Fachhochschule Aachen
Campus Jülich

Fachbereich 9
Medizintechnik und Technomathematik

Untersuchung der Signalqualität von ballistokardiographischen Signalen mittels Methoden des maschinellen Lernens

Bachelorarbeit
im Studiengang Scientific Programming

von

Cay Jakob Rahn
Matr.-Nr.: 3145495

12. November 2020

1. Prüfer: Prof. Dr. rer. nat. Alexander Voß
2. Prüfer: Dr.-Ing. Christoph Hoog Antink

Erklärung

Diese Arbeit ist von mir selbständig angefertigt und verfasst. Es sind keine anderen als die angegebenen Quellen und Hilfsmittel benutzt worden.

Ort, Datum	Unterschrift

Abstract

Ballistokardiographie (BKG) ist eine Messtechnik, bei der die durch den Herzschlag induzierten Massenverschiebungen des Körpers gemessen werden. Störungen, sogenannte Artefakte, die diese Signale überlagern, führen zu Fehlern in der Signalverarbeitung und so womöglich zu fehlerhaften Diagnosen. Die Beurteilung der Signalqualität und damit die zuverlässige Detektion dieser Artefakte ist ein für das BKG bis jetzt nicht hinreichend gelöstes Problem, besonders bei in Betten integrierten BKG-Systemen.
In dieser Arbeit werden die Grundlagen der Ballistokardiographie erarbeitet – der physiologische Ursprung, die Eigenschaften der Signale und Techniken der Signalverarbeitung. Messdaten werden aufbereitet und Methoden zur Beurteilung der Signalqualität, die sich für andere Aufnahmebedingungen, namentlich für in Stühlen integrierte Messsysteme und Bett-Aufnahmen bei gesunden, schlafenden Personen, erfolgreich zeigten, evaluiert. Aufbauend auf diesen Ergebnissen werden Merkmale entwickelt, die Informationen über die Signalqualität enthalten. Modelle maschinellen Lernens werden ausgewählt und für Besonderheiten der Daten erweitert.
Bei der Bewertung der Güte dieser Modelle zeigt sich, dass die Beurteilung der Signalqualität verbessert werden kann, d.ḣ. weniger Signal fälschlicherweise als nicht informativ ausgeschlossen und der Fehler der Herzratenschätzung verkleinert wird. Diese Ergebnisse lassen sich sogar auch auf andere Aufnahmesituationen übertragen.

Inhaltsverzeichnis

Abkürzungsverzeichnis

AUC Area under the ROC Curve

CART Classification And Regression Trees

CLIE Continuous Local Interval Estimator

BKG Ballistokardiographie

DT Decision Tree

EKG Elektrokardiographie

FPR False Positive Rate

HR Herzrate

HRV Herzratenvariabilität

LDA Linear Discriminant Analysis

MAE Mean Absolute Error

MSE Mean Squared Error

MLP Multilayer-Perzeptron

PCA Principal Component Analysis

PPG Photoplethysmographie

RF Random Forest

ROC Receiver Operating Characteristic

SKG Seismokardiographie

SQI Signal Quality Index

SVM Support Vector Machine

TPR True Positive Rate

XGB eXtreme Gradient Boosting

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Der derzeitige demographische Wandel stellt das Gesundheitssystem vor eine große Herausforderung: Es gibt immer mehr Patient*innen, die medizinische Überwachung und Versorgung im Alter benötigen. Eine kontinuierliche autonome Überwachung von Vitalparametern im Krankenhaus oder auch Zuhause erlaubt es, Erkrankungen frühzeitig zu erkennen und zu beobachten, ohne dass große Personalkapazitäten vonnöten sind.

Für diesen Anwendungszweck eignen sich vor allem Messmethoden, die die Patient*innen im Alltag nicht einschränken und wenig invasiv sind. Im Englischen wird dies mit dem Begriff *unobtrusive* bezeichnet. Da es keine zufriedenstellende deutsche Entsprechung gibt, wird dieser im Folgenden nicht übersetzt. Solche *unobtrusive* Messmethoden benötigen meist keinen direkten Körper- oder Hautkontakt, liefern aber Information über Atmung und Herzschlag. Die Herausforderung bei einem so ermittelten Signal besteht in der Signalverarbeitung, da Messungenauigkeiten und Alltagsbewegungen zu Störungen im Signal führen. Nicht informatives, also nicht für die Verarbeitung geeignetes Signal muss aber zwingend identifiziert werden, da die Ergebnisse ansonsten stark verfälscht würden.

Eine solche *unobtrusive* Messmethode ist die Ballistokardiographie (BKG). Sensoren lassen sich beispielsweise in Betten und Stühlen integrieren. Aufgezeichnet werden Aktivitäten des Herzens und der Lunge. Die Signalmorphologie variiert jedoch sowohl zwischen den Patient*innen als auch innerhalb einer Person sehr stark, wodurch die automatische Beurteilung der Signalqualität erschwert wird. Dies ist jedoch essentiell, um eine aussagekräftige Signalverarbeitung zu ermöglichen. Besonders bei in Betten aufgenommenen Signalen ist deren Variation in Kombination mit Artefakten durch z. B. Körperbewegungen problematisch.

1.1 Ziel der Arbeit

Die Möglichkeiten zur Beurteilung der Signalqualität von BKG-Signalen mittels maschinellen Lernens werden untersucht. Im besonderen Fokus liegen dabei Langzeitaufnahmen

von bettlägerigen Patient*innen, sowohl nachts als auch tagsüber. Diese haben sich in der Vergangenheit als besonders anfällig für geringe Signalqualität gezeigt.[1]

Dafür werden zunächst die vorliegenden Daten aufbereitet und existierende Verfahren der Artefakterkennung für diese getestet und bewertet. Anschließend wird das Signal auf mögliche Merkmale untersucht und Verfahren entwickelt, die Aussagen über die Signalqualität ermöglichen. Anhand dieses Wissens werden Modelle des maschinellen Lernens ausgewählt und getestet. Zusätzlich wird untersucht, welchen Einfluss Annotation und Eingabeform der Daten auf das Ergebnis haben.

Langfristig soll ermöglicht werden, Ballistokardiographie (BKG) im medizinischen Alltag auch in unkontrollierten Umgebungen anzuwenden. Ziel dieser Arbeit ist es, für diesen Anwendungszweck Merkmale und Modelle zu entwickeln, die eine erste Einschätzung der Signalqualität ermöglichen.

1.2 Gliederung

Zunächst wird in **Kapitel 2** ein allgemeiner Überblick über den medizinischen und technischen Hintergrund gegeben, der nötig ist, um die vorliegende Arbeit zu verstehen.

In **Kapitel 3** wird das Thema der Verarbeitung ballistokardiographischer Signale näher betrachtet. Dazu gehört die Detektion von Herzschlägen bei BKG-Signalen, existierende Verfahren zur Artefakterkennung und die Vorverarbeitung der vorliegenden Messdaten.

Anschließend werden in **Kapitel 4** die vorgestellten existierenden Verfahren mit den vorliegenden Messdaten getestet, evaluiert und eine Analyse der Daten durchgeführt.

Aufbauend darauf werden in **Kapitel 5** weitere Merkmale konstruiert, untersucht und reduziert und eine Modellauswahl und Basis für den Vergleich verschiedener Modelle getroffen.

In **Kapitel 6** werden die Modelle verglichen und evaluiert. Anhand der trainierten Modelle werden die Merkmale erneut betrachtet und der Einfluss der Segmentlänge und des Schwellwertes der Annotation untersucht.

Die Ergebnisse der Arbeit werden in **Kapitel 7** zusammengefasst und ein Ausblick gegeben, wie darauf aufgebaut werden kann.

[1] Vgl. Hoog Antink et al. 2020.

2 Grundlagen

Zum Verständnis dieser Arbeit ist grundlegendes Wissen nötig, welches hier in die drei Bereiche Maschinelles Lernen, Medizinische Grundlagen und Ballistokardiographie unterteilt ist.

2.1 Maschinelles Lernen

Da in dieser Arbeit Methoden des Maschinellen Lernens verwendet werden, wird im Folgenden eine Übersicht über seine Prinzipien, gängige Techniken und Evaluationsmetriken sowie verschiedene Lernmodelle und deren mathematischer Hintergrund gegeben.

2.1.1 Grundprinzipien

Maschinelles Lernen ist die „künstliche“ Generierung von Wissen auf Basis von Erfahrung: Aus Beispielen wird gelernt und dieses Wissen nach einer Trainingsphase verallgemeinert. Dafür wird mit Mustererkennung gearbeitet und ein statistisches Modell aufgebaut, das auf den Trainingsdaten beruht. Die Trainingsdaten X bestehen aus Merkmalsvektoren $x \in X \subseteq \mathbb{R}^n$. Gesucht ist eine Funktion $f : X \to Y$, die diese Daten abbildet. Unterschieden wird bei der Beschreibung dieser Daten zwischen Klassifikation und Regression. Bei einer Klassifikation werden die Eingabedaten in verschiedene Klassen unterteilt. Bei einer Regression dagegen werden stetige Werte vorhergesagt, es wird die Verteilung der Daten beschrieben. Sie kann z. B. dafür genutzt werden, Verkaufszahlen vorauszusagen. Bei einigen Klassifikationsverfahren wird außerdem die Wahrscheinlichkeit der Angehörigkeit zu einer Klasse benannt.

Es wird zwischen verschiedenen Arten des maschinellen Lernens unterschieden, dem überwachten und dem unüberwachten Lernen. Letzteres wird hier nicht betrachtet. Bei überwachtem Lernen bestehen die vorliegenden Daten aus Eingabe-Ausgabe-Paaren $(x_1, y_1), ..., (x_n, y_n)$ mit $x_i \in X$ und $y_i \in Y$. Demnach ist nur die Funktion $f : X \to Y$ unbekannt. Der Lernalgorithmus sucht aus einem Hypothesenset die Funktion g, die f möglichst gut approximiert. Ein Lernalgorithmus ermittelt diese Funktion aus einem Hypothesenset. Dieses Vorgehen ist in Abbildung 2.1 visualisiert. Bei unüberwachtem Lernen

dagegen ist Y unbekannt; die Eingabedaten haben die Form $x_1, ..., x_n$ mit $x_i \in X$. Hier ist ein g gesucht, das die Daten möglichst gut beschreibt und sie beispielsweise eigenständig in Kategorien einteilt. In dieser Arbeit wird allerdings ausschließlich überwachtes Lernen betrachtet.

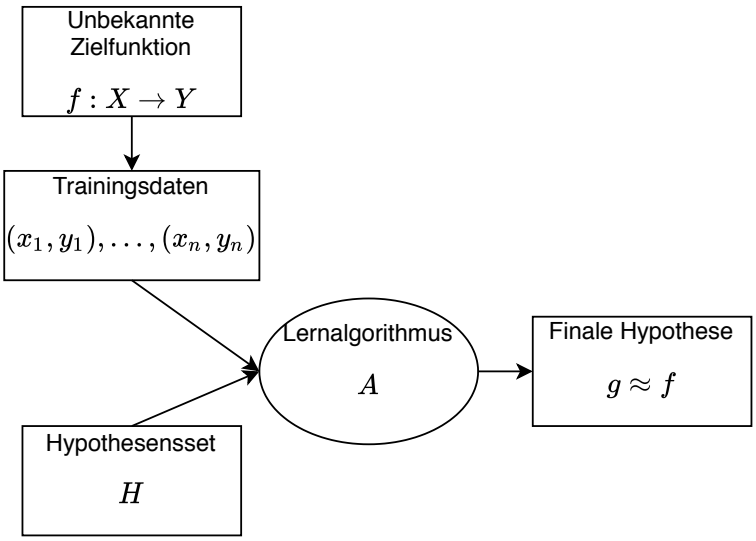


Abbildung 2.1: Darstellung von überwachtem Lernen.

Ein häufig beobachtetes Problem bei maschinellem Lernen ist das sogenannte Overfitting, auf Deutsch Überanpassung. Dabei approximiert das Modell zwar die Trainingsdaten gut, bildet aber keine gute Generalisierung für unbekannte Daten. Das Modell ist also zu gut an die Trainingsdaten angepasst. Gründe dafür können eine zu hohe Komplexität des Lernmodells, zu viel Training oder zu wenig Trainingsdaten sein. In beiden Fällen wird ungewollt ein Teil des Rauschens der Trainingsdaten in das Modell übernommen. Das Gegenteil von Overfitting ist Underfitting. Hier hat das Modell die Beziehung von Merkmalen und Ziel nicht ausreichend erfasst. Dies ist unter anderem der Fall, wenn die zum Training verwendete Stichprobe verzerrt ist.

Zum Prozess des maschinellen Lernens gehört ebenfalls das Sammeln der Daten und die Transformation in Merkmale, die dem Lernalgorithmus als Eingabe dienen. Dieser Prozess ist entscheidend für den Erfolg des Lernprozesses, da nur aus aussagekräftigen Daten ein gutes Modell erzeugt werden kann.

2.1.2 Evaluation und Validierung

Ein erzeugtes Modell muss in jedem Fall auf Daten validiert werden, mit denen nicht trainiert wurde, um den Fehler auf unbekannten Daten abschätzen zu können. Eine Möglichkeit, dies zu tun, bietet die Hold-Out-Validierung, bei der die Daten zufällig in Trainings-

und Testset aufgeteilt werden. Das Modell wird mit dem Trainingsset aufgebaut und auf dem Testset getestet und evaluiert. Eine andere übliche Technik ist die Kreuzvalidierung, bei der die Daten auf v gleich große Mengen, im Englischen *folds*, verteilt werden. Anschließend werden v Modelle trainiert, wobei jeweils eine Menge ausgelassen wird, auf der anschließend getestet wird. Bei einer extremen Variante, der Leave-One-Out Kreuzvalidierung entspricht v der Anzahl der Datenpunkte. Üblich ist jedoch v-fache Kreuzvalidierung mit $v = 5$ oder $v = 10$, auch abhängig von der Datenmenge und der verfügbaren Rechenleistung.

Neben der Validierung des finalen Modells müssen auch bei der Modellauswahl Entscheidungen getroffen werden, vor allem, welches Modell und welche Hyperparameter gewählt werden. Hyperparameter sind Parameter von Modellen maschinellen Lernens, die vor dem Training des Modells festgelegt werden und die Modellarchitektur bestimmen oder den Lernalgorithmus betreffen. Da auch die Wahl dieser validiert werden muss, werden insgesamt drei Datensets benötigt: Trainings-, Validierungs- und Testset, da sonst die Wahl der Hyperparameter durch das Testset beeinflusst würde und kein unabhängiger Test mehr möglich wäre. Das Testset wird also erst genutzt, nachdem alle Entscheidungen getroffen wurden. Typisch ist ein Hyperparameter-Tuning durch Kreuzvalidierung auf dem Trainingsset mit einem anschließenden Retraining mit den ermittelten Parametern auf dem gesamten Trainingsset. Der Ablauf ist in Abbildung 2.2 visualisiert.

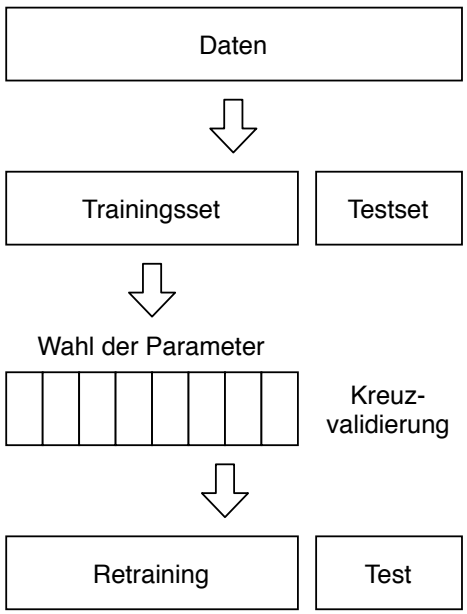


Abbildung 2.2: Ablauf von Training und Validierung.

Um die Auswahl für ein Modell treffen zu können, werden Evaluationsmetriken benötigt. Für diese werden im Folgenden die englischen Bezeichnungen verwendet, da diese auch im Deutschen geläufiger sind.

Vor allem für die Bewertung binärer Klassifikationen gibt es verschiedenste Metriken, die die richtigen und falschen Klassifikationen miteinander gewichten. Dafür wird bei positiv klassifizierten Datenpunkten zwischen Richtig-Positiven TP und Falsch-Positiven FP unterschieden; bei den negativ klassifizierten zwischen Richtig-Negativen TN und Falsch-Negativen FN. Eine Darstellung dieser vier nennt man Confusion Matrix. Diese ermöglicht eine erste Einschätzung der Fehlerverteilung. Ein häufig verwendetes und einfaches Gütemaß ist die Accuracy ACC $= \frac{TP+TN}{TP+TN+FN+FP}$, die allerdings bei ungleich großen Klassen problematisch ist, da eine hohe Genauigkeit erreicht wird, wenn immer die größere Klasse vorausgesagt wird. In diesem Fall kann auch die Balanced Accuracy verwendet werden, die dieses Ungleichgewicht einbezieht:

$$\text{balanced-accuracy} = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right).$$

Bei der Precision, auch Positive Predictive Value, PPV $= \frac{TP}{TP+FP}$ werden falsch-positive Klassifikationen „bestraft“; beim Recall, der True Positive Rate TPR $= \frac{TP}{TP+FN}$, auch Sensitivity genannt, falsch-negative Klassifikationen. Letztere werden oft in einem Maß zusammengefasst, dem F1-Score, der das harmonische Mittel aus beiden bildet:

$$F_1 = 2 \cdot \frac{\text{PPV} \cdot \text{TPR}}{\text{PPV} + \text{TPR}}.$$

Eine weitere Möglichkeit, die Qualität eines binären Klassifikators zu evaluieren, ist die Receiver Operating Characteristic (ROC) Kurve, bei der die True Positive Rate (TPR), also der Recall, und die False Positive Rate (FPR) gegeneinander aufgetragen werden und so den Kompromiss zwischen TPR und FPR abhängig vom Schwellwert zeigt. Die Fläche unter dieser Kurve, die Area under the ROC Curve (AUC), beschreibt, wie gut das Modell die beiden Klassen trennt. Eine perfekte Trennung führt zu AUC $= 1$.

Bei einer Regression werden üblicherweise der MAE oder der Mean Squared Error (MSE) betrachtet, die die Größe des Fehlers von dem vorhergesagten $\hat{y}$ im Vergleich zur Zielgröße y ausdrücken und wie folgt berechnet werden:

$$\text{MAE}(y,\hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} |y_i - \hat{y}_i|$$

$$\text{MSE}(y,\hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} (y_i - \hat{y}_i)^2$$

Die Auswahl der betrachteten Metriken ist von dem vorliegenden Problem abhängig.

2.1.3 Lernmodelle und deren mathematischer Hintergrund

Es gibt eine Vielzahl von Modellen des maschinellen Lernens, von denen eine Auswahl in dieser Arbeit betrachtet und vorgestellt wird. Alle verallgemeinern eine Verteilung von Trainingsdaten und haben das Ziel, eine Funktion g zu finden, die f approximiert und mit der die Zielgröße y vorhergesagt werden kann. Bei einer Klassifikation sollen die verschiedenen Klassen möglichst genau separiert und bei einer Regression der Wert möglichst genau vorhergesagt werden. Diese Funktion wird auch Entscheidungsfunktion genannt und minimiert eine gewählte Kostenfunktion, die beschreibt, wie sehr die Vorhersagen von der Zielgröße abweichen. Bei linearen Modellen beispielsweise ist die Vorhersage mit $\hat{y}_i = \theta x_i$ eine lineare Kombination der gewichteten Merkmale für Datenpunkt $x_i \in X$. $\hat{y}_i$ kann dabei entweder direkt die Vorhersage eines Wertes sein oder genutzt werden, um eine Klassenzugehörigkeit zu ermitteln. Bei einer binären Klassifikation kann dafür beispielsweise die Signum- oder die logistische Funktion mit einem Schwellwert genutzt werden. Die genutzte Funktion wird auch Aktivierungsfunktion genannt. Die Parameter der Entscheidungsfunktion, in diesem Fall die Koeffizienten des Merkmalsvektors, werden im Nachfolgenden allgemein θ genannt. Sie sind der Teil der Entscheidungsfunktion, der von den Daten gelernt werden muss.

Während des Trainings werden die besten Parameter θ ermittelt. Dafür wird eine Funktion benötigt, die ausdrückt, wie gut das Modell den Trainingsdaten angepasst ist. Diese besteht allgemein aus zwei Teilen: der Kostenfunktion L und der Regularisierung Ω:

$$C(\theta) = L(\theta) + \Omega(\theta)$$

Als Kostenfunktion wird vor allem für Regressionsprobleme häufig der MSE verwendet. Bei binären Klassifikationsproblemen kann aber beispielsweise auch die Anzahl der falschen Klassifikationen verwendet werden. Die Regularisierung kontrolliert die Komplexität des Modells, um Overfitting zu vermeiden. Das Ziel des Trainings ist es demnach, die Anpassung des Modells zu optimieren, also C zu minimieren. Dieses Minimierungsproblem wird abhängig von dem verwendeten Modell gelöst. Bei einer linearen Regression beispielsweise kann die Methode der kleinsten Quadrate verwendet werden, um die Fehlerquadrate zu minimieren, da eine Regularisierung bei einem linearen Modell nicht nötig ist. Der Gewichtungsvektor θ kann hier durch direktes Auflösen berechnet werden:

$$\underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} (\theta^T x_i - y_i)^2 \Rightarrow \theta = (X^T X)^{-1} X^T y$$

Bei komplexeren Minimierungsproblemen ist die Methode kleinster Quadrate nicht ausreichend und es werden andere Methoden zur Fehlerminimierung benötigt. Eine ist das Gradientenabstiegsverfahren, bei dem in jedem Schritt die Ableitung der Kostenfunktion nach jedem der Gewichte berechnet wird und der Schritt gewählt wird, der die Funktion am stärksten minimiert.

