arbeitsmyokard werden spezialisierte Myozyten bezeichnet, welche zur Kontraktion fähig sind [1]. Der Sinusknoten, welcher im rechten Vorhof in Höhe der Einmündung der oberen Hohlvene liegt, besteht aus Myozyten zur Erregungsbildung. Er dient als primäres Erregungszentrum des Myokards und bildet 60-80 Erregungen pro Minute aus. Die vom Sinusknoten ausgehenden Reize breiten sich in das Arbeitsmyokard der Vorhöfe

aus, welches nach ca. 90 ms vollständig erregt ist und kontrahiert [8]. [9]

Das Herzskelett trennt die Vorhöfe vollständig von den Kammern, sodass auch Erregungen nicht von den Vorhofmyozyten auf die Kammermyozyten übergehen können. Lediglich der Atrioventrikulknoten (AV-Knoten), welcher am Übergang zwischen Vorhofseptum und Ventrikelseptum lokalisiert ist, ist in der Lage, Erregungen von den Vorhöfen in die Kammern weiterzuleiten [9]. Die Myozyten des AV-Knotens besitzen mit 0,05 m/s eine geringere Reizweiterleitungsgeschwindigkeit als Vorhofmyozyten, bei welchen sie bei 0,5 m/s liegt. Durch die verzögerte Reizweiterleitung in die Kammern ist die Kontraktion der Vorhöfe beendet, bevor die Kammern kontrahieren. Auf diese Weise wird sichergestellt, dass die Kammern vor der Kontraktion ausreichend mit Blut gefüllt sind. Die verzögerte Reizweiterleitung sorgt bei pathologisch hohen Erregungsfrequenzen im Vorhof dafür, dass nicht jede Erregung in das Kammermyokard weitergeleitet wird, sodass sich die Kammern trotz dessen vor der Kontraktion ausreichend füllen können. Der AV-Knoten verfügt wie der Sinusknoten über die Fähigkeit, autonom Erregungen zu erzeugen. Kommt es durch pathophysiologische Bedingungen zu einem Ausfall des Sinusknotens, kann der AV-Knoten mit einer Frequenz von 40-50 Erregungen pro Minute als sekundärer Schrittmacher dienen. [8]

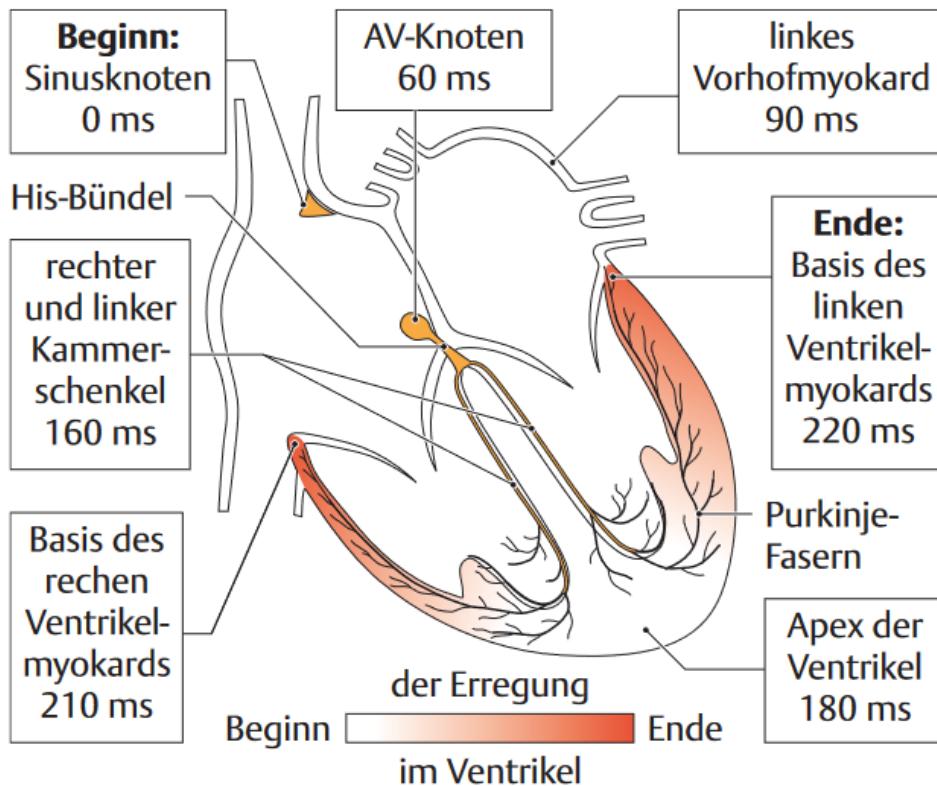


Abb. 2.1: Darstellung des Erregungsleitsystems des Herzens. Zu sehen ist die Erregungsausbreitung vom Beginn im Sinusknoten bis zum Ende im Myokard des linken Vorhofes. Zusätzlich angegeben ist die Dauer der Erregungsausbreitung. Entnommen aus [8].

Vom AV-Knoten aus wird die Erregung über das His-Bündel geleitet, über den linken und rechten Tawara-Schenkel am Kammerseptum entlang und in die Purkinje-Fasern hinein. Dieser Teil des Erregungsleitungssystems ist von den Kammermyozyten elektrisch isoliert. Erst die Purkinje-Fasern verzweigen sich ins Kammermyokard und übertragen den Reiz an die Kam-

mermyozyten. Die Zeitspanne von Erregungsbildung im Sinusknoten bis zur vollständigen Erregung der Kammermyozyten beträgt ca. 220 ms. Im Anschluss an die Kammererregung erfolgt die Kammerkontraktion. [8]

2.2 Funktionsweise der Kardiomyozyten

Wie bereits im vorherigen Abschnitt 2.1 erwähnt, gibt es verschiedene Typen von Kardiomyozyten, welche unterschiedliche Funktionen aufweisen: Erregungsbildung (Sinusknoten), Erregungsweiterleitung (AV-Knoten und Purkinjefasern) und Kontraktion (Arbeitsmyokard der Vorhöfe und Kammern). Die Myozyten zur Erregungsbildung und -weiterleitung besitzen außerdem die Fähigkeit, autonom einen Reiz auszulösen, die Myozyten des Arbeitsmyokards nicht. Die Ursache hierfür liegt in den unterschiedlichen Aktionspotentialen der Zelltypen begründet. Ein Aktionspotential ist eine starke Änderung des Membranpotentials einer Zelle und dient der Informationsübertragung. Der Verlauf der Aktionspotentiale (siehe Abb. 2.2) wird in die Phasen 0 bis 4 eingeteilt, welche im Folgenden näher erläutert werden. [8]

Eine Zelle des Arbeitsmyokards ist in Ruhe bei -90 mV polarisiert (Ruhemembranpotential), dies wird als Phase 4 bezeichnet. Erreicht die Zelle ein elektrischer Reiz, depolarisiert sie auf 0 mV, wobei Natrium-Ionen durch spezifische Kanäle in die Zelle hinein fließen. Dies ist Phase 0 und der Beginn des Aktionspotentials. Während der Depolarisation kommt es zu einem Overshoot, bei welchem die Zelle kurzzeitig bei 20 mV polarisiert ist (Phase 1). [1]

Auf den Overshoot folgt eine kurze Repolarisation mit anschließendem Plateau im Aktionspotential (Phase 2). Die initiale Repolarisation kommt durch das Fließen von Kalium-Ionen aus der Zelle heraus zustande. Im Laufe dessen öffnen sich Calcium-Ionen-Kanäle, durch die Calcium-Ionen in die Zelle hinein fließen und den Auswärtsstrom der Kalium-Ionen ausgleichen, wodurch das Plateau entsteht. Der Einwärtsstrom der Calcium-Ionen leitet die Kontraktion des Myozyts ein. [8]

In Phase 3 repolarisiert die Zelle durch einen Kalium-Ionen Auswärtsstrom bis sie sich wieder im Ruhemembranpotential, also in Phase 4, befindet. Die Zelle verbleibt in Phase 4, bis ein erneuter Reiz zur Depolarisation führt. Aufrechterhalten wird das Ruhemembranpotential durch einen kontinuierlichen Austausch von Natrium- und Kalium-Ionen zwischen Intra- und Extrazellulärraum. [1]

Übertragen wird ein Reiz von benachbarten Myozyten über sogenannte Gap Junctions, die aus einer Ansammlung unspezifischer Ionenkanäle bestehen, die Connexone genannt werden [8].

Im Gegensatz zu einem Myozyt des Arbeitsmyokards können Sinusknotenmyozyten und Myozyten des Reizleitungsgewebes autonom eine Erregung generieren. Diese Fähigkeit besitzen sie, da sie kein stabiles Ruhemembranpotential haben, sondern in Phase 4 langsam depolarisieren (Spontandepolarisation [10]). Diese Depolarisation wird hauptsächlich durch HCN-Kanäle (hyperpolarization activated cyclic nucleotide gated channels) verursacht, durch welche Ionen in die Zelle hineinfließen. Das maximale negative Potential eines Sinusknoten- oder Reizleitungsmozyts liegt bei -60 mV. Erreicht die Zelle bis zu einem Schwellwert von ca. -40 mV kein Reiz, depolarisiert sie autonom [10]. Sinusknoten- und AV-Knoten-Myozyten be-

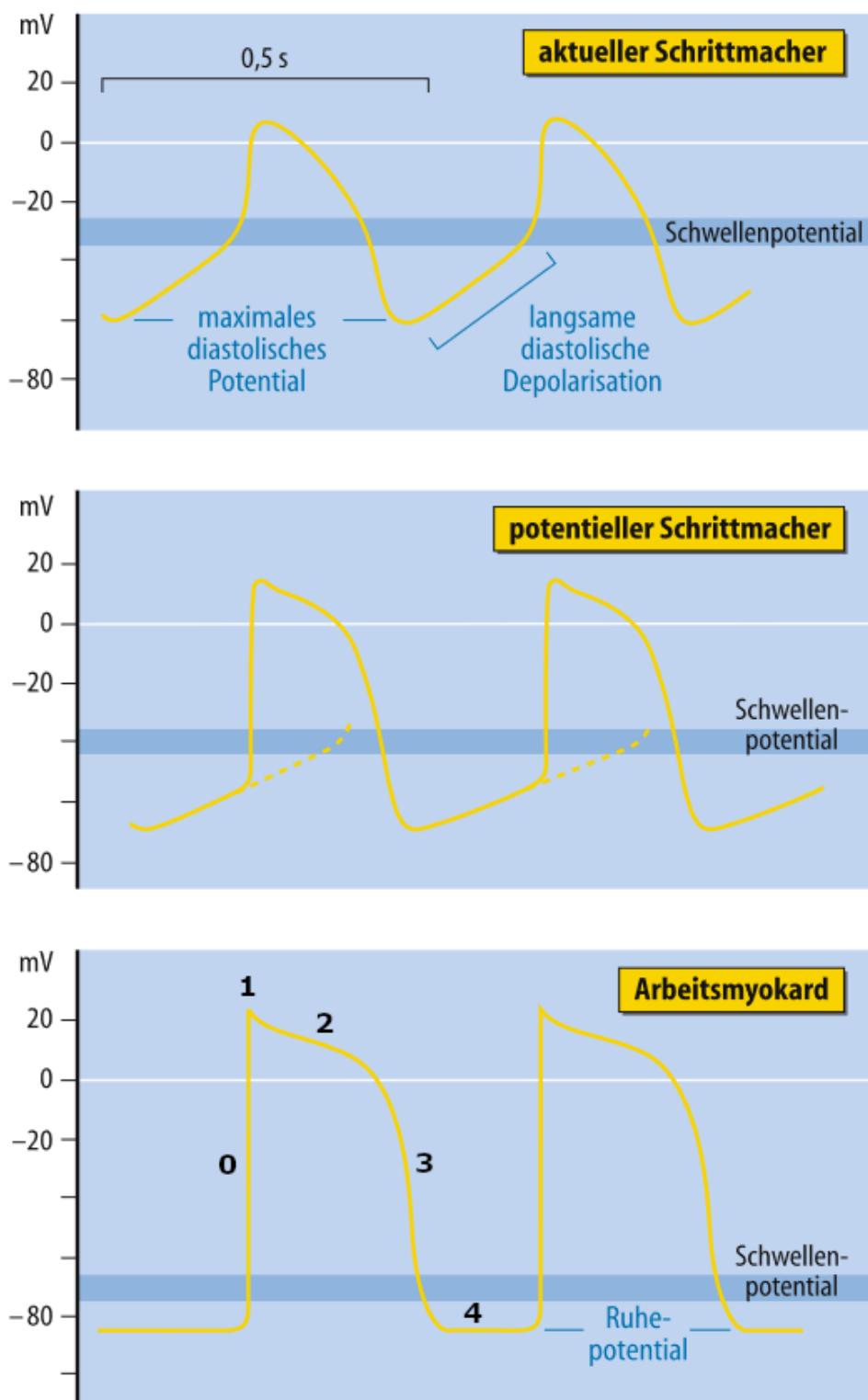


Abb. 2.2: Darstellung der Aktionspotentiale einer Sinusknotenzelle (oben), einer Erregungsleitungszelle (mitte) und einer Zelle des Arbeitsmyokards (unten), abgeändert entnommen aus [10]. Die Phasen 0 bis 4 des Aktionspotentialverlaufs sind eingetragen.

sitzen keine Phase 1 und 2 im Aktionspotential. Bei Myozyten des Sinusknotens wird dieser Schwellwert schneller erreicht, als bei Myozyten des Reizleitungsgewebes, weshalb der Sinusknoten als der primäre Schrittmacher fungiert. Fällt der Sinusknoten aus, kann der AV-Knoten

als sekundärer Schrittmacher dienen, fällt auch dieser aus, können Teile des Erregungsleitsystems der Ventrikel als tertiärer Schrittmacher einspringen. [8]

Kardiomyozyten sind für einige Zeit nach Beginn der Phase 0 nicht erregbar. Diese Zeit wird als absolute Refraktärphase bezeichnet. Erst wenn das Aktionspotential in Phase 3 wieder bei ca. -40 mV liegt, ist die Zelle wieder durch Reize erregbar. Dies ist die relative Refraktärphase. Um die Zelle in der relativen Refraktärphase zu erregen, ist jedoch ein erheblich stärkerer Reiz notwendig als vom Ruhemembranpotential aus. Da die Refraktärphase länger ist als die Dauer, die zur Depolarisation benötigt wird, wird durch sie eine rückläufige Erregungsausbreitung im Herzen verhindert. [8]

Das Herz muss in der Lage sein, sich dem Bedürfnis des Körpers nach einer erhöhten Sauerstoffversorgung, bspw. beim Sport, anzupassen. Dies wird unter anderem durch eine Erhöhung der Herzfrequenz und der Arbeitslast des Herzens erreicht. Dies ist möglich, da das Herz von Nerven des autonomen Nervensystems innerviert ist. Sympathikus und Vagusnerv haben Einfluss auf die Geschwindigkeit der Spontandepolarisation der Sinusknotenzellen, wodurch eine Frequenzsteigerung (positive Chronotropie) durch Stimulation durch den Sympathikus oder Frequenzsenkung (negative Chronotropie) durch Aktivität des Vagusnervs erzielt wird. Ebenso wird eine Erhöhung der Reizleitungsgeschwindigkeit im AV-Knoten (positive Dromotropie) durch Stimulation durch den Sympathikus bzw. Verringerung der Reizleitungsgeschwindigkeit (negative Dromotropie) durch Stimulation durch den Vagusnerv erzielt. [10]

2.3 Pathophysiologie von Vorhofflimmern

VHF ist eine Herzrhythmusstörung, bei der die Vorhoferregung gestört ist und die Vorhöfe somit schnell und unregelmäßig kontrahieren. Durch diese unkoordinierten Vorhofaktionen wird unter Belastung oder bei fortgeschrittenem VHF die Füllung der Ventrikel erheblich beeinträchtigt, sodass nicht bei jeder Kammerkontraktion genug Blut ausgestoßen wird. [4]

VHF wird unterteilt in paroxysmales, persistierendes und permanentes VHF, wobei diese Charakterisierung die Dauer der VHF-Episode beschreibt. Als paroxysmal wird VHF bezeichnet, wenn es unter einer Woche andauert und eigenständig endet (spontane Konversion [1]). Dauert das VHF länger als 7 Tage an, ist eine Kardioversion nötig und das VHF wird als persistierend bezeichnet. Persistierendes VHF kann in permanentes VHF übergehen, wenn die Kardioversion nicht erfolgreich (oder gar nicht) durchgeführt wurde. Im Falle des permanenten VHF ist dieses dauerhaft bestehend. [11]

Die Entstehung von VHF ist ein Zusammenspiel aus Auslösern und Aufrechterhaltungsmechanismus (siehe Abb. 2.3) durch verändertes Gewebe im Vorhof. Die Veränderung des Gewebes zeichnet sich durch Umbau der elektrischen Reizleitungsbahnen sowie strukturelle Veränderungen des Vorhofgewebes aus (auch als Remodelling bezeichnet). [12]

2.3.1 Auslöser von Vorhofflimmern

Paroxysmale VHF-Episoden werden durch spontane Depolarisationen von Gewebe mit erhöhter Automatik in oder außerhalb der Vorhöfe (bspw. in der Lungenvene) verursacht. Myozyten,

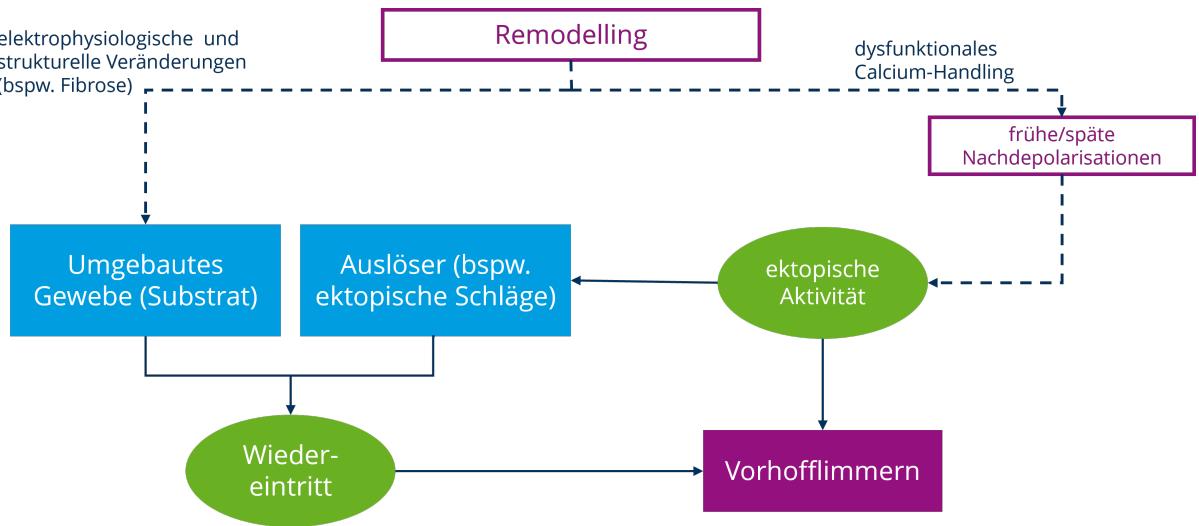


Abb. 2.3: Schematische Darstellung von Mechanismen, die zu Vorhofflimmern führen. Eine dysfunktionales Calcium-Handling führt zu frühen oder verspäteten Nachdepolarisationen, welche ektopische Aktivität auslösen. Findet die ektopische Aktivität in hoher Frequenz statt, kann sie Vorhofflimmern aufrecht erhalten. Sie kann auch eine kreisende Erregung verursachen, die Vorhofflimmern aufrecht erhält. Auch ein struktureller und elektrophysiologischer Umbau des Vorhofmyokards kann zu kreisenden Erregungen und damit zu Vorhofflimmern führen. Entnommen und übersetzt aus [12].

die eine erhöhte Automatik aufweisen, haben ein weniger niedriges maximales negatives Potential, die spontane Phase-4-Depolarisation läuft schneller ab und ein niedrigeres Schwellenpotential ist vorhanden. Nachdem ein Myozyt mit erhöhter Automatik einen Reiz generiert hat, kann es zu sogenannten frühen Nachdepolarisationen oder verzögerten Nachdepolarisationen kommen. Eine frühe Nachdepolarisation tritt während eines übermäßig verlängerten Aktionspotentials in Phase 3 auf, welches dafür sorgt, dass die deaktivierte Calcium-Kanäle sich erholen und erneut zu einem Calcium-Ionen-Fluss in den Myozyt hinein führen [13]. Eine verzögerte Nachdepolarisation tritt auf, nachdem ein Myozyt bereits zum Ruhemembranpotential zurückgekehrt ist. Sie wird verursacht durch einen Calcium-Überschuss innerhalb der Zelle, der durch Natrium-Calcium-Austauscher zu einem positiven Natrium-Ionen-Fluss in die Zelle hinein führt. Sind die Nachdepolarisationen stark genug, um das Schwellenpotential zu überschreiten, wird ein erneutes Aktionspotential ausgelöst. [14]

Die ektopischen Depolarisationen zeichnen sich mit 350 bis 600 Reizen pro Minute durch eine höhere Frequenz als die des Sinusknotens aus [15]. Hauptausgangspunkt der ektopischen Depolarisationen stellen in 94% der Fälle die Pulmonalvenen dar [16]. Dies liegt daran, dass das Gewebe der Pulmonalvenen andere elektrophysiologische Eigenschaften aufweist, als das Vorhofgewebe [17]. Es konnten Myozyten in den Pulmonalvenen nachgewiesen werden, die eine besondere Leitfähigkeit aufweisen, bspw. Schrittmacherzellen und Purkinje-Zellen [18]. Neben den Pulmonalvenen kann VHF auch durch anders verortete ektopische Depolarisationen ausgelöst werden, wie bspw. am Vorhofseptum oder an der posterioren Wand des linken Vorhofs [19].

Eine Unausgeglichenheit des autonomen Nervensystems kann als Auslöser der ektopischen Aktivität dienen. Hammer et al. [20] zeigen, dass VHF-Episoden eine erhöhte Aktivität des Sym-

pathikus vorausgeht, gefolgt von einer erhöhten Aktivität des Vagusnervs. Eine durch dysfunktionale Sympathikusaktivität erhöhte Ausschüttung von Calcium führt zu einer erhöhten Frequenz von Aktionspotentialen. Eine dysfunktionale Aktivität des Vagusnervs wiederum führt zu einer verkürzten Aktionspotentialdauer und Refraktärperiode. [21]

2.3.2 Aufrechterhaltung von Vorhofflimmern

VHF kann durch die im vorherigen Abschnitt erwähnten ektopischen Aktivitäten aufrechterhalten werden, wenn diese schnell und andauernd stattfinden. Mit länger anhaltendem VHF kommt es zu einem fortschreitendem Umbau des Vorhofmyokards, welcher wiederum VHF begünstigt [22]. Hat sich das Vorhofgewebe bereits umgebaut, kann es durch dieses anfällige Substrat zu einer Reentry-Störung im Vorhof kommen. Substrat, welches empfänglich für Reentry-Störungen ist, zeichnet sich durch strukturelle und elektrophysiologische Veränderungen aus. Es kommt zu kreisenden Erregungen, da einige Leitungsbahnen eine kürzere Refraktärperiode und veränderte Leitgeschwindigkeit aufweisen, sodass die Depolarisationswelle wiederholt auf bereits wieder erregbare Myozyten trifft. Elektrophysiologische Veränderungen können etwa eine Verkürzung der Refraktärperiode, beschleunigte Repolarisation oder eine abnormale Reizleitung durch Veränderungen in den die Myozyten verbindenden Connexonen sein. Die wichtigste strukturelle Veränderung ist eine Dilatation der Vorhöfe und damit einhergehende Fibrose. [12]

Es herrscht keine Einigkeit über den Mechanismus, welcher zu langanhaltendem VHF führt. Aktuelle Hypothesen bezüglich Reentry-Störungen bei VHF beschreiben unabhängige Wavelets, Reentrant Rotors und die Double Layer Hypothese [23].

Moe und Abildskov [24] haben die Hypothese aufgestellt, dass von einer ektopischen Aktivität eine Depolarisationswelle ausgeht, die sich durch das Vorhofmyokard fortpflanzt und sich aufgrund unterschiedlicher Refraktärperioden der Myozyten in mehrere voneinander unabhängige Wavelets aufspaltet. Diese Wavelets propagieren sich chaotisch weiter durch Gewebe der Vorhöfe, welches bereits erregbar ist, und können sich erneut an refraktärem Gewebe aufspalten oder mit anderen Wavelets zusammenführen.