Eine andere Möglichkeit ist, nicht linear separierbare Daten mit dem sogenannten Kerneltrick linear separierbar zu machen, statt nicht-lineare, komplexere Verfahren zu verwenden. Dabei wird durch Ersetzen des Skalarproduktes implizit der Variablenraum transformiert. Ein Beispiel ist der Gaußsche RBF-Kernel, wobei RBF für *radial basis function* steht. Sind x und x' Merkmalsvektoren, ist der RBF-Kernel wie folgt definiert:

$$K_{RBF}(x,x') := \exp(-\gamma\|x-x'\|^2) \text{ mit } \gamma > 0$$

In der Transformation werden alle Nicht-Linearitäten zusammengefasst, wodurch anschließend Linearität möglich ist.

Im Folgenden werden die in dieser Arbeit genutzten Modelle näher vorgestellt. Die Nachimplementierung von existierenden Algorithmen umfasst außerdem drei Modelle, die hier nicht detailliert vorgestellt werden: die Support Vector Machine (SVM), die Linear Discriminant Analysis (LDA) und das Multilayer-Perzeptron (MLP). Wichtig für die Einordnung ist, dass es sich bei LDA und SVM um zunächst lineare Modelle handelt und die SVM häufig mit dem Kerneltrick verwendet wird, um die Daten zu transformieren. Beim MLP handelt es sich um ein Neuronales Netz, das Nichtlinearität ermöglicht.

Classification And Regression Trees

Classification And Regression Trees (CART) sind nicht-lineare Modelle, die einem Binärbaum entsprechen. Sie können sowohl für Regression als auch für Klassifikation verwendet werden. Zwar sind sie hinsichtlich ihrer Performance oft anderen Lernmodellen unterlegen, werden aber als Teil von komplexeren Modellen genutzt und haben den Vorteil, sehr gut nachvollziehbar zu sein. Im Gegensatz zu vielen anderen Modellen benötigen sie keine Normalisierung der Merkmale. Es wird ein Pfad in einem Baum durchlaufen, indem an jedem Knoten unterschieden wird, ob ein bestimmtes Merkmal über oder unter einem Schwellwert liegt. Die Kostenfunktion bei einer Regression ist in der Regel der MSE. Für die Kostenfunktion einer Klassifikation wird häufig die „Reinheit“ eines Knoten m gemessen. Sie beschreibt, wie deutlich die Mehrheit einer einzelnen Klasse in m ist.

Für den Aufbau eines CART wird häufig das zu den greedy Algorithmen gehörende rekursive binäre Teilen verwendet. Hierbei wird ein Baum von der Wurzel zu den Blättern hin aufgebaut. In jedem Schritt wird das Merkmal gewählt, mit dem sich die Daten zu diesem Zeitpunkt am besten separieren lassen und der Schwellwert ermittelt. Abbruchkriterien sind unter anderem eine maximal erlaubte Tiefe oder eine Mindestanzahl von Datenpunkten in einem Blatt. Wenn diese zu viel Spielraum erlauben, neigen Bäume zum Overfitting.

Random Forest

Ein Zusammenschluss aus mehreren unkorrelierten Bäumen wird Random Forest (RF) genannt. Jeder dieser Bäume ist ein eigenständiges Modell, das ein Ergebnis liefert. Aus der Menge der Einzelergebnisse wird das endgültige Ergebnis ermittelt. Dadurch wird ein Teil der Nachvollziehbarkeit von Bäumen gegen eine bessere Generalisierung eingetauscht. Die einzelnen Bäume werden zufällig erzeugt, indem für jeden Baum n zufällige Datenpunkte ausgewählt werden. Von M Merkmalen werden nun $m \ll M$ ebenfalls zufällig gewählt, die als Kriterium für den Aufbau des Baumes verwendet werden. Der darauf folgende Aufbau des Baumes entspricht dem zuvor beschriebenen. Bei einer Klassifikation ist das endgültige Ergebnis je nach Implementierung ein Mehrheitsentscheid oder eine Kombination aller Wahrscheinlichkeiten; bei einer Regression der Durchschnitt aller Werte.

Ein Random Forest hat gegenüber anderen Modellen den Vorteil, dass er sehr schnell trainiert und somit sehr effizient für große Datenmengen ist. Gleichzeitig kann das Ergebnis relativ einfach nachvollzogen werden.

Gradient Boosted Trees

Auch ein Gradient Boosted Tree kombiniert mehrere einfache Bäume. Im Gegensatz zum RF hängt beim Gradient Boosted Tree jeder Baum von früheren Bäumen ab. Gestartet wird mit einem Baum, auf dessen Ergebnissen aufgebaut wird. Das endgültige Ergebnis wird auch hier aus der Menge an einzelnen Ergebnissen ermittelt, aber folgende Bäume konzentrieren sich jeweils darauf, die Schwächen der schon erzeugten Modelle auszugleichen.

Auch Gradient Boosted Trees ermöglichen eine nachvollziehbare Entscheidung. Meist erreichen sie etwas bessere Ergebnisse als die einfacheren Random Forests.

2.2 Medizinische Grundlagen

Die vorliegende Arbeit beschäftigt sich mit der Beurteilung der Signalqualität in ballistokardiographischen Signalen. Zum Verständnis der gemessenen Vorgänge und der Problematik in Bezug auf die Signalqualität und dessen Beurteilung ist grundlegendes medizinisches Wissen über die gemessenen Vorgänge und messtechnisches Verständnis nötig. Aufgrund dessen wird hier eine kurze Übersicht über die medizinischen Grundlagen gegeben.

2.2.1 Kardiorespiratorisches System

Das kardiorespiratorische System (zusammengesetzt aus *kardìa*, deutsch ‚Herz' und *respiratio*, deutsch ‚Atmung') setzt sich aus zwei Teilsystemen zusammen, dem kardiovaskulären und dem respiratorischen System, die zusammen die Versorgung der Organe mit Sauerstoff sicherstellen.

Das kardiovaskuläre System umfasst das Herz, die Arterien und die Venen. In einem Zyklus wird das sauerstoffreiche Blut von der linken Herzkammer durch die Arterien zu den Organen gepumpt, wo sich der Sauerstoff zur Versorgung dieser vom Blut löst. Die Venen transportieren das nun kohlenstoffdioxidreiche Blut in die rechte Herzkammer. Von dort wird es zur Lunge geführt, mit Sauerstoff angereichert und in die linke Herzkammer geleitet. Von dort beginnt der Vorgang von Neuem. Hierbei ist die Herzfrequenz ein relevanter messbarer Vitalparameter.

Ein Herzschlag selbst besteht aus zwei Phasen: einer füllenden und einer auswerfenden Phase. Während der Diastole, der Erschlaffungs- und Bluteinströmungsphase, füllen sich die Herzkammern mit Blut. Diese Phase endet mit dem Schließen der Herzklappen und die Systole beginnt. Die Systole ist die Anspannungs- und Blutausströmungsphase: Die Herzklappen öffnen sich durch Kontraktion des Herzmuskels und das Blut kann ausströmen.

Das respiratorische System umfasst die Lungen und den Lungenkreislauf. In einem Atemzyklus wird durch gezielte Muskelbewegungen Luft aus der Umgebung eingeatmet. Mit dem eingeatmeten Sauerstoff wird sauerstoffarmes Blut angereichert und anschließend die nun sauerstoffarme Luft ausgeatmet. In diesem Zusammenhang ist der Vitalparameter der Atemfrequenz messbar.

2.2.2 Übersicht Messtechniken

Zur Untersuchung der in dieser Arbeit betrachteten Ballistokardiographie (BKG) wird diese oft mit anderen Messmethoden als Referenz aufgenommen. Im Folgenden werden diese zur Einordnung kurz vorgestellt. BKG selbst wird im nächsten Abschnitt separat betrachtet.

Die Elektrokardiographie (EKG) zeichnet die elektrischen Aktivitäten des Herzmuskels auf, indem mit mehreren Elektroden die Spannungsänderung gemessen wird. Hier ist die Herzfrequenz sehr gut ablesbar.

Die Photoplethysmographie (PPG) ist ein optisches Messverfahren, bei dem die Menge des von der Haut reflektierten bzw. transmittierten Lichtes gemessen wird. Dadurch kann die Änderung des Blutvolumens gemessen werden; die Lichtmenge nimmt bei Durchlaufen einer Pulswelle durch die Arterie deutlich ab. Dieses Signal bietet Rückschluss auf Atmung und Herzschlag.

Oft gemeinsam mit dem BKG betrachtet wird die Seismokardiographie (SKG), bei der die Vibration der Wand des Brustkorbs, die durch den Herzschlag entsteht, aufgezeichnet wird. Aufgrund von fehlenden einheitlichen Definitionen wird in der Literatur teils auch der Begriff BKG für SKG genutzt.[1]

2.3 Ballistokardiographie

Im Folgenden wird in die Ballistokardiographie eingeführt. Diese Einführung beinhaltet den medizinischen und technischen Hintergrund, das Einsatzgebiet und die Signaleigenschaften.

2.3.1 Medizinischer und technischer Hintergrund

Ballistokardiographie (zusammengesetzt aus altgriechisch *ballein*, deutsch ‚werfen', *kardía*, deutsch ‚Herz' und *graphein*, deutsch ‚schreiben') ist die graphische Darstellung der wiederholten, durch den Herzschlag verursachten Bewegungen des menschlichen Körpers. Erstmals schon im 19. Jahrhundert beobachtet[2], ermöglicht der technische Fortschritt in der Sensortechnik heute aussagekräftige Messungen. Das BKG liefert durch

[1] Vgl. Inan et al. 2015.
[2] Vgl. Gordon 1877.

die Aufzeichnung von zirkulierendem Blut und mechanischer Herzaktivität Informationen über die Gesamtleistung des kardiovaskulären Systems.[3] Konkret gemessen wird eine Massenbewegung, die durch die schnelle Beschleunigung des Blutes entsteht, wenn es während des Herzschlages durch die großen Arterien bewegt wird: Bei der Verteilung des Blutes in die peripheren Blutgefäße verschiebt sich das Zentrum der Körpermasse in Richtung der Füße und während der atrialen Systole Richtung Körpermitte. Die BKG-Wellenform entsteht durch diese Schwerpunktverschiebung.

Die Messung dieser Bewegung ist mit verschiedenen Sensortypen, die z. B. hydraulisch oder elektromechanisch auf Druck reagieren, möglich. Sensoren können unter anderem in Waagen, Stühlen und Betten eingebaut werden. Besonders bei im Bett gemessenen Signalen kann oft nicht klar zwischen SKG und BKG unterschieden werden, da sich myokardiale Vibrationen und Masseverschiebungen durch den Blutfluss überlagern. Diese gemischten Signale werden in der Literatur teils auch als *cardiac vibration signals* bezeichnet.[4] Da im Bereich der Signalverarbeitung oft nicht zwischen reinem BKG und gemischten Signalen unterschieden wird, geschieht dies in der vorliegenden Arbeit ebenfalls nicht.

Verschiedene Studien kommen zu unterschiedlichen Ergebnissen bezüglich der Frage, welchen kardiovaskulären Ursprung die einzelnen Signalteile haben. Aufgrund dessen gestaltet sich die detaillierte Interpretation des BKG-Signals als schwierig. Da es neben Informationen zur Herzrate (HR) und Herzratenvariabilität (HRV) ein genauerer Indikator für das Alter des Herzens als Lebensalter ist, hat es trotzdem klinische Relevanz. Außerdem lassen sich durch abnormale Ballistokardiogramme Herzerkrankungen voraussagen, bevor Symptome auftreten. Besonders bei älteren Personen sind diese also eine wichtige Warnung.[5]

2.3.2 Einsatzgebiet

Durch diese Beschreibung wird deutlich, dass BKG große Unterschiede zu der sehr bekannten EKG zeigt. Der entscheidende Vorteil der BKGs liegt darin, dass kein einschränkender Körperkontakt durch z. B. aufgeklebte Elektroden nötig ist: Es lässt sich in Alltagsgegenständen wie Stühlen aber vor allem auch Betten integrieren, ohne dass es während der Messung zu Einschränkungen im alltäglichen Leben kommt oder medizinisches Fachpersonal anwesend sein muss. Damit gehört es zu den *unobtrusive* Messmethoden und eignet sich gut zur Langzeit- und Trendbeobachtung des Gesundheitszustandes - sowohl im klinischen Kontext als auch Zuhause. Besonders für Patient*innen mit chronischen Krankheiten und zur Früherkennung krankhafter Veränderungen bietet eine gesund-

[3] Vgl. Pinheiro et al. 2010.
[4] Vgl. Brüser, Winter et al. 2013.
[5] Vgl. zu diesem Absatz Pinheiro et al. 2010.

heitliche Überwachung von Zuhause aus großes Potential.[6] Je nach Aufbau des Messsystems verändert sich auch die Art der Informationen, die aus dem BKG-Signal gewonnen werden können. Sehr genaue, kontrolliert aufgenommene BKG-Signale ermöglichen eine aussagekräftige Analyse der Morphologie, wobei beispielsweise in Betten eingebautes BKG zunächst nur Aussagen zu Herzrate und Herzratenvariabilität bietet. Zusätzlich zu Informationen der Herzaktivitäten ermöglichen Bettsysteme aber auch Informationen über das allgemeine Aktivitätslevel und somit auch über die Schlafqualität.[7] In dieser Arbeit wird es um die Aufzeichnung von BKG-Signalen in Betten gehen. Der Aufbau eines solchen Bettsystems ist in Abbildung 2.3 gezeigt.

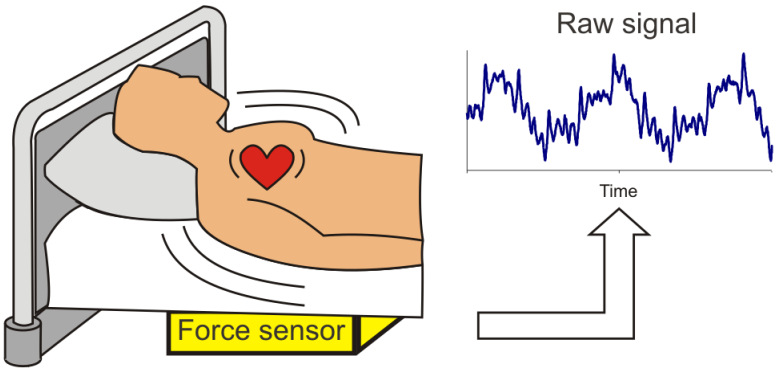


Abbildung 2.3: Übersicht über die Funktionsweise eines allgemeinen im Bett eingebetteten BKG-Systems.[8]

Allerdings ergeben sich neben diesen umfassenden Möglichkeiten auch Nachteile gegenüber konventionellen Messmethoden. Die größte Herausforderung ist eine stark variierende Signalqualität, die sich durch das unkontrollierte Umfeld und die Art der Messung ergibt.

2.3.3 Signaleigenschaften

Das gemessene BKG-Signal setzt sich aus Herzaktivitäten, Atmungsaktivitäten und Körperbewegungen zusammen. Gegebenenfalls wird es noch durch Störungen der Messung beeinflusst. Bei einer gesunden Person ohne Störeinflüsse wird die in Abbildung 2.4 abgebildete Wellenform erwartet. Diese Idealform lässt sich in 3 Gruppen unterteilen: Die präsystolische, wobei diese häufig nicht beachtet wird, die systolische und die diastolische Gruppe.[9] Die mit H bis K markierten Extremwerte gehören bei dieser Unterteilung

[6] Vgl. Inan et al. 2015.
[7] Vgl. Brüser, Stadlthanner et al. 2011.
[8] Entnommen aus Brüser, Stadlthanner et al. 2011
[9] Vgl. Pinheiro et al. 2010.

zur systolischen Gruppe, die Wellen L bis N zur diastolischen Gruppe. Die präsystolische Gruppe, die aus den Wellen F und G besteht, ist in hier nicht abgebildet. I und J werden auch als *ejection waves* bezeichnet. In Bezug auf andere Messmethoden ist zu bemerken, dass die H-Welle nahezu synchron mit dem ersten Herzgeräusch ist. Der Abstand des R-Peaks, des Hochpunkts eines EKGs, zur H-Welle variiert im Bereich von 0,2 bis 0,3 Sekunden.[10] Die Amplitude der Wellen ohne Störeinflüsse ist hauptsächlich abhängig von dem Herzzeitvolumen, der Herzkraft und der Geschwindigkeit des Auswurfs.[11]

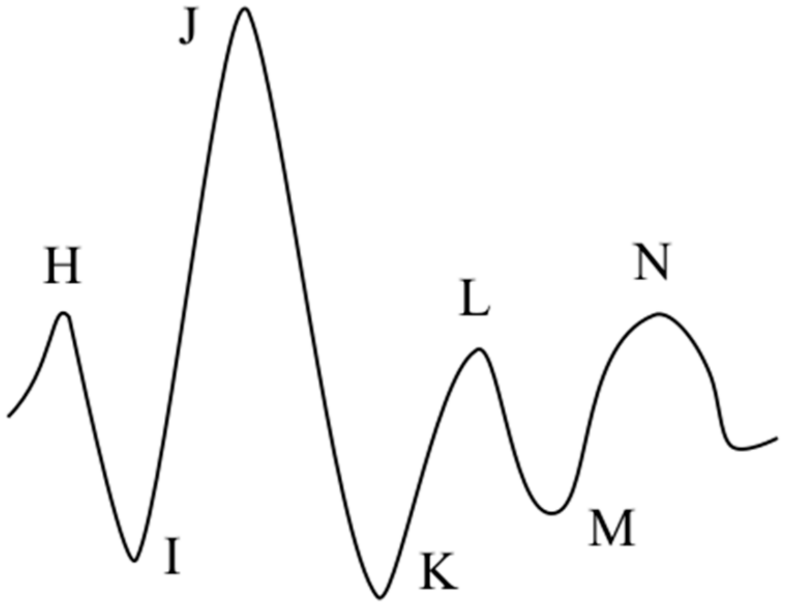


Abbildung 2.4: Beispiel eines typischen BKG-Signals mit Nomenklatur.[12]

Im Idealfall wird zwar die oben beschriebene Wellenform erwartet, bei der die Wellen H bis L eine deutliche W-Form bilden, allerdings ist es trotz dieser typischen Form selten, dass alle nicht-systolischen Komponenten sichtbar sind.[13] Es gibt starke Variationen der Signalmorphologie sowohl zwischen als auch innerhalb von Individuen. Der größte Einfluss ergibt sich durch die verwendeten Sensoren und die Position der Person, also zum Beispiel, ob im Stehen, Sitzen oder Liegen gemessen wird.[14] Es gibt Studien, die zeigen, dass die intraindividuelle Varianz über serielle Messungen hinweg niedrig ist.[15] Allerdings gilt das nicht, wenn sich die Position der Person verändert. Hierbei reicht es schon, wenn die Person in Rückenlage statt Seitenlage liegt.[16] Aufgrund dieser Variationen in der Signalmorphologie wurden schon in den 1950er Jahren drei Achsen für die

[10] Vgl. de Lalla et al. 1950.
[11] Vgl. Pinheiro et al. 2010.
[12] Entnommen aus Albukhari et al. 2019 nach Starr et al. 1939.
[13] Vgl. Pinheiro et al. 2010.
[14] Vgl. Sadek, Biswas und Abdulrazak 2019.
[15] Vgl. Inan et al. 2015.
[16] Vgl. Brüser, Stadlthanner et al. 2011.

Aufzeichnung des BKGs definiert: Die longitudinale (Kopf-Fuß), die transversale (Seite-Seite) und die dorsoventrale (Rücken-Brust).[17] Zu Beginn maßen die meisten Systeme entlang der longitudinalen Achse, die z. B. der Messung auf einer Waage entspricht. *Unobtrusive* Messsysteme, wie die hier betrachtete Messung in Betten, messen entlang einer Kombination der transversalen und der dorsoventralen Achse – abhängig von der Position der Person. Besonders diese Kombination sorgt für eine große intra- und individuelle Variation des Signals. Abbildung 2.5 verdeutlicht dies durch den direkten Vergleich von BKG-Aufzeichnungen zweier Herzschläge von 2 Personen. Bei jeder dieser beiden Personen wurde in zwei verschiedenen Positionen gemessen.[18] Auch der Ursprung des Signals ist abhängig von der Messachse. Bei longitudinal gemessenem BKG ist der Einfluss des Herzzeitvolumens schon seit 1939 beobachtet.[19] Im Gegensatz dazu ist der Ursprung des in Betten gemessenen BKG-Signals nicht genau bekannt. Das liegt unter anderem daran, dass mechanische Komponenten wie z. B. die Matratze einen schwer zu modellierenden Einfluss haben.

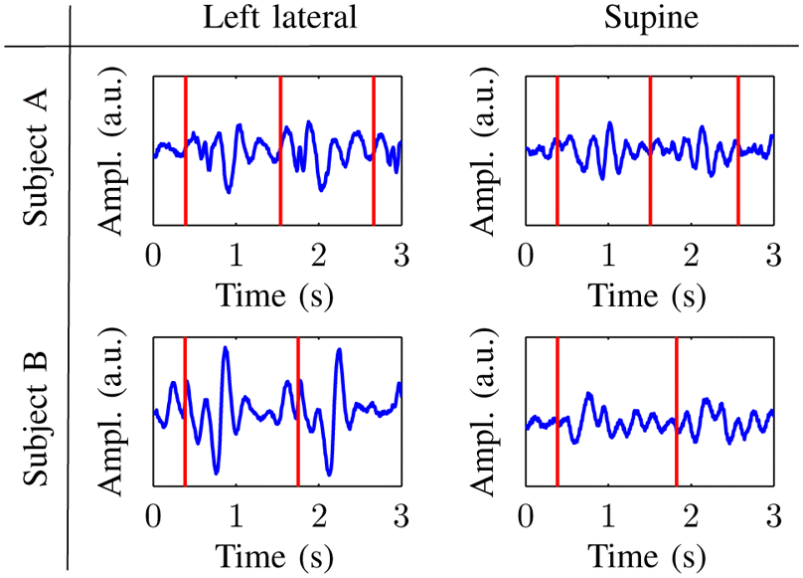


Abbildung 2.5: Hochpass-gefilterte BKG-Aufnahmen von zwei Herzschlägen zwei verschiedener Personen, jeweils in Rücken- und Seitenlage gemessen. Die vertikalen Linien markieren die R-Peaks der EKG-Referenz.[20]

Neben Einflüssen der verwendeten Messachse und der Körperposition beeinflusst auch die Atmung die Signalform. Normale Atmung beeinflusst die Amplitude der *ejection waves* I und J. Bei Atemstillstand dagegen werden die H und J Wellen verzerrt. Auch bei einer gesunden, sich nicht bewegenden Person, die ihre Atmung kontrolliert, wird kein

[17] Vgl. Brüser, Stadlthanner et al. 2011; Inan et al. 2015.
[18] Vgl. Brüser, Stadlthanner et al. 2011.
[19] Vgl. Starr et al. 1939.
[20] Entnommen aus Brüser, Stadlthanner et al. 2011.

exakt Schlag für Schlag reproduzierbares Signal erzeugt werden.[21] Von Zink et al. werden die Einflüsse der Atmung in der vertikalen Achse eines dorsoventralen BKGs als große Schwingungen einer Wellenlänge von fünf bis zehn Sekunden beschrieben. Innerhalb dieser sind kleinere Schwingungen mit höherer Frequenz sichtbar, die jedoch keiner bestimmten Sequenz folgen.[22] Zusätzlich zu dieser schon beschriebenen Variabilität kommt es sehr leicht zum Entstehen von Artefakten. Ursprung ist entweder das Messsystem selbst oder Körperbewegungen. Insgesamt führt Bewegung der Patient*innen, auch die der Atmung, zu einem *baseline drift*. Stärkere Bewegungen führen zu einer Massenverschiebung, die um ein Vielfaches größer als die gemessenen Vorgänge ist. Aufgrund dessen führt sie immer dazu, dass das Signal stark verzerrt oder sogar vollständig überlagert wird.