Unter einem Rotor wird eine um ein nicht erregtes Zentrum kreisende Erregungsausbreitung verstanden. Dieses Zentrum kann sowohl aus nicht erregbarem (bspw. fibrotischem) Gewebe bestehen, aber auch potentiell erregbar sein. Ebenso können Rotoren einen Drift aufweisen und sich somit durch das Vorhofmyokard bewegen. Die Wavelet- und Rotoren-Hypothesen schließen sich nicht gegenseitig aus, da sowohl Wavelets als auch lokalisierte ektopische Depolarisationen einen Rotor auslösen können. [25]

Die Double Layer Hypothese beschreibt eine longitudinale Dissoziation im Vorhofmyokard, ausgelöst durch parallel zu den Muskelfasern verlaufendes Gewebe mit blockierter Leitfähigkeit. Dies führt zu Bereichen im Vorhofmyokard, die unabhängig voneinander erregbar sind, sodass sich entlang dieser Bereiche Erregungen propagieren können. [26]

2.4 Diagnostik von Vorhofflimmern

Durch die eingeschränkte Pumpfunktion der Vorhöfe beim VHF kann es zu Thrombenbildung in den Vorhöfen kommen. Lösen sich diese und wandern durch den Körper, können sie einen ischämischen Schlaganfall verursachen. Eine frühzeitige Diagnose von VHF kann dieses Risiko vermindern. [11]

Die Symptome von VHF sind vielfältig und reichen von Herzrasen über Synkopen und Schmerzen im Brustkorb bis zu Erschöpfung. Da VHF jedoch auch asymptomatisch verlaufen kann, ist die Diagnose besonders schwierig. [27]

Das EKG ist ein diagnostisches Hilfsmittel, mit welchem Rückschlüsse auf die Reizbildung und -weiterleitung im Herzen gezogen werden können. Es beinhaltet außerdem Informationen über Herzlage, Herzfrequenz und Erregungsrhythmus. [28]

2.4.1 Charakteristika des gesunden Elektrokardiogramms

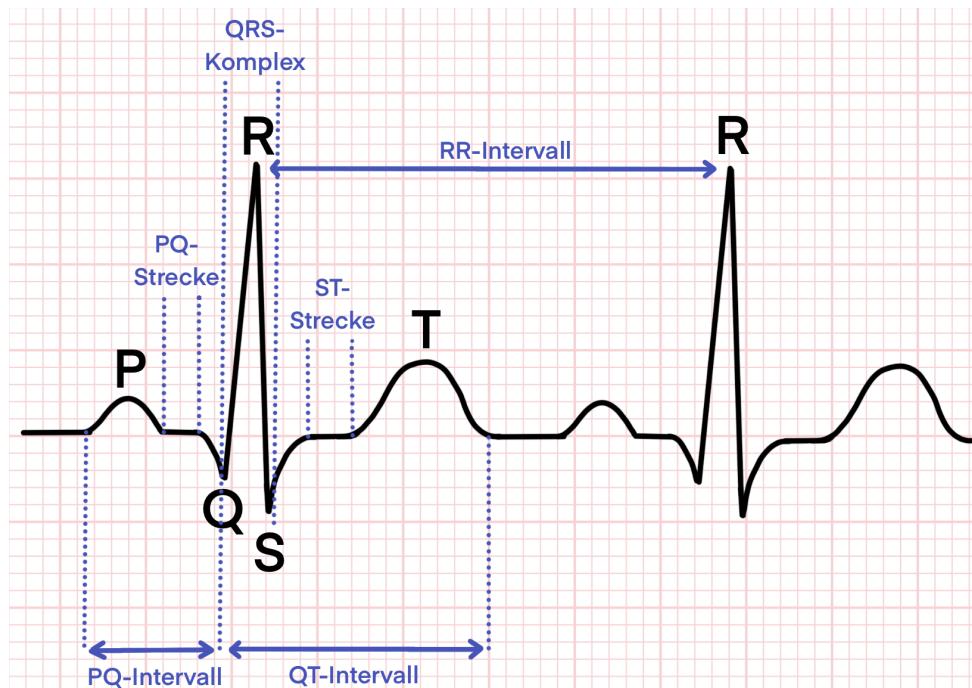


Abb. 2.4: Darstellung der Intervalle innerhalb eines Musterschlags im Elektrokardiogramm und eines RR-Intervalls, das als Abstand zwischen zwei Kammerdepolarisationen (R-Zacken) definiert ist. Abbildung mit Änderungen angelehnt an [4].

Die Depolarisation im Herzen führt zu einem elektrischen Feld an der Körperoberfläche. Zwischen verschiedenen Punkten am Körper entstehen so Potentialdifferenzen. Zwischen Elektroden, die auf der Körperoberfläche angebracht werden, wird dadurch eine Spannung erzeugt. Diese Messungen werden als Ableitungen bezeichnet und werden anhand der Elektrodenanordnung unterschieden. Zu den 12 Standardableitungen zählen die 6 Extremitätenableitungen (die bipolaren Ableitungen I, II und III nach Einthoven und die unipolaren Ableitungen aVR, aVL und aVF nach Goldberger), sowie die 6 unipolaren Thoraxableitungen V₁ bis V₆ nach Wilson. Die Extremitätenableitungen bilden die elektrischen Reize im Herzen auf die Frontalebene

ab, die Thoraxableitungen auf die Horizontalebene, wodurch verschiedene Blickwinkel auf die elektrischen Aktivitäten im Herzen ermöglicht werden. [28]

Die Aktivitäten im Herzen lassen sich wie folgt aus dem EKG ablesen (siehe Abb. 2.4) [4]:

- **P-Welle:** Vorhofdepolarisation
- **PQ-Strecke:** Atrioventrikuläre Überleitzeit (Erregung durchläuft den AV-Knoten, währenddessen sind die Vorhöfe voll erregt, die Kammern hingegen noch nicht)
- **QRS-Komplex:** Kammerdepolarisation
- **ST-Strecke:** Unterbrechung der elektrischen Aktivierung der Kammern vor Beginn der Repolarisation (Refraktärzeit, während der ST-Strecke sind die Kammern vollständig erregt)
- **T-Welle:** Kammerrepolarisation
- **PQ-Intervall:** Beginn der Vorhoferregung bis Beginn der Kammererregung [28]
- **QT-Intervall:** Zeit, die beide Ventrikel zur De- und Repolarisation benötigen, abhängig von der Herzfrequenz [28]
- **RR-Intervall:** Abstand zwischen zwei Kammerdepolarisationen, aus dem RR-Intervall lässt sich die Herzfrequenz berechnen

Die Refraktärperiode der Kammern erstreckt sich von der Q-Zacke bis zum Maximum der T-Welle. Anfang der T-Welle beginnt die vulnerable Phase. Währenddessen sind die Kardiomiozyten ungleichmäßig refraktär, sodass eine Erregung, die in die vulnerable Phase fällt, eine kreisende Erregung auslösen kann. [10]

2.4.2 Vorhofflimmern im Elektrokardiogramm

VHF zeichnet sich im EKG durch ein unregelmäßiges RR-Intervall aus. Dies kommt zustande, da der AV-Knoten nur 20-30% der chaotischen Vorhofaktionen in die Kammern weiterleitet. Diese Weiterleitung geschieht in chaotischen Intervallen, woraus eine absolute Kammerarrhythmie resultiert, die sich im EKG als unregelmäßiges RR-Intervall zeigt (siehe Abb. 2.5). Bei tachykard auftretendem VHF kann es zu einer sogenannten Pseudoregularisierung kommen, bei welcher das RR-Intervall regelmäßig erscheint. Auch wenn es zu einem AV-Block kommt und der AV-Knoten keine Reize in die Kammern weiterleitet ist das RR-Intervall regelmäßig, da ein untergeordneter Schrittmacher einspringt und einen Ersatzrhythmus erzeugt.

Aufgrund der chaotischen Vorhofaktionen treten anstelle der P-Wellen bei VHF Flimmerwellen (F-Wellen) auf, welche in grobe und feine F-Wellen unterschieden werden. Die Sichtbarkeit der F-Wellen ist jedoch ableitungs- und patientenabhängig. Auch bei einer Tachykardie sind die F-Wellen nicht zwingend im EKG sichtbar, weshalb das unregelmäßige RR-Intervall als primäres Diagnosekriterium gilt. [1]

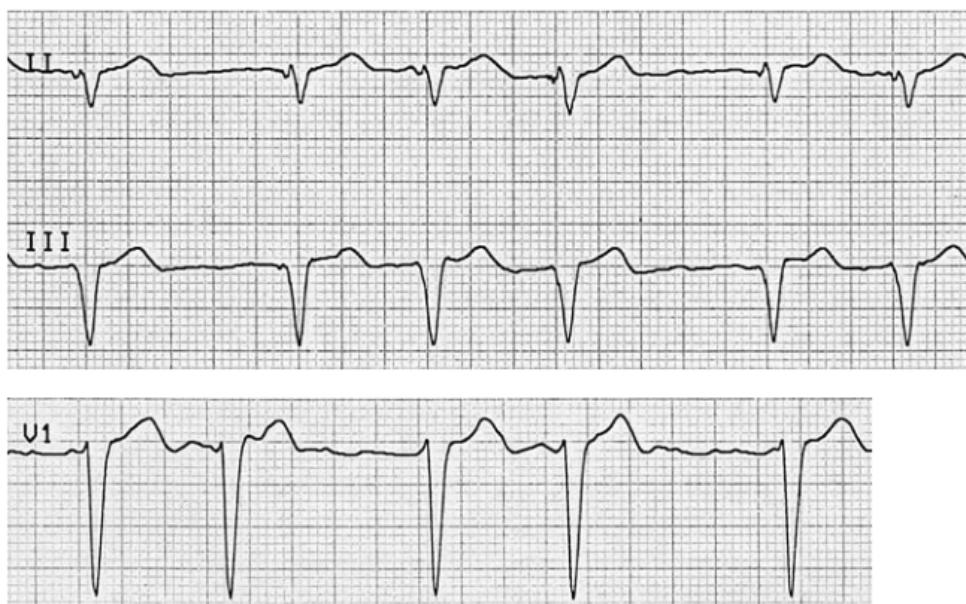


Abb. 2.5: Die Ableitungen II, III und V₁ eines Elektrokardiogramms. Zu sehen sind feine Flimmerwellen und eine absolute Kammerarrhythmie. Entnommen aus [1].

3 Machine Learning

In diesem Kapitel werden in Abschnitt 3.1 die für das weitere Verständnis der Arbeit nötigen Grundlagen zu ML und in Abschnitt 3.2 Grundlagen zu DL vermittelt. Metriken zur Bewertung von ML-Modellen werden in Abschnitt 3.3 eingeführt. Das Prinzip des Domain Shifts wird in Abschnitt 3.4 erklärt.

3.1 Grundlagen des Machine Learnings

Unter ML wird ein Prozess verstanden, bei welchem ein Computer aus gegebenen Daten ein statistisches Modell generiert, welches diese Daten abbildet und nach Beendigung der Lernphase gelernte Informationen auf andere Daten verallgemeinern kann. Angewendet wird so ein Modell bspw. für die Klassifikation von Daten in bestimmte Kategorien (bspw. ob ein EKG-Abschnitt einen Sinusrhythmus darstellt oder nicht) oder im Falle eines Regressionsproblems für die Vorhersage kontinuierlicher Werte (bspw. die Wahrscheinlichkeit, mit welcher ein Patient VHF entwickelt). [29]

Klassifikation ist der Prozess, ungekennzeichneten Daten automatisch ein Label zuzuweisen. Ein Modell, das dieses Problem löst, heißt Klassifikator. Um ein solches Modell zu erzeugen, nutzt ein Klassifikations-Lernalgorithmus eine Datenmenge mit gelabelten (Supervised Learning) oder ungelabelten (Unsupervised Learning) Beispielen als Eingabe. Die Label bei einer Klassifikationsaufgabe sind Teil einer endlichen Menge von Klassen. Eine Klassifikationsaufgabe mit zwei Klassen nennt sich binäre Klassifikation, mit mehr Klassen wird von einer Multiklassen-Klassifikation gesprochen. [30]

Beim Supervised Learning besteht die Datenmenge $\{(x_i, y_i)\}_{i=1}^N$ zur Modellerstellung aus gelabelten Beispielen. Ein gelabeltes Beispiel besteht aus einem Merkmalsvektor x_i und einem Label y_i . Jede Dimension $j = 1, \dots, D$ des Merkmalsvektors ist ein einzelnes Merkmal $x^{(j)}$, dass das Beispiel in irgendeiner Weise beschreibt. Bei allen Beispielen in der Datenmenge enthalten die Merkmalsvektoren an derselben Position j dieselbe Art Information (bspw. enthält $x^{(1)}$ das Geschlecht des jeweiligen Patienten). Das Label y_i des Beispiels beschreibt die Ausgabe, die bei Verwendung des Merkmalsvektors als Eingabe in das vom Lernalgorithmus erzeugte Modell erwartet wird. Im Falle eines Klassifikationsalgorithms ist das Label ein Element einer

endlichen Menge $\{1, 2, \dots, C\}$ von Klassen, wobei eine Klasse eine Kategorie beschreibt, der das Beispiel zugeordnet wird. Die beiden oben genannten Probleme Klassifikation und Regression gehören zum Bereich des Supervised Learning. [30]

Beim Unsupervised Learning besteht die Datenmenge $\{(x_i)\}_{i=1}^N$ aus einer Menge ungelabelter Beispiele, wobei ein Beispiel wieder einen Merkmalsvektor x_i enthält, jedoch kein Label [30]. Das Ziel des Lernalgorithmus besteht in diesem Fall darin, den Merkmalsvektor als Eingabe zu nehmen und Informationen über die Datenmenge zu erzeugen. Beim sogenannten Clustering teilt das System bspw. die Datenmenge anhand ähnlicher Eigenschaften in verschiedene zusammenhängende Gruppen ein. [29]

3.2 Deep Learning

DL ist ein Teilgebiet des ML und bezeichnet das Training von tiefen Neuronalen Netzen, also Neuronalen Netzen mit mehr als zwei Schichten zwischen Ein- und Ausgabeschicht. Algorithmen im klassischen ML erzeugen sogenannte flache Modelle, indem Parameter direkt anhand der Merkmale der Trainingsbeispiele angepasst werden. Neuronale Netze hingegen erlernen Parameter anhand der Ausgaben der vorhergehenden Schichten. Ein Vorteil von DL ist die selbstständige Merkmalsdetektion, sodass im Gegensatz zum klassischen ML die Merkmale nicht aufwändig per Hand ausgewählt werden müssen. Jedoch sind DL-Modelle im Gegensatz zu klassischen ML-Modellen schwer interpretierbar, da die große Anzahl an Schichten es schwierig macht nachzuvollziehen, wie das Modell Entscheidungen trifft. [30]

Goettling et al. [31] und Jo et al. [32] entwickeln Neuronale Netze mit dem Ziel, deren Entscheidungen nachvollziehbarer zu gestalten.

3.2.1 Artificial Neural Network

Ein Artificial Neural Network (ANN) besteht aus sogenannten künstlichen Neuronen, die in Schichten angeordnet sind. Ein Neuron

$$O = f\left(\sum_{j=1}^n x^{(j)} \cdot w^{(j)} + b\right) \quad (3.1)$$

nimmt einen Merkmalsvektor $x_i = [x^{(1)}, \dots, x^{(n)}]$, multipliziert ihn mit einem Gewichtsvektor $w_i = [w^{(1)}, \dots, w^{(n)}]$, summiert die erhaltenen Werte auf, addiert den Bias b , leitet das Ergebnis durch eine Aktivierungsfunktion f und gibt einen einzelnen Wert aus.

Die Werte der Gewichte (und des Bias) sind die Parameter, die beim Trainieren des Lernalgorithmus erlernt werden. Die Aktivierungsfunktion wird auf die gewichtete Summe aller Eingaben angewandt und fügt dem Neuron die Nichtlinearität hinzu. Sie wird vor dem Training festgelegt, häufig ist es eine der folgenden Funktionen:

$$\text{Sigmoid: } f(z) = \frac{1}{1 + e^{-z}} \quad (3.2)$$

$$\text{Tangens hyperbolicus: } f(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \quad (3.3)$$

$$\text{Rectified Linear Unit: } f(z) = \begin{cases} 0, & \text{wenn } z < 0 \\ z & \text{sonst} \end{cases} \quad (3.4)$$

$$\text{Softmax: } f(z) = [f^{(1)}, \dots, f^{(C)}], \text{ wobei } f^{(j)} = \frac{e^{z^{(j)}}}{\sum_{k=1}^C e^{z^{(k)}}} \quad (3.5)$$

Die Ausgabe der Softmax-Funktion beschreibt die Wahrscheinlichkeitsverteilung der verschiedenen Klassen C. Die Summe aller Ausgaben $f^{(1)}, \dots, f^{(C)}$ ergibt 1. Werden mehrere Neuronen zusammengeschaltet erhält man ein ANN. In einem ANN gibt es eine Eingabeschicht, eine Ausgabeschicht, sowie beliebig viele Schichten zwischen diesen, welche als Hidden Layer bezeichnet werden. Hat ein ANN viele Hidden Layer, wird es auch als Deep Neural Network (DNN) bezeichnet. [30]

Es gibt verschiedene Architekturen von ANNs. Die simpelste Architektur eines ANNs ist das Feed Forward Neural Network (FFNN) (auch Multilayer Perceptron) [33]. In einem FFNN existieren keine Schleifen und es gibt einen direkten Weg von Eingabe- zu Ausgabeschicht (siehe Abb. 3.1). Als fully connected bezeichnet man die Schichten, wenn alle Neuronen der Schicht voneinander unabhängig sind, jedes Neuron alle Ausgaben der vorherigen Schicht als Eingaben erhält und seine Ausgaben an alle Neuronen der nachfolgenden Schicht weitergibt. [34]

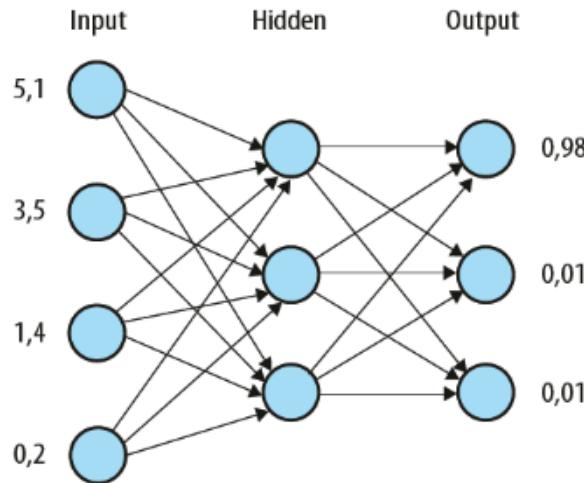


Abb. 3.1: Ein Feed Forward Neural Network mit einem fully connected Hidden Layer und beispielhaften Eingangs- und Ausgangswerten. Entnommen aus [29].

3.2.2 Convolutional Neural Network

Ein Convolutional Neural Network (CNN) ist ein FFNN, welches häufig zur Bildverarbeitung eingesetzt wird und gut für Mustererkennung geeignet ist. Es besitzt sogenannte Convolutional Layer, in denen Filter vorhanden sind. Diese Filter sind kleine Matrizen, die jeweils ein anderes Muster erkennen. Dies tun sie, in dem sie stückweise über die Bildmatrix geschoben werden und mit dem Bildausschnitt (Patch) gefaltet werden, den sie überlagern. Je ähnlicher sich Filtermatrix und Patch sind, desto größer ist der Wert, der bei der Convolution erhalten wird. Ein Convolutional Layer in einem CNN besteht aus mehreren Filtern, deren Ergebnis jeweils ein Bias-Parameter hinzugefügt wird, bevor die Aktivierungsfunktion angewendet wird. Die Filtermatrix und die Bias-Parameter sind trainierbare Werte in einem CNN. Convolutions besitzen eine Schrittweite, die angibt wie weit die Filtermatrix bei jedem Schritt geschoben wird, und ein Padding, welches dem Rand des Bildes hinzugefügt werden kann, damit dieser besser erkennbar ist. Als Aktivierungsfunktion wird üblicherweise die Rectified Linear Unit (ReLU)-Funktion genutzt. Je tiefer der Convolutional Layer ist, desto komplexer werden die Muster, die ein Filter erkennen kann. CNNs besitzen sogenannte Pooling Layer. Ein Pooling Layer folgt üblicherweise auf ein Convolutional Layer. Beim Max-Pooling wird ähnlich wie bei einer Convolution ein Fenster mit einer bestimmten Schrittweite über eine Eingabematrix geschoben, aber anstatt eine Convolution durchzuführen wird der größte Wert im Fenster ausgewählt und ausgegeben. [30]

Je tiefer ein DNN ist, desto schwieriger ist es zu trainieren, da das Degradationsproblem auftritt. Dies bedeutet, dass mit einer zunehmenden Anzahl an Schichten die Genauigkeit eines Modells ab einem bestimmten Punkt gesättigt ist und von da an beginnt zu sinken. Ein Modell mit weniger Schichten hat also eine bessere Leistung als ein Modell mit mehr Schichten, obwohl Letzteres in der Lage sein sollte, komplexere Muster zu erkennen. Ein Grund hierfür kann das *Vanishing Gradient Problem* sein. *Backpropagation* ist ein Algorithmus zur Aktualisierung der Parameterwerte von ANNs, mit welchem effizient der Gradient der ANNs berechnet werden kann [30]. Beim Vanishing Gradient Problem werden während der Backpropagation die Gradienten sehr klein und die Gewichte der vorderen Schichten ändern sich deshalb nicht mehr. Residual Networks (ResNets) lösen das Degradationsproblem, indem sie sogenannte Residualverbindungen (auch Skip oder Shortcut Connections genannt) nutzen, um Schichten zu überspringen. Ein ResNet besteht aus mehreren Residual Blocks, welche wiederum aus zwei oder mehr Convolutional Layern bestehen. Die Residual Blocks sind durch Shortcut Connections so verbunden, dass die Eingabe des einen Residual Blocks direkt mit seiner Ausgabe addiert wird und dies als Eingabe für den nächsten Block gilt. Die Shortcut Connections ermöglichen eine effektivere Gradientenweiterleitung während der Backpropagation und ermöglichen das Training tieferer Modelle. [35]

3.2.3 Training von Lernalgorithmen

Das Trainieren eines Lernalgorithmus ist im Grunde das Lösen einer Optimierungsaufgabe. Jeder Lernalgorithmus besitzt eine Verlustfunktion (engl. Loss Function), die auch als Fehler bezeichnet wird. Sie ist das Maß dafür, wie stark sich die Ausgabe des Lernalgorithmus von

dem tatsächlichen Label unterscheidet. Parameter sind Variablen im erzeugten Modell, deren Werte vom Lernalgorithmus während des Trainings erlernt werden. Sie werden auch als Gewicht und Bias bezeichnet. Gewichte stellen die Stärke der Verbindung zwischen den Neuronen dar und beeinflussen somit, wie stark oder schwach ein Eingangssignal das Neuron aktiviert und wie stark es zur Ausgabe beiträgt. Das Bias ist ebenfalls ein Gewicht, jedoch ist es eine Konstante und ermöglicht es dem Neuron, auch dann aktiviert zu werden, wenn alle Eingaben null sind. Die Optimierungsaufgabe besteht darin, die Parameter des Modells iterativ so anzupassen, dass der Wert der Loss Function minimiert wird, sich also die Ausgabe des Lernalgorithmus so gut wie möglich mit dem tatsächlichen Label deckt. [30]

Eines der am häufigsten angewandten Verfahren, welches zur Optimierung genutzt wird, ist der Gradientenabstieg (engl. Gradient Descent). Mit ihm kann iterativ das Minimum (oder Maximum) der Loss Function und damit die optimalen Parameterwerte gesucht werden. Eine Iteration wird als Epoche bezeichnet. Der Gradient Descent beginnt mit zufällig ausgewählten Werten für die Parameter. Anschließend wird mit Hilfe der Ableitung die Steigung der Kurve der Loss Function am Punkt der aktuellen Parameterwerte berechnet. Ist die Steigung positiv, werden die Parameterwerte in der nächsten Epoche reduziert, der Punkt wird also im Graphen nach links bewegt. Ist die Steigung negativ, werden die Parameterwerte in der nächsten Epoche erhöht, der Punkt wird also im Graphen nach rechts bewegt. Die Größe dieser Bewegung wird durch die Lernrate (engl. Learning Rate) definiert. Anschließend wird der Vorgang mit den neuen Parameterwerten so oft iteriert, bis ein Abbruchkriterium erfüllt ist. Sobald die Steigung null ist, wurde ein Minimum und somit optimale Parameterwerte gefunden. Allerdings kann es sich dabei um ein lokales Minimum handeln, welches nicht unbedingt bedeutet, dass die absolut optimalen Parameterwerte gefunden wurden. Lokale Minima lassen sich bspw. durch eine dynamische Anpassung der Learning Rate vermeiden. [29]

Der stochastische Gradientenabstieg (engl. Stochastic Gradient Descent) ist eine Version des Gradient Descent, der den Gradienten näherungsweise mit Teilmengen der Trainingsdatenmenge berechnet und so die Berechnung beschleunigt. Beim Training von ANNs werden häufig die Varianten *RMSprop* oder *Adam* des Stochastic Gradient Descent genutzt. [30]

3.2.4 Optimierung der Hyperparameter

Hyperparameter werden nicht durch den Lernalgorithmus selbst optimiert und müssen mit einem anderen Verfahren optimiert werden. Wenn die verfügbare Datenmenge groß genug, in der Validierungsmenge jede Klasse in einer ausreichenden Anzahl vertreten und die Anzahl der Hyperparameter sowie deren Wertebereich nicht zu groß ist, kann Rastersuche (engl. Grid Search) genutzt werden. Hierbei werden für jeden Hyperparameter Werte bspw. anhand einer logarithmischen Skala festgelegt. Anschließend werden verschiedene Kombinationen der Hyperparameterwerte zusammengestellt und es wird für jede dieser Kombinationen mit der Trainingsmenge ein Modell erstellt. Anschließend wird die Leistung der Modelle beurteilt. Wenn die optimalen Hyperparameterwerte gefunden wurden, können zusätzlich Werte in der Nähe ausprobiert werden, da sich unter Umständen noch bessere Hyperparameterwerte ergeben können. [30]

Bei der Zufallssuche (engl. Random Search) werden im Gegensatz zur Grid Search keine

diskreten Wertemengen für die Hyperparameter zusammengestellt, sondern es wird für jeden Hyperparameter eine statistische Verteilung vorgegeben, aus welcher Werte entnommen werden und festgelegt, wie viele Wertekombinationen ausprobiert werden sollen [30].