Besonders im Vergleich zu anderen kardiorespiratorischen Signalen wie dem EKG und PPG wird deutlich, dass BKG-Signale auch in konsekutiven Messungen deutlich variabler sind. Abbildung 2.6 zeigt dies am Beispiel von BKG-Aufnahmen eines im Bett integrierten Messsystems im Vergleich zum parallel aufgenommenen EKG. Es zeigt sich, dass selbst nach Entfernung von Überlagerungen von Atmung und Bewegung das BKG-Signal eine höhere Variabilität in Bezug auf Amplitudenhöhe, Reihenfolge der Extremwerte und der gesamten Form aufweist.[23] Es wird allerdings angenommen, dass aufeinander folgende Herzschläge einander ähneln.[24] Diese Eigenschaft wird Selbstähnlichkeit genannt. Brüser, Winter et al. nennen als eine mögliche Ausnahme den Fall, dass ein unregelmäßiger Herzschlag mit sehr niedrigem Schlagvolumen einem regulären Herzschlag folgt. In dem Fall ist es möglich, dass die Amplitude im Vergleich so klein ist, dass sie verdeckt wird. Dies ist z. B. bei Vorhofflimmern möglich. Eine Untersuchung von Rosales et al. zeigt dieses Verhalten der Selbstähnlichkeit nicht bei den kleineren Extremwerten, die den Hochpunkt J umgeben.[25] Dass die Ähnlichkeit um J am größten ist, zeigt auch Abbildung 2.6.

Zusammengefasst lässt sich sagen, dass es sich bei ballistokardiographischen Signalen um nichtstationäre Signale handelt, deren Ursprung nicht genau bekannt ist. Die Signalform wird von der Messachse, der Position und Körperhaltung der Proband*innen und dem Messsystem selbst beeinflusst. Besonders bei dem hier im Fokus liegenden Anwendungsfall Bett kommt es sowohl durch die unkontrollierbare Umgebung als auch die Signaleigenschaften selbst zu einer starken Variation der Morphologie und vielen Artefakten im Signal. Trotz dieser Einschränkungen ist die Ballistokardiographie eine Messtechnik, die

[21] Vgl. Pinheiro et al. 2010.
[22] Vgl. Zink et al. 2017.
[23] Vgl. Zink et al. 2017.
[24] Vgl. Brüser, Winter et al. 2013.
[25] Vgl. Rosales et al. 2012.
[26] Entnommen aus Zink et al. 2017.

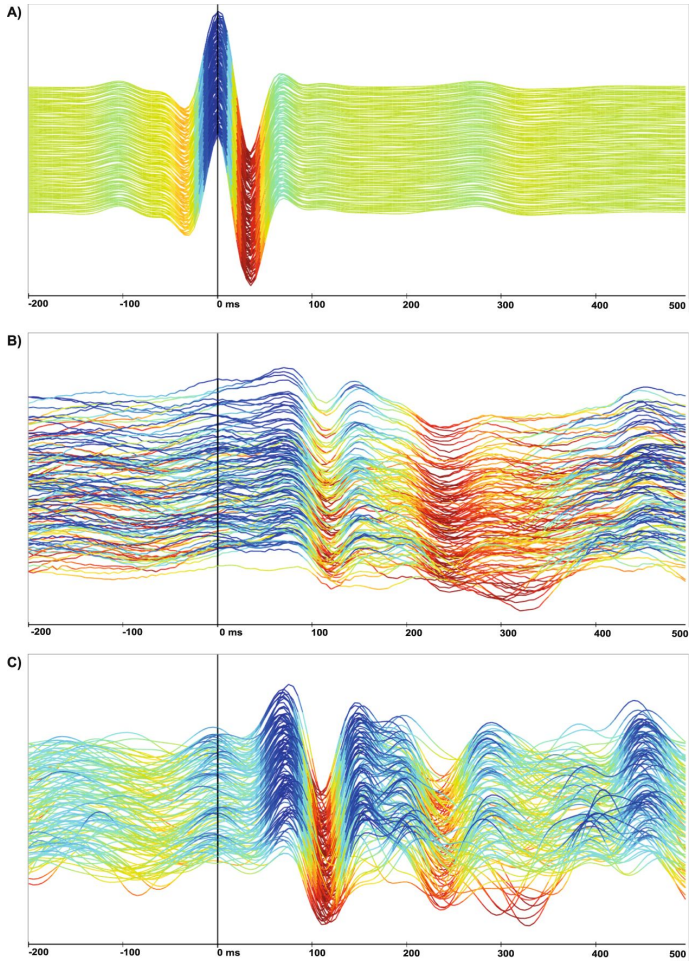


Abbildung 2.6: Diagramm aus 128 konsekutiven Herzschlagen im EKG (A) und BKG (B,C), segmentiert durch das EKG. Die Farben dienen der besseren Visualisierung der Amplituden. (A) EKG-Signal; (B) BKG-Signal mit Überlagerungen durch Atmung und Bewegung; (C) BKG-Signal ohne Bewegungsartefakte und Atmung.[26]

sich einfach *unobtrusive* in den Alltag integrieren lässt und Aussagen über die Herzrate und die Herzratenvariabilität ermöglicht.

3 Signalverarbeitung bei ballistokardiographischen Signalen

In diesem Kapitel wird eine Einführung in das Thema der Signalverarbeitung und besonders Artefakterkennung bei ballistokardiographischen Signalen gegeben. Hierzu wird zunächst Grundlegendes zu der Verarbeitung kardiorespiratorischer Signale erläutert. Ein in dieser Arbeit verwendeter Algorithmus zur Detektion von Herzschlägen wird vorgestellt. Anschließend wird die Thematik der Artefakterkennung eingeführt und der aktuelle Stand der Wissenschaft bei BKG-Signalen vorgestellt. Hierfür werden drei Verfahren im Detail betrachtet. Darüber hinaus werden die in dieser Arbeit untersuchten Daten vorgestellt und die durchgeführte Vorverarbeitung beschrieben.

3.1 Grundsätzliches

Kardiorespiratorische Signale sind durch die quasiperiodische Natur des Herzens und der Lunge selbst quasiperiodisch. Zwei zyklische Vorgänge werden gleichzeitig gemessen, lassen sich aber durch eine Bandpass-Filterung nach ihren unterschiedlichen Frequenzen filtern. Der Normbereich für die Atemfrequenz liegt bei 12 bis 25 Atemzügen pro Minute, alles ober- und unterhalb wird als abnormal betrachtet. Bei der Herzfrequenz wird ein Bereich von 30 bis 200 Schlägen pro Minute erwartet. Dabei entsprechen 30 Schläge pro Minute der Pulsabsenkung in der Nacht, der Ruhepuls selbst ist höher. Grundsätzlich gibt es verschiedene Arten der Signalverarbeitung: Algorithmen, die im Zeitbereich arbeiten, solche, die im Frequenzbereich arbeiten und Algorithmen, die beides kombinieren. Dabei ist zu beachten, dass bei frequenzbasierten Algorithmen durch die Analyse von spektralen Eigenschaften zunächst nur durchschnittliche Frequenzen ermittelt werden. Dies ist für einige medizinische Anwendungen ausreichend, für andere wie z. B. die Ermittlung der HRV allerdings nicht. Algorithmen, die auf dem Zeitbereich arbeiten, basieren oft auf Wissen über die Morphologie des physiologischen Signals. Durch die Eigenschaften der BKG-Signale, vor allem durch die variable Morphologie, ist die Nutzung von Wissen über die Morphologie schwieriger als bei anderen kardiorespiratorischen Signalen. Paalasmaa et al. sagen dazu:

> *The properties of the BCG signal vary so much in practice that no simple filtering rule can be devised for an accurate and reliable beat-to-beat interval detection.*[1]

Diese Aussage gilt sowohl für die Detektion von Schlag-zu-Schlag-Intervallen als auch für die Beurteilung der Signalqualität bzw. die Artefakterkennung.

3.2 Detektion von Herzschlägen

In dieser Arbeit wird der von Brüser, Winter et al.[2] entwickelte Algorithmus, der Continuous Local Interval Estimator (CLIE) verwendet und im Folgenden hier vorgestellt. Der Algorithmus beruht auf der in Kapitel 2.3 vorgestellten Annahme, dass aufeinander folgende Herzschläge sich ähneln, und schätzt die Herzrate anhand der Selbstähnlichkeit des Signals.

Der Algorithmus iteriert mit einem *Moving window* über das mit einem Bandpass gefilterte Signal. Es werden zwei Schwellwerte für die Intervalllänge T genutzt, $T_{\min}$ und $T_{\max}$, basierend auf dem bekanntem Bereich der Herzrate von 30 bis 200 Schlägen pro Minute. Die Länge des Analysefensters w_i entspricht $2 \cdot T_{\max}$, sodass mindestens zwei vollständige Herzschläge enthalten sind. f_s beschreibt dabei die Abtastrate des untersuchten Signals.

$$w_i[v] = x[n_i + v], v \in \{-T_{\max} * f_s, \ldots, T_{\max} * f_s\}.$$

In jedem Fenster wird die lokale Intervalllänge T_i geschätzt und anschließend das Zentrum des Fensters n_i weiterbewegt:

$$n_{i+1} = n_i + \Delta t * f_s.$$

Die Schätzung der Intervalllänge beruht auf drei Selbstähnlichkeitsmaßen, die wie folgt definiert sind:

[1] Paalasmaa et al. 2015.
[2] Vgl. Brüser, Winter et al. 2013.

$$E_{\text{Corr}}[N] = \frac{1}{N}\sum_{v=0}^{N} w[v]w[v-N],$$
$$E_{\text{AMDF}}[N] = (\frac{1}{N}\sum_{v=0}^{N} |w[v]-w[v-N]|)^{-1},$$
$$E_{\text{MAP}}[N] = \max_{v\in\{0,\ldots,n\}} (w[v]+w[v-N]).$$

Dabei berechnet E_{Corr} eine modifizierte Autokorrelationsfunktion, mit E_{AMDF} wird die Differenz des Signals bei verschiedenen Abständen miteinbezogen und mit E_{MAP} werden die maximalen Amplituden von zwei beliebigen Samples über das ganze Fenster berechnet. AMDF steht für *modified average magnitude difference function* und MAP für *maximum amplitude pairs*. Diese Schätzer entsprechen jeweils einer Wahrscheinlichkeitsfunktion, die beschreibt, wie wahrscheinlich es ist, dass n dem tatsächlichen Schlag-zu-Schlag-Intervall entspricht. Durch Skalierung können sie in Wahrscheinlichkeitsdichtefunktionen verwandelt werden. Durch Kombination dieser drei Funktionen wird nun der wahrscheinlichste Wert für N ermittelt:

$$N_{\text{opt}} = \operatorname*{argmax}_N p(N|E_{\text{Corr}}, E_{\text{AMDF}}, E_{\text{MAP}})$$

Nach dem Satz von Bayes kann die Wahrscheinlichkeit, dass N der tatsächlichen Intervalllänge entspricht, auch wie folgt ausgedrückt werden:

$$p(N|E_{\text{Corr}}, E_{\text{AMDF}}, E_{\text{MAP}}) = \frac{p(E_{\text{Corr}}, E_{\text{AMDF}}, E_{\text{MAP}}|N)p(N)}{p(E_{\text{Corr}}, E_{\text{AMDF}}, E_{\text{MAP}})}$$

Da $p(E_{\text{Corr}}, E_{\text{AMDF}}, E_{\text{MAP}})$ unabhängig von N ist, kann es für die Ermittlung der wahrscheinlichsten Intervalllänge T_{opt} vernachlässigt werden. Unter den Annahmen, dass die Ergebnisse der drei Schätzer nicht voneinander, sondern nur von N abhängen und dass N gleichverteilt ist, kann T_{opt} wie folgt ermittelt werden:

$$E_f[N] = E_{\text{Corr}}[N] \cdot E_{\text{AMDF}}[N] \cdot E_{\text{MAP}}[N],$$
$$N_{\text{opt}} = \operatorname*{argmax}_n E_f[N]$$

Durch lineare Skalierung erhält man auch hier eine Dichtefunktion. Abbildung 3.1 zeigt

die 3 einzelnen Dichtefunktionen und die fusionierte Funktion. In letzterer zeigt ein deutlicher Hochpunkt N_{opt}.

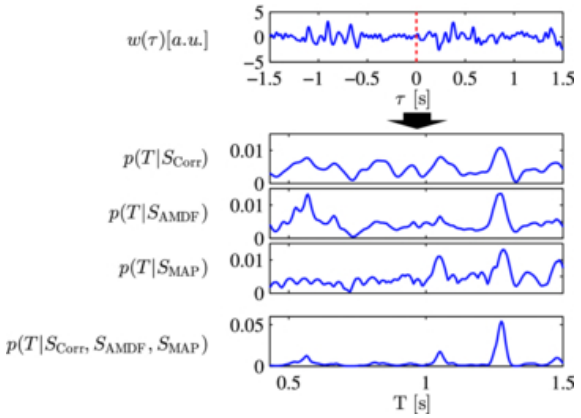


Abbildung 3.1: Die drei Intervallschätzer und ihre Fusion.[3]

Nun gibt es für jeden Punkt im Signal eine Schätzung der Intervalllänge. Mit Hilfe der Fenstergröße und dieser Länge können die zu einem Herzschlag gehörenden Hochpunkte ermittelt werden, nämlich die, die die größte kombinierte Amplitude mit dem durch die Intervalllänge gegebenen Abstand besitzen. Für jeden dieser Punkte P_k existiert nun eine Menge an Schätzungen T_k, mit der eine robuste Intervallschätzung $\overline{T_k} = \text{median}(T_k)$ ermittelt werden kann.[4]

3.3 Artefakterkennung

Artefakte sind irrelevante Signalteile mit variierender Amplitude, Frequenz und Dauer, die das physiologische Signal stören.[5] Das Ziel der Artefakterkennung ist es, nur die Teile des Signals zu verarbeiten, die Vitalparameter enthalten. Bewegungsartefakte, Sensorstörungen und ähnliches, die diese Parameter überlagern, sollen die Verarbeitung nicht beeinflussen. Dabei ist die Quelle der Störung selbst irrelevant, allerdings sollten keine medizinisch induzierten Abnormalitäten als gestörtes Signal klassifiziert werden. Insgesamt ist es für aussagekräftige Messungen wünschenswert, eine hohe Coverage zu erreichen, das heißt einen möglichst hohen Anteil an Signal zu verwenden. Sadek, Biswas, Yongwei et al. unterscheidet zwischen informativem und nicht-informativem Signal. Informatives Signal enthält mit *noise* gemischtes Signal von guter Qualität, aus dem Vitalparameter ohne weiteres extrahiert werden können.[6] Im Gegensatz dazu sind die Informationen bei

[3] Entnommen aus Brüser, Winter et al. 2013.
[4] Vgl. zu diesem Kapitel Brüser, Winter et al. 2013.
[5] Vgl. Nizami et al. 2013.
[6] Vgl. Sadek, Biswas, Yongwei et al. 2016.

nicht-informativem Signal so mit Artefakten und *noise* vermischt, dass vor Extraktion der Vitalparameter weitere individuelle Verarbeitung nötig ist oder die Extraktion von physiologischen Eigenschaften unmöglich ist. Teils wird die Signalqualität auch mit so genannten Signal Quality Indices (SQIs) gemessen, die je nach SQI und Anwendungsfall verschiedene Aussagen haben. Im klinischen Kontext genutzte Artefakterkennung verwendet oft relativ einfaches Preprocessing.[7] Außerdem sind in den meisten Algorithmen bestimmte Informationen *hard coded*. Das kann zum einen etwas wie Typ oder Frequenz der Daten sein, aber auch demographische Informationen über die Patient*innen wie Alter, Gewicht oder medizinischer Zustand.[8]

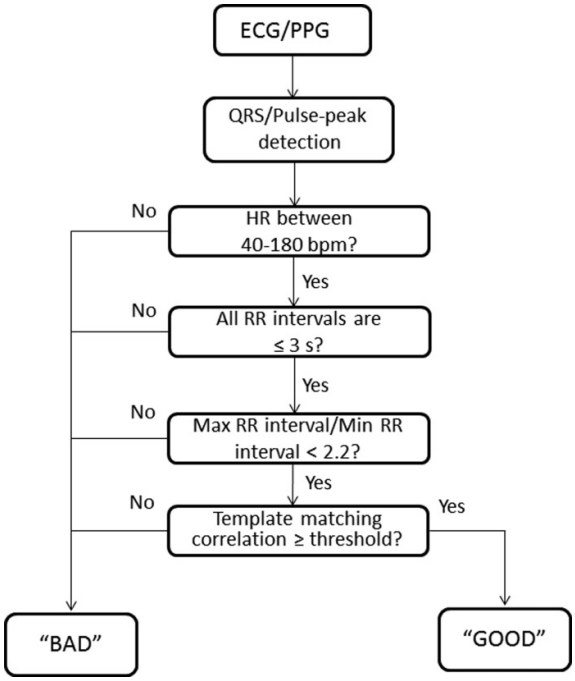


Abbildung 3.2: Flussdiagramm eines Algorithmus zur Beurteilung der Signalqualität.[9]

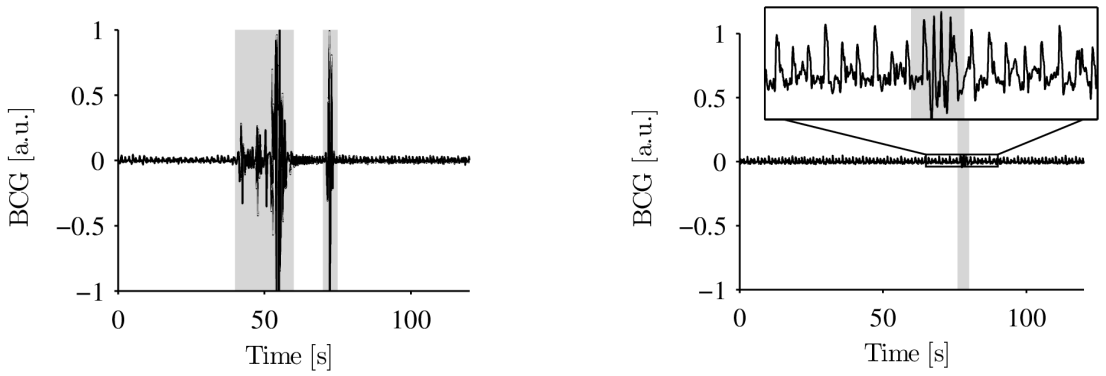
Um einen Eindruck über übliche Beurteilung der Signalqualität bei anderen kardiorespiratorischen Signalen zu bekommen, wird in Abbildung 3.2 ein Beispiel für EKG und PPG gezeigt. In jeweils 10-Sekunden Fenstern wird zunächst eine Segmentierung der Herzschläge durchgeführt und anschließend vier Kriterien überprüft, die jeweils ausreichend sind, um die Signalqualität als schlecht zu klassifizieren. Im ersten Kriterium wird geprüft ob die Herzrate zwischen 40 und 180 Schlägen pro Minute liegt. Im zweiten Schritt wird sichergestellt, dass kein Schlag fehlt, indem geprüft wird, ob alle Intervalle kürzer als drei Sekunden sind. Eine Intervalllänge von 3 Sekunden entspräche über eine Minute einer Herzrate von 20 Schlägen pro Minute. Anschließend wird geprüft, ob die Herzrate nur in

[7] Vgl. Nizami et al. 2013.
[8] Vgl. Nizami et al. 2013.
[9] Entnommen aus Orphanidou et al. 2015.

einem begrenzten Bereich variiert; das Verhältnis der maximalen zur minimalen Intervalllänge muss kleiner als 2,2 sein. Abschließend wird geprüft, ob die Korrelation mit einem erstellten Template einen gewissen Schwellwert nicht unterschreitet. Dieser Algorithmus enthält übliche Techniken der Signalbeurteilung: Eine Begrenzung der akzeptierten Herzrate und Herzratenvariabilität und dem Vergleich mit einem zuvor erstellten Template.

Bei ballistokardiographischen Signalen gestaltet sich auch die Beurteilung der Signalqualität schwieriger als bei anderen kardiorespiratorischen Signalen. Zusätzlich zu der gegebenen Variabilität durch die Atmung ist eine Problematik, dass plötzliche Veränderungen der Signalmorphologie bei Positionsänderungen zuvor erstellte Templates obsolet machen und auch andere Schwellwerte nicht mehr angemessen sind.

Bei Artefakten in BKG-Signalen kann zwischen Artefakten mit hoher und mit niedriger Energie unterschieden werden. Artefakte mit hoher Energie, wie in Abbildung 3.3a gezeigt entstehen, weil Bewegungen stärkere Krafteinwirkungen verursachen als Atmung und Herzschlag. Diese Tatsache wird häufig zur Erkennung von Bewegungen verwendet. Artefakte mit niedriger Energie, wie in Abbildung 3.3b gezeigt, sind weniger auffällig.