3.3 Bewertung von Klassifikatoren

Die Datenmenge kann in drei Teilmengen aufgeteilt werden: Trainingsmenge, Validierungs menge und Testmenge. Um den Klassifikator zu validieren, gibt es zwei geläufige Ansätze: Die Holdout-Validierung und die Kreuzvalidierung (engl. Cross Validation). Für die Validierung werden Daten benötigt, die nicht zum Training beigetragen haben. Bei der Holdout-Validierung ist die Trainingsmenge im Idealfall die größte dieser Teilmengen und wird für das Training genutzt. Validierungs- und Testmenge sind sogenannte Holdout-Datenmengen. Sie sind im Idealfall wesentlich kleiner und tragen nicht zum Training bei. Zu beachten ist, dass bei einer insgesamt kleinen Validierungs- und Testmenge die zugrundeliegende Verteilung der Daten unter Umständen nicht vollständig abgebildet wird, sodass die Ergebnisse aus Validierung und Test weniger aussagekräftig sind. Die Validierungsmenge wird verwendet, um den besten Lernalgorithmus auszuwählen und die besten Werte für die Hyperparameter zu ermitteln. Bei der Cross Validation wird die Datenmenge nur in Trainings- und Testmenge aufgeteilt. Die Trainingsmenge wird wiederum in n Mengen aufgeteilt. Anschließend werden $n - 1$ Modelle trainiert, wobei jeweils $n - 1$ Mengen zum Training und die n -te Menge zur Validierung genutzt wird. In beiden Fällen wird die Testmenge genutzt, um das finale Modell zu beurteilen, bevor es produktiv eingesetzt wird. [30]

Bei dem Erstellen eines Modells können verschiedene Probleme auftreten. Wenn das Modell nicht genug an die Daten angepasst ist, kommt es zur sogenannten Unteranpassung (engl. Underfitting), das Modell hat ein hohes Bias. Das Modell macht also sowohl bei der Vorhersage von Labeln der Trainingsdaten als auch bei denen der Testdaten viele Fehler. Ursache hierfür kann u.A. sein, dass ein zu einfaches Modell gewählt wurde oder die erstellten Merkmale nicht aussagekräftig genug sind. Sagt das Modell die Label der Trainingsdaten sehr gut vorher, aber die der Testdaten nur sehr schlecht, besteht eine Überanpassung (engl. Overfitting), das Modell hat eine hohe Varianz. Dies kann durch ein zu komplex gewähltes Modell auftreten oder dadurch, dass zu viele Merkmale bei nur wenigen Trainingsbeispielen vorliegen. Um die Generalisierbarkeit des Modells zu gewährleisten, sollten die Trainingsdaten eine ausreichende Diversität aufweisen. Dadurch wird vermieden, dass das Modell sich nur an spezifische Eigenheiten der Trainingsdaten anpasst, was ebenfalls zu Overfitting führen kann. Zusätzlich müssen die Testdaten möglichst unabhängig von den Trainingsdaten sein, um eine realistische Einschätzung der Modellleistung auf bisher unbekannten Daten zu ermöglichen. [30]

3.3.1 Metriken zur Bewertung der Klassifikationsgüte

Um festzustellen, ob das Modell Klassen von Daten gut vorhersagen kann, die dem Lernalgorithmus unbekannt sind, gibt es bestimmte Metriken, die in diesem Abschnitt vorgestellt werden.

Confusion Matrix

In einer binären Klassifikationsaufgabe mit den Labels *positiv* und *negativ* kann man im allgemeinen vier Fälle unterscheiden:

- True Positives (TP): Anzahl der korrekt als positiv vorhergesagten Label
- False Positives (FP): Anzahl der fälschlicherweise als positiv vorhergesagten Label
- False Negatives (FN): Anzahl der fälschlicherweise als negativ vorhergesagten Label
- True Negatives (TN): Anzahl der korrekt als negativ vorhergesagten Label

Die Matrix in der sie eingetragen werden nennt sich Confusion Matrix (Wahrheitsmatrix). Ziel ist, die Werte für TP und TN zu maximieren. Diese Fälle können auch auf Multiklassen-Klassifikationen erweitert werden. [29]

Accuracy

Die Accuracy (Korrektklassifikationsrate) gibt das Verhältnis der Anzahl richtiger Vorhersagen der Labels zur Menge aller untersuchten Beispiele als

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (3.6)$$

[29] an.

Precision-Recall

Die Precision (Genauigkeit) wird angegeben als

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3.7)$$

und stellt das Verhältnis der richtigen positiven Vorhersagen zu den positiven Vorhersagen insgesamt dar.

Recall (auch Sensitivity, Trefferquote oder True-positive-rate) ist das Verhältnis der richtigen positiven Vorhersagen zu den insgesamt in der Datenmenge vorhandenen positiven Beispielen:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3.8)$$

Es kann eine Precision-Recall-Kurve generiert werden und anhand der Fläche unter der Kurve (Area Under Curve (AUC)) beurteilt werden, wie gut Precision und Recall gleichzeitig sind. Um gute Werte sowohl für Precision als auch für Recall auszuwählen, wird der F1-Score genutzt. Der F1-Score wird entsprechend

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3.9)$$

berechnet und ist das harmonische Mittel zwischen Precision und Recall. [29]

Specificity

Die Specificity oder auch True-negative-rate gibt das Verhältnis der richtigen negativen Vorhersagen zu den insgesamt in der Datenmenge vorhandenen negativen Beispielen als

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (3.10)$$

an.

Receiver-Operating-Characteristic-Kurve

In der Receiver-Operating-Characteristic (ROC) Kurve wird die Fallout-Rate (False-positive-rate) $fp/(fp + tn)$ gegen den Recall eingetragen. Die Fläche unter der Kurve (AUC) ist ein Maß für die Leistung des Modells. [30]

3.4 Domain Shift

In der Praxis trifft ein Modell oft auf veränderte Bedingungen: Die Daten, auf denen es trainiert wurde, unterscheiden sich häufig von denen, die es später in der Anwendung sieht. Dieses Phänomen wird Domain Shift genannt [36]. Ein Domain Shift tritt auf, wenn die Datenverteilung, auf denen ein Modell trainiert wurde (Quelldomäne), unterschiedlich zur Verteilung der Daten, die das Modell in der Praxis sieht (Zieldomäne), ist. Dies führt zu einem Leistungsabfall des Modells. Der Begriff Datenverteilung bezieht sich hierbei auf die statistischen Eigenschaften und Merkmale der Daten insgesamt, nicht nur auf die Häufigkeit der Klassen. [37]

Mathematisch lässt sich eine Domäne \mathcal{S} beschreiben als eine Kombination von Merkmalen x aus einem Eingaberaum \mathcal{X} und Labeln y aus einem Ausgaberaum \mathcal{Y} , mit einer zugehörigen Datenverteilung P_{XY} . Dies wird als

$$\mathcal{S} = (x_i, y_i)_{i=1}^n \sim P_{XY} \quad (3.11)$$

notiert. X und Y beschreiben die zugehörigen Zufallsvariablen. Die bekanntesten Lösungsansätze für das Domain Shift-Problem sind Transfer Learning, Domain Adaptation (DAp) und DG. [37]

3.4.1 Domain Adaptation und Transfer Learning

Beim Transfer Learning können die Domänen sich in Merkmalsraum, Labelraum und/oder Datenverteilung unterscheiden. DAp beschreibt den konkreteren Fall, bei welchem sich die Domänen nur in ihrer Datenverteilung unterscheiden. Da dieser Fall beim Transfer Learning mit inbegriffen ist, kann DAp mittels Transfer Learning erreicht werden. [38]

Bei der DAp wird ein Modell bspw. mittels Pretraining-Finetuning zuerst auf eine Quelldomäne \mathcal{S}^{src} trainiert und anschließend auf eine Zieldomäne \mathcal{S}^{tar} angepasst. Sowohl beim Transfer Learning als auch bei der DAp besteht während des Trainings Zugriff auf Daten aus der Zieldomäne \mathcal{S}^{tar} . [37]

3.4.2 Domain Generalization

Das Ziel der DG ist, ein Modell auf eine oder mehrere unterschiedliche, aber verwandte Domänen $\mathcal{S}^1, \dots, \mathcal{S}^n$ zu trainieren, sodass es gut auf unbekannte Domänen $\mathcal{S}^{n+1}, \mathcal{S}^{n+2}, \dots$ verallgemeinern kann. Im Gegensatz zur DAp besteht während des Trainings kein Zugriff auf Daten aus den Zieldomänen, sodass DG zwar ein komplexerer, aber auch ein Ansatz näher an der tatsächlichen Praxis ist. [37]

Methoden für DG lassen sich aufteilen in Data Manipulation-basierte Methoden, Representation Learning-basierte Methoden und allgemeine Lernstrategien, die zur DG beitragen. Diese Methoden lassen sich kombinieren, um eine noch bessere DG zu erreichen. [37]

Data Manipulation-basierte Methoden

Die Kategorie der Data Manipulation-basierten Methoden konzentriert sich auf die Veränderung der Eingabedaten, um die Diversität und Quantität der existierenden Trainingsdaten zu erhöhen, wodurch ein Overfitting reduziert und damit die Fähigkeit des Modells zu Generalisieren erhöht wird. Bei der Data Augmentation werden vorhandene Daten mit Rauschen erweitert, rotiert, zugeschnitten, skaliert oder anderweitig transformiert. Bei der Data Generation werden Daten bspw. mittels eines Generative Adversarial Networks künstlich erzeugt. [37]

Representation Learning-basierte Methoden

Beim Representation Learning geht es darum, dass Modelle Merkmale erlernen, die die Eingabedaten möglichst gut repräsentieren. Es gibt dazu zwei Lernstrategien, die gut zur DG geeignet sind, nämlich Domain Invariant Representation Learning und Feature Disentanglement.

Beim Feature Disentanglement werden Merkmale in domäneninvariante und domänen-spezifische Merkmale getrennt. Beim Domain Invariant Representation Learning ist das Ziel, domänenunabhängige Merkmale zu extrahieren. Dies kann bspw. durch Domain Adversarial Learning (siehe Abschnitt 3.4.3) oder Explicit Feature Alignment erreicht werden. Bei Letzterem wird darauf abgezielt, Merkmalsverteilungen zwischen verschiedenen Domänen explizit anzugeleichen. [37]

Allgemeine Lernstrategien zur DG

Es gibt allgemeine ML Paradigmen, die sich gut für DG eignen. Eine gut geeignete Lernstrategie ist das Ensemble Learning. Hierbei werden mehrere Klassifikatoren trainiert und das Ergebnis bspw. anhand einer Mehrheitsentscheidung aller Modelle bestimmt. Ein Ensemble hat den Vorteil, dass durch die Diversität der Modelle die Gesamtleistung erhöht wird. [37]

Der in dieser Arbeit vorgestellte Ansatz vereint die zwei DG Methoden Domain Adversarial Learning und Ensemble Learning.

3.4.3 Domain Adversarial Learning

Ganin et al. [7] haben den Ansatz des Domain Adversarial Learnings ursprünglich als Methode zur DAp entwickelt. Sie stützen sich hierbei auf die Hypothese, dass die Vorhersagen eines

Modells auf Merkmalen getroffen werden müssen, die unabhängig von Quell- und Zieldomäne sind, um den Domain Shift zu überwinden. Der Ansatz von Ganin et al. nutzt gelabelte Daten aus der Quelldomäne und nicht-gelabelte Daten aus der Zieldomäne und zielt darauf ab, dass Merkmale erlernt werden, die differenzierend für die Hauptklassifikationsaufgabe fungieren, jedoch unempfindlich gegenüber dem Domain Shift sind.

Ganin et al. führen aus, dass ihr Ansatz mit beinahe jedem FFNN funktioniert und das Training mit den Standardmethoden Backpropagation und Gradientenabstieg durchgeführt werden kann. Die Architektur (siehe Abb. 3.2) eines DANNs besteht aus einem *Feature Extractor* und einem *Label Predictor*, die zusammen ein Standard-FFNN bilden, sowie einem *Domain Classifier*, der mit dem Feature Extractor durch einen Gradient Reversal Layer (GRL) verbunden ist. Der GRL multipliziert den Gradienten während der Backpropagation mit einem konstanten negativen Wert. Die Gradientenumkehr führt dazu, dass das Modell lernt, die Domänen möglichst schlecht zu unterscheiden. [7]

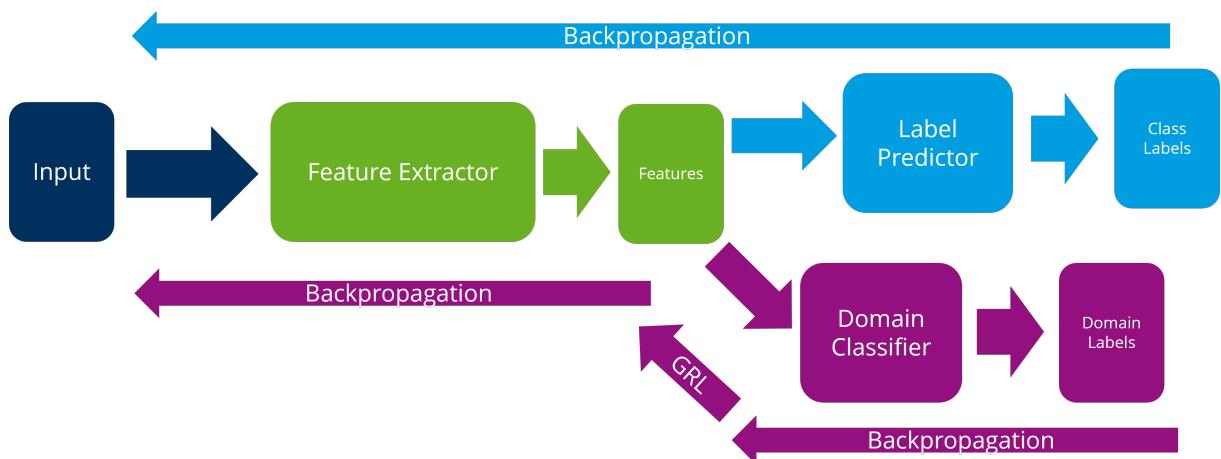


Abb. 3.2: Ein Domain Adversarial Neural Network nach [7]. Es besteht aus einem Feature Extractor, einem Label Predictor und einem Domain Classifier. Der Gradient des Domain Classifiers wird während der Backpropagation durch den Gradient Reversal Layer (GRL) invertiert.

4 Stand der Technik

In diesem Kapitel wird in Abschnitt 4.1 zunächst eine Übersicht über aktuelle Ansätze in der Zeitreihenklassifikation gegeben. Anschließend werden in Abschnitt 4.2 Ansätze zur Klassifikation von VHF mittels DL zusammengefasst. Zuletzt werden in Abschnitt 4.3 Vorgehen und die Ergebnisse einer systematischen Literaturrecherche zur DG in der EKG-Klassifikation präsentiert.

4.1 Ansätze zur Zeitreihenklassifikation

Da EKG-Daten Zeitreihendaten sind, muss für diese Arbeit eine DNN-Architektur gewählt werden, die die zeitlichen Abhängigkeiten dieser Daten gut verarbeiten kann. Fawaz et al. [39] haben in ihrem Review die gängisten Ansätze zur Zeitreihenklassifikation in einem einheitlichen Rahmen miteinander verglichen. Dazu wurden die gewählten Modelle auf insgesamt 97 uni- und multivariaten Zeitreihendatensätzen trainiert. Die 85 univariaten Zeitreihendatensätze stammen aus dem *University of California, Riverside (UCR)/University of East Anglia (UEA) time series archive* [40] [41], die 12 multivariaten Datensätze stammen aus *Baydogan's archive* [42], welches 2015 auf der Website heruntergeladen werden konnte, aktuell jedoch nicht. Für univariate Zeitreihendaten, welche in dieser Arbeit als Ein-Kanal-EKGs zu finden sind, zeigte beim paarweisen Vergleich von neun DNN-Klassifikatoren das ResNet die höchste Klassifikationsgüte. Dies ist dargestellt im Critical Difference Diagramm in Abb. 4.1. Je niedriger der Rang im Diagramm, desto höher ist die durchschnittliche Klassifikationsgüte des Modells. Mit einer Querlinie verbundene Modelle unterscheiden sich statistisch nicht signifikant in ihrer Klassifikationsgüte. Zu sehen ist, dass das ResNet einen Score von 2 besitzt und sich somit signifikant von den übrigen Modellen abhebt. Fawaz et al. veröffentlichten das im Rahmen des Reviews entwickelte Framework inklusive der implementierten Klassifikatoren³. [39]

Nachträglich wurde von Fawaz et al. eine Website [43] erstellt, die eine Übersicht über die Ergebnisse des Vergleichsframeworks der im Review genutzten Klassifikatoren sowie neuerer Ansätze bietet. Ein neuerer Ansatz ist *InceptionTime*, der eine bessere Leistung bezüglich Accuracy erzielt als ein ResNet. Zu sehen in Abb. 4.2 ist außerdem H-InceptionTime, welches ein hybri-

³<https://github.com/hfawaz/dl-4-tsc>

der Ansatz ist. Da sich diese Arbeit mit reinen DL-Ansätzen beschäftigt, wird H-InceptionTime nicht weiter berücksichtigt.

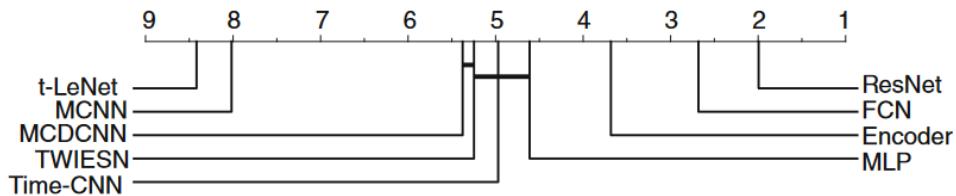


Abb. 4.1: Das Critical Difference Diagramm zeigt die Klassifikationsgüte von neun Deep Neural Networks über das UCR/UEA-Archive im paarweisen Vergleich. Je niedriger der Rang, desto höher ist die durchschnittliche Klassifikationsgüte des Modells. Mit einer Querlinie verbundene Modelle unterscheiden sich statistisch nicht signifikant in ihrer Leistung. Entnommen aus [39].

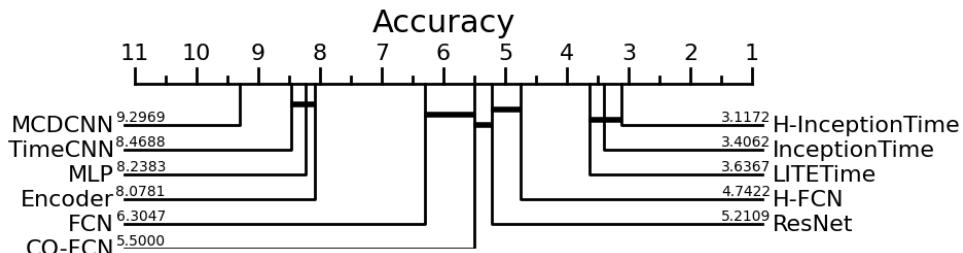


Abb. 4.2: Das Critical Difference Diagramm zeigt die Klassifikationsgüte von Deep Neural Networks über das UCR/UEA-Archive im paarweisen Vergleich. Die Rangordnung der Klassifikatoren wurde basierend auf der Accuracy als Metrik erstellt. Je niedriger der Rang, desto höher ist die durchschnittliche Klassifikationsgüte des Modells. Mit einer Querlinie verbundene Modelle unterscheiden sich statistisch nicht signifikant in ihrer Leistung. Entnommen aus [43].

InceptionTime [6] ist ein Ensemble aus CNNs zur Zeitreihenklassifikation, welches auf Inception-v4 basiert, einem CNN zur Bildklassifikation. Als Eingabe können sowohl univariate als auch multivariate Zeitreihen dienen. Ein einzelner InceptionTime Klassifikator in der Standardkonfiguration ist aufgebaut aus zwei Residual Blocks, welche wiederum aus drei Inception Modulen gebildet werden. Es gibt, wie beim ResNet, lineare Shortcut Verbindungen, die den Input des ersten Residual Blocks direkt zum Input des nächsten Residual Blocks hinzufügen. Auf die Residual Blocks folgt ein Global Average Pooling (GAP) Layer, die den Durchschnitt der entstandenen multivariaten Zeitreihendaten über die gesamte Zeitdimension berechnet. Die letzte Schicht ist ein Fully Connected (FC) Layer mit einer Softmax-Aktivierungsfunktion. Die Anzahl der Neuronen dieser Schicht entspricht der Anzahl der Klassen im Datensatz. In Abb. 4.3 ist der Aufbau dargestellt. [6]

In Abb. 4.4 ist ein einzelnes Inception Modul dargestellt. Wird eine Zeitreihe der Dimension M in ein Inception Modul eingegeben, so wird sie im ersten Schritt durch ein Bottleneck Layer mit m Filtern mit Wahrnehmungsbereich und Schrittweite 1 auf eine Dimension von $m \ll M$ reduziert. Dies reduziert die Komplexität des Modells und somit auch das Overfitting, während gleichzeitig größere Filter als beim ResNet ermöglicht werden. Im zweiten Schritt werden Convolutionen mit verschiedenen großen Wahrnehmungsbereichen auf die Ausgabe des Bottleneck Layers angewendet.

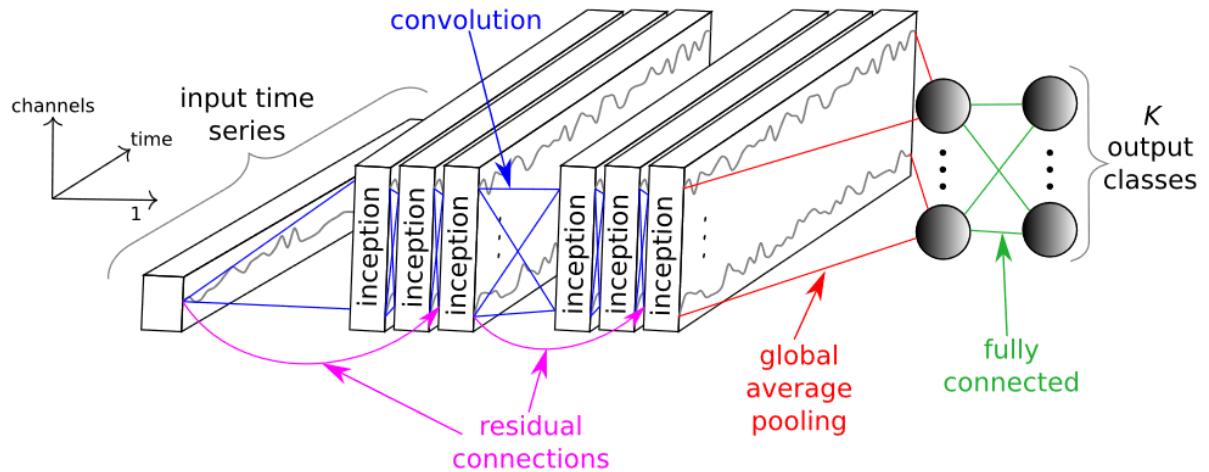


Abb. 4.3: Darstellung eines einzelnen InceptionTime Klassifikators mit 6 Inception Modulen und einer univariaten Zeitreihe als Eingabe. Entnommen aus [6].

eck Layers angewandt. Parallel zu diesen Convolutions gibt es einen MaxPooling-Pfad, der die maximalen Werte innerhalb eines Fensters zusammenfasst und dadurch die Robustheit des Modells gegenüber Rauschen erhöht. Die Ausgabe der MaxPooling-Operation wird ebenfalls mittels eines Bottleneck Layers in ihrer Dimension reduziert. Die Ausgaben der unabhängigen Convolutions sowie der MaxPooling-Operation werden konkateniert und bilden die Ausgabe des Inception Moduls. Die Standardkonfiguration eines Inception Moduls besteht aus 3 Filtersets mit jeweils 32 Filtern der Länge $l \in \{10, 20, 40\}$ und MaxPooling. Die Anzahl der Filter pro Schicht entspricht also $32 \times 4 = 128$. Die Größe des Bottleneck Layers liegt standardmäßig bei $m = 32$. [6]

Fawaz et al. empfehlen, InceptionTime als Ensemble zu verwenden, da es laut ihnen eine hohe Standardabweichung in der Accuracy zwischen einzelnen InceptionTime Klassifikatoren gibt. Ein InceptionTime Ensemble von Fawaz et al. besteht aus 5 unabhängig voneinander trainierten Modellen. Jedes Modell trägt mit demselben Gewicht zur finalen Vorhersage bei. Die Anzahl der einzelnen Modelle im Ensemble wurde bei 5 festgelegt, da keine signifikante Verbesserung bei mehr als 5 Modellen im Ensemble zu beobachten war. Das InceptionTime Ensemble wurde mit einem ResNet Ensemble der Größe 5 verglichen, da dies das bisher beste DL-Ensemble für Zeitreihendaten war. Beide Ensembles wurden mit den 85 Zeitreihendatenbanken des UCR-Archives trainiert und evaluiert. In einem paarweisen Vergleich erzielte das InceptionTime Ensemble bessere Ergebnisse als das ResNet Ensemble mit einem Win/Tie/Loss von 54/8/23. [6]

4.2 Ansätze zur Deep Learning-basierten Vorhofflimmern-Detektion in EKGs

Es gibt bereits Ansätze zur erfolgreichen VHF-Detektion in 12-Kanal-EKGs mit Hilfe von DNNs [44]. Auch mit einer reduzierten Kanalanzahl kann VHF erfolgreich mittels DNNs detektiert werden [44]. Ein Auszug an Ansätzen zu VHF-Detektion sowohl mit voller als auch reduzierter

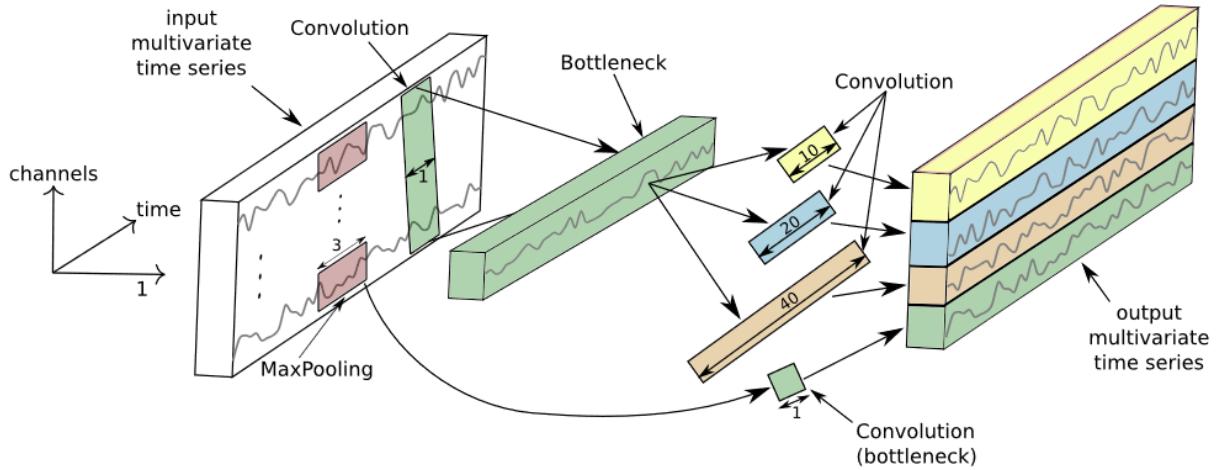


Abb. 4.4: Darstellung eines einzelnen Inception Moduls mit einem Bottleneck Layer von $m = 1$. Entnommen aus [6].

Kanalanzahl ist in Tab. 4.1 zu sehen. Da in dieser Arbeit 10-Sekunden-EKGs verwendet werden, wurden nur Ansätze ausgewählt, welche ebenfalls 10-Sekunden-EKGs nutzen. Es werden F1-Scores von bis zu 0,991 mit 12-Kanal-EKGs erreicht, sowie von bis zu 0,954 bei Nutzung eines Kanals.

Tab. 4.1: Auswahl von Veröffentlichungen die Deep Learning zur Detektion von Vorhofflimmern mit 10-Sekunden-EKGs nutzen. Wenn nur eine Ableitung genutzt wurde, wurde die genutzte Ableitung mit angegeben. Bei Lai et al. [45] handelt es sich um einen Ein-Kanal-EKG-Patch mit einer Platzierung, die mit Ableitung II korrespondiert.

Referenz	Ableitungen	Algorithmus	Spec.	Rec.	Acc.	F1
Attia et al., 2019 [46]	12	CNN	0,834	0,823	0,833	0,454
Cai et al., 2020 [47]	12	DDNN	0,994	0,992	0,994	0,991
Ribeiro et al., 2020 [48]	12	ResNet	1	0,769	-	0,870
Lai et al., 2020 [45]	1 (modified II)	CNN	0,934	0,931	0,931	-
Baalman et al., 2020 [49]	1 (II)	RNN	-	-	0,960	0,940
Goettling et al., 2024 [31]	1 (II)	CNN	0,958	0,949	0,953	0,954

4.3 Verwendung von Domain Generalization in der EKG-Klassifikation

DG ist ein relevantes Problem in der EKG-Klassifikation, da es eine Vielzahl an möglichen Domain Shifts gibt, jedoch in der Praxis selten bereits ein Testdatensatz während des Trainings vorliegt, um DAp-Methoden wie Transfer Learning anzuwenden [50]. Im Rahmen dieser Arbeit wurde eine systematische Literaturanalyse zur DG in der EKG-Klassifikation durchgeführt. Es wurden vier Suchstrings definiert (zu finden in Anhang A.1), mit welchen die Literaturdatenbanken *IEEE Xplore*, *PubMed* und *Clarivate Web of Science* durchsucht wurden, resultierend in insgesamt 367 Suchergebnissen. Durch Titel- und Abstrakt-Screening wurden 39 relevante Veröffentlichungen gefunden. Nach dem Volltext-Screening dieser Veröffentlichungen wurden

7 als tatsächlich relevant identifiziert.

Stammen Aufnahmen aus verschiedenen Krankenhäusern, tritt ein Domain Shift auf, da die EKGs unter verschiedenen Bedingungen aufgenommen wurden [50]. Ballas & Diou [51] [52] beschäftigen sich mit diesem Domain Shift und entwickeln eine DL-Architektur, welche auf ein ResNet-18 aufsetzt und aus mittleren Convolutional Layern Merkmale extrahiert. Durch diese Extraktion aus mittleren Schichten soll die Verschmelzung von domäneninvarianten mit domänenspezifischen Merkmalen so gering wie möglich gehalten werden.

Hasani et al. [50] verwenden als DG-Methoden Data Augmentation und Domain Adversarial Learning. Dabei wurden bei der Datenvorverarbeitung zufällig Frequenzen aus den Aufnahmen herausgefiltert oder hinzugefügt, um die statistischen Eigenschaften der Trainingsdaten variabel zu gestalten. Zusätzlich wurden zufällig einzelne Ableitungen durch Nulllinien oder Rauschen ersetzt, Ableitungen vertauscht, Aufnahmen einzelner Ableitungen invertiert, ein Bandpass angewandt oder die Aufnahmen skaliert. Beim Training des Modells wurde ein Domain Adversarial Head mit einem GRL und zwei FC-Layer mit Dropout hinzugefügt, um die Generalisierung der Modelle über unbekannte Zieldomänen hinweg zu verbessern. Der Feature Extractor des Modells besteht aus zwei parallelen CNNs mit unterschiedlich großen Wahrnehmungsbereichen, sodass sowohl grobe als auch feinere Merkmale extrahiert werden können. Auf die CNNs folgen zwei Long Short-Term Memory (LSTM) Layer, um zeitabhängige Merkmale zu extrahieren. Der Label Predictor besteht aus zwei FC-Layern mit Dropout.

Shang et al. [53] entwickeln ebenfalls ein DANN und nutzen als Domain die jeweilige Datenbank aus der die Aufnahmen stammen. Der Feature Extractor des DANNs ist ein modifiziertes ResNet mit einem Convolutional Layer mit großen Wahrnehmungsbereich und 8 Residual Blocks. Der Domain Classifier besteht aus einem GRL und drei FC-Layern mit Dropout, wobei der Label Predictor aus zwei FC-Layern ohne Dropout besteht.

Shin et al. [54] nennen ihre Methode zur DG „Denoise and Contrast Attention Module“, welche auf ein U-Net aufsetzt. Im ersten Schritt werden die Aufnahmen mit Hilfe eines CNNs entrauscht. Der Attention-Mechanismus analysiert im Anschluss die Unterschiede in der Morphologie und die zeitlichen Intervalle der Herzschläge, um ventrikuläre Extrasystolen zu detektieren.

Die bisher genannten Ansätze betrachten zwar den Domain Shift zwischen verschiedenen 12-Kanal-EKG-Datenbanken, jedoch nicht den Domain Shift, der auftritt, wenn Signale morphologisch verändert sind. Solche Signale treten beim Domain Shift zwischen EKG- und Photoplethysmographie (PPG)-Signalen oder zwischen 12-Kanal-EKG- und EKG-Patch-Signalen mit reduzierter Kanalanzahl auf.

Mit diesem Problem beschäftigen sich Shashikumar et al. [55] und Ramesh et al. [56], indem sie Transfer Learning nutzen, um ihre auf EKGs trainierten Modelle für PPG-Signale anzupassen. Transfer Learning ist jedoch, wie bereits erwähnt, ein Ansatz aus der DAp und setzt voraus, dass ein Trainingsdatensatz aus der Zieldomäne vorliegt. Shashikumar et al. [55] extrahieren zeit-frequenzbezogene Merkmale aus den EKGs und PPGs und nutzen diese als Merkmale für ihren Klassifikator. Ramesh et al. [56] nutzen nur Merkmale der Herzratenvariabilität (HRV) zur VHF-Klassifikation, da diese sowohl aus EKG- als auch aus PPG-Signalen extrahiert werden können. Sie evaluieren die Fähigkeit zur DG ihres Modells sowohl mit als auch ohne Transfer Learning. Sowohl Shashikumar et al. [55] als auch Ramesh et al. [56] nutzen Ansätze, um

auf morphologisch veränderten Signalen als Zieldomäne zu klassifizieren. Jedoch tun sie dies, indem sie Merkmale nutzen, in denen keine Informationen zur Signalmorphologie enthalten sind. In dieser Arbeit wird daher ein Ansatz verfolgt, der die Signalmorphologie berücksichtigt.

Eine Übersicht über verwendete Algorithmen, Methoden zur DG, sowie Quell- und Zieldomänen ist in Tab. 4.2 zu finden.

Tab. 4.2: Use Cases von Deep Learning Algorithmen zur Klassifikation von Elektrokardiogrammen (EKGs) mit Nutzung von Domain Generalization Methoden. Wenn möglich, wurde der F1-Score angegeben. HRV: Herzratenvariabilität, PPG: Photoplethysmographie, VES: Ventrikuläre Extrasystolen, VHF: Vorhofflimmern

Referenz	Algorithmus	Generalization Methode	Merkmale	Domänen	Performance Quelldomäne	Performance Zieldomäne
Ballas & Diou, 2022 [51]	ResNet-18	Feature Disentanglement	Deep Features	12-Kanal-EKG-DBs aus verschiedenen Krankenhäusern	F1 0,94 (VHF)	F1 0,91 (VHF)
Ballas & Diou, 2024 [52]	ResNet-18 S-ResNet	Feature Disentanglement	Deep Features	12-Kanal-EKG-DBs aus verschiedenen Krankenhäusern	ResNet-18 F1 0,88 (VHF) S-ResNet F1 0,87 (VHF)	ResNet-1 F1 0,64 (VHF) S-ResNet F1 0,63 (VHF)
Hasani et al., 2020 [50]	CNN-LSTM	Data Augmentation, Domain Adversarial Learning	Alter, Geschlecht, Deep Features	12-Kanal-EKG-DBs aus verschiedenen Krankenhäusern	CinC 2020 score 0,63	CinC 2020 score 0,44
Shang et al., 2021 [53]	SE-ResNet	Domain Adversarial Learning	Alter, Geschlecht, HRV Features, Deep Features	12-Kanal-EKG-DBs aus verschiedenen Krankenhäusern	CinC 2020 score 12-Kanal 0,72 6-Kanal 0,68	CinC 2020 score 12-Kanal 0,44 6-Kanal 0,49
Shin et al., 2023 [54]	U-Net + DCAM	Domain Invariant Representation Learning with Attention Model	Deep Features	12-Kanal-EKG-DBs aus verschiedenen Krankenhäusern	F1 0,99 (VES)	gemittelt über 6 Zieldomänen F1 0,89 (VES)
Shashikumar et al., 2018 [55]	CNN + BRNN + Attention Model	Explicit Feature Alignment, Transfer Learning	Deep Features, Time Series Covariates, Signal Quality Indices	EKG, PPG	AUC 0,97	AUC 0,94
Ramesh et al., 2021 [56]	CNN	Explicit Feature Alignment, Transfer Learning	HRV Features	EKG, PPG	F1 0,93	Ohne Transfer Learning F1 0,72 Mit Transfer Learning F1 0,89

5 Methodik

In diesem Kapitel werden in Abschnitt 5.1 zunächst die zum Training und zur Evaluation verwendeten Datensätze und in Abschnitt 5.2 deren Vorverarbeitung beschrieben. Anschließend wird in Abschnitt 5.3 der in dieser Arbeit verwendete Ansatz für ein DL-Modell beschrieben. Zuletzt wird in Abschnitt 5.4 auf die Umsetzung dieses Modells genauer eingegangen.

5.1 Genutztes Datenmaterial

In dieser Arbeit wurden vier Datenbanken genutzt: die xECGArch-Datenbank [31], Icentia11k [57] [58], Saitama Heart Database Atrial Fibrillation (SHDB-AF) [59] [60] und eine Datenbank, welche aus dem TIMELY-Datensatz [5] erstellt wurde.

xECGArch-Datenbank

Für das Training der Modelle wurde die von Goettling et al. [31] für xECGArch erstellte Datenbank verwendet. Die xECGArch-Datenbank wurde zusammengesetzt aus den Datenbanken PTB-XL [61] [62], Chapman-Shaoxing [63] [64], Georgia-12-Lead [65] [66] und China Physiological Signal Challenge 2018 (CPSC2018) [67]. Sie enthält 4 927 als VHF annotierte, sowie 4 927 als nicht-VHF annotierte Aufnahmen. 429 der nicht-VHF-Aufnahmen enthalten einen normalen Sinusrhythmus, Sinus Tachykardie oder Sinus Bradykardie. Die genaue Zusammensetzung der Datenbank aus den Ursprungsdatenbanken kann in Tab. 5.1 nachgelesen werden. Die xECGArch-Datenbank ist aufgeteilt in ein Trainingsset mit 4 420 VHF- und 4 448 nicht-VHF-Aufnahmen, welches für das Training der Modelle in dieser Arbeit genutzt wurde, sowie ein Testset mit 507 VHF- und 479 nicht-VHF-Aufnahmen, welches für die Evaluation in Kapitel 6 genutzt wurde. Die Aufnahmen haben eine Abtastrate von 500 Hz. [31]

Icentia11k

Da das Ziel dieser Arbeit die Entwicklung eines DL-Modells ist, welches robust gegen signalmorphologische Veränderungen bei Aufnahmen mobiler EKG-Patches ist, werden zur Evaluation Datenbanken mit Aufnahmen solcher Patches benötigt. Die Icentia11k-Datenbank [57] [58] ist

Tab. 5.1: Anzahl und Ursprungsdatenbank der EKG-Aufnahmen in der xECGArch-Datenbank. AF steht für Vorhofflimmern, NSR für normalen Sinusrhythmus, ST für Sinustachykardie ohne andere Annotation, SB für Sinusbradykardie ohne andere Annotation und Other umfasst alle anderen Rhythmusstörungen. Entnommen aus [31].

Klasse	Datenbank				
	PTB-XL	Georgia	CPSC2018	Chapman-Shaoxing	Total
NSR, ST, SB	290	67	12	123	492
OTHER	169	2 032	1 921	313	4 435
AF	1 497	553	1 097	1 780	4 927

zusammengesetzt aus Aufnahmen von 11 000 unterschiedlichen Patienten, welche mit dem CardioSTAT⁴ EKG-Patch [68] aufgezeichnet wurden. Dabei handelt es sich um einen Ein-Kanal-Patch, welches mittig auf dem Brustkorb angebracht wird und eine maximale Tragedauer von zwei Wochen hat. Die Aufnahmen der Patienten wurden in ca. einstündige Segmente zerlegt und von Technologen der Icentia inc. (Québec, Canada) annotiert. Mögliche Rhythmus-Label sind NSR für normalen Sinusrhythmus, AFib für VHF und AFLutter für Vorhofflimmern. Einige Abschnitte sind nicht annotiert. Insgesamt gibt es 542 157 Segmente mit einer Abtastrate von 250 Hz. [57]

SHDB-AF

Die SHDB-AF [59] [60] ist eine japanische Datenbank, die entwickelt wurde, um die Generalisierfähigkeit von DL-Modellen über verschiedene Domain Shifts hinweg zu testen und eignet sich deshalb, um auch das in dieser Arbeit entwickelte Modell zu evaluieren. Die Datenbank setzt sich zusammen aus Aufnahmen von 100 unterschiedlichen Patienten, welche unter paroxysmalem VHF leiden. Die Aufnahmen wurden mit dem Holter-Monitor der Fukuda Denshi Co., Ltd. (Tokyo, Japan) aufgezeichnet und enthalten zwei Kanäle, bei denen es sich um modifizierte CC5- und NASA-Ableitungen handelt. Die EKGs sind ca. 24 Stunden lang und wurden mit einer Abtastrate von 125 Hz aufgezeichnet, vor der Veröffentlichung mit einem Bandpass mit den Grenzfrequenzen 0,67 und 100 Hz gefiltert und auf 200 Hz resampelt. Annotiert sind Abschnitte mit VHF als AFIB, Vorhofflimmern als AFL, Vorhoftachykardie als AT, Vorhofbradykardie als AB und andere supraventrikuläre Tachykardien als PAT oder NOD. Die übrigen Abschnitte erhielten das Label N. [59]

TIMELY-Datensatz

Die Daten des TIMELY-Datensatzes wurden im Rahmen des TIMELY Projektes erhoben. Ziel des Projekts ist die Entwicklung einer eHealth-Plattform, welche durch künstliche Intelligenz unterstützt wird und die Ärzten und Patienten als Hilfsmittel bei der Behandlung von Herz-erkrankungen dient. Die EKGs wurden mit dem net_ECG-Patch der SEMDATEX GmbH (Berlin,

⁴<https://www.cardiostat.com/>

Deutschland)⁵, welches drei EKG-Ableitungen aufnimmt und über dem Herzen platziert wird, aufgezeichnet. Die Aufnahmen haben eine durchschnittliche Dauer von 24 Stunden und eine Abtastrate von 800 Hz. Die Daten sind nicht annotiert, sodass im Rahmen dieser Arbeit ein Datensatz erstellt und annotiert wurde. [5] [69]

5.2 Vorverarbeitung der Datensätze

Alle Datensätze wurden mit der Python Toolbox Neurokit2 [70] vorverarbeitet. Gefiltert wurde mit einem Butterworth-Bandpass 4. Ordnung und den Grenzfrequenzen 0.5 und 150 Hz. Diese Grenzfrequenzen wurden ausgewählt, da die diagnostisch relevanten Informationen eines EKGs sich innerhalb dieser Frequenzen befinden [71]. Genutzt wurde die Implementierung butterworth_ba von Neurokit2, da diese weniger Randartefakte eingeführt hat als die alternative Implementierung butterworth. Anschließend wurde mit einem Powerline-Filter die Frequenz 50 Hz herausgefiltert. Zuletzt wurden die Daten z-normalisiert.

xECGArch-Datenbank

Das Datenmaterial wurde von der PhysioNet-Website [72] zur *PhysioNet/Computing in Cardiology Challenge 2020* [65] [66] heruntergeladen. Genutzt wurden nur Aufnahmen mit mindestens 10 Sekunden Länge. Längere Aufnahmen wurden gekürzt, indem ein 10-Sekunden-Segment aus der Mitte der Aufnahme entnommen wurde. Von den 12 Kanälen der Aufnahmen wurden die Brustwandableitungen nicht verwendet, da sie morphologisch starke Unterschiede zu den Aufnahmen der EKG-Patches aufweisen. Die Aufnahmen der meisten EKG-Patches enthalten keine Brustwandableitungen, da sie auf Mobilität und einfache Platzierbarkeit ausgelegt sind. Das Einbeziehen der Brustwandableitungen in das Training des DL-Modells kann daher kontraproduktiv sein, da das Modell lernen würde, sich auf Signaleigenschaften zu verlassen, die in den mobilen Aufnahmen nicht vorhanden sind. Die übrigen 6 Kanäle I, II, III, aVR, aVL und aVF wurden als einzelne Kanäle für das Training genutzt.

Icentia11k

Der Datensatz wurde von der PhysioNet-Website [72] heruntergeladen. Zu 363 Segmenten fehlte entweder die Header- oder Annotations-Datei, sodass diese Segmente nicht weiter betrachtet wurden. Um die Menge an Daten zu verringern und die Diversität des Datensatzes dennoch zu erhalten wurde von jedem Patienten ein zufälliges Segment ausgewählt. Aus diesem Segment wurde ein zufälliges 10-Sekunden-Fenster ausgewählt, wobei nicht-annotierte Abschnitte ignoriert wurden. Ist ein gesamtes Segment nicht annotiert, sodass kein 10-Sekunden-Fenster extrahiert werden kann oder ist das einzige annotierte Intervall kürzer als 10 Sekunden, wird ein anderes Segment ausgewählt. Nach Auswahl der Fenster wurden diese auf 500 Hz mit linearer Interpolation resampelt und anschließend wie zuvor erwähnt gefiltert. Insgesamt gibt es 513 Fenster mit dem Label VHF und 10 350 Fenster, die nicht mit VHF gelabelt wurden. Die genaue Klassenzusammensetzung lässt sich in Tab. 5.2 nachlesen.

⁵<https://www.semdatex.com/de/externe-geraete>

Tab. 5.2: Anzahl und Label der 10-Sekunden-Fenster in dem für die Evaluation genutzten Ausschnitt aus der Icentia11k-Datenbank. NSR steht für normalen Sinusrhythmus, AFL für Vorhofflimmern und AF für Vorhofflimmern.

Klasse	Anzahl in Icentia11k
NSR	10 176
AFIB	513
AFL	174

SHDB-AF

Die SHDB-AF wurde von der PhysioNet-Website [72] heruntergeladen. Jede Aufnahme wurde auf 500 Hz mit linearer Interpolation resampelt, vollständig in 10-Sekunden-Fenster aufgeteilt und anschließend gefiltert. Insgesamt gibt es 167 288 Fenster mit dem Label VHF und 688 090 Fenster, die nicht mit VHF gelabelt wurden. Die genaue Klassenzusammensetzung lässt sich in Tab. 5.3 nachlesen.

Tab. 5.3: Anzahl und Label der zur Evaluation genutzten 10-Sekunden-Fenster in der SHDB-AF-Datenbank. AFIB bedeutet VHF, AFL Vorhofflimmern, AT Vorhoftachykardie, AB Vorhofbradykardie und andere supraventrikuläre Tachykardien besitzen die Label PAT oder NOD. Alle übrigen Fenster wurden mit N gelabelt.

Klasse	Anzahl in SHDB-AF
N	672 084
NOD	89
PAT	117
AT	2562
AB	366
AFL	12 872
AFIB	167 288

TIMELY-Datensatz

Die Rohdaten des TIMELY-Datensatzes wurden am Institut für Biomedizinische Technik der TU Dresden zur Verfügung gestellt. Die Verarbeitung, welche durchgeführt wurde, um aus den Rohdaten einen annotierten Datensatz zu erstellen, wird im Folgenden erklärt. Es war den Patienten möglich, das EKG-Patch bspw. zum Duschen abzunehmen. Für die Zeitspanne, in der der Patch nicht angebracht war, wurde ein Rechtecksignal und anschließend eine Nulllinie aufgezeichnet. Mit Hilfe einer Discard-Mask wurden diese Rechtecksignale, Nulllinien, sowie zu niedramplitudige Signale entfernt. Die übrigen Segmente wurden mit Hilfe eines Ensembles aus einem Decision Tree Ensemble von Hammer et al. [73], xECGArch [31] und einem von Ribeiro et al. entwickelten Modell [48] klassifiziert. Dafür wurden die Segmente entsprechend den drei Ansätzen vorverarbeitet und in 30- bzw. 10-Sekunden-Fenster zerlegt. Für die 10-Sekunden-Fenster wurden die Aufnahmen von 800 Hz auf 500 Hz mit Nutzung des Mittelwerts als Aggregationsverfahren resamplet.

Die Patienten sind nicht gleichmäßig im Datensatz vertreten, da alle als VHF klassifizierten Fenster aus Aufnahmen von 5 Patienten bestehen. Insgesamt gab es 2 007 als VHF klassifi-

zierte Fenster. Um den Datensatz zu erhalten wurden zufällig 2 100 als nicht-VHF klassifizierte Fenster ausgewählt.

Anschließend wurden die automatische Annotation manuell durch zwei Personen unabhängig überprüft. Ein Fenster wurde als VHF annotiert, wenn zwei der drei zur Annotation berücksichtigten Parameter *fehlende P-Welle*, *F-Wellen* und *absolute Arrhythmie* mit *ja* beantwortet wurden. Wurden mindestens zwei der Parameter mit *nein* beantwortet, wurde das Fenster mit nicht-VHF annotiert. Im Falle von Uneinigkeit der Annotationen wurde das Fenster entfernt. Fenster, bei denen die Ableitung durch extrem starkes Rauschen unbrauchbar war, wurden entfernt. Dadurch entsteht eine leicht unterschiedliche Anzahl an verfügbaren Fenstern pro Ableitung. Die genaue Anzahl lässt sich in Tab. 5.3 nachlesen. Für diese Arbeit wurde der so erhaltene Datensatz aus 10-Sekunden-Fenstern wie zuvor beschrieben mit Neurokit2 gefiltert.

Tab. 5.4: Anzahl und Label der 10-Sekunden-Fenster in dem für die Evaluation genutzten Timely-Datensatz. AF steht für Vorhofflimmern, OTHER für nicht-Vorhofflimmern.

Klasse	Anzahl im Timely-Datensatz		
	Ableitung 1	Ableitung 2	Ableitung 3
OTHER	2 303	2 284	2 245
AF	1 660	1 660	1 661

5.3 Modellarchitektur

Das in dieser Arbeit entwickelte Modell soll auf 12-Kanal-EKGs trainiert werden, jedoch anschließend in der Lage sein, die EKGs von mobilen EKG-Patches zu klassifizieren. Da sich die Elektrodenplatzierung mobiler EKG-Patches zu der Elektrodenplatzierung beim Goldstandard, dem 12-Kanal-EKG, unterscheidet, wird die elektrische Aktivität im Herzen aus einer anderen Perspektive abgebildet. Dies führt dazu, dass die aufgezeichneten EKGs signalmorphologische Unterschiede zu den 12-Kanal-EKGs enthalten. Es gibt also in der Signalmorphologie einen Domain Shift von der Quelldomäne, dem 12-Kanal-EKG, hin zur Zieldomäne, den Aufnahmen der mobilen EKG-Patches.

Um diesem Domain Shift entgegenzuwirken, werden zwei Methoden aus der DG angewandt: Domain Adversarial Learning und Ensemble Learning.