(a) Hohe Energie

(b) Niedrige Energie

Abbildung 3.3: Artefakte in BKG-Signalen.[10]

Bei der Betrachtung verschiedener Ansätze zur Signalverarbeitung und Artefakterkennung bei BKG-Signalen ist auffällig, dass Proband*innen oft angewiesen werden, sich möglichst wenig zu bewegen. Das ist bei der Messung im Alltag, insbesondere bei Messungen in Betten nicht realistisch. Dazu kommt, dass beim Einsatz im Alltag auch Kontakt mit der Person nicht immer gewährleistet ist, z. B. wenn diese sich aufrichtet. Hoog Antink et al. haben festgestellt, dass bei Messsystemen in Betten besonders Messungen tagsüber große Signalteile von schlechter Qualität aufweisen, deutlich mehr als nachts. Dies ist durch gesteigerte Aktivität tagsüber erklärbar. Ebenfalls im *unobtrusive* Kontext nicht zielführende Methoden zur Artefakterkennung verwenden eine EKG-Referenz.

[10] Graphik von C. Brüser

Im Folgenden werden drei verschiedene Methoden zur Beurteilung der Signalqualität vorgestellt.

3.3.1 Schwellwertbasierte Artefakterkennung

In „Noninvasive ambulatory measurement system of cardiac activity“ präsentieren Pino et al.[11] einen Ansatz für die Erkennung von Körperbewegungen für ein in einen Stuhl eingebettetes BKG-Messsystem. Dafür werden über ein *moving window* Maximum, Minimum, Standardabweichung und Mittelwert ermittelt und daraus 2 Schwellwerte berechnet:

$$\begin{aligned} T_1 &= \frac{\max + \min}{2}, \\ T_2 &= \text{mean} + 1{,}1 \cdot \text{std}. \end{aligned}$$

Ein Segment wird als informativ klassifiziert, wenn $T_1 \leq T_2$ ist. Die Länge des *moving window* ist mit einer Sekunde bei einer Abtastrate von 200 Hz benannt. Untersucht wurden sowohl Freiwillige im Labor, als auch im Krankenhauswartezimmer für eine sehr kurze Messdauer von ein bis zwei Minuten. Mit diesem Ansatz wurde bei mehr als 50 % der Laborgruppe eine Coverage zwischen 87 % und 95 % erreicht. Die Coverage der im Krankenhaus aufgenommenen Gruppe war bedeutend niedriger; hier lagen 50 % der Messungen zwischen 48 % und 95 % Coverage. Zu der Genauigkeit der Herzschlagdetektion auf den akzeptierten Signalteilen wird keine Aussage in Zahlen getroffen, sondern nur der folgende Bland-Altman Graph gezeigt.

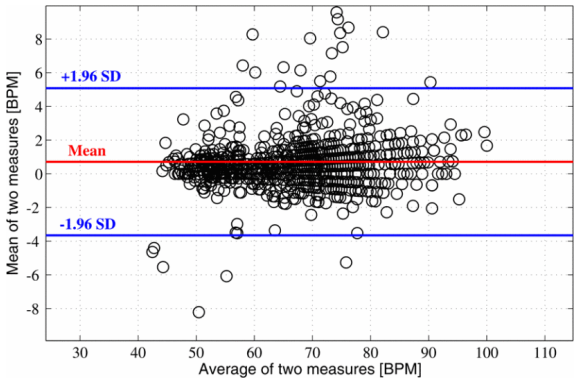


Abbildung 3.4: Bland-Altman Graph zwischen von EKG und BKG berechneter HR.[12]

[11] Pino et al. 2015.
[12] Entnommen aus Pino et al. 2015

Hier zeigt sich, dass beim Großteil des hier betrachteten Signals die HR größtenteils mit einer Genauigkeit von ± 3 Schläge pro Minute bestimmt werden konnte. Allerdings handelt es sich hier um im Sitzen gemessenes Signal, bei dem die Variabilität geringer ist als bei in Betten aufgenommenem BKG.

3.3.2 Maschinelles Lernen mit statistischen Merkmalen

Ein Algorithmus zur Beurteilung der Signalqualität mittels maschinellen Lernens wird von Sadek, Biswas, Yongwei et al. in „Sensor data quality processing for vital signs with opportunistic ambient sensing“ beschrieben.[13] Betrachtet werden BKG-Signale, die in einem Massagesessel aufgenommen werden, also ebenfalls im Sitzen aufgenommenes Signal, bei dem eine geringere Variabilität als in dem hier untersuchten Anwendungsfall erwartet wird.

Die vorliegenden Daten wurden manuell von Expert*innen als informativ oder nicht informativ klassifiziert und in 10-Sekunden-Segmente, die sich nicht überlappen, aufgeteilt. Insgesamt waren 58 % der Daten als informativ und 42 % als nicht-informativ gelabelt. Von diesen Segmenten wurden nach einer Bandpass-Filterung auf 1 bis 12 Hz 13 statistische Merkmale berechnet:

- Minimun
- Maximum
- Mittelwert
- Standardabweichung
- Schiefe
- Kurtosis[14]
- Spannweite
- Interquartilspannweite
- mittlere absolute Abweichung
- Anzahl der Nulldurchgänge
- Varianz der lokalen Minima
- Varianz der lokalen Maxima

[13] Vgl. Sadek, Biswas, Yongwei et al. 2016.
[14] Die Kurtosis ist eine Maßzahl für die Steilheit einer Wahrscheinlichkeitsfunktion

- Mittelwerte der Signalhüllkurve[15]

Für fünf verschiedene Modelle des maschinellen Lernens wurden jeweils die besten Hyperparameter über Kreuzvalidierung auf den Trainingsdaten ermittelt und die Modelle anschließend mit diesen Hyperparametern trainiert. Anschließend wurden die Modelle auf unbekannten Daten getestet. Das Training und Testen wurde mit getauschten Gruppen wiederholt. Das beste Ergebnis wurde mit einem Random Forest erreicht: Die durchschnittliche Accuracy der Kreuzvalidierung betrug 98,13 % bzw. 100 % bei getauschten Gruppen. Auf dem Testset wurde eine Genauigkeit von 92,3 % bzw. 97,99 % erreicht. Weitere Evaluationsmetriken außer einer Confusion Matrix für den besten Klassifikator sind nicht gegeben.

Diese Ergebnisse sind sehr gut, allerdings muss bei der Einordnung beachtet werden, dass bei der Kreuzvalidierung die Segmente zufällig verteilt wurden und nicht beachtet wurde, dass der Algorithmus für aussagekräftige Validierung einzelne Personen nicht kennen sollte. Dadurch ist die Performance auf gänzlich unbekannten Daten weiterhin nicht bekannt und vermutlich schlechter, als die Zahlen es hier vermuten lassen.

3.3.3 Ähnlichkeit der Intervallschätzer des CLIE-Algorithmus

Ein weiteres Maß für die Signalqualität basiert auf dem in Kapitel 3.2 vorgestellten Algorithmus zur Intervallschätzung. Dieser Signal Quality Index misst, wie einig sich die drei Intervallschätzer sind. Wenn diese sich uneinig sind, ist der SQI bei 0, je ähnlicher sich die Schätzungen sind, desto höher ist er. Für jedes Fenster i wird er wie folgt berechnet:

$$q = \frac{E_f[n_{\text{opt}}, i]}{\sum E_f[n, i]}$$

Bei der Ermittlung der konkreten Schlag-zu-Schlag Intervalle wird von dem SQI q äquivalent zu den geschätzten Intervalllängen der Median berechnet. Schätzungen dessen Qualität unter einem Schwellwert q_{th} werden verworfen. Die Wahl von q_{th} hängt von der gewünschten Genauigkeit und Coverage ab, je höher die gewünschte Genauigkeit, desto niedriger die Coverage.

Dieser SQI wird in der Literatur[16] mit unterschiedlich gewähltem q_{th} genutzt und erreicht damit bei der BKG-Messung in Betten während der Nacht bei gesunden Patient*innen gute Ergebnisse. Brüser, Winter et al.[17] erreichen zum Beispiel bei acht gesunden Proband*innen durchschnittlich eine Coverage von 85 % bei einem Fehler zur

[15] Die Signalhüllkurve ist eine glatte Kurve, die die Extrema des Signals umreißt.
[16] Bspw. Hoog Antink et al. 2020; Zink et al. 2017.
[17] Brüser, Winter et al. 2013.

EKG-Referenz von 0,61 % Schlägen pro Minute.[18] Bei Patient*innen im Krankenhaus zeigte sich, dass sowohl die Ergebnisse bezüglich der Coverage als auch der Genauigkeit bedeutend schlechter sind. So erreichen Hoog Antink et al.[19] bei 14 Patient*innen durchschnittlich 40 % Coverage bei einem Fehler von 2,16 % Schlägen pro Minute. Werden nur die nächtlichen Aufnahmen betrachtet, wird eine Coverage von 52 % bei einem Fehler von 1,29 % Schlägen pro Minute erreicht. Bei der Berechnung letzterer Ergebnisse wurden zusätzlich Intervallschätzungen verworfen, deren relative Abweichung vom Median der anderen Schätzungen einen bestimmten Schwellwert überschreitet.[20]

Dieser Algorithmus ist der einzige der drei betrachteten, der im Bett aufgenommenes BKG-Signal untersucht. Die Ergebnisse sind bei gesunden, schlafenden Patient*innen sehr gut, im tatsächlichen Einsatz im Krankenhaus wird allerdings eine deutlich geringere Coverage bei höherem Fehler erreicht.

3.4 Messdaten

Die vorliegenden Messdaten wurden in der Gefäßstation des Universitätskrankenhauses in Tampere (Finnland) aufgenommen und wurden im Paper „Ballistocardiography can estimate beat-to-beat heart rate accurately at night in patients after vascular intervention“[21] bereits untersucht. Insgesamt wurden 14 Patient*innen, zwei weiblich und zwölf männlich, bis zu 24 h überwacht. Alle hatten sich verschiedenen gefäßchirurgischen Eingriffen unterzogen. Das Durchschnittsalter der Patient*innen betrug 69,57 Jahre und die durchschnittliche Messdauer 17,7 h, wobei sie zwischen 4,46 h und 22,96 h variierte.

Gemessen wurde mit einem EMFit QS Bettsensor, der zwischen der Matratze des Krankenhausbettes und dem Bettgestell positioniert wurde. Die Abtastrate des BKG betrug 100 Hz. Das Referenz-EKG wurde mit einem Faros 360 Patientenmonitor mit einer Abtastrate von 1 kHz aufgenommen und hat drei Kanäle. Durch die unterschiedlichen Systemzeiten der beiden Geräte kam es zu einem Drift zwischen den beiden Signalen. Bei den langen Messungen zeigte sich, dass dieser zeitvariant ist.

Es liegen also BKG-Daten und ein 3-kanaliges EKG-Signal vor. Außerdem liegen schon detektierte und nach ihrer Qualität gefilterte Herzschläge mitsamt ihrer Qualität nach dem in Kapitel 3.3.3 beschriebenen SQI, der geschätzten Länge und der Länge des Referenzsignals vor. Zusätzlich gibt es zu jedem Paar von EKG und BKG-Daten einen Vektor,

[18] Vgl. Brüser, Winter et al. 2013.
[19] Hoog Antink et al. 2020.
[20] Vgl. Hoog Antink et al. 2020.
[21] Hoog Antink et al. 2020.

der den zeitvarianten Drift berücksichtigt. Zu jeder Sekunde des BKG-Signals wird die entsprechende Sekunde der EKG-Referenz benannt.

3.4.1 Vorverarbeitung

Die verschiedenen Daten liegen in unterschiedlichen Formaten und Dateien vor, die zunächst zusammengeführt und vorverarbeitet werden müssen. Dieser Vorgang sowie die entwickelte Datenstruktur werden im Folgenden beschrieben.

Für jede*n Patient*in existiert eine Matlab-Datei, die das rohe BKG-Signal und die bereits detektierten Intervalle mit Position, Qualität und Referenzlänge enthält. Die Vektoren zum Ausgleich des Drifts liegen ebenfalls in Matlab-Dateien vor. Das Referenzsignal ist im European Data Format gespeichert. Da die Nummerierung von BKG-Daten und EKG-Daten variiert, existiert außerdem eine Zuordnung dieser. Das Ziel der entwickelten Datenstruktur ist es, diese Daten zusammenzuführen. Die Klassenstruktur ist in Abbildung 3.5 visualisiert. Die Klasse *DataSeries* enthält alle Daten, die zu einer Person gehören, also das BKG-Signal, das EKG-Signal und den Vektor, um den Drift auszugleichen. Alle Patient*innen sind in *Data* gesammelt. Dort werden auch alle Daten eingelesen und zugeordnet.

Um die Herzraten auf Intervallen bestimmen zu können, müssen sowohl das BKG- als auch das EKG-Signal vorverarbeitet werden. Da diese Vorverarbeitung durch die große Datenmenge sehr aufwändig ist, werden die Ergebnisse als .csv-Dateien serialisiert. Beim Laden der Daten kann so geprüft werden, ob die Berechnung wiederholt werden muss oder geladen werden kann.

Bei der Vorverarbeitung des BKG-Signals wird der in Kapitel 3.2 vorgestellte CLIE-Algorithmus angewendet und die ermittelten Indizes der Intervalle, die geschätzten Intervallängen und der in Kapitel 3.3.3 beschriebene Qualitätsindex gespeichert. Für das EKG-Signal wird für alle drei Kanäle eine Detektion der QRS-Komplexe durchgeführt, mit denen die Schlag-zu-Schlag Intervalle ermittelt werden können. Dafür wird die Python-Implementierung von Howell et al.[22] des von Elgendi et al. beschriebenen Algorithmus[23] genutzt. Dieser zeigt in einer Untersuchung von Porr et al. gute Ergebnisse.[24]

[22] Howell et al. 2019.
[23] Vgl. Elgendi et al. 2010.
[24] Vgl. Porr et al. 2019.

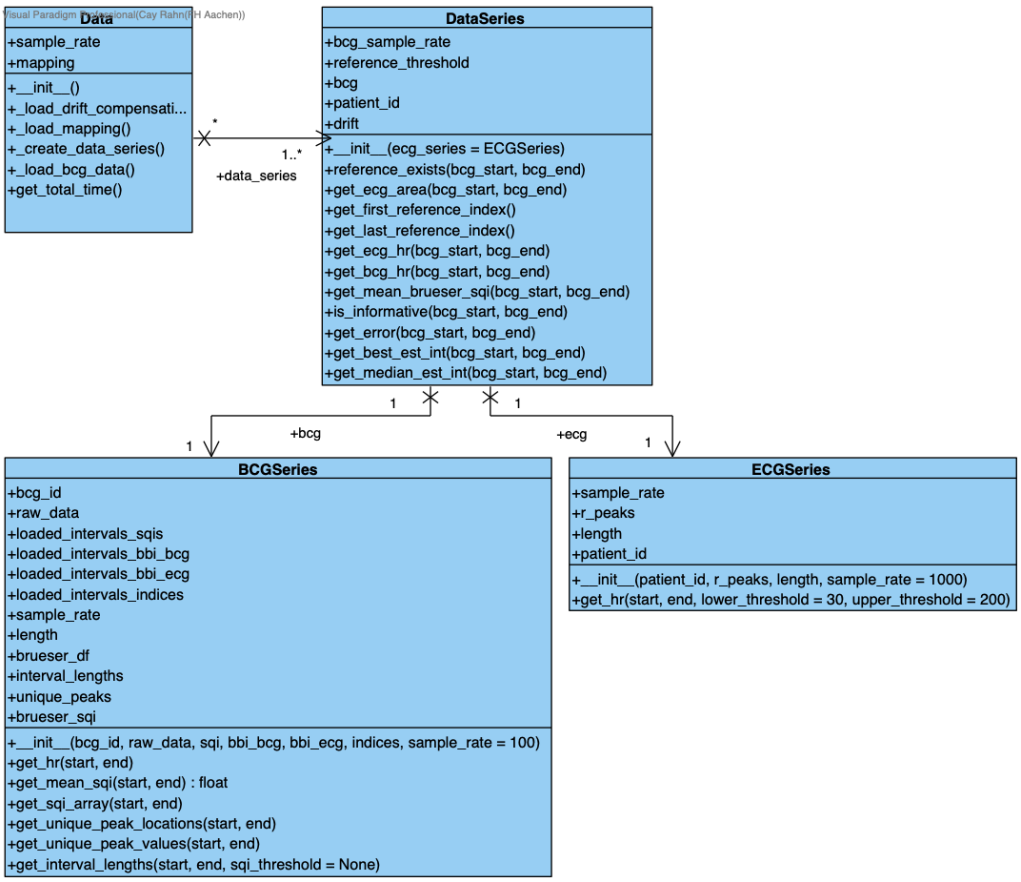


Abbildung 3.5: Klassendiagramm der Datenstruktur für die Messdaten.

3.4.2 Annotation der Daten

Die vorliegenden Daten sind nicht annotiert. Es ist im Rahmen dieser Arbeit nicht möglich, die Annotation durch Expert*innen durchführen zu lassen, weshalb auf das parallel aufgenommene EKG zurückgegriffen wird. Aufgrund des nicht-linearen Drifts der Daten ist eine herzschlaggenaue Synchronisierung schwierig. Aus diesem Grund wurde entschieden, die Annotation bereichsweise vorzunehmen. Für jeden Bereich wird zunächst geprüft, ob ein aussagekräftiges Referenzsignal existiert. Falls nicht, wird der entsprechende Bereich ausgeschlossen und nicht weiter verwendet. Ansonsten wird die durchschnittliche Herzrate sowohl für BKG- als auch EKG-Signal berechnet. Für die Berechnung der Herzfrequenz im BKG wird der Median der geschätzten Intervalllängen verwendet. So werden Ausreißer weniger stark gewichtet. Wenn mit den geschätzten Intervalllängen eine Coverage von unter 80 % erreicht wird, wird die Herzrate verworfen, da in diesem Fall nur kleine Teile des Signals aussagekräftig sind.

Auch bei dem Referenzsignal wird die Herzrate über den Median der Intervalllängen ermittelt. Die Intervalllängen entsprechen hier den Abständen der ermittelten R-Peaks. Da drei EKG-Kanäle vorliegen, wird die Herzrate zunächst für jeden Kanal einzeln ge-

schätzt. Falls die geschätzte Herzrate außerhalb des erwarteten Bereichs von 30 bis 200 Schlägen pro Minute liegt, wird der Kanal verworfen. Die Auswahl des Kanals gestaltet sich schwieriger, da auch im Referenzsignal teilweise schlechtes Signal vorliegt. Um dies auszugleichen, wird zur Auswahl des Kanals die Herzrate ebenfalls auf einem doppelt so großen Fenster berechnet. Von diesen drei Herzraten wird die ausgewählt, deren Kanal die geringste Spannweite an Intervalllängen aufweist. Mit Hilfe der geschätzten Herzrate des größeren Fensters wird nun der Kanal für die endgültige Schätzung ausgewählt, der die geringste Abweichung zu der Schätzung des größeren Fensters hat. Der Ablauf ist in Abbildung 3.6 visualisiert.

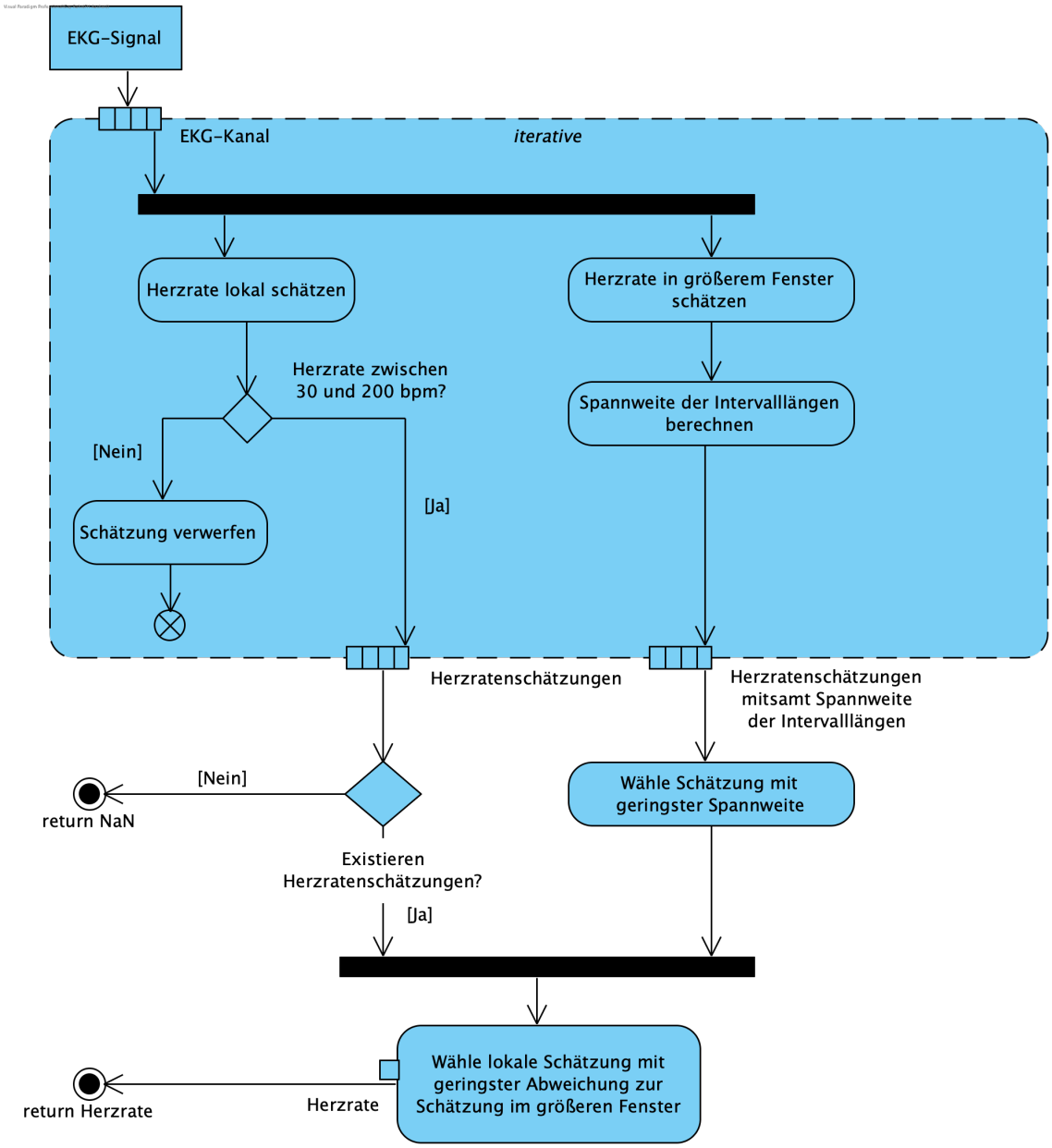


Abbildung 3.6: Aktivitätsdiagramm der Herzratenschätzung auf dem EKG-Referenzsignal.