Domain Adversarial Neural Network

Die Architektur des DANNs ist in Abb. 5.1 dargestellt. Als Eingabe dienen 1-Kanal-EKGs mit den Domain Labels I, II, III, aVR, aVL und aVF und den Zielklassen Labels VHF und nicht-VHF. Als Feature Extractor des DANNs wurde ein einzelner InceptionTime Klassifikator im klassischen InceptionTime Design (siehe Kapitel 4.1) gewählt.

In jedem Inception Modul sind somit 3 Filtersets mit jeweils 32 Filtern mit den unterschiedlichen Filtergrößen $l \in \{10, 20, 40\}$ enthalten. Filtergrößen mit geringerem Wahrnehmungsbereich dienen zur Erkennung feinerer Muster im EKG wie P-Welle und F-Wellen, wobei sich

Filter mit größerem Wahrnehmungsbereich besser zur Erkennung von längerfristigen Merkmalen wie einer absoluten Arrhythmie eignen. Die Anzahl der Inception Module wurde als Hyperparameter in der Grid Search ermittelt.

An den GAP-Layer des Feature Extractors wurden angelehnt an die Modelle von Ganin et al. [7] ein Domain Classifier und ein Label Predictor angebracht. Der Domain Classifier besteht aus zwei FC-Layern mit Rectified Linear Unit (ReLU) als Aktivierungsfunktion und einem FC-Layer mit 6 Neuronen als Ausgabeschicht mit einer Softmax-Aktivierungsfunktion. Der Label Predictor besteht ebenso aus zwei FC-Layern mit einer ReLU-Aktivierungsfunktion. Die Ausgabeschicht ist ein FC-Layer mit 2 Neuronen und einer Softmax-Aktivierungsfunktion.

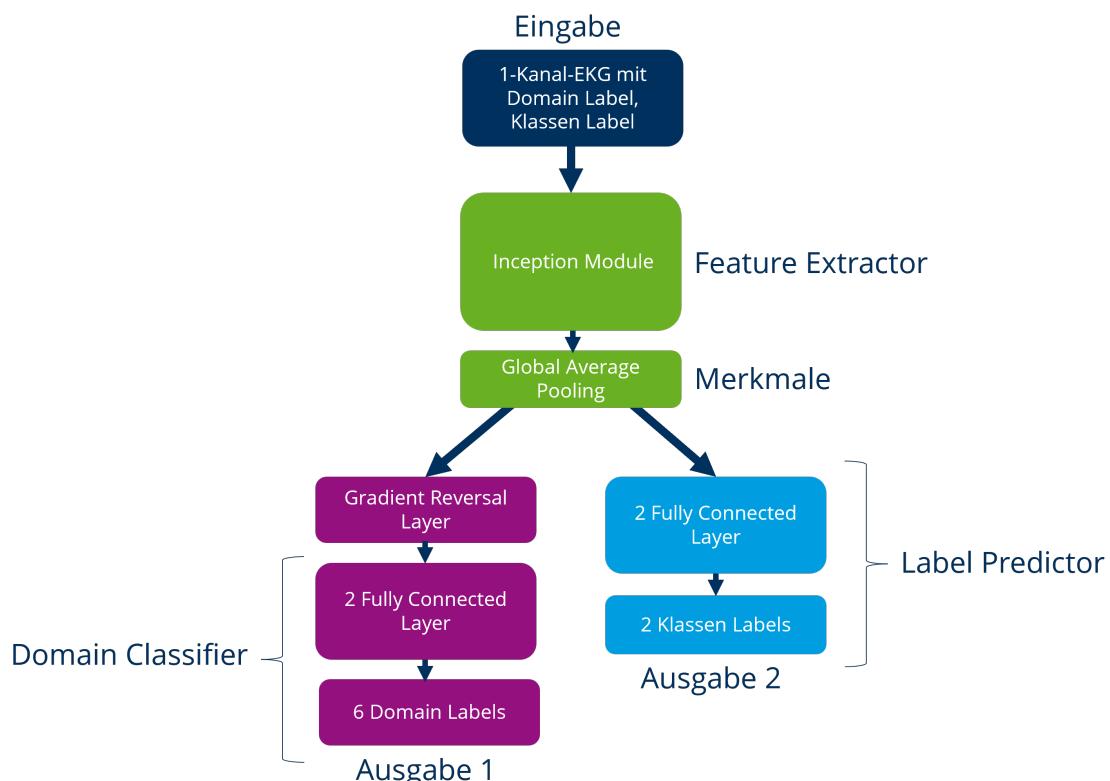


Abb. 5.1: Architektur des DANNs. Als Eingabe dienen 1-Kanal-EKGs. Als Feature Extractor dient InceptionTime. Für den Domain Classifier wird an den GAP-Layer der GRL angeschlossen, worauf zwei FC-Layer folgen. Die Ausgabeschicht des Domain Classifiers besitzt 6 Neuronen für die 6 Domänen. Der Label Predictor besitzt 2 FC-Layer angeschlossen an den GAP-Layer und 2 Neuronen als Ausgabeschicht für die zwei Zielklassen VHF und nicht-VHF.

DANN mit Direktausgabe der Merkmale

Zur Untersuchung des Einflusses der Feature-Transformation, die in den FC-Layern stattfindet, auf die Generalisierungsfähigkeit des Modells wird ein DANN ohne FC-Layer zwischen GAP-Layer und Ausgabeschicht im Label Predictor trainiert. Dadurch werden die Merkmale aus dem Feature Extractor ohne Zwischenrepräsentationen direkt ausgegeben. Diese Version des DANNs wird im weiteren Verlauf der Arbeit als DANNdirect bezeichnet und ist in Abb. 5.2 dargestellt.

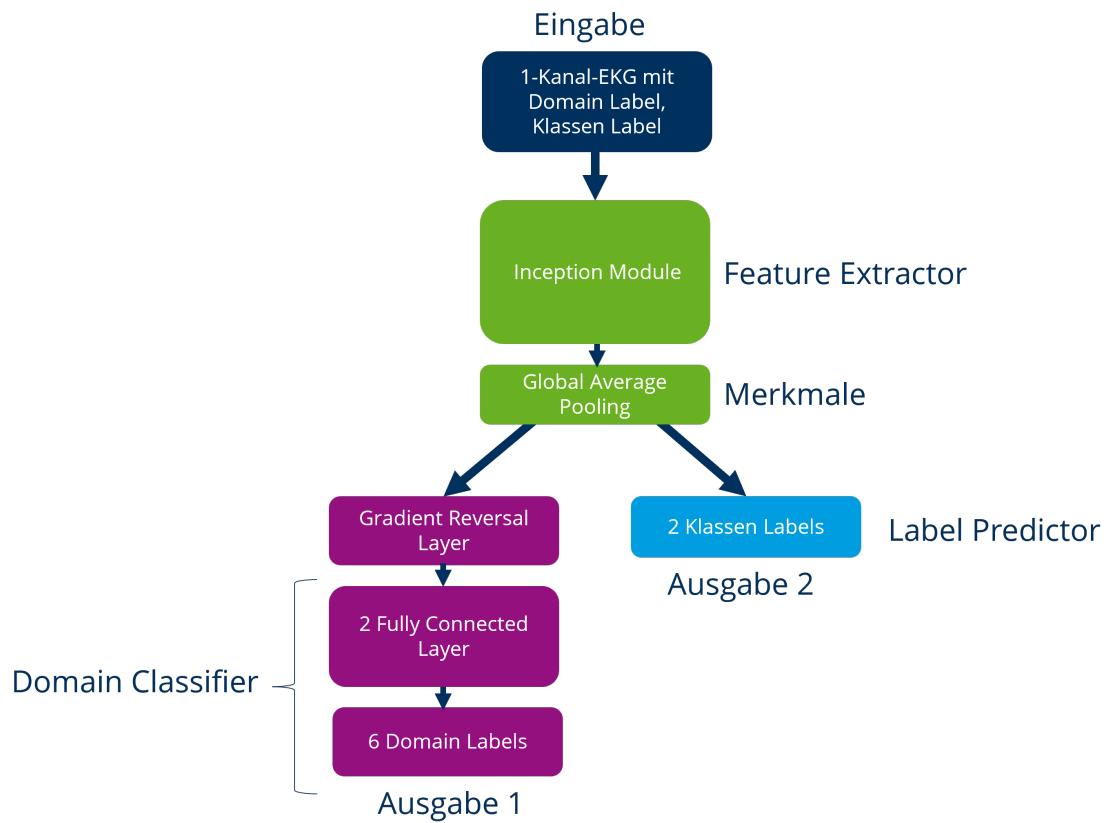


Abb. 5.2: Architektur des DANNdirect Modells. Als Eingabe dienen 1-Kanal-EKGs. Als Feature Extractor dient InceptionTime. Für den Domain Classifier wird an den GAP-Schicht den GRL angeschlossen, worauf zwei FC-Layer folgen. Die Ausgabeschicht des Domain Classifiers besitzt 6 Neuronen für die 6 Domänen. Der Label Predictor besitzt 2 Neuronen als Ausgabeschicht für die zwei Zielklassen VHF und nicht-VHF.

Vergleichsmodell InceptionTime

Als Vergleichsmodell wurde ein wie in Kapitel 4.1 beschriebenes klassisches InceptionTime-Ensemble gewählt. Wie auch beim Feature Extractor des DANNs sind in den Inception Modulen jeweils 3 Filtersets mit 32 Filtern mit den Filtergrößen $l \in \{10, 20, 40\}$ enthalten. Auch beim Vergleichsmodell wurde die Anzahl der Inception Module per Grid Search ermittelt.

Für alle drei genannten Architekturen wird das Ensemble aus 5 Modellen erstellt, welche gleich gewichtet eine Mehrheitsentscheidung für die VHF-Klassifikation treffen, wie es von Fawaz et al. [6] für InceptionTime empfohlen wird. Zusätzlich wird ein Ensembleergebnis gewichtet bestimmt, indem die jeweiligen Sicherheiten gemittelt werden und mittels argmax die Klasse bestimmt wird, sodass Modelle, die sich sicherer in ihrer Entscheidung sind, mehr zum Ergebnis beitragen. Die 5 Modelle stammen aus der 5-fold Cross Validation (siehe Abschnitt 5.4) und sind somit auf teilweise unterschiedlichen Daten trainiert.

5.4 Trainings- und Optimierungsprozess

Das Training der Modelle fand auf dem High Performance Computer (HPC) der Technischen Universität Dresden statt. Genutzt wurde das GPU-Cluster Alpha Centauri, welches AMD

EPYC CPUs und NVIDIA A100-SXM4 GPUs besitzt. Die lokale Entwicklung und die Evaluation der Modelle fand auf einer NVIDIA GeForce RTX 3070 Laptop GPU statt. Programmiert wurde in Python 3.10, genutzte Frameworks sind aeon 0.11.1, tensorflow-gpu 2.9.0 und keras 2.9.0. Aeon wird zum Import der IndividualInceptionTimeClassifier genutzt.

Für das Training des DANNs und DANNDirects wurde der Adam Optimizer genutzt. Der beim Training beobachtete Loss ist der ungewichtete kombinierte Loss aus dem Label Predictor Loss und dem Domain Classifier Loss. Das Training wurde per Early Stopping beendet, wenn für 20 Epochen keine weitere Verringerung im Loss des Label Predictors auftrat. Für das Training wurde ein absoluter Grenzwert von 500 Epochen gewählt. Die Hyperparameteroptimierung wurde per Grid Search durchgeführt. Optimierte wurden die Hyperparameter batch_size = [32, 64, 128], learning_rate = [0.01, 0.001, 0.0001] und Anzahl der Inception-Module, bezeichnet als depth = [3, 6, 9]. Die Konstante des GRLs wird auf 1 gesetzt, sodass der Gradient des Domain Classifiers während der Backpropagation vollständig invertiert wird.

Obwohl nach Ganin et al. [7] für Domain Adversarial Learning keine Domain Labels nötig sind, wird in diesem Ansatz vollständig supervised Learning genutzt, da Domain Labels in Form der jeweiligen Ableitungen I, II, III, aVR, aVL und aVF der EKG-Signale vorhanden sind. Die Zielklassen des Label Predictors sind VHF und nicht-VHF. Zur Validierung während des Trainings wird 5-fold Cross Validation mit balancierten Klassen (siehe Tab. 5.5) sowohl in den Zielklassen als auch in den Domain-Klassen genutzt. Die Modelle, die während der 5-fold Cross Validation trainiert wurden, werden für das Ensemble genutzt.

Für das Training der InceptionTime-Modelle werden dieselben Werte in der Hyperparameteroptimierung angepasst. Das Training wird ebenfalls per Early Stopping beendet, sobald der Loss der InceptionTime-Ausgabe nach 20 Epochen nicht weiter gesunken ist. Für das Training wurde ein absoluter Grenzwert von 500 Epochen gewählt. Zur Validierung wird ebenfalls 5-fold Cross Validation genutzt. Die Modelle, die während der Cross Validation trainiert wurden, werden für das Ensemble genutzt. Trainingsklassen sind VHF und nicht-VHF.

Tab. 5.5: Verteilung der Trainings- und Validierungslabels für Vorhofflimmern (VHF) und Domains in der Cross Validation.

Fold	Datensatz	Klasse					
Fold 1		nicht-VHF	VHF	-	-	-	-
	Training VHF	21367	21199	-	-	-	-
	Validation VHF	5321	5321	-	-	-	-
		I	II	III	aVR	aVL	aVF
	Training Domain	7077	7127	7141	7103	7057	7061
	Validation Domain	1791	1741	1727	1765	1811	1807
Fold 2		nicht-VHF	VHF	-	-	-	-
	Training VHF	21394	21172	-	-	-	-
	Validation VHF	5294	5348	-	-	-	-
		I	II	III	aVR	aVL	aVF
	Training Domain	7090	7104	7033	7063	7148	7128
	Validation Domain	1778	1764	1835	1805	1720	1740
Fold 3		nicht-VHF	VHF	-	-	-	-
	Training VHF	21316	21250	-	-	-	-
	Validation VHF	5372	5270	-	-	-	-
		I	II	III	aVR	aVL	aVF
	Training Domain	7086	7073	7079	7153	7054	7121
	Validation Domain	1782	1795	1789	1715	1814	1747
Fold 4		nicht-VHF	VHF	-	-	-	-
	Training VHF	21294	21273	-	-	-	-
	Validation VHF	5394	5247	-	-	-	-
		I	II	III	aVR	aVL	aVF
	Training Domain	7090	7117	7134	7144	7042	7040
	Validation Domain	1778	1751	1734	1724	1826	1828
Fold 5		nicht-VHF	VHF	-	-	-	-
	Training VHF	21381	21186	-	-	-	-
	Validation VHF	5307	5334	-	-	-	-
		I	II	III	aVR	aVL	aVF
	Training Domain	7129	7051	7085	7009	7171	7122
	Validation Domain	1739	1817	1783	1859	1697	1746

6 Ergebnisse

In diesem Kapitel werden in Abschnitt 6.1 zunächst die Ergebnisse der Hyperparameteroptimierung und somit die jeweils optimale Hyperparameterkombination der Modelle vorgestellt. Anschließend werden die Modelle auf Daten der Quelldomäne (Abschnitt 6.2) und auf Daten der Zieldomäne (Abschnitt 6.3) evaluiert. Zuletzt werden in Abschnitt 6.4 Ergebnisse von Modellen, welche mit nicht-normalisierten Daten trainiert wurden, präsentiert.

6.1 Ergebnisse der Hyperparameteroptimierung

Die Hyperparameteroptimierungen wurde ausgewertet, indem für die Metriken der 5 Modelle der Cross Validation der Mittelwert gebildet und von diesem die Standardabweichung σ zwischen den 5 Modellen abgezogen wurde. Die beste Hyperparameterkombination wurde anhand des F1-Scores ausgewählt. Optimierte Hyperparameter sind Anzahl der Inception Module, genannt `depth d`, `learning rate l` und `batch size b`.

DANN

Alle Hyperparameterkombinationen und deren Klassifikationsgüte sind in Tab. 6.1 zu sehen. Die beste Hyperparameterkombination ist diejenige mit 9 Inception Modulen, einer `learning rate` von 0,0001 und einer `batch size` von 32. Sie erreicht einen F1-Score von 0,954 (Mittelwert - σ), einen Recall von 0,943 (Mittelwert - σ) und eine Specificity von 0,964 (Mittelwert - σ). In Abb. 6.1 ist als Beispiel der Trainings- und Validierungsverlauf von Loss und Accuracy des Label Predictors und des Domain Classifiers des DANN Modells aus Fold 1 der Cross Validation zu sehen.

DANNdirect

Für das DANNdirect wurde keine separate Grid Search durchgeführt, sondern es wurde diejenige Hyperparameterkombination für das Training genutzt, die bei der Grid Search des DANN als am besten identifiziert wurde. Es erreicht mit 9 Inception Modulen, einer `learning rate` von 0,0001 und einer `batch size` von 32 einen F1-Score von 0,937 (bereinigt um $\sigma = 0,003$),

Tab. 6.1: Ergebnisse der Grid Search für das DANN. Optimierte Hyperparameter sind depth d (Anzahl der Inception Module), learning rate l und batch size b. Die Metriken wurden berechnet, indem vom Durchschnitt aus allen 5 Folds der Cross Validation die Standardabweichung σ abgezogen wurde. σ ist für die jeweilige Metrik rechts neben der betreffenden Metrik angegeben. Negative Werte kommen zustande, wenn σ größer als der Durchschnitt ist. Die beste Hyperparameterkombination ist fett hervorgehoben.

Hyperparameterkombination	F1	σ	Recall	σ	Specificity	σ
d=3 l=0,0001 b=128	0,931	0,005	0,941	0,003	0,915	0,012
d=3 l=0,0001 b=64	0,930	0,005	0,919	0,012	0,931	0,008
d=3 l=0,0001 b=32	0,929	0,004	0,928	0,008	0,923	0,006
d=3 l=0,001 b=128	0,369	0,303	0,312	0,365	0,650	0,206
d=3 l=0,001 b=64	0,333	0,365	0,380	0,382	0,383	0,385
d=3 l=0,001 b=32	0,765	0,113	0,921	0,005	0,439	0,335
d=3 l=0,01 b=128	-0,134	0,268	-0,200	0,400	0,400	0,400
d=3 l=0,01 b=64	-0,134	0,268	-0,200	0,400	0,400	0,400
d=3 l=0,01 b=32	-0,134	0,268	-0,200	0,400	0,400	0,400
d=6 l=0,0001 b=128	0,951	0,002	0,943	0,003	0,957	0,005
d=6 l=0,0001 b=64	0,949	0,003	0,935	0,007	0,959	0,005
d=6 l=0,0001 b=32	0,950	0,003	0,943	0,005	0,953	0,007
d=6 l=0,001 b=128	0,950	0,003	0,939	0,005	0,962	0,003
d=6 l=0,001 b=64	0,387	0,377	0,380	0,374	0,959	0,012
d=6 l=0,001 b=32	0,381	0,381	0,377	0,377	0,956	0,016
d=6 l=0,01 b=128	-0,134	0,268	-0,200	0,400	0,400	0,400
d=6 l=0,01 b=64	-0,134	0,268	-0,200	0,400	0,400	0,400
d=6 l=0,01 b=32	-0,134	0,268	-0,200	0,400	0,400	0,400
d=9 l=0,0001 b=128	0,952	0,003	0,940	0,005	0,959	0,008
d=9 l=0,0001 b=64	0,952	0,003	0,940	0,007	0,961	0,005
d=9 l=0,0001 b=32	0,954	0,002	0,943	0,004	0,964	0,003
d=9 l=0,001 b=128	0,952	0,001	0,944	0,003	0,958	0,004
d=9 l=0,001 b=64	0,954	0,002	0,945	0,004	0,958	0,004
d=9 l=0,001 b=32	0,950	0,002	0,937	0,006	0,961	0,003
d=9 l=0,01 b=128	-0,083	0,406	-0,088	0,476	0,395	0,396
d=9 l=0,01 b=64	-0,134	0,268	-0,200	0,400	0,400	0,400
d=9 l=0,01 b=32	-0,134	0,268	-0,200	0,400	0,400	0,400

einen Recall von 0,932 (bereinigt um $\sigma = 0,009$) und eine Specificity von 0,933 (bereinigt um $\sigma = 0,007$).

Vergleichsmodell InceptionTime

Die vollständigen Ergebnisse der Grid Search des InceptionTime Modells sind im Anhang A.1 zu finden. Die beste Hyperparameterkombination ist diejenige mit 9 Inception Modulen, einer learning rate von 0,0001 und einer batch size von 64. Die durchschnittlichen Werte der Modelle aus der Cross Validation sind ein F1-Score von 0,939 (bereinigt um $\sigma = 0,002$), einen Recall von 0,935 (bereinigt um $\sigma = 0,006$) und eine Specificity von 0,939 (bereinigt um $\sigma = 0,003$).

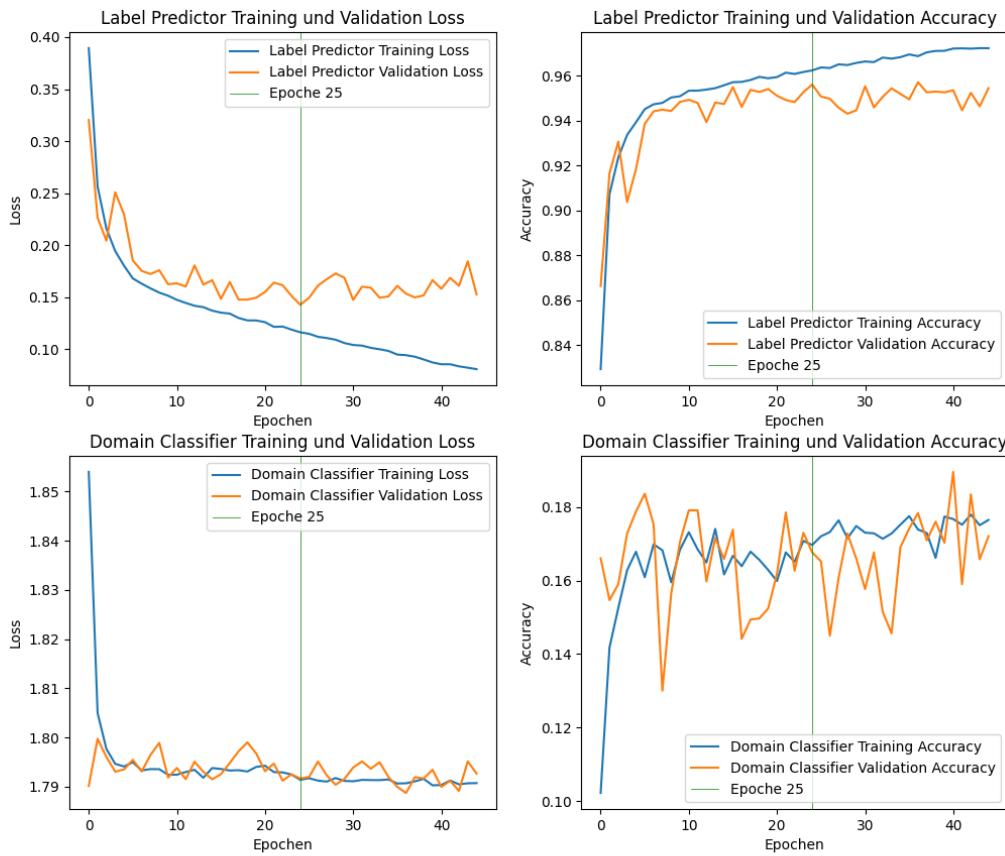


Abb. 6.1: Trainings- und Validierungsverlauf des Label Predictors (oben) und des Domain Classifiers (unten) des DANN Modells aus Fold 1 der Cross Validation. Epoche 25 ist die beste Epoche des Trainings.

Tab. 6.2: Evaluation der Modelle mit dem Testdatensatz der xECGArch-Datenbank. Globaler F1-Score der gewichteten und ungewichteten Ensembles, sowie Standardabweichung σ und Durchschnittswert der Einzelmodelle. Der beste Wert ist hervorgehoben.

Modell	xECGArch-Datenbank			
	F1 \varnothing Modelle	σ	F1 Ensemble gewichtet	F1 Ensemble ungewichtet
DANN	0,947	0,003	0,951	0,950
DANNdirect	0,931	0,003	0,940	0,938
InceptionTime	0,930	0,002	0,938	0,938

6.2 Evaluierung auf Daten der Quelldomäne

Die Modelle wurden auf dem Testdatensatz der xECGArch-Datenbank evaluiert. Auf diesen Datensatz hatten die Modelle während des Trainings keinen Zugriff. Zur Klassifikation wurden für alle EKGs die Ableitungen I, II, III, aVR, aVL und aVF als eine gemischte Datenmenge ein-

Tab. 6.3: Evaluation der Modelle mit dem Testdatensatz der xECGArch-Datenbank. Ergebnisse des gewichteten und ungewichteten Ensembles. Der jeweils beste F1-Score pro Ableitung ist hervorgehoben.