Die Auswirkungen dieses Verfahrens sind in Abbildung 3.7 sichtbar. Eine Fusion nur auf dem aktuellen Bereich führt zu einigen Fehlern in der Herzratenschätzung, die die Umgebung einbeziehende Fusion führt zu deutlich besseren Ergebnissen. Auch ist hier sichtbar, wie groß die Abweichung einzelner Kanäle ist.

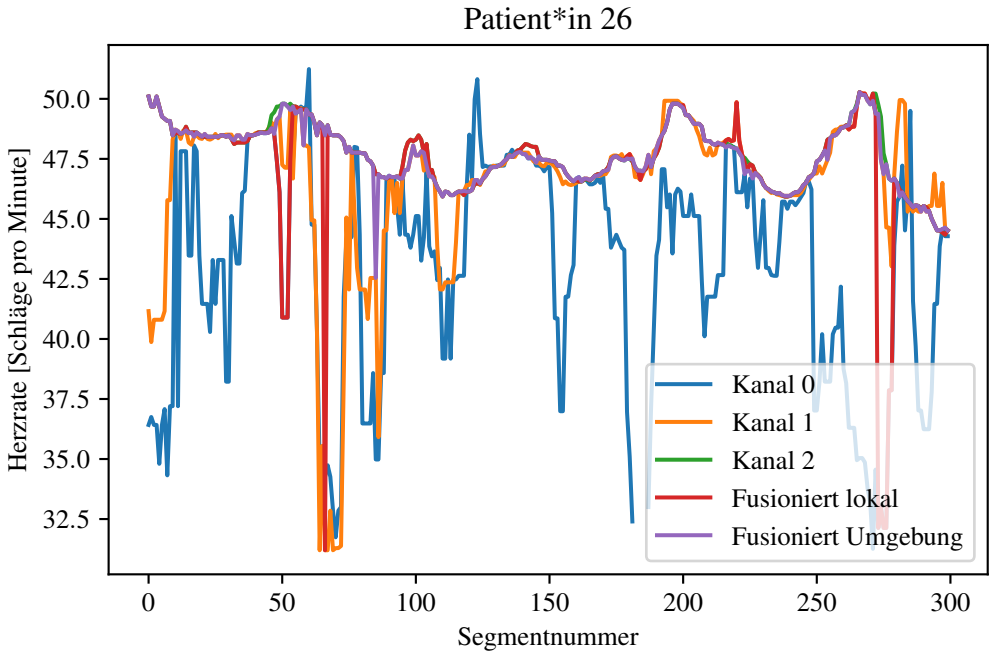


Abbildung 3.7: EKG-Herzratenschätzungen auf sich überlappenden 10-Sekunden-Segmenten für Patient*in 26. Fusioniert lokal zeigt die Fusion ohne Miteinbeziehung der Umgebung, Fusioniert Umgebung die oben beschriebene Fusion.

Auf Basis dieser beiden Herzraten HR_{EKG} und HR_{BKG} wird nun die Annotation vorgenommen. Dabei wurde sich an der Norm für EKG-Patientenmonitoren orientiert. Diese schreibt vor, dass die Abweichung der Herzrate maximal 10 % oder 5 Schläge pro Minute betragen darf, je nachdem welcher der beiden Werte größer ist,[25] und wurde hier so übernommen. Wenn eine höhere oder niedrigere Genauigkeit gefordert ist, kann dies auch auf 15 % und 7,5 Schläge pro Minute oder 5 % und 2,5 Schläge pro Minute angepasst werden. Die entsprechenden Verteilungen der Labels bei einer binären Klassifikation für 10 Sekunden lange Segmente mit 90 % Überlappung für 5, 10 und 15 % maximale Abweichung sind in Abbildung 3.8 zu sehen.

[25] *Medizinische elektrische Geräte - Teil 2-27: Besondere Festlegungen für die Sicherheit einschließlich der wesentlichen Leistungsmerkmale von Elektrokardiographie-Überwachungsgeräten* 2015.

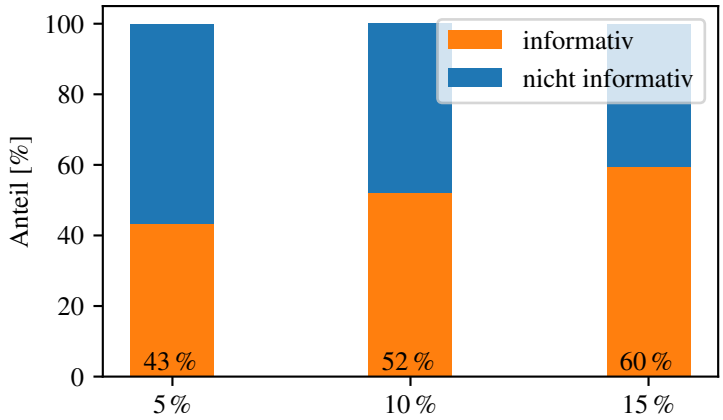


Abbildung 3.8: Verteilung der Labels je nach maximal zulässiger Abweichung.

Es ist auffällig, dass die Anteile der informativen Segmente stark patient*innenabhängig sind. In Abbildung 3.9 wird dies für 10 % maximale Abweichung visualisiert.

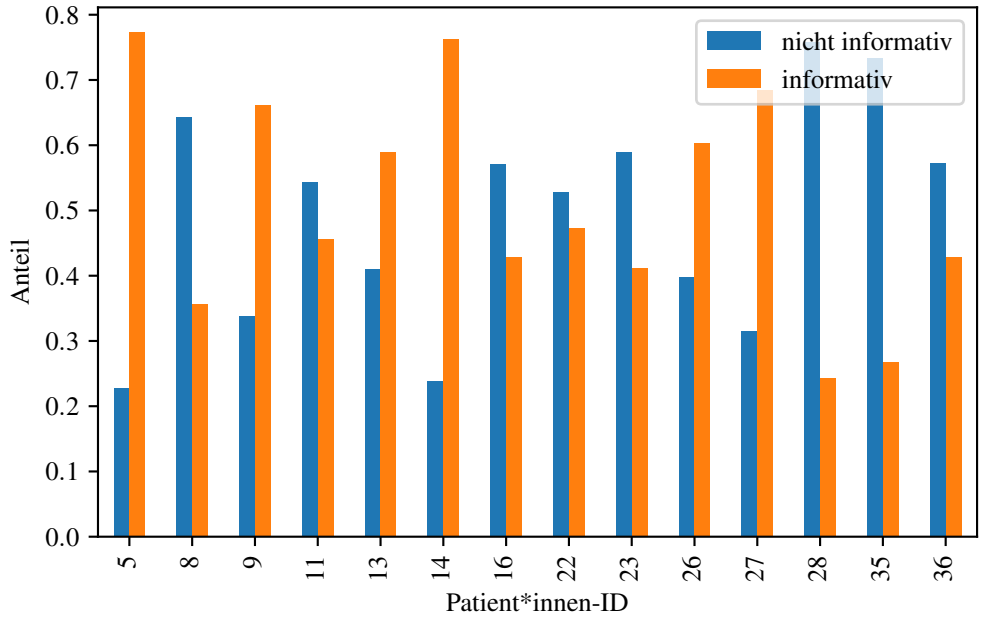


Abbildung 3.9: Verteilung der Label pro Patient*in für maximal 10 % Abweichung.

Neben einer binären Klassifikation ist es auch möglich, mit Hilfe dieses Fehlermaßes, im Folgenden E_{HR} genannt, einen Wertebereich für eine Regression zu erzeugen. Dafür ist es nötig, ein einheitliches Maß für die Unterscheidung von relativer und absoluter Abweichung zu finden, das heißt den absoluten Fehler so umzurechnen, dass er mit der relativen Abweichung vergleichbar ist. Bei Herzraten über 50 Schlägen pro Minute entspricht E_{HR} dem relativen Fehler $\Delta\mathrm{HR}_{\mathrm{relativ}}$, bei Herzraten darunter ermöglicht die absolute Abweichung den größeren Spielraum und wird daher für die Annotation verwendet. Eine

Umrechnung dieser erfolgt auf Basis der Überlegung, dass der Schwellwert von 10 % relativer Abweichung einer absoluten Abweichung von 5 Schlägen pro Minute entspricht. Daraus lässt sich folgern, dass eine Abweichung von 1 Schlag pro Minute einer relativen Abweichung von 2 % entspräche. Analog lässt sich E_{HR} bei Herzraten unter 50 Schlägen pro Minute durch die folgende Formel berechnen:

$$E_{\mathrm{HR}} \hat{=} \Delta \mathrm{HR}_{\mathrm{abs}} \cdot 2.$$

Wenn keine Herzrate aus dem BKG-Signal ermittelt werden konnte, wird E_{HR} auf den maximalen Wert 667 gesetzt, da das der maximal möglichen relativen Abweichung von Herzraten zwischen 30 und 200 Schlägen pro Minute entspricht. Für E_{HR} wird im Folgenden die Einheit Fehlereinheit FE verwendet. Die Verteilung von E_{HR}, dargestellt in Abbildung 3.10, zeigt, dass Segmente mit Fehlern von mehr als 20 FE und solche mit Fehlern unter 5 FE jeweils die beiden größten Gruppen bilden.

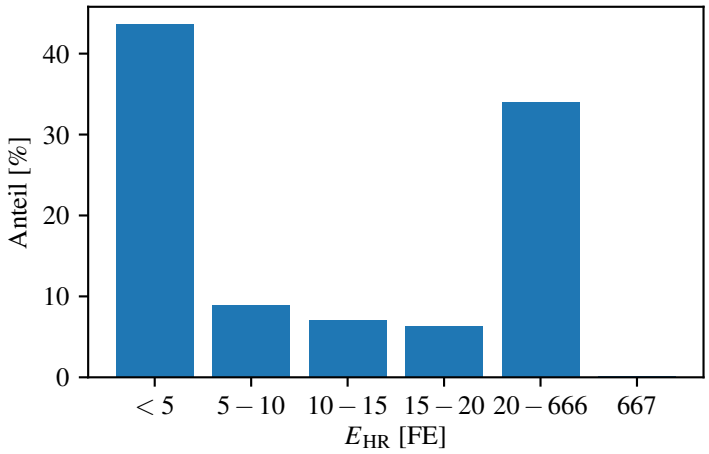


Abbildung 3.10: Verteilung von E_{HR}.

4 Analyse

In diesem Kapitel werden die vorgestellten Daten näher untersucht. Das beinhaltet sowohl die Anwendung der in Kapitel 3.3 beschriebenen existierenden Verfahren als auch eine Analyse der verwendeten Merkmale.

4.1 Aufbau und Evaluation der Verfahren

Um Modelle zur Beurteilung der Signalqualität anzuwenden und zu untersuchen, müssen sowohl die verwendeten Merkmale extrahiert als auch die Ergebnisse evaluiert werden. In den hier durchgeführten Untersuchungen wurde die Segmentlänge standardmäßig auf 10 Sekunden festgelegt und der Abstand der Segmente auf 1 Sekunde, wodurch sich die Segmente zu je 90 % überlappen. Diese Segmentlänge wird zum einen häufig verwendet[1], zum anderen bietet sie eine ausreichend große Robustheit für die Annotation der Daten. Die für die Verfahren jeweils benötigten Merkmale werden segmentweise extrahiert und serialisiert. Zusätzlich zu den verwendeten Merkmalen werden allgemeine Informationen über die Segmente gespeichert. Dazu gehören Patient*innen-ID, EKG-Herzrate, BKG-Herzrate, absoluter und relativer Fehler, E_{HR} und die binäre Annotation. Für letztere wird als Standard $E_{\mathrm{HR}} = 10$ als Schwellwert verwendet.

Diese Daten werden anschließend in Trainings- und Testset aufgeteilt. Um diese beiden Gruppen inhaltlich vollständig zu trennen und auszuschließen, dass auf teilweise bekannten Daten validiert wird, geschieht die Trennung anhand der Patient*innen-IDs, sodass Segmente einer Person lediglich in einem der beiden Sets verwendet werden. Das Testset in dieser Arbeit entspricht einem Drittel der Patient*innen, die restlichen zwei Drittel werden zur Datenexploration, zur Merkmalskonstruktion und zum Training verwendet. Die Verteilung von E_{HR} im Testset ist in Abbildung 4.1 gezeigt. Von den existierenden Verfahren benötigen zwei keine Trainingsphase; um vergleichbare Ergebnisse zu erhalten, wird auch bei diesen bei der Evaluation nur das Testset betrachtet.

Wie schon in Kapitel 3.3 beschrieben, sind bei der Verarbeitung von medizinischen Signalen zwei Kenngrößen wichtig: Die Coverage und die Qualität des Signals, in diesem Fall

[1] Vgl. Orphanidou et al. 2015; Sadek, Biswas, Yongwei et al. 2016; Yu et al. 2020.

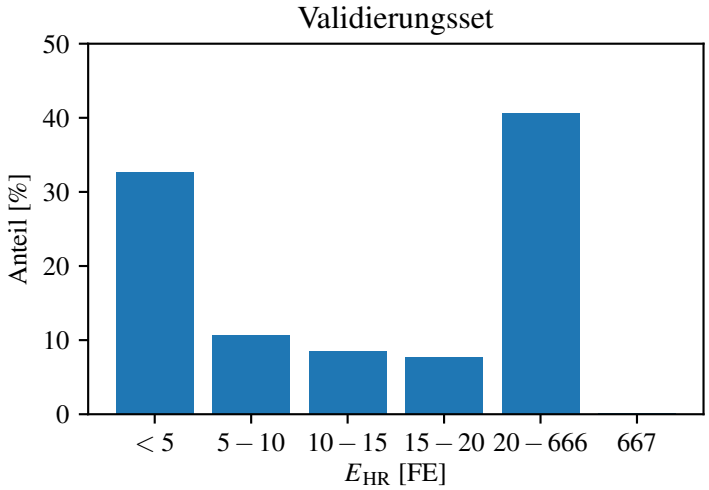


Abbildung 4.1: Verteilung von E_{HR} auf dem Testset, wobei 667 FE dem maximalen Fehler entspricht.

also die Genauigkeit der geschätzten Herzrate im Vergleich zur Referenz. Diese beiden müssen gegeneinander aufgewogen werden und werden aus diesem Grund beide betrachtet. Die Qualität der Modelle kann zunächst anhand der binären Klassifikation beurteilt werden. Da diese aber keine Auskunft darüber enthält, wie nah ein klassifiziertes Signal an dem Schwellwert für E_{HR} liegt, wird das in einer tiefergehenden Evaluation ebenfalls betrachtet. Insbesondere falsch klassifiziertes Signal, also Falsch-Negative und Falsch-Positive, ist interessant, um zu beurteilen, ob Fehlklassifikationen lediglich im Grenzbereich $E_{HR} \approx 10$ FE oder allgemein vorkommen. Im Zuge der tiefergehenden Evaluation wird sowohl der Fehler auf dem als informativ klassifizierten Signal betrachtet als auch die Coverage in Bezug auf das ganze Signal für verschiedene Fehlergrößen.

4.2 Anwendung existierender Verfahren

Zunächst werden die in Kapitel 3.3 beschriebenen existierenden Artefakterkennungsverfahren mit den in dieser Arbeit untersuchten Daten wie oben beschrieben getestet und ihre Leistungsfähigkeit untersucht und bewertet.

4.2.1 Ähnlichkeit der Intervallschätzer des CLIE-Algorithmus

Der SQI, der die Ähnlichkeit der Intervallschätzer des CLIE-Algorithmus angibt, wird im Normalfall herzschlagweise angewendet. Da die Datenannotion nur bereichsweise vorgenommen werden kann, wurde entschieden, ein Segment als informativ zu klassifizieren,

wenn mit den Herzschlägen, deren SQI über einem Schwellwert q_{th} liegt, eine Coverage über einem gegebenen Schwellwert c_{th} auf dem Segment erreicht wird. Getestete Schwellwerte für die Coverage sind 50 %, 75 % und 100 %. Zum Testen des Algorithmus wurde unter anderem $q_{th} = 0{,}4$ gewählt, da dieser Wert auch von Zink et al. verwendet wird.[2] Zusätzlich wurden $q_{th} = 0{,}3$ und $q_{th} = 0{,}2$ untersucht, um den Einfluss von q_{th} einzuordnen. Bei der Berechnung der Merkmale werden für jedes Segment die detektierten Herzschläge extrahiert, deren SQI über q_{th} liegt. Auf Basis dieser Intervalllängen wird wie in Kapitel 3.4.2 beschrieben die Herzrate, im Folgenden HR_{SQI} genannt, ermittelt. Extrahierte Merkmale sind damit in diesem Fall HR_{SQI} und die Coverage C_{SQI}. Für die Auswertung muss beachtet werden, dass sich die ermittelte Herzrate HR_{SQI} von der zur Annotation verwendeten Herzrate unterscheiden kann. Aufgrund der Vergleichbarkeit wird HR_{SQI} nicht für die Berechnung des MAE des Algorithmus verwendet.

Bei einer ersten Betrachtung von MAE in FE, Coverage und Accuracy wird sichtbar, dass die Wahl der Schwellwerte großen Einfluss auf das Ergebnis hat und in jedem Fall Aussagekraft des Signals gegen Coverage eingetauscht wird. Die genauen Ergebnisse sind in Tabelle 4.1 zu finden. Auch ist deutlich, dass die Accuracy der Klassifikation mit mit Werten knapp über 0,6 nicht gut ist. Des Weiteren ist die Coverage bei verhältnismäßig kleineren MAE sehr niedrig.

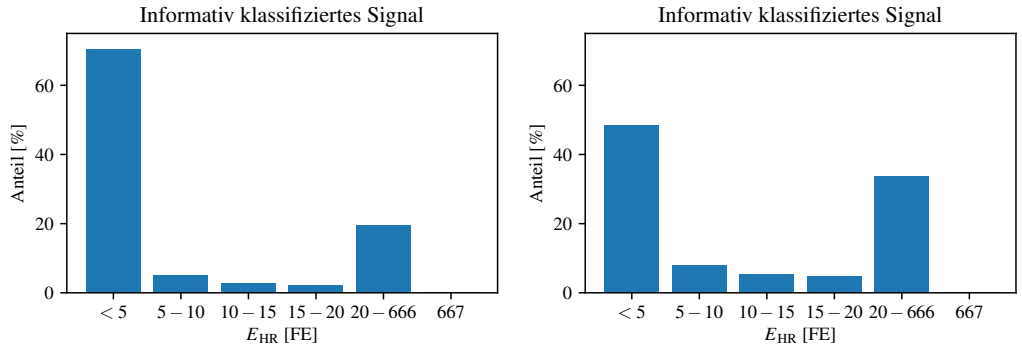
	MAE [FE]	Coverage [%]	Accuracy
insgesamt	21,85	-	-
annotiert	3,28	43,22	-
$q_{th} = 0{,}4, c_{th} = 50$	12,30	20,95	0,65
$q_{th} = 0{,}4, c_{th} = 75$	9,57	11,78	0,63
$q_{th} = 0{,}4, c_{th} = 100$	8,77	5,26	0,60
$q_{th} = 0{,}3, c_{th} = 50$	21,36	58,73	0,55
$q_{th} = 0{,}3, c_{th} = 75$	17,97	37,87	0,62
$q_{th} = 0{,}3, c_{th} = 100$	13,45	19,25	0,63
$q_{th} = 0{,}2, c_{th} = 75$	21,39	96,43	0,44

Tabelle 4.1: Fehler und Coverage für verschiedene Schwellwerte im Vergleich zum gesamten Signal und der Annotation.

Die detailliertere Evaluation wird beispielhaft für die Schwellwerte $q_{th} = 0.4$ und $q_{th} = 0.3$ mit $c_{th} = 75$ vorgestellt. Positiv hervorzuheben ist, dass bei beiden kein Signal als informativ klassifiziert wird, bei denen E_{HR} maximal ist. Allerdings ist bei über 20 % der mit $q_{th} = 0.4$ als informativ klassifizierten Segmenten E_{HR} größer als 20, mit $q_{th} = 0.3$ bei sogar mehr als 35 %, wie auch in Abbildung 4.2 abgebildet. Das bedeutet, dass Falschklassifikationen nicht nur im Randbereich vorkommen. Dies wird noch deutlicher, wenn man sich den MAE auf jeweils auf den Falsch-Positiven und Falsch-Negativen anschaut.

[2] Zink et al. 2017.

So beträgt der MAE bei $q_{th} = 0.4$ für falsch-negative Segmente 3,65 FE, ist also nah an dem MAE aller informativen Segmente von 3,28 FE.



(a) $q_{th} = 0.4$ (b) $q_{th} = 0.3$

Abbildung 4.2: Verteilung von E_{HR} bei den als informativ klassifizierten Segmenten.

Außerdem wurde untersucht, wie hoch die Coverage unter einem bestimmten Fehler E_{HR} auf dem gesamten Signal ist, wenn nur die als informativ klassifizierten Segmente verwendet werden. Hier wird besonders sichtbar, wie niedrig die erreichte Coverage ist. So werden mit $q_{th} = 0{,}4$ und $c_{th} = 75$ nur 8,29 % Coverage für $E_{\mathrm{HR}} < 5$ erreicht werden, obwohl das ganze Signal 32,61 % enthält. Mit $q_{th} = 0{,}3$ und $c_{th} = 75$ sind es immerhin 18,32 %, aber auch das liegt deutlich unter dem tatsächlichen Wert. Die Verteilung ist in Tabelle 4.2 gezeigt.

	insgesamt	$q_{th} = 0{,}3, c_{th} = 75$	$q_{th} = 0{,}4, c_{th} = 75$
$E_{\mathrm{HR}} < 5$ FE	32,61 %	18,32 %	8,29 %
$E_{\mathrm{HR}} < 10$ FE	43,21 %	21,34 %	8,89 %
$E_{\mathrm{HR}} < 15$ FE	51,64 %	23,35 %	9,22 %
$E_{\mathrm{HR}} < 20$ FE	59,38 %	25,12 %	9,47 %

Tabelle 4.2: Coverage unter bestimmten Fehlern E_{HR} vor und nach Klassifikation.