Modell	Ensemble gewichtet			Ensemble ungewichtet		
	F1	Recall	Specificity	F1	Recall	Specificity
Ableitung I						
DANN	0,947	0,933	0,960	0,945	0,935	0,954
DANNdirect	0,931	0,935	0,923	0,928	0,931	0,921
InceptionTime	0,929	0,929	0,925	0,930	0,931	0,925
Ableitung II						
DANN	0,954	0,933	0,975	0,954	0,933	0,975
DANNdirect	0,947	0,943	0,950	0,941	0,931	0,952
InceptionTime	0,946	0,941	0,950	0,946	0,937	0,954
Ableitung III						
DANN	0,946	0,929	0,962	0,943	0,923	0,962
DANNdirect	0,939	0,937	0,937	0,937	0,937	0,933
InceptionTime	0,935	0,929	0,939	0,935	0,929	0,939
Ableitung aVR						
DANN	0,962	0,955	0,969	0,961	0,951	0,971
DANNdirect	0,944	0,939	0,948	0,942	0,935	0,948
InceptionTime	0,943	0,937	0,948	0,943	0,933	0,952
Ableitung aVL						
DANN	0,944	0,931	0,956	0,943	0,927	0,958
DANNdirect	0,930	0,935	0,921	0,926	0,929	0,919
InceptionTime	0,926	0,929	0,919	0,921	0,925	0,910
Ableitung aVF						
DANN	0,956	0,939	0,973	0,957	0,941	0,973
DANNdirect	0,950	0,945	0,952	0,953	0,949	0,956
InceptionTime	0,950	0,949	0,948	0,950	0,947	0,952

gegeben, analog zur Struktur des Trainingsdatensatzes. Der globale F1-Score wurde aus der Gesamtmenge der Vorhersagen gesammelt berechnet. Das dynamisch nach Klassifikations- sicherheit der einzelnen Modelle gewichtete DANN Ensemble erreicht mit 0,951 den besten globalen F1-Score. Die globalen F1-Scores der Ensembles, sowie den Durchschnitt der einzelnen Modelle der Ensembles und die Standardabweichung dieser lassen sich in Tab. 6.2 nachlesen.

In den Abbildungen 6.2 und 6.3 sind als Beispiel die ROC-Kurven des Label Predictors und des Domain Classifiers des DANN Modells aus Fold 1 der Cross Validation dargestellt. Dabei fällt auf, dass die ROC-Kurve des Label Predictors des DANN Modells eine große Fläche von 0,982 aufspannt, während die ROC-Kurven des Domain Classifiers des DANN Modells nahe an der Winkelhalbierenden liegen und Flächen zwischen 0,477 bis 0,564 aufspannen.

Aufgeteilt nach Ableitung (siehe Tab. 6.3) wird der beste F1-Score mit 0,962 auf Ableitung aVR

vom gewichteten DANN Ensemble erreicht. Die durchschnittlichen F1-Scores der Ensembles über alle Ableitungen betragen für das DANN 0,952 im gewichteten und 0,951 im ungewichteten Fall. Das gewichtete DANNDirect Ensemble erreicht einen durchschnittlichen F1-Score von 0,940 und das ungewichtete einen von 0,938. Das gewichtete und ungewichtete InceptionTime Ensemble erreichen einen durchschnittlichen F1-Score von 0,938.

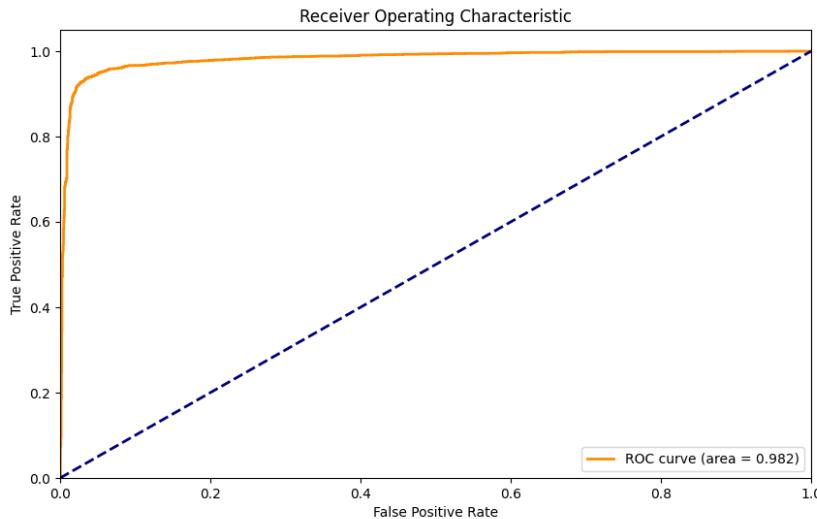


Abb. 6.2: Receiver-Operating-Characteristic-Kurve des Label Predictors des DANN Modells aus Fold 1 der Cross Validation.

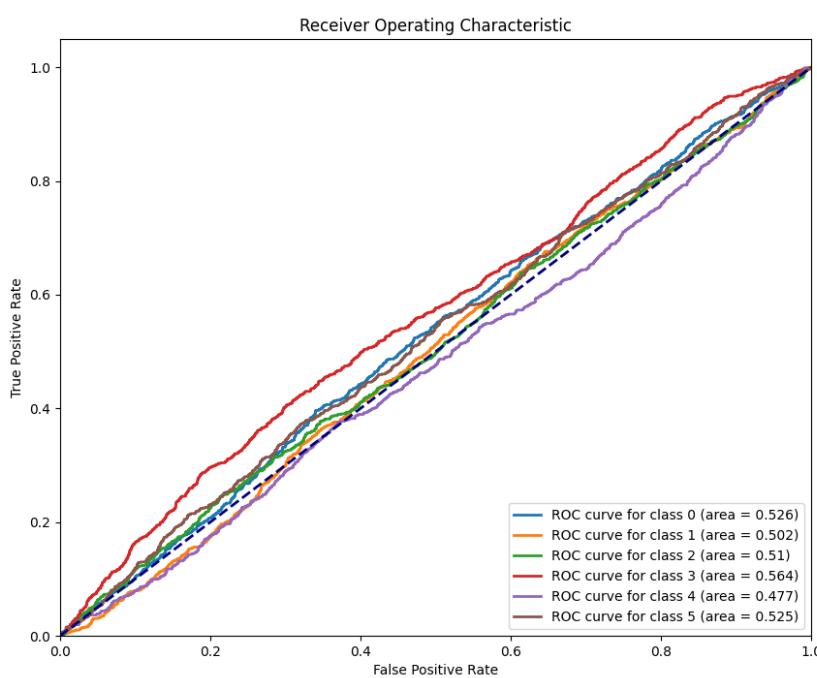


Abb. 6.3: Receiver-Operating-Characteristic-Kurve des Domain Classifiers des DANN Modells aus Fold 1 der Cross Validation.

6.3 Evaluierung auf Daten der Zieldomänen

Um einzuschätzen, wie die Modelle mit einem Domain Shift umgehen, wurden die Modelle auf Daten von mobilen EKG-Patches evaluiert, deren Signale eine andere Morphologie im Vergleich zu 12-Kanal-EKG-Signalen aufweisen.

Icentia11k

Bei der Icentia11k-Datenbank erreicht das ungewichtete DANN Ensemble mit 0,666 den besten F1-Score (siehe Tab. 6.4) und den besten Recall mit 0,979. Die beste Specificity wird mit 0,952 sowohl vom ungewichteten als auch vom gewichteten DANN Ensemble erreicht (siehe Tab. 6.5).

Tab. 6.4: Evaluation der Modelle auf der Zieldomäne Icentia11k. F1-Score der gewichteten und ungewichteten Ensembles, sowie Standardabweichung σ und Durchschnittswert der Einzelmodelle.

Modell	Icentia11k			
	F1 Ø Modelle	σ	F1 Ensemble gewichtet	F1 Ensemble ungewichtet
DANN	0,641	0,018	0,665	0,666
DANNdirect	0,498	0,032	0,518	0,515
InceptionTime	0,520	0,013	0,543	0,541

Tab. 6.5: Evaluation der Modelle auf der Zieldomäne Icentia11k. Recall und Specificity des gewichteten und ungewichteten Ensembles.

Modell	Ensemble gewichtet		Ensemble ungewichtet	
	Recall	Specificity	Recall	Specificity
DANN	0,973	0,952	0,979	0,952
DANNdirect	0,947	0,915	0,943	0,915
InceptionTime	0,936	0,925	0,936	0,924

TIMELY-Datensatz

Die durchschnittliche Klassifikationsgüte der Ensembles auf dem TIMELY-Datensatz beträgt für das DANN 0,973 (F1-Score) im gewichteten und 0,969 (F1-Score) im ungewichteten Fall. Das gewichtete DANNdirect Ensemble erreicht einen durchschnittlichen F1-Score von 0,948 und das ungewichtete einen von 0,947. Das gewichtete und das ungewichtete InceptionTime Ensemble erreichen jeweils einen durchschnittlichen F1-Score von 0,949.

Die Ergebnisse des TIMELY-Datensatzes wurden nach Ableitung aufgeschlüsselt. Auf allen drei Ableitungen erreicht jeweils das gewichtete DANN Ensemble den besten F1-Score mit 0,952 auf Ableitung 1, 0,986 auf Ableitung 2 und 0,981 auf Ableitung 3 (siehe Tab. 6.6). Recall und Specificity der Modelle lässt sich aus Tab. 6.7 entnehmen.

Tab. 6.6: Evaluation der Modelle mit den drei Ableitungen des TIMELY-Datensatzes. F1-Score der gewichteten und ungewichteten Ensembles, sowie Standardabweichung σ und Durchschnittswert der Einzelmodelle.

Modell	F1 \varnothing Modelle	σ	F1 Ensemble	F1 Ensemble
			gewichtet	ungewichtet
TIMELY Ableitung 1				
DANN	0,946	0,022	0,952	0,942
DANNdirect	0,908	0,020	0,903	0,905
InceptionTime	0,908	0,026	0,904	0,905
TIMELY Ableitung 2				
DANN	0,972	0,022	0,986	0,985
DANNdirect	0,969	0,004	0,972	0,972
InceptionTime	0,971	0,004	0,976	0,975
TIMELY Ableitung 3				
DANN	0,972	0,006	0,981	0,980
DANNdirect	0,959	0,005	0,969	0,965
InceptionTime	0,957	0,006	0,968	0,968

Tab. 6.7: Evaluation der Modelle auf der Zieldomäne TIMELY. Recall und Specificity des gewichteten und ungewichteten Ensembles.

Modell	Ensemble gewichtet		Ensemble ungewichtet	
	Recall	Specificity	Recall	Specificity
TIMELY Ableitung 1				
DANN	1,00	0,927	1,00	0,912
DANNdirect	0,997	0,848	0,995	0,853
InceptionTime	0,998	0,850	0,998	0,851
TIMELY Ableitung 2				
DANN	1,00	0,979	1,00	0,978
DANNdirect	0,981	0,973	0,981	0,972
InceptionTime	0,981	0,979	0,981	0,978
TIMELY Ableitung 3				
DANN	1,00	0,971	1,00	0,969
DANNdirect	0,977	0,971	0,974	0,967
InceptionTime	0,978	0,969	0,980	0,967

SHDB-AF

Da 23 Patienten der SHDB-AF während ihrer Aufnahme keine VHF-Episode hatten, konnte für diese Patienten der Recall und somit der F1-Score nicht berechnet werden, sodass die Evaluierung nur für die übrigen 77 Patienten durchgeführt wurde. Bei der SHDB-AF erreicht das DANN Ensemble einen durchschnittlichen F1-Score von 0,815 im gewichteten und einen durchschnittlichen F1-Score von 0,810 im ungewichteten Fall. Das gewichtete DANNdirect Ensemble erreicht einen durchschnittlichen F1-Score von 0,755 und das ungewichtete einen

durchschnittlichen F1-Score von 0,753. Für das gewichtete InceptionTime Ensemble beträgt der durchschnittliche F1-Score 0,757 und für das ungewichtete beträgt der durchschnittliche F1-Score 0,756. Die Standardabweichung im F1-Score des gewichteten DANN Ensembles zwischen den 77 Patienten mit VHF beträgt 0,220.

In Tab. 6.8 und Tab. 6.9 ist die Klassifikationsgüte der Modelle nach Ableitung aufgeschlüsselt. Auf den einzelnen Ableitungen erreicht das gewichtete DANN den besten F1-Score mit 0,777 auf der CC5-Ableitung und 0,852 auf der NASA-Ableitung.

Tab. 6.8: Evaluation der Modelle mit den zwei Ableitungen der SHDB-AF-Datenbank. F1-Score der gewichteten und ungewichteten Ensembles, sowie Standardabweichung σ und Durchschnittswert der Einzelmodelle.

Modell	F1 \varnothing Modelle	σ	F1 Ensemble gewichtet	F1 Ensemble ungewichtet
SHDB-AF CC5-Ableitung				
DANN	0,763	0,014	0,777	0,776
DANNdirect	0,694	0,028	0,717	0,715
InceptionTime	0,717	0,012	0,731	0,730
SHDB-AF NASA-Ableitung				
DANN	0,824	0,012	0,852	0,844
DANNdirect	0,753	0,018	0,793	0,790
InceptionTime	0,753	0,012	0,783	0,781

Tab. 6.9: Evaluation der Modelle auf der Zieldomäne SHDB-AF. Recall und Specificity des gewichteten und ungewichteten Ensembles.

Modell	Ensemble gewichtet		Ensemble ungewichtet	
	Recall	Specificity	Recall	Specificity
CC5-Ableitung				
DANN	0,949	0,947	0,949	0,947
DANNdirect	0,948	0,904	0,946	0,903
InceptionTime	0,951	0,912	0,948	0,914
NASA-Ableitung				
DANN	0,933	0,944	0,932	0,944
DANNdirect	0,925	0,916	0,922	0,915
InceptionTime	0,920	0,920	0,918	0,920

6.4 Untersuchung des Einflusses der Normalisierung der Daten

Im Rahmen dieser Arbeit wurden ebenfalls Modelle trainiert, ohne zuvor die Trainingsdaten zu z-normalisieren. Die beste Hyperparameterkombination dieser Modelle wurde ebenfalls per Grid Search ermittelt. Für das DANN wurde als optimale Hyperparameterkombination 9 Inception Module, eine learning rate von 0,0001 und eine batch size von 64 ermittelt. Der

durchschnittliche F1-Score abzüglich der Standardabweichung aus den Modellen der Cross Validation beträgt 0,956. Die Standardabweichung beträgt 0,002. Das DANNdirect wurde anschließend mit derselben Hyperparameterkombination trainiert.

Für InceptionTime wurde als optimale Hyperparameterkombination eine Anzahl von 9 Inception Modulen, eine learning rate von 0,001 und eine batch size von 128 ermittelt. Der durchschnittliche F1-Score abzüglich der Standardabweichung der InceptionTime Modelle aus der Cross Validation beträgt 0,941. Die Standardabweichung beträgt 0,003.

Die Modelle wurden auf dem nicht normalisierten xECGArch-Testdatensatz evaluiert. Die höchste Klassifikationsgüte erreichen das gewichtete und ungewichtete DANN Ensemble mit einem globalen F1-Score von 0,953 (siehe Tab. 6.10). Anschließend wurden die Modelle auf dem nicht-normalisierten TIMELY-Datensatz angewendet, um zu evaluieren, wie sie sich bei einem Domain Shift verhalten. In Tab. 6.11 ist die Klassifikationsgüte der Modelle für jede Ableitung des Datensatzes eingetragen. Die durchschnittlich höchste Klassifikationsgüte erreicht das DANN Ensemble mit einem F1-Score von 0,946.

Tab. 6.10: Evaluation der Modelle mit dem nicht-normalisierten Testdatensatz der xECGArch-Datenbank. Globaler F1-Score der gewichteten und ungewichteten Ensembles, sowie Standardabweichung σ und Durchschnittswert der Einzelmodelle. Der beste Wert ist hervorgehoben.

Modell	xECGArch-Datenbank			
	F1 \otimes Modelle	σ	F1 Ensemble gewichtet	F1 Ensemble ungewichtet
DANN	0,948	0,003	0,953	0,953
DANNdirect	0,932	0,003	0,937	0,938
InceptionTime	0,934	0,003	0,942	0,942

Tab. 6.11: Evaluation der Modelle mit den drei nicht-normalisierten Ableitungen des TIMELY-Datensatzes. F1-Score der gewichteten und ungewichteten Ensembles, sowie Standardabweichung σ und Durchschnittswert der Einzelmodelle.

Modell	F1 \otimes Modelle	σ	F1 Ensemble gewichtet	F1 Ensemble ungewichtet
			TIMELY Ableitung 1	
DANN	0,935	0,014	0,949	0,947
DANNdirect	0,892	0,022	0,891	0,892
InceptionTime	0,902	0,025	0,916	0,915
TIMELY Ableitung 2				
DANN	0,938	0,036	0,921	0,917
DANNdirect	0,931	0,039	0,940	0,961
InceptionTime	0,952	0,007	0,967	0,966
TIMELY Ableitung 3				
DANN	0,936	0,022	0,968	0,962
DANNdirect	0,920	0,023	0,949	0,945
InceptionTime	0,903	0,022	0,936	0,931

7 Diskussion

Ziel dieser Arbeit war die Entwicklung eines Modells zur VHF-Detektion im EKG, das robust gegen veränderte Signalmorphologie ist. Dazu wurden 3 Modelle (DANN, DANNDirect und InceptionTime) trainiert und auf Daten der Quell- und der Zieldomäne getestet. Die entsprechenden Ergebnisse werden in diesem Kapitel eingeordnet und diskutiert. In Abschnitt 7.1 wird zunächst die Klassifikationsgüte der Modelle auf der Quelldomäne eingeordnet. In Abschnitt 7.2 werden die Ergebnisse der Hyperparameteroptimierung und in Abschnitt 7.3 die Ergebnisse auf den Testdatensätzen diskutiert. Zuletzt wird in Abschnitt 7.4 der Einfluss von Normalisierung diskutiert.

7.1 Allgemeine Einordnung der Klassifikationsgüte der Modelle

Der Trainingsverlauf in Abb. 6.1 zeigt, dass das DANN wie erwartet trainiert. Der Loss des Label Predictors sinkt im Verlauf des Trainings bis Epoche 25, nach welcher per Early Stopping das Training beendet wurde, während der Loss des Domain Classifiers konstant hoch blieb. Die ROC-Kurven von Label Predictor (siehe Abb. 6.2) und Domain Classifier (siehe Abb. 6.3) zeigen ebenfalls, dass das Modell in der Lage ist, die Hauptklassifikationsaufgabe auszuführen, während der Domain Classifier nicht in der Lage ist, die einzelnen Ableitungen zu differenzieren. Die Fläche, die durch die ROC-Kurve des Label Predictors aufgespannt wird, ist mit 0,982 sehr groß, was laut [74] als sehr gut eingeordnet werden kann. Die Teilhypthesen a) und b) konnten somit bewiesen werden.

Werden die einzelnen Modelle zu einem ungewichteten Ensemble kombiniert, erhöht sich die Gesamtgüte (siehe Tab. 6.2). Gewichtet man das Ensemble, erhöht sich die Klassifikationsgüte noch einmal. Teilhypothese d) wurde somit bestätigt.

Teilhypothese e) besagt, dass das DANN auf dem xECGArch-Testdatensatz eine geringere Klassifikationsgüte erzielen sollte als das Vergleichsmodell, da es aufgrund des Domain Adversarial Learnings weniger stark overfittet. Dies konnte nicht nachgewiesen werden, da sowohl die einzelnen Modelle als auch beide DANN Ensembles auf dem Testdatensatz der 12-Kanal-EKGs eine höhere Klassifikationsgüte erzielen als das Vergleichsmodell. Dies lässt sich durch die Tatsache erklären, dass die 12-Kanal-EKGs nicht insgesamt, sondern jeweils nur ein einzel-

ner Kanal als Eingabe der Modelle genutzt wurde. Zwischen den einzelnen Kanälen existieren morphologische Unterschiede, sodass Domain Adversarial Learning auch hier dazu beiträgt, universelle Merkmale zu erlernen.

Das von Goettling et al. [31] entwickelte Modell xECGArch, dessen Trainings- und Testdatensatz in dieser Arbeit verwendet wurde, nutzt Ableitung II für Training und Evaluation. Es erreicht bei der Detektion von VHF einen F1-Score von 0,954, sowie einen Recall von 0,949 und eine Specificity von 0,958. Das gewichtete DANN Ensemble, welches mit den 6 Extremitätenableitungen des xECGArch-Datensatzes trainiert wurde, erreicht auf Ableitung II ebenfalls einen F1-Score von 0,954, sowie einen etwas geringeren Recall von 0,933 und eine etwas höhere Specificity von 0,975. Aus dem Vergleich der F1-Scores lässt sich schließen, dass das gewichtete DANN Ensemble eine vergleichbare Klassifikationsgüte wie xECGArch aufweist. Der leicht niedrigere Recall, sowie die leicht höhere Specificity deuten darauf hin, dass das gewichtete DANN Ensemble minimal weniger wahre VHF-Episoden detektiert und dabei weniger falsch positive Ergebnisse erreicht. In der klinischen Praxis ist wünschenswert, einen höheren Recall zu erreichen und alle tatsächlichen VHF-Fälle zu erkennen und dafür falsch positive Detektionen in Kauf zu nehmen. Recall und Specificity von über 0,900 können nach [75] dennoch als sehr gut eingeordnet werden.

Ribeiro et al. [48] haben im Rahmen ihrer Arbeit die Klassifikationsgüte von Menschen zusätzlich zur Klassifikationsgüte ihres DNNs erhoben. Assistenzärzte für Kardiologie im vierten Jahr erreichten einen F1-Score von 0,769 in der Erkennung von VHF in 12-Kanal-EKGs. Das gewichtete DANN Ensemble erreicht auf 12-Kanal-EKGs einen F-Score von 0,951. Es erzielt also eine höhere Klassifikationsgüte, als die von Ribeiro et al. getesteten Ärzte, wobei beachtet werden muss, dass es sich um verschiedene 12-Kanal-EKG-Datenbanken handelt.

Insgesamt muss angemerkt werden, dass es sich bei dem in dieser Arbeit genutzten Ansatz nicht um einen erklärbaren Ansatz handelt. Dies bedeutet, dass zwar Vermutungen angestellt werden können, welche Merkmale vom Modell zur Vorhersage genutzt werden (bspw. Rhythmus-Merkmale im Fall von Filtern mit großem Wahrnehmungsbereich und morphologische Merkmale im Fall von Filtern mit kleinerem Wahrnehmungsbereich), jedoch keine Methode umgesetzt wurde, um dies zu überprüfen.

Weiterhin muss angemerkt werden, dass bei der Vorverarbeitung der Datensätze nicht bedacht wurde, dass die Netzfrequenz in Nordamerika bei 60 Hz liegt und somit alle Datensätze mit einem Powerline Filter mit der Frequenz 50 Hz gefiltert wurden. Jedoch hat dies auf die Klassifikationsgüte vermutlich nur einen geringen Einfluss, da sowohl Daten aus Ländern mit einer Netzfrequenz von 60 Hz, als auch Daten aus Ländern mit einer Netzfrequenz von 50 Hz im Trainingsdatensatz vertreten sind.

7.2 Ergebnisse der Hyperparameteroptimierung

Einige Hyperparameterkombinationen aus der DANN Grid Search, insbesondere jene mit einer learning rate von 0,01 besitzen eine besonders geringe Klassifikationsgüte (siehe Tab. 6.1). Dabei fällt auf, dass diese Modelle alle denselben F1-Score von -0,134 (Mittelwert - σ) besitzen und sich auch in den restlichen Metriken nicht unterscheiden. Die negativen Werte in

der Tabelle kommen zustande, wenn die Standardabweichung zwischen den 5 Modellen der Cross Validation größer ist als der Durchschnittswert. Es kann vorkommen, dass bspw. durch Probleme mit einzelnen Knoten Jobs auf einem HPC nicht korrekt ausgeführt werden. Um sicherzugehen, dass die schlechten Werte nicht durch einen Fehler während der Berechnung auf dem HPC zustande kamen, wurde ein zusätzliches Modell mit einer dieser Hyperparameterkombinationen auf einem anderen Knoten trainiert, welches dasselbe Ergebnis lieferte. Die schlechten Werte entstehen durch Modelle, die am Ende des Trainingsvorgangs nur in der Lage sind, einen einzigen Wert vorherzusagen (entweder nur 0 oder nur 1). Vermutlich durch die geringe learning rate ausgelöst, hängen diese Modelle in einem lokalen Minimum fest. Sind weniger Inception Module im Modell vorhanden, ist auch eine learning rate von 0,001 zu gering, um die Modelle erfolgreich zu trainieren. Eine Anzahl von 9 Inception Modulen gleicht 0,001 als learning rate aus.

Das DANNdirect Ensemble hat keine wesentlich höhere Klassifikationsgüte als das InceptionTime Ensemble. Dabei sollte berücksichtigt werden, dass für das Training der DANNdirect Modelle keine zusätzliche Grid Search durchgeführt wurde, sondern die Ergebnisse der Hyperparameteroptimierung des DANNs genutzt wurden. Dadurch könnten die Hyperparameter des DANNdirect Ensembles suboptimal sein, was die Klassifikationsgüte der Modelle beeinträchtigt haben könnte. Das Entfernen der FC-Layer des Label Predictors könnte ebenfalls die Klassifikationsgüte beeinträchtigt haben, da die FC-Layer dazu beitragen, die Features aus dem Feature Extractor zu kombinieren und für die Klassifikation zu optimieren.

7.3 Diskussion der Ergebnisse auf unterschiedlichen Testdatensätzen

Icentia11k

Auf dem Icentia11-Datensatz liegt der F1-Score des gewichteten DANN Ensembles bei 0,665 und ist somit im Vergleich zu den F1-Scores auf dem xECGArch Datensatz (0,944-0,956) deutlich schlechter. Recall und Specificity liegen bei 0,973 und 0,952 und sind damit relativ hoch. Der hohe Recall zeigt, dass das Ensemble fast alle VHF-Fälle erkennt, eine hohe Specificity bedeutet, dass das Ensemble auch die meisten gesunden Patienten korrekt klassifiziert. Der dennoch niedrige F1-Score lässt sich durch die große Anzahl an falsch-positiven Klassifikationen erklären. Von insgesamt 513 VHF-Fällen erkennt das Ensemble 499 korrekt, jedoch werden zusätzlich 489 negative EKGs fälschlicherweise als VHF klassifiziert. Der Icentia11k-Datensatz ist sehr unausgeglichen und enthält mit 10 689 viele negative und mit 513 nur sehr wenige positive VHF-Fälle. Bei unausgeglichenen Datensätzen mit vielen negativen Fällen ist die Specificity keine zuverlässige Metrik, da es viele negative Fälle gibt, die korrekt negativ erkannt werden können.

Ein Grund für die falsch positiven Klassifikationen kann die Auflösung der Icentia11k-EKG-Signale sein. Sie ist teilweise sehr gering, sodass davon ausgegangen werden kann, dass die Modelle nur die R-Zacken und somit nur Rhythmusmerkmale extrahieren und zur Klassifikation nutzen konnten und keine morphologischen Merkmale, was den Abfall in der Klassifikations-

leistung erklären kann. Ein Beispiel für eine Aufnahme mit geringer Auflösung ist in Abb. 7.1 zu sehen. Zusätzlich gibt es eine große Anzahl an verrauschten Signalen wie in Abb. 7.2 dargestellt, die die Klassifikationsgüte beeinflusst haben könnte.

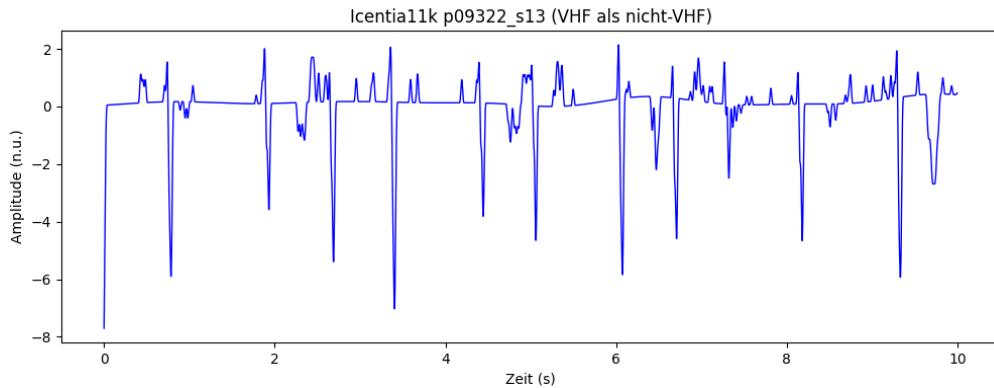


Abb. 7.1: Eine Aufnahme entnommen von Patient 09322 aus der lcentia11k-Datenbank. Sie wurde als Vorhofflimmern annotiert und vom DANN Ensemble falsch klassifiziert. Zu sehen ist, dass die Aufnahme sehr gering aufgelöst ist.

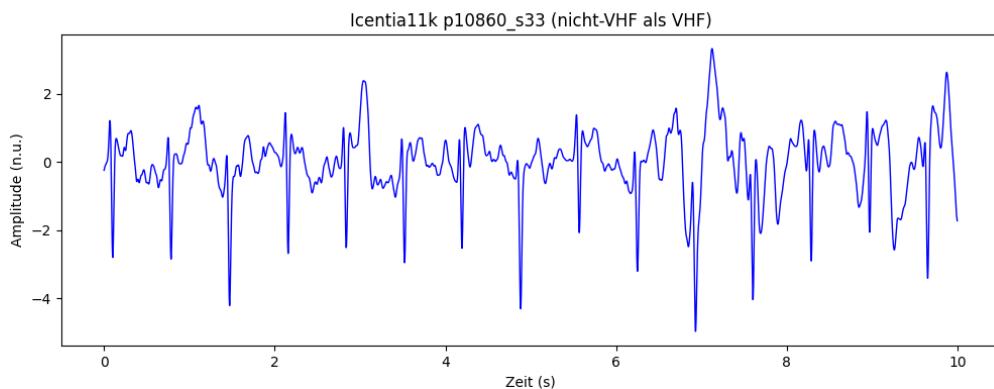


Abb. 7.2: Eine Aufnahme entnommen von Patient 10860 aus der lcentia11k-Datenbank. Sie wurde als normaler Sinusrhythmus annotiert und vom DANN Ensemble falsch klassifiziert. Zu sehen ist, dass die Aufnahme verrauscht ist.

TIMELY-Datensatz

Das gewichtete DANN Ensemble erzielt auf allen drei Ableitungen des TIMELY-Datensatzes einen F1-Score von 0,952-0,986, der laut [75] als sehr gut eingeordnet werden kann. Damit erzielt das Ensemble eine höhere Klassifikationsgüte auf einer signalmorphologisch veränderten Zieldomäne als das Modell von Ramesh et al. [56]. Dieses Modell erreicht auf der Quelldomäne EKG einen F1-Score von 0,93 und auf der Zieldomäne PPG, die ein morphologisch verändertes Signal darstellt, mit Transfer Learning einen F1-Score von 0,89.

Des Weiteren erreicht das gewichtete DANN Ensemble den bestmöglichen Recall von 1,00. Dies bedeutet, dass alle VHF-Fälle korrekt als VHF klassifiziert wurden. Zusammen mit einer hohen durchschnittlichen Specificity von 0,959 und einem hohen F1-Score lässt sich die Aussage treffen, dass es keine übermäßig hohen falsch positiven Klassifikationen gibt. Anzumerken ist

hier, dass, obwohl die TIMELY-Ableitungen bei richtiger Positionierung der Elektroden modifizierte Einthoven-Ableitungen sind, dies in der Praxis nicht zutrifft. Die Morphologie der einzelnen Ableitungen variiert stark mit der Platzierung und Rotation des EKG-Patches. Da die Patienten selbst den Patch angebracht haben, kam es vor, dass der Patch nicht über dem Herzen saß oder unter Umständen rotiert angebracht wurde, wodurch Ableitungen morphologisch verändert oder gar vertauscht sein können. Daraus lässt sich schließen, dass das Aufschlüsseln der Klassifikationsgüte nach Ableitung nicht sinnvoll ist und die Beurteilung der Klassifikationsgüte des Modells auf dem Durchschnittswert stattfinden sollte. Da auch der durchschnittliche F1-Score mit 0,973 sehr gut ist, kann Teilhypothese c) als bestätigt angesehen werden. Die Haupthypothese wurde bewiesen, da die Klassifikationsgüte des gewichteten DANN Ensembles höher ist als die Klassifikationsgüte des gewichteten InceptionTime Ensembles (F1-Score von 0,949).

SHDB-AF

Es fällt auf, dass die Klassifikationsgüte der Modelle auf der SHDB-AF Datenbank auf der CC5-Ableitung wesentlich schlechter ist, als auf der NASA-Ableitung (F1-Scores von 0,717-0,777 vs. F1-Scores von 0,783-0,852). Die Ursache hierfür liegt darin, dass es sich bei der CC5-Ableitung um eine Brustwandableitung handelt. Für das Training der Modelle wurden nur die Ableitungen I, II, III, aVR, aVL und aVF genutzt, jedoch nicht die 6 Brustwandableitungen. Somit ist der Domain Shift zwischen den Trainingsdaten und der CC5-Ableitung der SHDB-AF Datenbank größer als der zwischen den Trainingsdaten und der NASA-Ableitung der SHDB-AF Datenbank und ein Abfall der Klassifikationsleistung der Modelle zu erwarten.

Des Weiteren ist die Standardabweichung der F1-Scores für VHF innerhalb der 77 Patienten, die während der Aufzeichnung eine VHF-Episode hatten, relativ hoch. Als Beispiel werden nun die Ergebnisse des gewichteten DANN Ensembles auf der NASA-Ableitung betrachtet. Hier beträgt die Standardabweichung für die VHF-Klasse 0,220. Die Standardabweichung der F1-Scores für die nicht-VHF-Klasse innerhalb derselben 77 Patienten ist mit 0,191 vergleichbar hoch.

In Abb. 7.3 ist der F1-Score für die VHF-Klasse gegen die durchschnittliche Sicherheit des Ensembles pro Patient aufgetragen. Zusätzlich ist die Standardabweichung der Sicherheit pro Patient in Form von T-Balken eingetragen. Zu sehen ist, dass der Großteil der Teil Aufnahmen mit einem guten F1-Score und einer hohen durchschnittlichen Sicherheit klassifiziert werden konnte. Jedoch gibt es einige Patienten, deren Aufnahmen einen sehr geringen F1-Score und auch eine geringere Sicherheit in der Klassifikation aufweisen. Auf einige dieser Aufnahmen wird nun eingegangen.

Die Klassifikation von Patient 003 erfolgte mit einem F1-Score von 0,264 für VHF bei einer durchschnittlichen Sicherheit von 0,846. Bei diesem Patienten wurde ein mit Vorhofflimmern annotierter Abschnitt vom Ensemble als VHF klassifiziert, sowie ein mit VHF annotierter Abschnitt nicht als VHF detektiert. In Abb. 7.4 ist ein mit VHF annotiertes Fenster dargestellt (welches als nicht-VHF klassifiziert wurde), welches regelmäßige Vorhoferregungen aufweist, was auf Vorhofflimmern oder Sinusrhythmus hinweist und nicht auf VHF. In Abb. 7.5 ist ein mit Vorhofflimmern annotiertes Fenster dargestellt, welches als VHF klassifiziert wurde. Hier ist keine regelmäßige

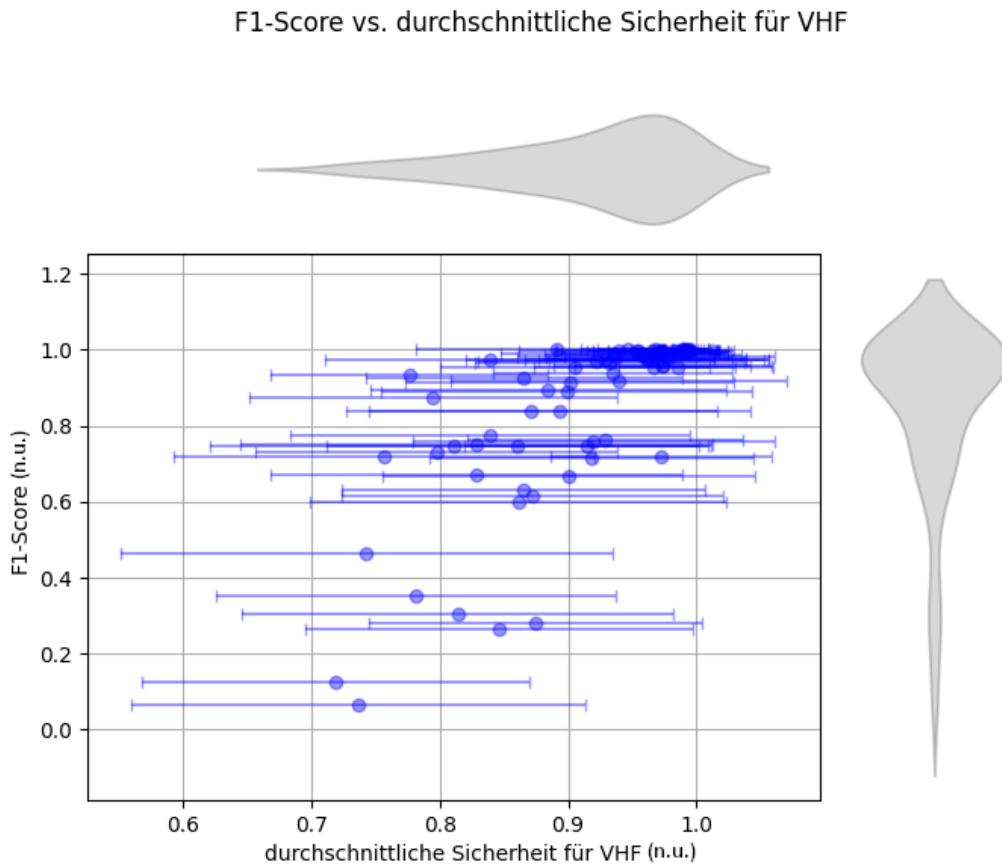


Abb. 7.3: F1-Score für die VHF-Klasse gegen die durchschnittliche Sicherheit des gewichteten DANN Ensembles pro Patient aufgetragen. Zusätzlich ist die Standardabweichung der Sicherheit pro Patient in Form von T-Balken eingetragen. Anhand des Violinenplots lässt sich die Verteilung der F1-Scores bzw. der durchschnittlichen Sicherheit für die Klassifikation über die Aufnahmen ablesen.

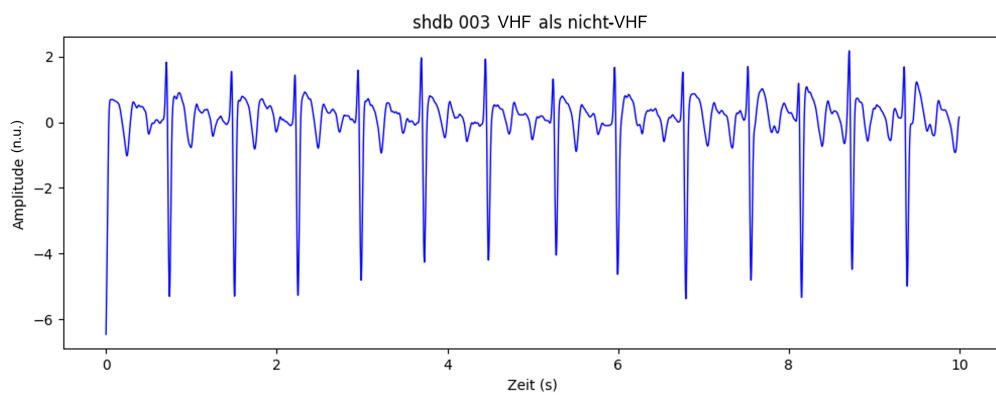


Abb. 7.4: Eine Aufnahme von Patient 003 aus der SHDB-AF Datenbank, welche mit VHF Annotiert und als nicht-VHF vom gewichteten DANN Ensemble klassifiziert wurde. Zu sehen ist eine regelmäßige Vorhoferregung.

Vorhoferregung zu sehen, es sind Flimmerwellen und eine leichte absolute Arrhythmie zu erkennen, sodass dies auf VHF hinweist. Die Korrektheit der Annotationen dieser Abschnitte ist somit fraglich.

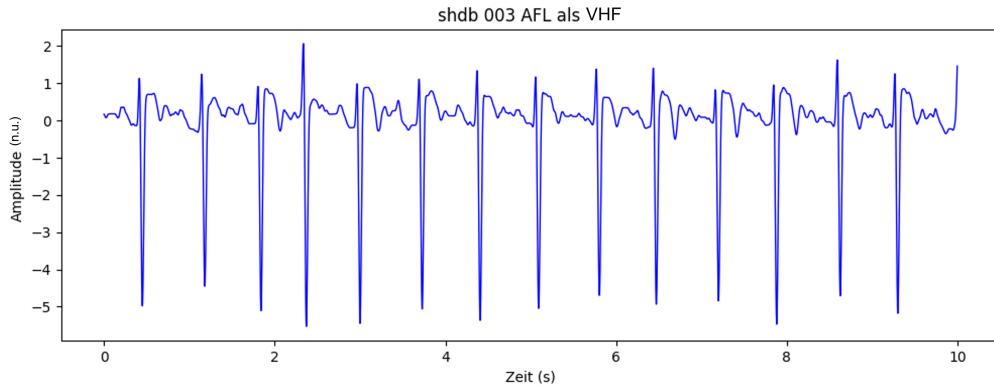


Abb. 7.5: Eine Aufnahme von Patient 003 aus der SHDB-AF Datenbank, welche mit Vorhofflimmern annotiert und als VHF vom gewichteten DANN Ensemble klassifiziert wurde. Es ist keine regelmäßige Vorhoferregung zu sehen.

Die Klassifikation von Patient 050 erfolgte mit einem F1-Score für VHF von 0,127 bei einer durchschnittlichen Sicherheit von 0,719. Dieser Patient besitzt insgesamt 8575 10-Sekunden-Fenster. Von diesen 8575 Fenstern sind 4 Fenster mit VHF annotiert. Diese 4 Fenster hat das DANN Ensemble korrekt erkannt. Auch die nicht-VHF-Fenster wurden größtenteils korrekt klassifiziert, denn von 8571 negativen Fenstern wurden nur 55 falsch positiv klassifiziert. Durch diese starke Unebenheit der Klassen fallen jedoch selbst im Verhältnis wenige falsch positive Klassifikationen sehr stark ins Gewicht, sodass der F1-Score sehr schlecht ist. Daraus lässt sich schließen, dass der F1-Score bei unbalancierten Datensätzen keine geeignete Metrik ist.

Die Klassifikation von Patient 111 erfolgte mit einer Genauigkeit von $F1 = 0,066$ auf der VHF-Klasse bei einer durchschnittlichen Sicherheit von 0,737. Bei diesem Patienten wurden Abschnitte mit Vorhoftachykardie annotiert, in denen das Ensemble VHF detektiert. In Abb. 7.6 ist ein solches Fenster abgebildet. Auch hier ist die Annotation fraglich, da keine regelmäßige Vorhoferregung sichtbar und eine absolute Arrhythmie vorhanden ist.

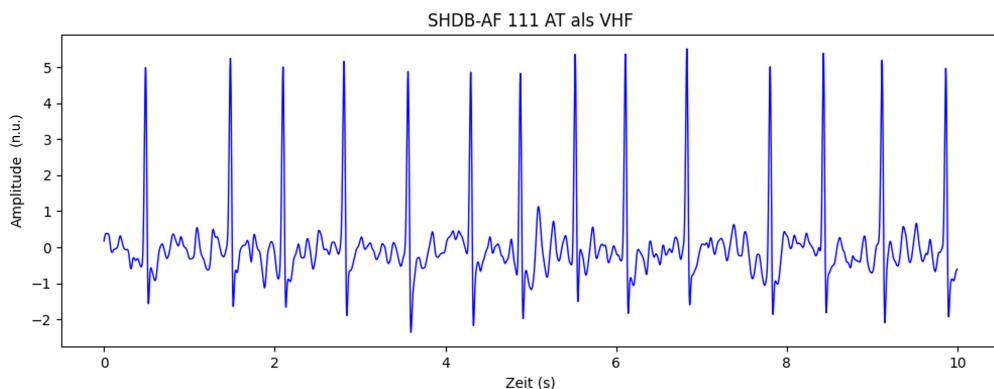


Abb. 7.6: Eine Aufnahme von Patient 111 aus der SHDB-AF Datenbank, welche mit Vorhoftachykardie annotiert und als VHF vom gewichteten DANN Ensemble klassifiziert wurde. Es ist keine regelmäßige Vorhoferregung zu sehen, jedoch eine absolute Arrhythmie.

Abb. 7.7 ist der F1-Score für die nicht-VHF-Klasse gegen die durchschnittliche Sicherheit des Ensembles pro Patient aufgetragen. Zusätzlich ist die Standardabweichung der Sicherheit pro

Patient in Form von T-Balken eingetragen. Zu sehen ist, dass auch hier der Großteil der Aufnahmen mit einem guten F1-Score und einer hohen durchschnittlichen Sicherheit klassifiziert werden konnte.

Die Klassifikation von Patient 024 besitzt einen F1-Score von 0,094 für die nicht-VHF-Klasse mit einer durchschnittlichen Sicherheit von 0,606. Hier tritt derselbe Fall auf, wie bei der Klassifikation von Patient 050, mit dem Unterschied, dass Patient 024 nur drei Fenster besitzt, welche der nicht-VHF-Klasse angehören.

Es lässt sich abschließend die Aussage treffen, dass die Klassifikation auf der SHDB-AF Datenbank zum Großteil gut funktioniert, jedoch die Gesamtgüte des Ensembles durch Aufnahmen mit einer starken Uneinheitlichkeit in der Klassenverteilung und durch eventuelle falsche Annotationen reduziert wird. Aufnahmen mit einer unausgeglichenen Klassenverteilung mit vielen negativen Fällen tragen dazu bei, dass die Specificity hoch ist, da viele negative Fälle korrekt als negativ klassifiziert werden und gleichzeitig der F1-Score durch verhältnismäßig wenig falsch positive Klassifikationen verringert wird.

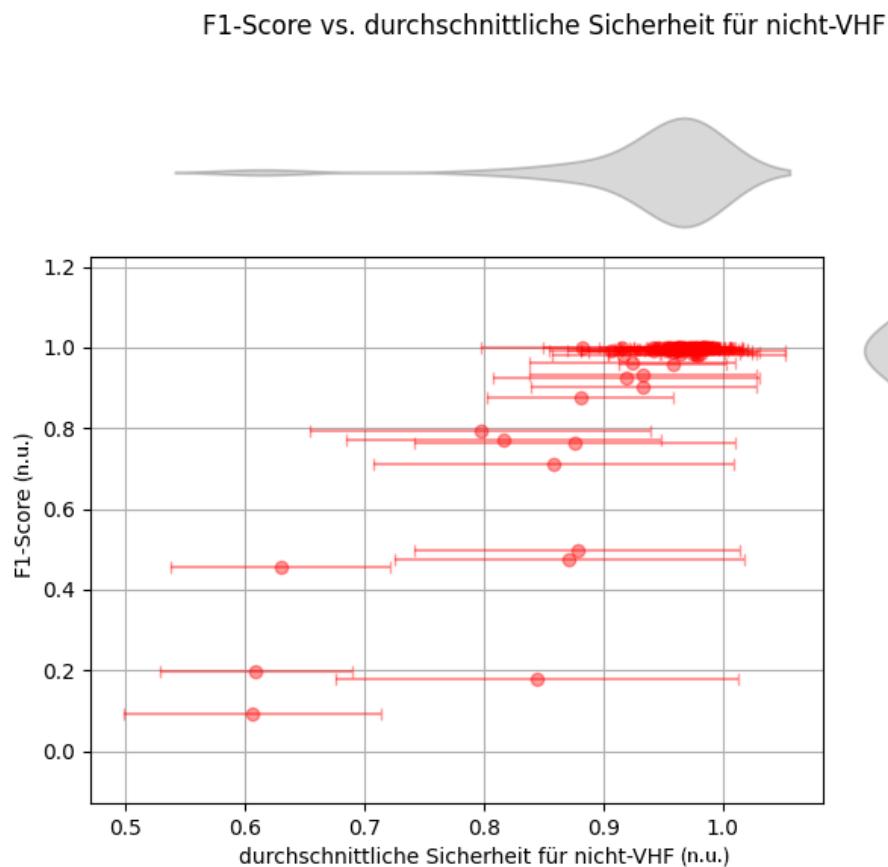


Abb. 7.7: F1-Score für die nicht-VHF-Klasse gegen die durchschnittliche Sicherheit des gewichteten DANN Ensembles pro Patient aufgetragen. Zusätzlich ist die Standardabweichung der Sicherheit pro Patient in Form von T-Balken eingetragen. Anhand des Violinenplots lässt sich die Verteilung der F1-Scores bzw. der durchschnittlichen Sicherheit für die Klassifikation über die Aufnahmen ablesen.

7.4 Einfluss von Normalisierung

Bei der Nutzung von nicht-normalisierten Daten erzielt das gewichtete DANN Ensemble mit einem F1-Score von 0,953 (siehe Tab. 6.10) eine minimal höhere Klassifikationsgüte auf dem Testdatensatz der Quelldomäne, als das Ensemble, welches normalisierte Daten nutzt (F1-score von 0,951, siehe Tab. 6.2). Auf dem Timely Datensatz erreicht das gewichtete DANN Ensemble ohne Normalisierung der Daten schlechtere F1-Scores (je nach Ableitung F1-Scores von 0,949-0,968 ohne Normalisierung (Tab. 6.11) vs. 0,952-0,986 mit Normalisierung (Tab. 6.6)). Daraus lässt sich schließen, dass das Unterlassen der z-Normalisierung ein Overfitting an die Trainingsdaten begünstigt.

8 Zusammenfassung und Ausblick

In dieser Arbeit wurde ein robuster Ansatz zur Detektion von VHF entwickelt, welcher mit EKGs der Extremitätenableitungen aus Standard-12-Kanal-EKGs trainiert wurde und anwendbar auf morphologisch veränderte Signale mobiler EKG-Patches ist.

Dazu wurden zuerst physiologische Grundlagen zu VHF, sowie Grundlagen des DL erarbeitet. Es wurden Ansätze zu DG identifiziert, welche genutzt werden können, um DL-basierte Ansätze auf unbekannte und morphologisch veränderte Signale zu übertragen. Einer dieser Ansätze ist Domain Adversarial Learning, welcher genutzt wurde, um ein DANN auf Basis von InceptionTime zu entwickeln. Das entwickelte Modell wurde zusammen mit einem Vergleichsmodell ohne Domain Adversarial Learning auf Daten aus der Quelldomäne 12-Kanal-EKG, sowie Daten aus der Zieldomäne mobiler EKG-Patches angewendet. Dabei konnte gezeigt werden, dass durch Domain Adversarial Learning eine Steigerung der Klassifikationsgüte gegenüber des Vergleichsmodells sowohl auf Daten der Quelldomäne als auch auf Daten verschiedener Zieldomänen erreicht werden kann.

Eine Möglichkeit, die in der Diskussion angesprochene fehlende Erklärbarkeit zu verbessern, ist die Verbindung mit Methoden der explainable AI. Explainable AI zielt darauf ab, sogenannte Black-Box Algorithmen zu vermeiden, indem i.d.R nachträglich geschätzt wird, wie ein ANN seine Entscheidungen trifft.

Die Architektur des DANNs kann weiter angepasst werden, um eine noch höhere Klassifikationsgüte zu erzielen. Bspw. können für eine bessere Regularisierung im Label Predictor und Domain Classifier Dropout Layer eingefügt werden. Diese Schichten reduzieren Overfitting, indem zufällige Eingaben für eine bestimmte Anzahl von Neuronen auf 0 gesetzt werden.