Alles in allem zeigt sich, dass trotz einer sehr niedrigen erreichten Coverage der Fehler in der Klassifikation verhältnismäßig hoch ist. Wenn q_{th} und c_{th} nicht zu niedrig gewählt werden, kann ein Teil des nicht informativen Signals markiert und somit der Fehler in weiterer Verarbeitung insgesamt reduziert werden.

4.2.2 Schwellwertbasierte Artefakterkennung

Für das zweite getestete Verfahren werden lediglich zwei Schwellwerte T_1 und T_2 benötigt, die auf Standardabweichung, Minimum, Maximum und Durchschnitt des Signals beruhen. Die Klassifikation ist ein einfacher Vergleich. Pino et al. verwenden sehr

kleine Fenster, die hier aufgrund fehlender Robustheit der Annotation nicht verwendet werden können.[3] Stattdessen wird der Algorithmus sowohl auf 10 Sekunden- als auch 4 Sekunden-Segmenten getestet, um schlechte Ergebnisse bedingt durch eine bestimmte Segmentlänge auszuschließen.

Die Tests mit beiden Segmentlängen zeigen, dass dieses Verfahren für die in dieser Arbeit untersuchten Daten nicht nutzbar ist, da bei beiden über 95 % der Daten als informativ klassifiziert werden, darunter auch Segmente, bei denen E_{HR} maximal ist. Weiterführende Evaluation bietet hier keine weiteren Erkenntnisse. Dieses Verfahren ist damit nicht weiter nutzbar.

4.2.3 Maschinelles Lernen mittels statistischer Merkmale

Für das maschinelle Lernen mittels statistischer Merkmale werden die in Kapitel 3.3.2 aufgezählten Merkmale extrahiert. Als Bibliothek für die Modelle des maschinellen Lernens wird *scikit-learn*[4] verwendet. Da die Daten sich grundlegend von den von Sadek, Biswas, Yongwei et al. untersuchten unterscheiden, werden die Hyperparameter der Modelle im Zuge dieser Arbeit erneut optimiert. Bei der dafür durchgeführten Kreuzvalidierung wird wie schon bei der Unterteilung in Trainings- und Testset anhand der Patient*innen-ID geteilt, damit auch diese aussagekräftige Ergebnisse liefert.

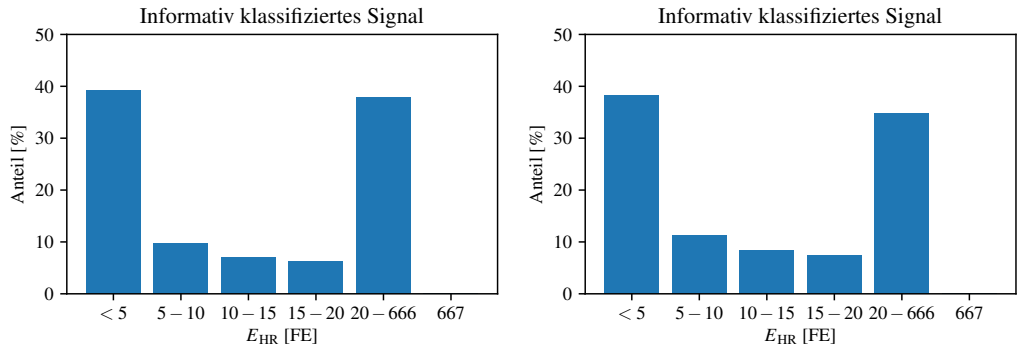
Insgesamt zeigen bereits Accuracy und MAE, siehe Tabelle 4.3, eindeutig, dass die Klassifikation von keinem der Modelle den Ergebnissen des Papers von Sadek, Biswas, Yongwei et al.[5] entspricht. Die Accuracy und AUC von je unter 0,6 zeigen, dass die Klassifikation nur minimal besser als reines Raten ist. Teils ist der MAE der als informativ klassifizierten Segmente sogar größer als der auf dem gesamten Signal. In der folgenden tiefergehenden Evaluation werden lediglich RF und MLP betrachtet. Zwar sind die Ergebnisse des Decision Tree (DT) auf einem ähnlichen Niveau, aber da ein RF lediglich ein Zusammenschluss mehrerer DT ist, ist eine zusätzliche Analyse nicht lohnenswert.

[3] Pino et al. 2015.
[4] Pedregosa et al. 2011.
[5] Sadek, Biswas, Yongwei et al. 2016.

	MAE [FE]	Coverage [%]	Accuracy (Testset)	Accuracy (Kreuzvalidierung Trainingsset)	AUC
insgesamt	21,84	-	-	-	-
annotiert	3,28	43,22 %	-	-	-
LDA	21,84	84,81 %	0,49	0,51	0,53
SVM	22,06	88,14 %	0,47	0,55	0,53
DT	19,47	70,39 %	0,55	0,55	0,59
RF	20,14	62,02 %	0,55	0,58	0,59
MLP	18,41	42,04 %	0,56	0,54	0,58

Tabelle 4.3: Fehler und Coverage für die verschiedenen Modelle im Vergleich zum gesamten Signal und der Annotation.

Bei einer ähnlich hohen, beziehungsweise höheren Coverage als die der Annotation ist der deutlich höhere MAE der Klassifikation in falsch-positiven Segmenten begründet. Der Anteil von Segmenten mit einem Fehler E_{HR} größer als 20 FE ist bei beiden Modellen mit 35 bis 40 % ähnlich hoch. Zusätzlich sind 0,002 % der vom RF als informativ klassifizierten Segmente solche, bei den E_{HR} maximal ist. Das bedeutet, dass die falsch-positiv klassifizierten Segmente bei beiden Klassifikatoren keine Randfälle sind. Der MAE der falsch-negativen Klassifikationen beträgt beim RF 4,07 FE und beim MLP 3,39 FE, ist also leicht höher als der MAE aller informativen Segmente, aber nicht relevant. Auch bei diesen handelt es sich demnach nicht um Randfälle.



(a) RF

(b) MLP

Abbildung 4.3: Verteilung von E_{HR} bei den als informativ klassifizierten Segmenten.

Auch für diese Modelle wird untersucht, wie hoch die Coverage unter einem bestimmten Fehler E_{HR} bei der Betrachtung des als informativ klassifizierten Signals auf dem ganzen Signal ist. Auffallend ist, dass die Werte beim MLP knapp halb so hoch wie die tatsächliche Coverage sind. Das zeigt ergänzend zu der gemessen Accuracy, dass ca. die Hälfte der Segmente, unabhängig von ihrer tatsächlichen Qualität, als informativ klassifiziert werden. Der RF ist näher an der tatsächlichen Coverage, aber es wird auch mehr

nicht-informatives Signal falsch-positiv klassifiziert. Die Werte sind in Tabelle 4.4 aufgelistet.

	insgesamt	RF	MLP
$E_{\mathrm{HR}} < 5\,\mathrm{FE}$	32,61 %	24,34 %	16,07 %
$E_{\mathrm{HR}} < 10\,\mathrm{FE}$	43,22 %	30,35 %	20,82 %
$E_{\mathrm{HR}} < 15\,\mathrm{FE}$	51,66 %	34,66 %	24,34 %
$E_{\mathrm{HR}} < 20\,\mathrm{FE}$	59,40 %	38,52 %	27,45 %

Tabelle 4.4: Coverage unter bestimmten Fehlern E_{HR} vor und nach Klassifikation.

Insgesamt zeigt sich, dass die Klassifikation mittels statistischer Merkmale bei vorliegenden Daten und Annotation nicht sehr zuverlässig ist. Zwar wurden auf anderen Daten sehr gute Ergebnisse erreicht, aber es bestätigt sich die Vermutung, dass sich dies, vermutlich aufgrund der unterschiedlichen Aufnahmesituation der Daten und der nicht aussagekräftigen Validierung der Ergebnisse für die im Paper untersuchten Daten, nicht übertragen lässt.

4.3 Analyse der Merkmale

Obwohl keines der Verfahren sehr gute Ergebnisse für die vorliegenden Daten liefert, ist besonders bei den Modellen maschinellen Lernens interessant, welchen Einfluss und welchen Informationsgewinn die jeweiligen Merkmale haben. Aus diesem Grund wird im Folgenden eine explorative Datenanalyse durchgeführt.

Zunächst wird betrachtet, ob die Merkmale untereinander und mit E_{HR} und der Annotation korreliert sind. Da die Merkmale das Segment gemeinschaftlich statistisch beschreiben, ist eine hohe Korrelation untereinander erwartet und zeigt sich auch. Eine Ausnahme bilden Kurtosis, Mittelwert und Schiefe. Die Anzahl der Nulldurchgänge ist ebenfalls nur schwächer mit den restlichen Mermalen korreliert. Außerdem zeigt sich bei der Betrachtung des in Abbildung 4.4 abgebildeten Korrelationsdiagramms, dass die Korrelation zu E_{HR} und der Annotation nicht signifikant ist. Auch bei paarweiser Visualisierung zeigt sich, dass sich anhand keines untersuchten Merkmalspaares informative von nicht informativen Segmenten unterscheiden lassen.

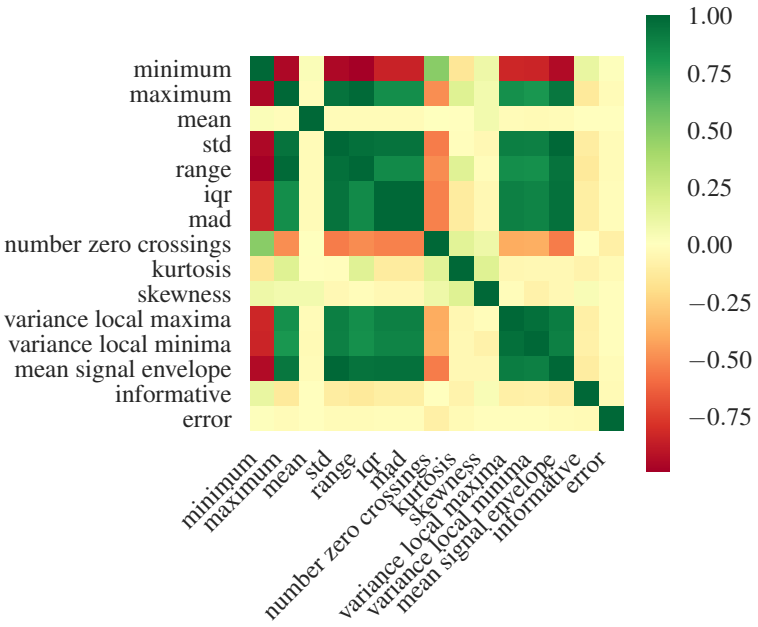


Abbildung 4.4: Korrelationsdiagramm der statistischen Merkmale, E_{HR} und der binären Annotation.

Weiteren Einblick ermöglicht die Analyse des RF, da bei diesem ermittelt werden kann, wie wichtig die einzelnen Merkmale jeweils für die Entscheidungsfindung sind. Das Ergebnis zeigt, dass der Durchschnitt des Signals deutlich weniger Einfluss als die anderen betrachteten Merkmale hat. Die Schiefe (skewness) und die Anzahl der Nulldurchgänge sind, wie in Abbildung 4.5 sichtbar, die beiden wichtigsten Merkmale.

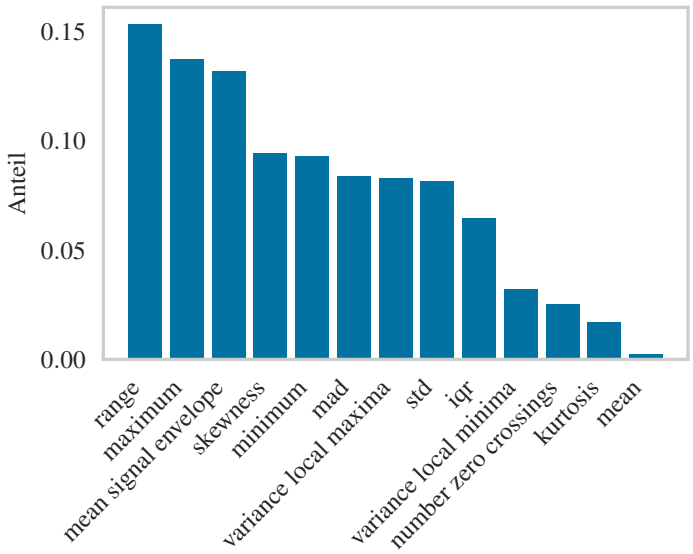


Abbildung 4.5: Wichtigkeit der Merkmale für den RF-Klassifikator mit statistischen Merkmalen.

Da die Visualisierung hochdimensionaler Daten schwierig ist und auch Modelle maschinellen Lernens bei einer hohen Merkmalszahl aufwändiger zu trainieren sind, wurde ebenfalls untersucht, wie sich die Daten bei einer Transformation in einen niedriger dimensionierten Raum verhalten. Untersuchte Transformationen umfassen sowohl unüberwachte Verfahren wie eine PCA mit verschiedenen Kerneln als auch überwachte Verfahren wie eine Dimensionsreduktion mit einer LDA. Auch die transformierten Daten lassen sich nicht voneinander trennen, wie auch in Abbildung 4.6 beispielhaft für eine PCA mit linearem Kernel gezeigt ist.

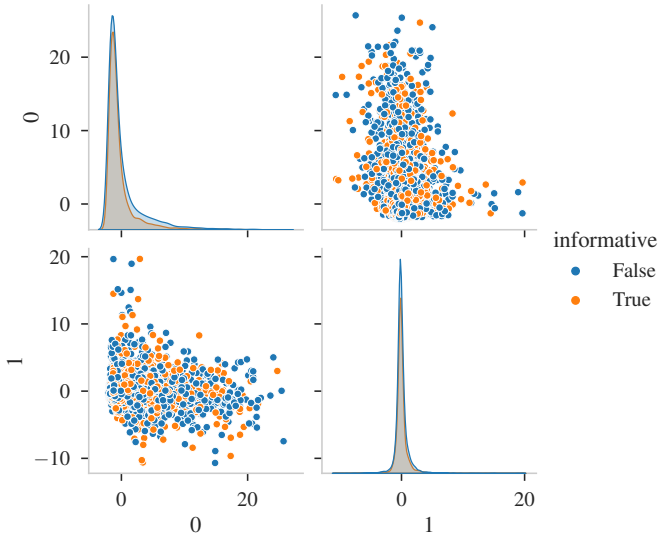


Abbildung 4.6: Dimensionsreduktion der statistischen Merkmale mit einer PCA mit linearem Kernel.

Da eine hohe Korrelation zwischen einzelnen Merkmalen, wie sie auch bei den statistischen Merkmalen vorliegt, dazu führen kann, dass die Interpretation der Wichtigkeit der Merkmale und die Qualität der Modelle generell eingeschränkt ist[6], wird außerdem die Abhängigkeit der Merkmale voneinander untersucht. Hierzu wird das Python-Paket `rfimp` verwendet, das ermittelt, welche Merkmale durch andere vorhergesagt werden können. Nach Reduktion der Merkmalsmenge um diese bleiben lediglich fünf Merkmale: die Standardabweichung, der Durchschnitt, die Anzahl der Nulldurchgänge, Schiefe und Kurtosis. Betrachtet man für diese die Wichtigkeit der einzelnen Merkmale bei einem Random Forest, zeigt sich, dass weiterhin der Durchschnitt am wenigsten Einfluss besitzt, wobei die Standardabweichung jetzt am wichtigsten ist. Die gesamte Verteilung ist in Abbildung 4.7 sichtbar.

[6] Harrison 2019, Kapitel 8.

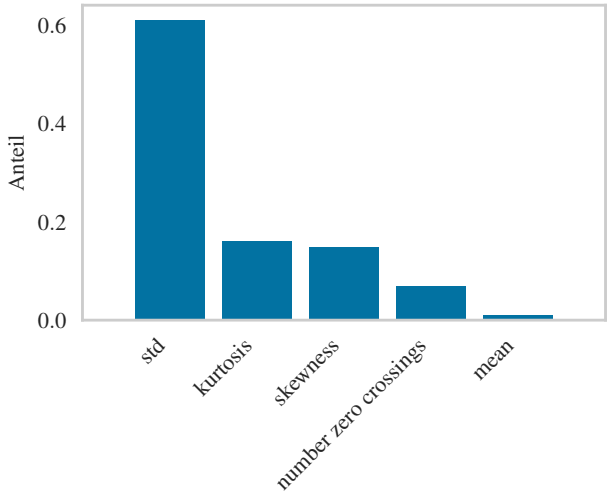


Abbildung 4.7: Wichtigkeit der Merkmale für den RF-Klassifikator mit reduziertem Merkmalsset statistischer Merkmale.

Die Ergebnisse dieses Random Forests mit reduzierter Merkmalsanzahl können mit dem in Kapitel 4.2.3 untersuchten Random Forest verglichen werden. Der MAE und die Coverage sind bei reduzierter Merkmalszahl geringfügig niedriger, die AUC deutlich. Insgesamt sind die Klassifikatoren auf einem ähnlichen Niveau, auch wenn dieser Vergleich bei der geringen Qualität schwierig ist. In Tabelle 4.5 sind die beiden Random Forests gegenübergestellt.

	insgesamt	RF mit 13 Merkmalen	RF mit 5 Merkmalen
$E_{\mathrm{HR}} < 5\,\mathrm{FE}$	32,61 %	24,34 %	23,43 %
$E_{\mathrm{HR}} < 10\,\mathrm{FE}$	43,22 %	30,35 %	28,98 %
$E_{\mathrm{HR}} < 15\,\mathrm{FE}$	51,66 %	34,66 %	32,94 %
$E_{\mathrm{HR}} < 20\ \mathrm{FE}$	59,40 %	38,52 %	36,45 %
MAE	21,85 FE	20,14 FE	19,48 FE
Coverage	43,22 %	62,02 %	57,58 %
Accuracy (Testset)	-	0,55	0,57
Accuracy (Kreuzvalidierung Trainingsset)	-	0,52	0,51
AUC	-	0,59	0,50

Tabelle 4.5: Random Forest mit reduzierter Merkmalszahl im Vergleich zu allen 13 statistischen Merkmalen.

Anhand der statistischen Merkmale kann also keine zuverlässige Klassifikation für den Schwellwert $E_{\mathrm{HR}} = 10\,\mathrm{FE}$ vorgenommen werden. Vermutlich sorgt die große Variation

der in Betten aufgenommenen BKG-Signale dafür, dass sich die Signale nicht statistisch verallgemeinern lassen. Auch die reine Betrachtung der Ähnlichkeit der Intervallschätzer führt zu einem im Verhältnis zur Coverage hohen MAE. Es müssen also weitere Merkmale gefunden werden, die allein oder ergänzend zu den bereits betrachteten eine bessere Aussage zu der Signalqualität ermöglichen.

5 Synthese

Die Untersuchungen in Kapitel 4 haben gezeigt, dass die existierenden untersuchten Verfahren sich nur bedingt eignen, die Qualität von BKG-Signal aus Langzeitaufnahmen von Patient*innen zu beurteilen. Deshalb werden weitere Möglichkeiten zu diesem Zweck untersucht und der Fokus dabei vor allem auf die Konstruktion und Analyse der Eingabemerkmale gelegt. Es wird eine Auswahl von Modellen zum Testen getroffen und ein Basisklassifikator zum Vergleich dieser Modelle entwickelt.

5.1 Merkmalskonstruktion

Grundsätzlich muss zwischen zwei Eingabeformen unterschieden werden: Der bisher betrachteten Eingabe von Merkmalen und der Eingabe des Signals selbst. Letzteres hat den Vorteil, dass keine Informationen verloren gehen können. Allerdings ist das Training so sehr rechen- und damit auch zeitaufwändig und die Merkmale, die zur Beurteilung der Signalqualität genutzt werden, sind nur schwer nachvollziehbar. Aus diesen Gründen wird in dieser Arbeit die Eingabe von Merkmalen untersucht.

Neben der Konstruktion von neuen Merkmalen können ebenfalls die Ergebnisse aus Kapitel 4 verwendet werden, das bedeutet das reduzierte Set statistischer Merkmale und der SQI des CLIE-Algorithmus, bzw. die Coverage durch Intervalle, deren SQI über einem Schwellwert q_{th} liegt. Das Vorgehen bei der Konstruktion neuer Merkmale besteht darin, diese zunächst zu sammeln und anschließend zu untersuchen, Zusammenhänge zu ermitteln und mit den gewonnenen Erkenntnissen die Merkmale zu reduzieren und in Relation zueinander zu setzen.