Die Klassifikationsgüte des Ensembles auf langen, zusammenhängenden Aufnahmen wie die der SHDB-AF-Datenbank oder Teilen des TIMELY-Datensatzes kann unter Umständen verbessert werden, indem ein zusätzlicher Zeitfilter eingebaut wird, der nur dann ein Fenster mit VHF klassifiziert, wenn das vorhergehende und das nachfolgende Fenster ebenso mit VHF klassifiziert werden. Solch ein Filter würde die Robustheit gegenüber verrauschten Fenstern oder anderweitig unsicheren Detektionen erhöhen.

Abschließend lässt sich sagen, dass Domain Adversarial Learning ein sinnvoller Ansatz ist, um ein ANN zu entwickeln, welches robust gegenüber signalmorphologischer Veränderungen ist. Das in dieser Arbeit entwickelte DANN kann sowohl erfolgreich auf EKGs der Extremitä-

tenableitungen von Standard-12-Kanal-EKGs, als auch auf unbekannte Daten mobiler EKG-Patches angewendet werden. Somit kann es einen wertvollen Beitrag leisten, den Arbeitsaufwand von medizinischem Fachpersonal bei der Analyse von Langzeit-EKG-Aufnahmen zu reduzieren und dazu beitragen, dass VHF frühzeitig erkannt und damit Folgeerkrankungen vorbeugt wird.

Quellenverzeichnis

- [1] Marc Gertsch. *Das EKG : auf einen Blick und im Detail*. 1. Aufl. Heidelberg: Springer, 2007. ISBN: 978-3-540-34371-4.
- [2] Sumeet S. Chugh u. a. „Worldwide Epidemiology of Atrial Fibrillation“. In: *Circulation* 129.8 (25. Feb. 2014), S. 837–847. DOI: 10.1161/CIRCULATIONAHA.113.005119.
- [3] Junqing Liang u. a. „Global trends and epidemiological impact of metabolic risk factors on atrial fibrillation and atrial flutter from 1990 to 2021“. In: *Scientific Reports* 15.1 (7. Feb. 2025), S. 4561. DOI: 10.1038/s41598-025-88744-4.
- [4] Ali Rizwan u. a. „A Review on the State of the Art in Atrial Fibrillation Detection Enabled by Machine Learning“. In: *IEEE reviews in biomedical engineering* 14 (2021), S. 219–239. DOI: 10.1109/RBME.2020.2976507.
- [5] B. Schmitz u. a. „Patient-centered cardiac rehabilitation by AI-powered lifestyle intervention – the timely approach“. In: *Atherosclerosis* 355 (1. Aug. 2022), S. 251. DOI: 10.1016/j.atherosclerosis.2022.06.959.
- [6] Hassan Ismail Fawaz u. a. „InceptionTime: Finding AlexNet for Time Series Classification“. In: *Data Mining and Knowledge Discovery* 34.6 (Nov. 2020), S. 1936–1962. DOI: 10.1007/s10618-020-00710-y.
- [7] Yaroslav Ganin u. a. „Domain-Adversarial Training of Neural Networks“. In: *Journal of Machine Learning Research* 17.59 (2016), S. 1–35.
- [8] Michael Gekle, Erhard Wischmeyer und Stefan Gründer. *Taschenlehrbuch Physiologie*. 2. Aufl. Stuttgart; New York: Georg Thieme Verlag, 2015. 826 S. ISBN: 978-3-13-144982-5.
- [9] Karl Zilles und Bernhard Tillmann. *Anatomie: mit 121 Tabellen*. Springer-Lehrbuch. Berlin Heidelberg: Springer, 2010. 1022 S. ISBN: 978-3-540-69481-6.
- [10] Robert F. Schmidt, Florian Lang und Gerhard Thewst, Hrsg. *Physiologie des Menschen*. Springer-Lehrbuch. Berlin, Heidelberg: Springer, 2005. ISBN: 978-3-540-21882-1 978-3-540-26416-3. DOI: 10.1007/b137547.
- [11] Erland Erdmann, Hrsg. *Klinische Kardiologie*. Berlin, Heidelberg: Springer, 2011. ISBN: 978-3-642-16480-4 978-3-642-16481-1. DOI: 10.1007/978-3-642-16481-1.

- [12] Rohan S. Wijesurendra und Barbara Casadei. „Mechanisms of atrial fibrillation“. In: *Heart* 105.24 (1. Dez. 2019), S. 1860–1867. DOI: 10.1136/heartjnl-2018-314267.
- [13] Stanley Nattel u. a. „Molecular Basis of Atrial Fibrillation Pathophysiology and Therapy“. In: *Circulation Research* 127.1 (19. Juni 2020), S. 51–72. DOI: 10.1161/CIRCRESAHA.120.316363.
- [14] R. Khan. „Identifying and understanding the role of pulmonary vein activity in atrial fibrillation“. In: *Cardiovascular Research* 64.3 (1. Dez. 2004), S. 387–394. DOI: 10.1016/j.cardiores.2004.07.025.
- [15] Daniel C. Sigg u. a., Hrsg. *Cardiac Electrophysiology Methods and Models*. Boston, MA: Springer US, 2010. ISBN: 978-1-4419-6657-5 978-1-4419-6658-2. DOI: 10.1007/978-1-4419-6658-2.
- [16] Michel Haïssaguerre u. a. „Spontaneous Initiation of Atrial Fibrillation by Ectopic Beats Originating in the Pulmonary Veins“. In: *New England Journal of Medicine* 339.10 (3. Sep. 1998), S. 659–666. DOI: 10.1056/NEJM199809033391003.
- [17] Shih-Ann Chen u. a. „Initiation of Atrial Fibrillation by Ectopic Beats Originating From the Pulmonary Veins“. In: *Circulation* 100.18 (2. Nov. 1999), S. 1879–1886. DOI: 10.1161/01.CIR.100.18.1879.
- [18] Alejandro Perez-Lugones u. a. „Evidence of Specialized Conduction Cells in Human Pulmonary Veins of Patients with Atrial Fibrillation“. In: *Journal of Cardiovascular Electrophysiology* 14.8 (2003), S. 803–809. DOI: 10.1046/j.1540-8167.2003.03075.x.
- [19] Pasquale Santangeli und Francis E. Marchlinski. „Techniques for the provocation, localization, and ablation of non-pulmonary vein triggers for atrial fibrillation“. In: *Heart Rhythm* 14.7 (1. Juli 2017), S. 1087–1096. DOI: 10.1016/j.hrthm.2017.02.030.
- [20] Alexander Hammer, Hagen Malberg und Martin Schmidt. „Cardiovascular Reflections of Sympathovagal Imbalance Precede the Onset of Atrial Fibrillation“. In: *2023 Computing in Cardiology (CinC)*. 2023 Computing in Cardiology (CinC). Bd. 50. Okt. 2023, S. 1–4. DOI: 10.22489/CinC.2023.399.
- [21] Ahsan A. Khan, Gregory Y. H. Lip und Alena Shantsila. „Heart rate variability in atrial fibrillation: The balance between sympathetic and parasympathetic nervous system“. In: *European Journal of Clinical Investigation* 49.11 (2019), e13174. DOI: 10.1111/eci.13174.
- [22] Maurits C.E.F. Wijffels u. a. „Atrial Fibrillation Begets Atrial Fibrillation“. In: *Circulation* 92.7 (Okt. 1995), S. 1954–1968. DOI: 10.1161/01.CIR.92.7.1954.
- [23] Laila Staerk u. a. „Atrial Fibrillation: Epidemiology, Pathophysiology, and Clinical Outcomes“. In: *Circulation Research* 120.9 (28. Apr. 2017), S. 1501–1517. DOI: 10.1161/CIRCRESAHA.117.309732.
- [24] G. K. Moe und J. A. Abildskov. „Atrial fibrillation as a self-sustaining arrhythmia independent of focal discharge“. In: *American Heart Journal* 58.1 (1. Juli 1959), S. 59–70. DOI: 10.1016/0002-8703(59)90274-1.

- [25] María S. Guillem u. a. „Presence and stability of rotors in atrial fibrillation: evidence and therapeutic implications“. In: *Cardiovascular Research* 109.4 (1. Apr. 2016), S. 480–492. DOI: 10.1093/cvr/cvw011.
- [26] Maurits A. Allessie u. a. „Electropathological Substrate of Long-Standing Persistent Atrial Fibrillation in Patients With Structural Heart Disease“. In: *Circulation: Arrhythmia and Electrophysiology* 3.6 (Dez. 2010), S. 606–615. DOI: 10.1161/CIRCEP.109.910125.
- [27] Bianca J. J. M. Brundel u. a. „Atrial fibrillation“. In: *Nature Reviews. Disease Primers* 8.1 (7. Apr. 2022), S. 21. DOI: 10.1038/s41572-022-00347-9.
- [28] Adolf Faller, Michael Schünke und Gabriele Schünke. *Der Körper des Menschen : Einführung in Bau und Funktion*. 14. Aufl. Stuttgart; New York: Thieme, 2004. ISBN: 3-13-329714-7.
- [29] Chi Nhan Nguyen und Oliver Zeigermann. *Machine Learning - kurz & gut*. 1. Aufl. Heidelberg: dpunkt.verlag, 2018. ISBN: 978-3-96009-052-6.
- [30] Andriy Burkov. *Machine Learning kompakt*. 1. Aufl. Frechen: mitp Verlag, 2019. ISBN: 978-3-95845-996-0.
- [31] Marc Goettling u. a. *xECGArch: A trustworthy deep learning architecture for interpretable ECG analysis considering short-term and long-term features*. 19. Apr. 2024. DOI: 10.21203/rs.3.rs-3654418/v1.
- [32] Yong-Yeon Jo u. a. „Explainable artificial intelligence to detect atrial fibrillation using electrocardiogram“. In: *International Journal of Cardiology* 328 (1. Apr. 2021), S. 104–110. DOI: 10.1016/j.ijcard.2020.11.053.
- [33] Kevin P. Murphy. *Probabilistic Machine Learning: An introduction*. MIT Press, 2022. (Besucht am 29.07.2024).
- [34] Michael A. Nielsen. *Neural Networks and Deep Learning*. Determination Press, 2015.
- [35] Kaiming He u. a. *Deep Residual Learning for Image Recognition*. 10. Dez. 2015. DOI: 10.48550/arXiv.1512.03385.
- [36] Theekshana Dissanayake u. a. „Domain Generalization in Biosignal Classification“. In: *IEEE Transactions on Biomedical Engineering* 68.6 (Juni 2021), S. 1978–1989. DOI: 10.1109/TBME.2020.3045720.
- [37] Jindong Wang u. a. „Generalizing to Unseen Domains: A Survey on Domain Generalization“. In: *IEEE Transactions on Knowledge and Data Engineering* 35.8 (Aug. 2023), S. 8052–8072. DOI: 10.1109/TKDE.2022.3178128.
- [38] Wouter M. Kouw und Marco Loog. *An introduction to domain adaptation and transfer learning*. 14. Jan. 2019. DOI: 10.48550/arXiv.1812.11806.
- [39] Hassan Ismail Fawaz u. a. „Deep learning for time series classification: a review“. In: *Data Mining and Knowledge Discovery* 33.4 (1. Juli 2019), S. 917–963. DOI: 10.1007/s10618-019-00619-1.
- [40] Hoang Anh Dau u. a. „The UCR time series archive“. In: *IEEE/CAA Journal of Automatica Sinica* 6.6 (Nov. 2019), S. 1293–1305. DOI: 10.1109/JAS.2019.1911747.

- [41] Anthony Bagnall u. a. „The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances“. In: *Data Mining and Knowledge Discovery* 31.3 (1. Mai 2017), S. 606–660. DOI: 10.1007/s10618-016-0483-9.
- [42] Mustafa Baydogan. *Symbolic Representations for Multivariate Time Series Classification (SMTS)* - Mustafa Baydogan. 19. Apr. 2022. URL: <https://www.mustafabaydogan.com/research/time-series-data-mining/symbolic-representations-for-multivariate-time-series-classification-smts/> (besucht am 31.01.2025).
- [43] Ali Ismail Fawaz u. a. *Deep Learning for Time Series Classification*. URL: <https://msd-irimas.github.io/pages/dl4tsc/> (besucht am 31.01.2025).
- [44] Paul Murat et al. „Review of Deep Learning-Based Atrial Fibrillation Detection Studies“. In: *International Journal of Environmental Research and Public Health* 1.1 (2021), S. 1–2. DOI: 10.3390/ijerph200401001.
- [45] Dakun Lai u. a. „Non-Standardized Patch-Based ECG Lead Together With Deep Learning Based Algorithm for Automatic Screening of Atrial Fibrillation“. In: *IEEE Journal of Biomedical and Health Informatics* 24.6 (Juni 2020), S. 1569–1578. DOI: 10.1109/JBHI.2020.2980454.
- [46] Zachi I. Attia u. a. „An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction“. In: *The Lancet* 394.10201 (7. Sep. 2019), S. 861–867. DOI: 10.1016/S0140-6736(19)31721-0.
- [47] Wenjuan Cai u. a. „Accurate detection of atrial fibrillation from 12-lead ECG using deep neural network“. In: *Computers in Biology and Medicine* 116 (1. Jan. 2020), S. 103378. DOI: 10.1016/j.compbiomed.2019.103378.
- [48] Antônio H. Ribeiro u. a. „Automatic diagnosis of the 12-lead ECG using a deep neural network“. In: *Nature Communications* 11.1 (9. Apr. 2020), S. 1760. DOI: 10.1038/s41467-020-15432-4.
- [49] Sarah W. E. Baalman u. a. „A morphology based deep learning model for atrial fibrillation detection using single cycle electrocardiographic samples“. In: *International Journal of Cardiology* 316 (1. Okt. 2020), S. 130–136. DOI: 10.1016/j.ijcard.2020.04.046.
- [50] Hosein Hasani, Adeleh Bitarafan und Mahdieh Soleymani Baghshah. „Classification of 12-lead ECG Signals With Adversarial Multi-Source Domain Generalization“. In: *2020 Computing in Cardiology*. 2020 Computing in Cardiology. Sep. 2020, S. 1–4. DOI: 10.22489/CinC.2020.445.
- [51] Aristotelis Ballas und Christos Diou. „A Domain Generalization Approach for Out-Of-Distribution 12-lead ECG Classification with Convolutional Neural Networks“. In: *2022 IEEE Eighth International Conference on Big Data Computing Service and Applications (BigDataService)*. 2022 IEEE Eighth International Conference on Big Data Computing Service and Applications (BigDataService). Aug. 2022, S. 9–13. DOI: 10.1109/BigDataService55688.2022.00009.

- [52] Aristotelis Ballas und Christos Diou. „Towards Domain Generalization for ECG and EEG Classification: Algorithms and Benchmarks“. In: *IEEE Transactions on Emerging Topics in Computational Intelligence* 8.1 (Feb. 2024), S. 44–54. ISSN: 2471-285X. DOI: 10.1109/TETCI.2023.3306253.
- [53] Zuogang Shang u. a. „Deep Discriminative Domain Generalization with Adversarial Feature Learning for Classifying ECG Signals“. In: *2021 Computing in Cardiology (CinC)*. 2021 Computing in Cardiology (CinC). Bd. 48. Sep. 2021, S. 1–4. DOI: 10.23919/CinC53138.2021.9662844.
- [54] Keewon Shin u. a. „Enhancing the performance of premature ventricular contraction detection in unseen datasets through deep learning with denoise and contrast attention module“. In: *Computers in Biology and Medicine* 166 (Nov. 2023), S. 107532. DOI: 10.1016/j.combiomed.2023.107532.
- [55] Spreeth P. Shashikumar u. a. „Detection of Paroxysmal Atrial Fibrillation using Attention-based Bidirectional Recurrent Neural Networks“. In: *KDD’18: PROCEEDINGS OF THE 24TH ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY & DATA MINING*. 24th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD). New York: Assoc Computing Machinery, 2018, S. 715–723. DOI: 10.1145/3219819.3219912.
- [56] Jayroop Ramesh u. a. „Atrial Fibrillation Classification with Smart Wearables Using Short-Term Heart Rate Variability and Deep Convolutional Neural Networks“. In: *Sensors* 21.21 (Jan. 2021), S. 7233. DOI: 10.3390/s21217233.
- [57] Shawn Tan u. a. „Icentia11K: An Unsupervised Representation Learning Dataset for Arrhythmia Subtype Discovery“. In: *Computing in Cardiology Conference (CinC)* (2021).
- [58] Shawn Tan u. a. *Icentia11k Single Lead Continuous Raw Electrocardiogram Dataset (version 1.0)*. PhysioNet. Version 1.0. 2022. DOI: 10.13026/KK0V-R952.
- [59] Kenta Tsutsui u. a. *SHDB-AF: a Japanese Holter ECG database of atrial fibrillation*. 22. Juni 2024. DOI: 10.48550/arXiv.2406.16974.
- [60] Kenta Tsutsui u. a. *SHDB-AF: a Japanese Holter ECG database of atrial fibrillation (version 1.0.0)*. PhysioNet. 2024. DOI: <https://doi.org/10.13026/10mk-y852>.
- [61] Patrick Wagner u. a. „PTB-XL, a large publicly available electrocardiography dataset“. In: *Scientific Data* 7.1 (25. Mai 2020), S. 154. DOI: 10.1038/s41597-020-0495-6.
- [62] Patrick Wagner u. a. *PTB-XL, a large publicly available electrocardiography dataset (version 1.0.3)*. PhysioNet. DOI: 10.13026/KFZX-AW45.
- [63] Jianwei Zheng u. a. „A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients“. In: *Scientific Data* 7.1 (12. Feb. 2020), S. 48. DOI: 10.1038/s41597-020-0386-x.
- [64] Jianwei Zheng, Hangyuan Guo und Huimin Chu. *A large scale 12-lead electrocardiogram database for arrhythmia study (version 1.0.0)*. PhysioNet. DOI: 10.13026/WGEX-ER52.
- [65] Erick A Perez Alday u. a. „Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020“. In: *Physiological Measurement* 41.12 (1. Dez. 2020), S. 124003. DOI: 10.1088/1361-6579/abc960.

- [66] Erick Andres Perez Alday u. a. *Classification of 12-lead ECGs: The PhysioNet/Computing in Cardiology Challenge 2020 (version 1.0.2)*. PhysioNet. DOI: 10.13026/M77N-SX13.
- [67] Feifei Liu u. a. „An Open Access Database for Evaluating the Algorithms of Electrocardiogram Rhythm and Morphology Abnormality Detection“. In: *Journal of Medical Imaging and Health Informatics* 8.7 (1. Sep. 2018), S. 1368–1373. DOI: 10.1166/jmihi.2018.2442.
- [68] Pierre Paquet, David Levesque und Pierre Fecteau. „Adhesive extender for medical electrode and use thereof with wearable monitor“. US-Pat. 11304660B2. Icentia Inc. 19. Apr. 2022.
- [69] *A patient-centered early risk prediction, prevention, and intervention platform to support the continuum of care in coronary artery disease (CAD) using eHealth and artificial intelligence | TIMELY Project | Fact Sheet | H2020*. CORDIS | European Commission. DOI: 10.3030/101017424. URL: <https://cordis.europa.eu/project/id/101017424> (besucht am 11.02.2025).
- [70] Dominique Makowski u. a. „NeuroKit2: A Python toolbox for neurophysiological signal processing“. In: *Behavior Research Methods* 53.4 (1. Aug. 2021), S. 1689–1696. DOI: 10.3758/s13428-020-01516-y.
- [71] Paul Kligfield u. a. „Recommendations for the Standardization and Interpretation of the Electrocardiogram“. In: *Circulation* 115.10 (13. März 2007), S. 1306–1324. DOI: 10.1161/CIRCULATIONAHA.106.180200.
- [72] Ary L. Goldberger u. a. „PhysioBank, PhysioToolkit, and PhysioNet“. In: *Circulation* 101.23 (13. Juni 2000), e215–e220. DOI: 10.1161/01.CIR.101.23.e215.
- [73] Alexander Hammer u. a. „Automatic Classification of Full- and Reduced-Lead Electrocardiograms Using Morphological Feature Extraction“. In: *2021 Computing in Cardiology (CinC)*. 2021 Computing in Cardiology (CinC). Bd. 48. Sep. 2021, S. 1–4. DOI: 10.23919/CinC53138.2021.9662797.
- [74] Jayawant N. Mandrekar. „Receiver Operating Characteristic Curve in Diagnostic Test Assessment“. In: *Journal of Thoracic Oncology* 5.9 (Sep. 2010), S. 1315–1316. DOI: 10.1097/JTO.0b013e3181ec173d.
- [75] Elena Plante und Rebecca Vance. „Selection of preschool language tests: a data-based approach.“ In: *Language, Speech & Hearing Services in Schools* 25.1 (1. Jan. 1994), S. 15. DOI: 10.1044/0161-1461.2501.15.

Anhang

A Anhang

A.1 Suchstrings der systematischen Literaturrecherche

String 1

```
( "ECG" OR "electrocardiogra*" ) AND  
( "Out-of-Distribution Generalization" OR  
"Out-of-Distribution Generalisation" OR "Domain Generalization" OR  
"Domain Generalisation" OR "OOD Generalization" OR "OOD Generalisation" )
```

String 2

```
( "ECG" OR "electrocardiogra*" ) AND  
( "Out-of-Distribution Generalization" OR  
"Out-of-Distribution Generalisation" OR "Domain Generalization" OR  
"Domain Generalisation" OR "OOD Generalization" OR  
"OOD Generalisation" OR "Out-of-Distribution General*" OR  
"OOD General*" OR "General* Across Domains" OR  
"Cross-Domain General*" OR "Robustness to Domain Shift" OR  
"Multi-Domain Learning" OR "Domain-Invariant Learning" OR  
"Out-of-Distribution Attributes" OR "General* Target Domain" OR  
"Meta-Learning" OR "Domain Invarian*") AND  
( "Neural Network" OR "Deep Learning" ) AND  
( classif* OR detect*)
```

String 3

("ECG" OR "EEG" OR "electrocardiogra*" OR
"electroencephalogr*" OR "Biosignal") AND
("Out-of-Distribution Generalization" OR
"Out-of-Distribution Generalisation" OR "Domain Generalization" OR
"Domain Generalisation" OR "OOD Generalization" OR
"OOD Generalisation" OR "Out-of-Distribution General*" OR
"OOD General*" OR "General* Across Domains" OR
"Cross-Domain General*" OR "Robustness to Domain Shift" OR
"Multi-Domain Learning" OR "Domain-Invariant Learning" OR
"Out-of-Distribution Attributes" OR "General* Target Domain" OR
"Meta-Learning" OR "Domain Invarian*") AND
("Neural Network" OR "Deep Learning")

String 4

"Medical Imaging" AND
("Domain General*" OR "Out-of-Distribution General*" OR "OOD General*") AND
("Neural Network" OR "Deep Learning" OR "DL")

A.2 Grid Search Ergebnisse für InceptionTime

Tab. A.1: Ergebnisse der Grid Search für InceptionTime. Optimierte Hyperparameter sind depth d (Anzahl der Inception Module), learning rate l und batch size b. Die Metriken wurden berechnet, indem vom Durchschnitt aus allen 5 Folds der Cross Validation die Standardabweichung abgezogen wurde. Die Standardabweichung σ für die jeweilige Metrik rechts neben der betreffenden Metrik angegeben. Negative Werte kommen zustande, wenn die Standardabweichung größer als der Durchschnitt ist. Die beste Hyperparameterkombination ist hervorgehoben.

Hyperparameterkombination	F1	σ	Sensitivität	σ	Spezifität	σ
d=3 l=0,0001 b=128	0,919	0,004	0,928	0,007	0,903	0,008
d=3 l=0,0001 b=64	0,916	0,005	0,925	0,008	0,901	0,007
d=3 l=0,0001 b=32	0,919	0,003	0,925	0,006	0,910	0,004
d=3 l=0,001 b=128	0,923	0,003	0,930	0,007	0,909	0,007
d=3 l=0,001 b=64	0,921	0,004	0,930	0,007	0,900	0,011
d=3 l=0,001 b=32	0,921	0,004	0,927	0,006	0,913	0,003
d=3 l=0,01 b=128	0,920	0,004	0,928	0,008	0,905	0,008
d=3 l=0,01 b=64	0,917	0,005	0,917	0,010	0,911	0,008
d=3 l=0,01 b=32	0,918	0,005	0,919	0,015	0,897	0,015
d=6 l=0,0001 b=128	0,930	0,004	0,935	0,009	0,913	0,012
d=6 l=0,0001 b=64	0,933	0,002	0,935	0,006	0,923	0,008
d=6 l=0,0001 b=32	0,932	0,004	0,937	0,003	0,926	0,006
d=6 l=0,001 b=128	0,933	0,004	0,939	0,003	0,924	0,008
d=6 l=0,001 b=64	0,935	0,003	0,935	0,007	0,932	0,004
d=6 l=0,001 b=32	0,935	0,003	0,932	0,005	0,931	0,009
d=6 l=0,01 b=128	0,934	0,004	0,935	0,005	0,932	0,003
d=6 l=0,01 b=64	0,934	0,002	0,931	0,005	0,930	0,009
d=6 l=0,01 b=32	0,934	0,003	0,935	0,007	0,925	0,008
d=9 l=0,0001 b=128	0,938	0,002	0,933	0,006	0,933	0,008
d=9 l=0,0001 b=64	0,939	0,002	0,935	0,006	0,939	0,003
d=9 l=0,0001 b=32	0,934	0,004	0,926	0,013	0,935	0,006
d=9 l=0,001 b=128	0,938	0,004	0,938	0,003	0,934	0,009
d=9 l=0,001 b=64	0,939	0,004	0,938	0,005	0,937	0,007
d=9 l=0,001 b=32	0,936	0,006	0,937	0,005	0,933	0,009
d=9 l=0,01 b=128	0,936	0,004	0,932	0,005	0,938	0,006
d=9 l=0,01 b=64	0,938	0,004	0,937	0,002	0,939	0,008
d=9 l=0,01 b=32	0,934	0,006	0,924	0,012	0,939	0,007