Da die Untersuchung in Kapitel 4.2.1 gezeigt hat, dass die Ergebnisse stark abhängig von der Auswahl des Schwellwerts q_{th} für den SQI sind, wurde die Coverage für die Schwellwerte $q_{\text{th}} = 0.3$, $q_{\text{th}} = 0.4$ und $q_{\text{th}} = 0.5$ als Merkmal ausgewählt. Da die Verteilung des SQI auf dem Segment womöglich weitere Erkenntnisse ermöglicht, wurden außerdem Merkmale zu der Verteilung der SQI aller ermittelten Intervalle berechnet:

- Minimum SQI_{min}
- Maximum SQI_{max}

- Standardabweichung SQI_{std}
- Mittelwert SQI_{mean}
- Median $\text{SQI}_{\text{median}}$

Auch die geschätzten Intervalllängen können womöglich Aufschluss über die Signalqualität geben. Hier muss allerdings auch beachtet werden, dass dadurch auch die Gefahr von physiologischen Einschränkungen besteht, was Herzrate und Herzratenvariabilität betrifft, wenn die Trainingsdaten nicht variabel genug sind, sodass medizinische Abnormalitäten als Signal schlechter Qualität eingeordnet werden. Es muss also bei einer Verwendung geprüft werden, wie diese Werte einbezogen werden. Die serialisierten Merkmale mit Bezug auf die geschätzten Intervalllängen sind:

- Spannweite IL_{range}
- Standardabweichung IL_{std}

Bei PPG-Signalen verwenden Yu et al. erfolgreich herzratenbezogene Merkmale, indem die maximale Frequenz des Spektogramms der Autokorrelation über das Segment berechnet wird und in Verhältnis zur geschätzten Herzrate gesetzt wird.[1] Sei f_{ACF} die maximale Frequenz und f_{HR} die Frequenz der geschätzten Herzrate, werden zwei Merkmale daraus abgeleitet:

- $\text{ratio}_{\text{ACF}} = \frac{f_{\text{HR}}}{f_{\text{ACF}}}$
- $\text{diff}_{\text{ACF}} = f_{\text{HR}} - f_{\text{ACF}}$

Da auch das Spektogramm der gefilterten Daten des Segments Informationen liefern kann, wurden diese Merkmale analog mit der maximalen Frequenz der Daten f_{data} berechnet:

- $\text{ratio}_{\text{data}} = \frac{f_{\text{HR}}}{f_{\text{data}}}$
- $\text{diff}_{\text{data}} = f_{\text{HR}} - f_{\text{data}}$

Bei weiteren Merkmalen wird versucht, die Eigenschaft der Selbstähnlichkeit zu beschreiben. Dafür werden zunächst die Hochpunkte betrachtet, an denen der CLIE-Algorithmus die Herzschläge verortet. Von diesen werden folgende Merkmale serialisiert:

- Spannweite P_{range}
- Mittelwert P_{mean}
- Standardabweichung P_{std}

[1] Vgl. Yu et al. 2020.

In weiteren Merkmalen wird versucht, die Selbstähnlichkeit durch statistische Merkmale zu erfassen. Hierzu wird für die vom CLIE-Algorithmus erkannten Herzschläge jeweils Mittelwert, Standardabweichung und Spannweite berechnet. Von dieser Menge an Intervallen wird jeweils die Standardabweichung als Merkmal extrahiert, um die Variation über das Segment einzufangen:

- Standardabweichung alle Mittelwerte mean_{std}
- Standardabweichung aller Spannweiten $\text{range}_{\text{std}}$
- Standardabweichung aller Standardabweichungen std_{std}

Häufig werden für die Beurteilung von Signalqualität Templates verwendet. Bei BKG-Signalen werden diese beispielsweise durch Positionsänderungen obsolet, allerdings entspricht ein Segment nur einem sehr kurzen Zeitraum, weshalb Templates in diesem Fall verwendet werden können. Es werden zwei verschiedene Templates betrachtet: Zunächst das geschätzte Schlag-zu-Schlag-Intervall mit dem höchsten SQI, T_{SQI} genannt, und der Herzschlag mit der mittleren Intervalllänge, T_{median}, dessen Länge auch für die Schätzung der Herzrate verwendet wird. Es werden beide betrachtet, da ein sehr hoher SQI auch bei rhythmischen Artefakten auftreten kann. Zu diesen beiden Templates wird für jeden geschätzten Herzschlag die Kreuzkorrelation berechnet. Von dieser Menge an Korrelationen wird jeweils Mittelwert und Standardabweichung als Merkmal verwendet:

- $\text{mean}_{\text{T}_{\text{median}}}$
- $\text{std}_{\text{T}_{\text{median}}}$
- $\text{mean}_{\text{T}_{\text{SQI}}}$
- $\text{std}_{\text{T}_{\text{SQI}}}$

Außerdem wird die absolute Energie des Segmentes berechnet:

- $E_{\text{abs}} = \sum_t s(t)^2$

Da die Berechnung aller Merkmale für die Menge der Daten zeit- und rechenaufwändig ist, werden diese einmalig berechnet und anschließend in einer csv-Datei gespeichert.

5.2 Explorative Datenanalyse und Merkmalsreduktion

Durch die Extraktion mehrerer Merkmale pro betrachteter Eigenschaft entstehen korrelierte Merkmale, bei denen eine Reduktion der Merkmale zu einer verbesserten Performance führen kann. Außerdem werden durch korrelierte Merkmale Analysen der Wichtigkeit der Merkmale verzerrt. Eine Reduktion wird im Rahmen einer explorativen Daten-

analyse durchgeführt. Die Korrelationen sind in dem erzeugten Korrelationsdiagramm, das in Abbildung 5.1 gezeigt ist, deutlich sichtbar.

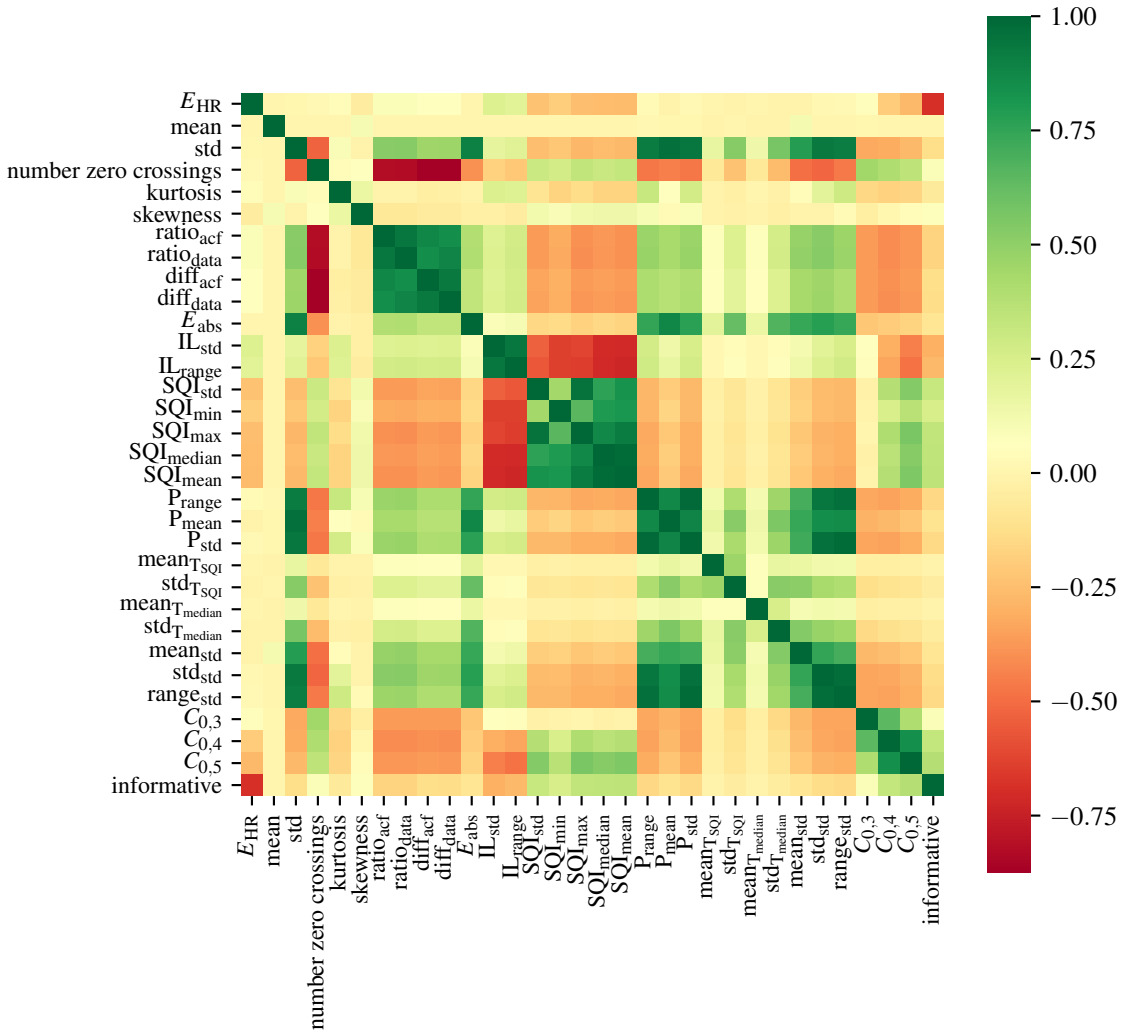


Abbildung 5.1: Korrelationsdiagramm aller entwickelten Merkmale, E_{HR} und der binären Annotation.

Auch bei diesen Merkmalen wird mit dem Python-Paket `rfimp` ermittelt, welche Merkmale durch andere vorhergesagt werden können. Im Zuge der Merkmalskonstruktion werden Merkmale verworfen, die sich durch ein einzelnes anderes Merkmal vorhersagen lassen. Um jeweils das Merkmal zu verwerfen, dass weniger Informationen beiträgt, wird außerdem mit der Bibliothek `sklearn` die Mutual Information zwischen den Merkmalen und der binären Annotation berechnet. Visualisiert ist sie in Abbildung 5.2.

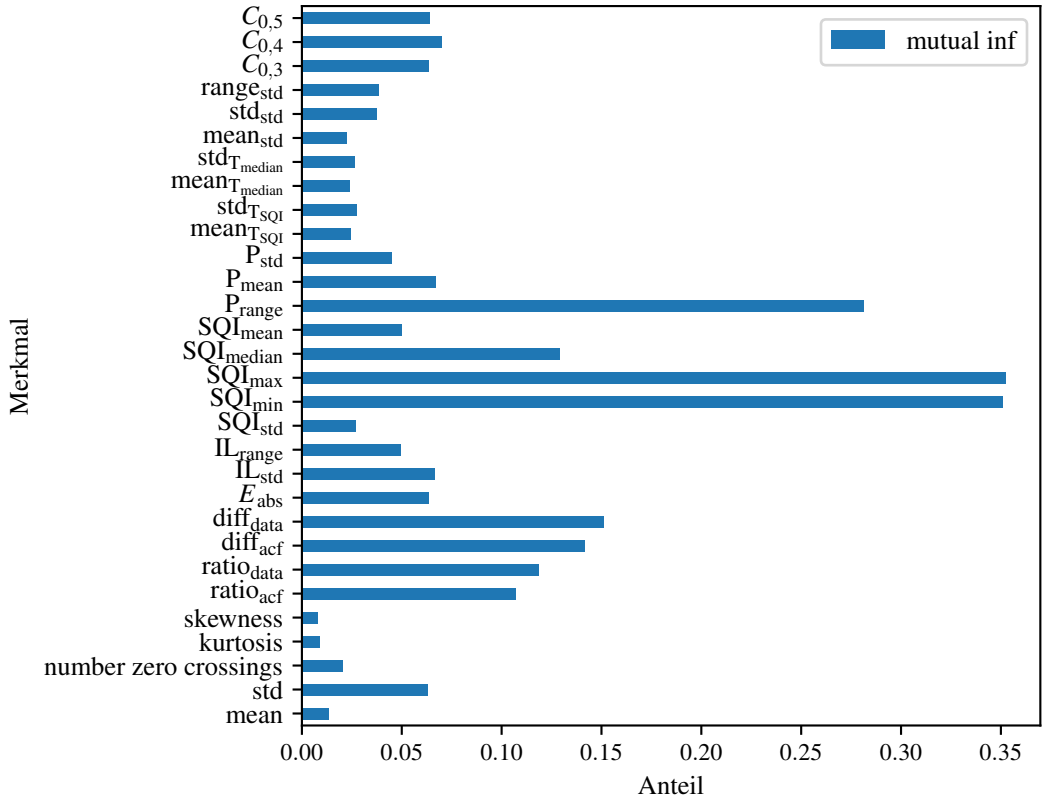


Abbildung 5.2: Visualisierung der Mutual Information zwischen allen Merkmalen und der binären Annotation.

Es zeigt sich, dass Standardabweichung und Spannweite der Werte jeweils die Vorhersage des anderen Wertes ermöglichen. Aus diesem Grund werden P_{std}, std_{std} und IL_{range} als Merkmale verworfen. Das gleiche gilt für SQI_{max}. Auch bietet E_{abs} keinen Mehrgewinn zu der Standardabweichung des Segments. Die Merkmale $\text{ratio}_{\text{acf}}$ und diff_{acf} sind durch ihre Definition stark miteinander korreliert und da die Untersuchung zeigt, dass diff_{acf} schlechter vorhergesagt werden kann und mehr Mutual Information mit der Zielgröße hat, wird $\text{ratio}_{\text{acf}}$ verworfen. Das Gleiche gilt für $\text{ratio}_{\text{data}}$. Wie zu erwarten war, korrelieren $\text{SQI}_{\text{median}}$ und SQI_{mean} stark miteinander. Die mit `rfimp` berechneten Abhängigkeiten zeigen, dass der Median schlechter vorhergesagt werden kann und mehr Information über die Zielgröße enthält, weshalb SQI_{mean} verworfen wird.

Nach Reduktion um diese Merkmale wird die Berechnung der Abhängigkeiten wiederholt. Es zeigen sich weitere Abhängigkeiten jeweils zwischen P_{mean} und der Standardabweichung und $\text{range}_{\text{std}}$ und P_{range}, weshalb Standardabweichung und $\text{range}_{\text{std}}$ verworfen werden. Keine der erkannten Abhängigkeiten ist unerwartet. Dieser Prozess erlaubt jedoch, vermutete Abhängigkeiten zu bestätigen und die Merkmale zu verwenden, die einen höheren Informationsgehalt haben.

Damit verbleiben insgesamt 20 Merkmale. Die Reduktion der Merkmale ermöglicht schnelleres Training und durch die weniger stark korrelierten Merkmale stabilere Modelle. Dennoch muss bei der Evaluation der Modelle untersucht werden, wie sich die Reduktion der Merkmale auswirkt.

Eine andere Möglichkeit der Merkmalsreduktion ist die Transformation in einen neuen Merkmalsraum, beispielsweise durch eine PCA. Die graphische Darstellung einer Reduktion in einen zweidimensionalen Merkmalsraum, siehe Abbildung 5.3, zeigt, dass die Daten so nicht vollständig linear separierbar sind.

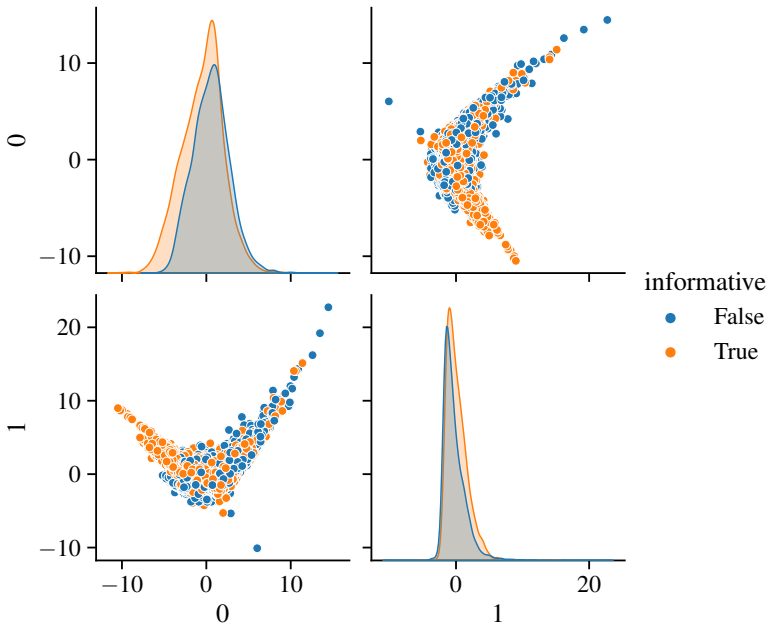


Abbildung 5.3: Transformation der Merkmale mittels PCA in einen zweidimensionalen Merkmalsraum.

Eine Untersuchung der Mutual Information der verbleibenden Merkmale (Abbildung 5.4) zeigt, dass $\mathrm{SQI}_{\mathrm{min}}$ mit Abstand am meisten davon besitzt. Auch die anderen Merkmale, die Informationen über den SQI beinhalten, zählen zu den Merkmalen mit mehr Mutual Information mit der Annotation der Daten. Zu diesen zählen auch die beiden herzratenbezogenen Merkmale $\mathrm{diff}_{\mathrm{data}}$ und $\mathrm{diff}_{\mathrm{acf}}$. Die übernommenen noch verbleibenden vier statistischen Merkmale sind die, die am wenigsten Mutual Information mit dem Ziel teilen.

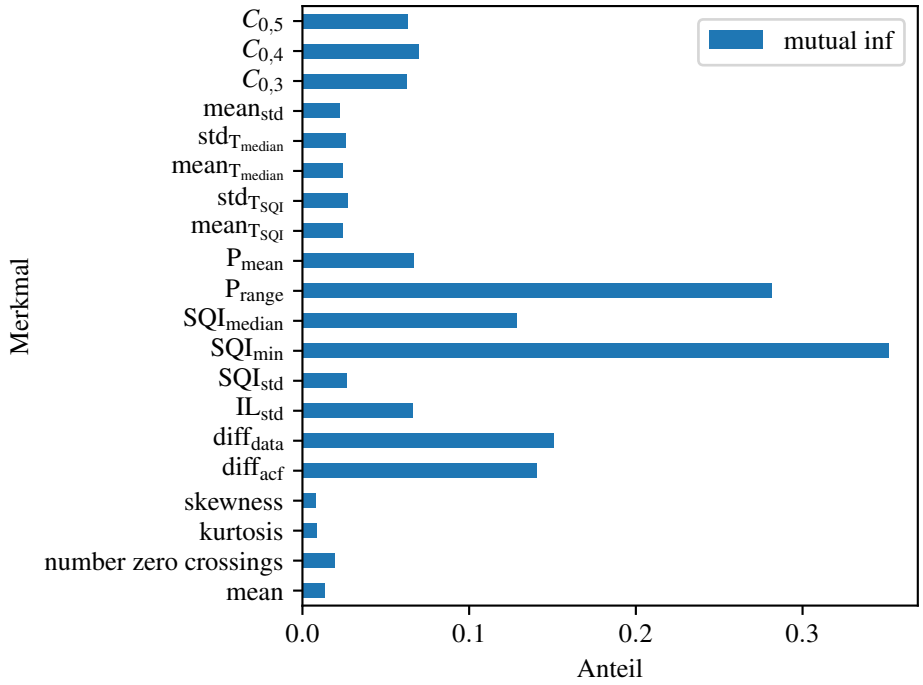


Abbildung 5.4: Mutual Information des reduzierten Merkmalssets mit der binären Annotation.

5.3 Auswahl der Modelle und Aufbau eines Basisklassifikators

Nachdem die Merkmale entwickelt sind, werden nun Modelle ausgewählt, mit denen die Untersuchung durchgeführt wird.

Die Auswahl beschränkt sich auf zwei Modelle, den Random Forest und den Gradient Boosted Tree. Da die durchgeführte PCA vermuten lässt, dass die Daten nicht linear separierbar sind, wurden diese nicht-linearen Modelle gewählt. Beide sind ein Ensemble aus schwächeren CARTs und ermöglichen schnell trainierende, robuste Modelle, deren Entscheidung sich leicht nachvollziehen lässt, und erzielen meist gute Ergebnisse. Dadurch wird eine Untersuchung der Merkmale, anhand derer die Signalqualität beurteilt werden kann, möglich.

Beide Modelle eignen sich sowohl für Regression als auch Klassifikation. Dies wird genutzt, um die Unterschiede der Ergebnisse und der Wichtigkeit der einzelnen Merkmale für beide Arten von Algorithmen zu vergleichen. Bei einer Regression kann der Fehler E_{HR} vorhergesagt werden und mit einem Schwellwert in eine binäre Klassifikation umgewandelt werden. Dies ermöglicht einen einfachen Vergleich von Klassifikation und Re-

gression. Damit ergeben sich insgesamt vier Lernmodelle: Jeweils Regression und Klassifikation mit einem RF und einem Gradient Boosted Tree.

Ein mögliches Problem bei Regressionsmodellen für die Vorhersage von E_{HR} ist, dass kleine Werte relevanter für die spätere Klassifikation sind, das Modell aber den Fehler allgemein minimiert. Das ist besonders problematisch, wenn bei großen Werten von E_{HR} auch der Fehler größer wird und diese den Lernprozess stark beeinflussen, obwohl sie für die Klassifikation nicht relevant sind. Idealerweise wird die Fehlerfunktion in diesem Fall gewichtet. Da das im Rahmen dieser Arbeit nicht möglich ist, wird – um diese Problematik zumindest zu verkleinern – eine Möglichkeit implementiert, die Zielgröße E_{HR} mit der Wurzelfunktion zu skalieren. So wächst die Zielgröße langsamer. Der Wurzelexponent wird dabei zum Hyperparameter des Modells.

Um die Evaluation der verschiedenen Modelle zu vereinfachen, wird eine Wrapper-Klasse entwickelt, die das Laden der Daten, das Hyperparameter-Tuning und die Evaluation der Ergebnisse bündelt. Außerdem werden im Zuge dessen die Modelle nach dem Training serialisiert, um sie für eine spätere Verwendung zu speichern. Dem Konstruktor der Wrapper-Klasse werden Lernmodell, Dateiname zum Laden oder Speichern des Modells und die Hyperparameter, von denen die optimale Auswahl getroffen wird, übergeben. Wenn kein Lernmodell übergeben wird, wird automatisch versucht, die Datei mit dem übergebenen Namen zu laden. Optional kann eine Merkmalsauswahl angegeben werden, die für das Modell verwendet werden soll. Des Weiteren werden die Eigenschaften des Datensets übergeben; das bedeutet die verwendete Segmentlänge, der Abstand, in dem die Segmente erzeugt werden, und der verwendete Threshold. Wurde das Datenset noch nicht erzeugt, wird dies nachgeholt. Sind Hyperparameter angegeben, wird das Hyperparameter-Tuning mit einer Kreuzvalidierung auf dem Trainingsset durchgeführt, wobei jeweils ein*e Patient*in zum Testen ausgelassen wird. Bei dem Hyperparameter-Tuning für die statistischen Merkmale wurde die Accuracy optimiert. Da diese bei nicht balancierten Datensets an Aussagekraft verliert und nicht zwischen Falsch-Positiven und Falsch-Negativen unterscheidet, wird hier die AUC optimiert. Diese bietet außerdem den Vorteil, dass untersucht wird, wie gut sich die Klassen durch das Modell voneinander trennen lassen - unabhängig von der Qualität der Klassifikation selbst.

Da bei der Erzeugung der Merkmale Lücken entstehen können, wenn für das Segment keine Intervallschätzungen existieren, werden diese Segmente unabhängig von dem verwendeten Lernmodell als nicht informativ klassifiziert. Auch beim Training werden Datenpunkte, deren Merkmale lückenhaft sind, ausgeschlossen. Um diese Besonderheit in der Evaluation berücksichtigen zu können, wird ein eigener Klassifikator `OwnClassifier` entwickelt, der von dem Basisklassifikator der Bibliothek `sklearn` erbt. Diesem wird ein anderes Modell übergeben, dass er um oben genannte Eigenschaften erweitert. Für die

Berechnung der AUC ist es nötig, für jeden getesteten Datenpunkt die Wahrscheinlichkeiten der Klassenzugehörigkeit berechnen zu können. Für lückenhafte Datenpunkte wird zurückgegeben, dass die Wahrscheinlichkeit, dass diese nicht-informativ sind, 1 ist. Für alle anderen wird die Wahrscheinlichkeit des darin eingebundenen Modells zurückgegeben.

Bei den Regressionsmodellen muss zusätzlich die Umwandlung in eine binäre Klassifikation durchgeführt werden. Auch hierfür wird eine Unterklasse des Basisklassifikators von `sklearn` erzeugt. Da dieser nur im Zusammenhang mit dem oben beschriebenen Klassifikator genutzt wird, ist eine Filterung der lückenhaften Datenpunkte nicht nötig. Auch hier wird das eingebundene Modell zunächst unabhängig trainiert, sodass es E_{HR} vorhersagen kann. Der vorhergesagte Fehler wird anschließend mit dem gewählten Schwellwert verglichen und das Segment anhand dieses Vergleiches klassifiziert. Da das Modell eine Regression durchführt, gibt es keine Wahrscheinlichkeiten zu den Klassenzugehörigkeiten der Vorhersagen. Hier muss also eine Funktion gefunden werden, die eine Berechnung dieser möglich macht. Die Schwierigkeit liegt darin, dass die Verteilung nicht symmetrisch ist, da der Schwellwert der Klassifikation nicht mittig in den möglichen Werten für E_{HR} liegt. Voraussetzung für eine mögliche Funktion f ist, dass die Wahrscheinlichkeit beider Klassen 0,5 ist, wenn das vorausgesagte E_{HR} gleich dem genutzten Schwellwert E_{th} ist. Außerdem muss die Wahrscheinlichkeit, dass das Segment informativ ist, 1 sein, wenn der vorhergesagte Fehler 0 ist:

$$f(0) = 1$$
$$f(E_{\mathrm{th}}) = 0{,}5.$$

Gleichzeitig muss sie der erwarteten Verteilung entsprechen, also dass die Wahrscheinlichkeit, dass das Segment informativ ist, mit steigendem vorausgesagten E_{HR} stark sinkt und gegen 0 geht. Aufgrund dieser Eigenschaften wurde eine Exponentialfunktion gewählt, deren Parameter abhängig von dem gewählten Schwellwert E_{th} leicht berechnet werden können. Aus den zuvor definierten Voraussetzungen ergeben sich die Parameter der Funktion:

$$f(E_{\mathrm{HR}}) = \mathrm{e}^{\frac{\ln(0{,}5)}{E_{\mathrm{th}}}}$$

Die Wahrscheinlichkeit, dass das Segment nicht informativ ist, ergibt sich durch $1 - f(E_{\mathrm{HR}})$.

Die entwickelten Modelle ermöglichen also eine Voraussage trotz lückenhafter Daten und

eine vollständige Umwandlung einer Regression in eine Klassifikation unter Beachtung der Schnittstellen von `sklearn`. Die entwickelte Wrapper-Klasse ermöglicht eine Bündelung von Hyperparameter-Tuning, Training, Vorhersage und Evaluation. Damit können auch andere Modelle und andere Daten einfach untersucht werden.

6 Evaluation der Ergebnisse

Im Folgenden werden die entwickelten Modelle untersucht und evaluiert. Für die beiden Random Forest-Modelle wird die Implementierung von `sklearn` genutzt, für die Gradient Boosted Trees die Bibliothek `XGBoost`, da das dort verwendete eXtreme Gradient Boosting (XGB) besser ist als die Implementierung in `sklearn`.[1] Die Ergebnisse der Modelle werden sowohl für das reduzierte als auch das vollständige Merkmalsset berechnet, verglichen und aufbauend darauf die Merkmalsauswahl optimiert. Außerdem wird der Einfluss der gewählten Segmentlänge und des gewählten Schwellwertes der Annotation untersucht. Für alle Untersuchungen wird das gleiche Testset wie in Kapitel 4 verwendet.

6.1 Evaluation der Modelle

Zunächst werden alle Modelle sowohl mit dem vollständigen Merkmalsset als auch dem reduzierten Merkmalsset verglichen; die Ergebnisse sind in Tabelle 6.1 gezeigt. Insgesamt zeigt sich bereits, dass eine deutlich höhere Coverage als bei der reinen Betrachtung der Intervallschätzer des CLIE-Algorithmus erreicht wird. Diese lag beispielsweise für $q_{\text{th}} = 0.3$ bei 20,93 % mit einem MAE von 13,90 FE. Die Ergebnisse der Klassifikation mit Gradient Boosted Trees und Random Forests sind zunächst ähnlich. Eine Ausnahme bildet das Regressionsmodell mit Gradient Boosted Trees: Es erreicht eine sehr hohe Coverage von über 80 % mit einem zu den anderen Modellen verhältnismäßig hohen MAE von über 16 FE.

Des Weiteren zeigt sich, dass das reduzierte Merkmalsset entgegen der Erwartung zu etwas schlechteren Ergebnissen führt. Eine Betrachtung der Wichtigkeit der Merkmale für die Modelle mit vollständigem Merkmalsset zeigt, dass hier jeweils $\text{ratio}_{\text{acf}}$ und $\text{ratio}_{\text{acf}}$ zu den wichtigsten Merkmalen gehören, welche beide kein Teil des reduzierten Merkmalssets sind. In Abbildung 6.1 ist die Verteilung der Relevanz der Merkmale exemplarisch für den RF-Klassifikator mit vollständigem Merkmalsset gezeigt. Ein Test mit einem RF-Klassifikator mit dem reduzierten Merkmalsset zuzüglich der oben genannten Merkmalen

[1] Harrison 2019, Kapitel 10.

	Merkmalsset	Modell	MAE [FE]	Coverage [%]	F1-Score	AUC
annotiert			3,28	43,21	-	-
Klassifikation	reduziert	RF	13,87	32,09	0,54	0,69
		XGB	13,14	29,79	0,53	0,69
	alle	RF	12,11	36,81	0,61	0,75
		XGB	11,38	35,93	0,62	0,75
Regression	reduziert	RF	14,34	41,21	0,57	0,69
		XGB	17,79	80,77	0,63	0,69
	alle	RF	12,56	46,27	0,64	0,75
		XGB	16,04	81,80	0,65	0,74

Tabelle 6.1: Vergleich aller Modelle mit reduziertem und vollständigem eigenem Merkmalsset.

zeigt, dass die Performance vergleichbar mit dem vollständigen Merkmalsset ist und ein MAE von 11,97 FE bei einer Coverage von 36,51 % erreicht wird.

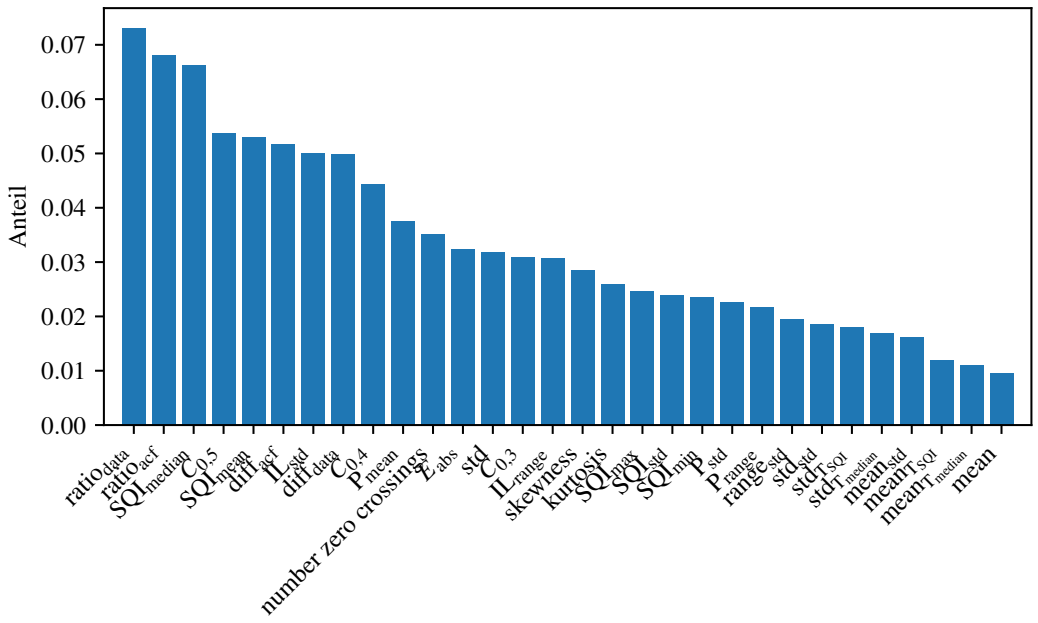


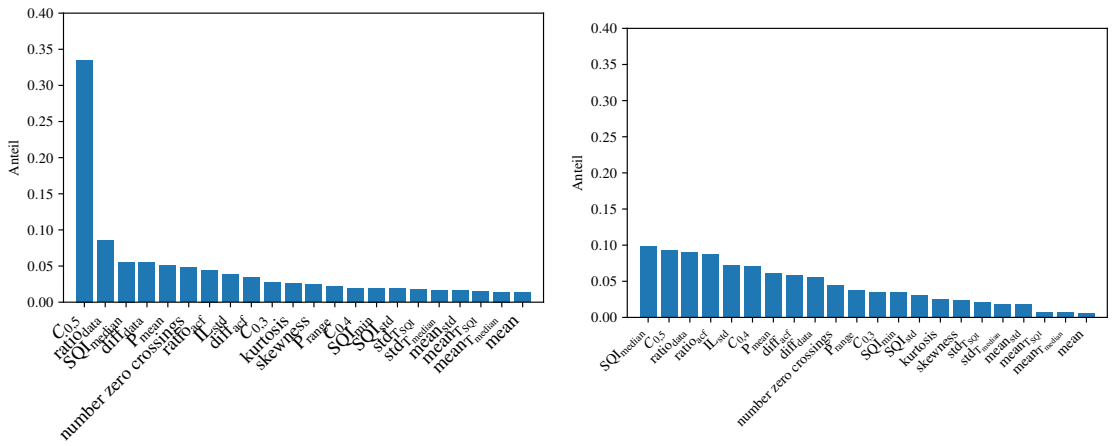
Abbildung 6.1: Wichtigkeit der Merkmale des RF-Klassifikators mit vollständigem Merkmalsset.

Bei den folgenden Untersuchungen wird aus diesem Grund das erweiterte reduzierte Merkmalsset verwendet. Die Ergebnisse der vier Modelle sind in Tabelle 6.2 abgebildet. Alle erzielen ähnliche Ergebnisse, wobei die Regressionsmodelle eine etwas höhere Coverage erreichen. Es gibt im Gegensatz zu den vorherigen Ergebnissen keine Ausreißer. Die AUC zeigt, dass alle Modelle bei den gegebenen Daten in der Lage sind, informatives und nicht informatives Signal zu separieren.

	Modell	MAE [FE]	Coverage [%]	F1-Score	AUC
annotiert		3,28	43,21	-	-
Klassifikation	RF	11,86	34,90	0,60	0,75
	XGB	12,44	41,99	0,63	0,75
Regression	RF	12,57	46,36	0,64	0,75
	XGB	12,66	47,59	0,64	0,74

Tabelle 6.2: Vergleich aller Modelle mit finalem Merkmalsset.

Betrachtet man die Wichtigkeit der Merkmale der beiden Klassifikationsmodelle, fällt auf, dass bei dem XGB-Klassifikator ein Merkmal allein deutlich wichtiger als alle anderen ist; sowohl beim reduzierten als auch beim vollständigen Merkmalsset. Bei den RF-Modellen dagegen ist die Wichtigkeit gleichmäßiger verteilt. Der direkte Vergleich ist in Abbildung 6.2 zu sehen. Trotz der unterschiedlichen Gewichtung der Merkmale erzielen beide Modelle ähnliche Ergebnisse. Allerdings ist der XGB-Klassifikator weniger stabil, falls das mit Abstand wichtigste Merkmal $C_{0,5}$ gestört wird. Es zeigt aber auch, dass die Ähnlichkeit der Intervallschätzer des CLIE-Algorithmus ein gutes Kriterium zur Beurteilung der Signalqualität ist.



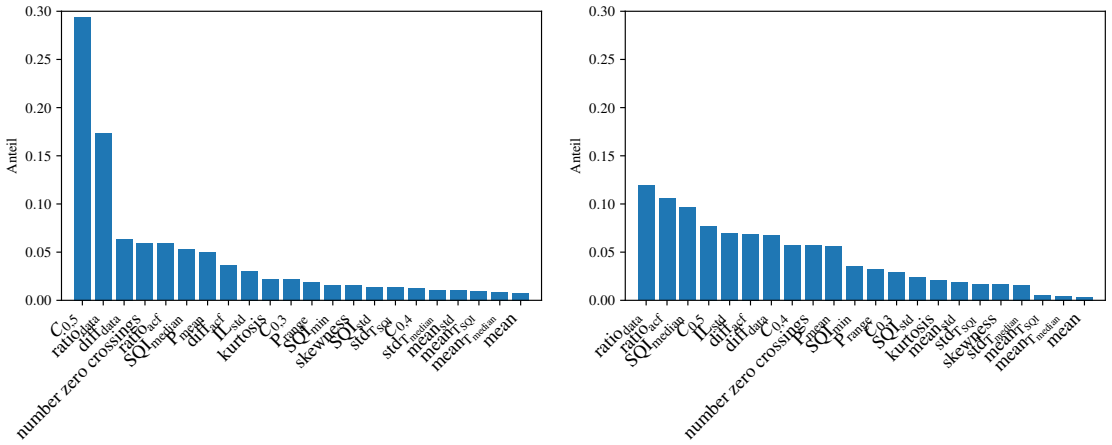
(a) XGB-Klassifikator

(b) RF-Klassifikator

Abbildung 6.2: Vergleich der Wichtigkeit der Merkmale zwischen RF-Klassifikator und XGB-Klassifikator.

Bei der Betrachtung der Regressionsmodelle zeigt sich, dass es bei dem XGB-Regressor zwei Merkmale gibt, die bedeutend wichtiger als der Rest sind: Erneut $C_{0,5}$ und $\text{ratio}_{\text{data}}$. Das RF-Modell zeigt auch hier eine gleichmäßigere Verteilung der Wichtigkeit der Merkmale. Die Wichtigkeit der übrigen Merkmale ist, wie in Abbildung 6.3 zu sehen, ähnlich verteilt.

Die RF-Modelle gewichten die einzelnen Merkmale bei den vorliegenden Daten ähnlicher als die XGB-Modelle. Dennoch sind die Ergebnisse der ersten Evaluation ähnlich. Auch



(a) XGB-Regressor

(b) RF-Regressor

Abbildung 6.3: Vergleich der Wichtigkeit der Merkmale zwischen RF-Regressor und XGB-Regressor.

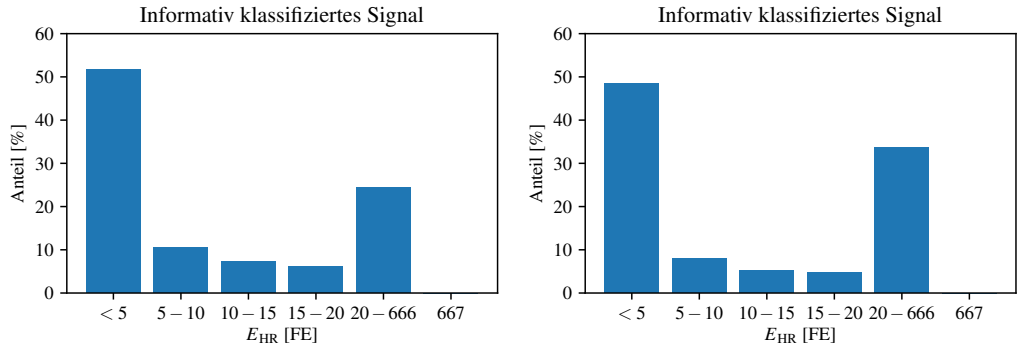
ist die Wichtigkeit der Merkmale zwischen den Klassifikations- und Regressionsmodellen ähnlich. Insgesamt zeigen die Regressionsmodelle aber eine überlegene Performance. Eine genauere Evaluation der Coverage und der Verteilung von E_{HR} wird beispielhaft für den RF-Regressor durchgeführt. Zum Vergleich werden die Ergebnisse der Klassifizierung anhand der Intervallschätzer des CLIE-Algorithmus für $q_{\mathrm{th}} = 0{,}3$ und $c_{\mathrm{th}} = 75$ gezeigt, da mit diesen Schwellwerten ebenfalls eine Reduzierung des MAE mit einer vergleichbaren Coverage von 37,87 % erreicht wurde. Der Vergleich der erreichten Coverage unter einem bestimmten E_{HR}, sichtbar in Tabelle 6.3, zeigt, dass die Coverage für geringe Fehler deutlich erhöht werden konnte.

	insgesamt	RF-Regressor	CLIE-Intervallschätzer
$E_{\mathrm{HR}} < 5\,\mathrm{FE}$	32,61 %	23,93 %	18,32 %
$E_{\mathrm{HR}} < 10\,\mathrm{FE}$	43,21 %	28,80 %	21,34 %
$E_{\mathrm{HR}} < 15\,\mathrm{FE}$	51,64 %	32,17 %	23,35 %
$E_{\mathrm{HR}} < 20\,\mathrm{FE}$	59,38 %	35,03 %	25,12 %

Tabelle 6.3: Coverage unter bestimmten Fehlern E_{HR} nach Klassifikation mittels RF-Regressor.

Auch eine Untersuchung der Verteilung von E_{HR} auf den als informativ klassifizierten Segmenten zeigt im Vergleich, dass der Anteil des Signals mit einem Fehler $E_{\mathrm{HR}} > 20\,\mathrm{FE}$ stark gesenkt werden konnte. Auch liegt der durchschnittliche MAE der falsch-negativen Segmente bei 4,36 FE, was über dem Durchschnitt aller als informativ klassifizierten Segmente von 3,28 FE liegt.

Mit den im Rahmen dieser Arbeit entwickelten Modellen kann also die Signalqualität zuverlässiger als bisher beurteilt werden. Die Coverage durch die Klassifikation konnte



(a) RF-Regressor

(b) Ähnlichkeit der Intervallschätzer

Abbildung 6.4: Verteilung von E_{HR} bei den als informativ klassifizierten Segmenten im Vergleich.

erhöht und der MAE der als informativ klassifizierten Segmente gesenkt werden.

6.2 Einfluss des Schwellwertes der Annotation

Nachdem die generelle Performance der Modelle untersucht wurde, wird nun der Einfluss des verwendeten Schwellwertes E_{th} der Annotation untersucht. Dafür werden die vier Modelle jeweils für vier Schwellwerte E_{th} trainiert: $E_{th} \in \{5, 10, 15, 20\}$.

Die Ergebnisse der Klassifikationsmodelle sind in Tabelle 6.4 gezeigt. Für diese Modelle ergibt sich für niedrigere Schwellwerte E_{th} jeweils eine niedrigere Coverage und ein niedrigerer MAE. Die Variation des MAE beträgt zwischen $E_{th} = 5\,\text{FE}$ und $E_{th} = 20\,\text{FE}$ allerdings nur gut 2 FE beim RF-Klassifikator bzw. 2,5 FE beim XGB-Klassifikator. Die Coverage kann dabei um ca. 15 Prozentpunkte für das RF-Modell und um ca. 18 Prozentpunkte für das XGB-Modell erhöht werden. Wie wichtig die Genauigkeit der Herzratenschätzung im Vergleich zur Coverage ist, hängt vom Anwendungsfall ab, allerdings ist der Gewinn durch die deutlich erhöhte Coverage vermutlich in den meisten Fällen größer. Für alle Schwellwerte ist weiterhin eine Trennung der beiden Klassen möglich, wobei jeweils die Modelle mit $E_{th} = 5\,\text{FE}$ die beste Trennung ermöglichen. Es kann also auch sinnvoll sein, lediglich den Schwellwert des Klassifikationsmodells anzupassen. Eine genauere Untersuchung der Auswirkungen durch eine solche Variation ist im Rahmen dieser Arbeit allerdings nicht möglich. In den Untersuchungen zeigt sich ebenfalls, dass der XGB-Klassifikator tendenziell eine höhere Coverage bei gleichzeitig höherem MAE erreicht als das RF-Modell.

	$E_{\text{th}} = 5\,\text{FE}$	$E_{\text{th}} = 10\,\text{FE}$	$E_{\text{th}} = 15\,\text{FE}$	$E_{\text{th}} = 20\,\text{FE}$
Coverage annotiert [%]	32,61	43,21	51,64	59,45
Coverage klassifiziert [%]	29,30	34,90	42,46	53,78
MAE [FE]	11,39	11,86	12,44	13,27
F1-Score	0,59	0,60	0,64	0,70
AUC	0,78	0,75	0,73	0,73

(a) RF-Klassifikator

	$E_{\text{th}} = 5\,\text{FE}$	$E_{\text{th}} = 10\,\text{FE}$	$E_{\text{th}} = 15\,\text{FE}$	$E_{\text{th}} = 20\,\text{FE}$
Coverage annotiert [%]	32,61	43,21	51,64	59,45
Coverage klassifiziert [%]	31,51	41,99	51,15	59,73
MAE [FE]	11,39	12,44	13,23	13,86
F1-Score	0,59	0,63	0,67	0,73
AUC	0,77	0,75	0,73	0,73

(b) XGB-Klassifikator

Tabelle 6.4: Variation des Schwellwerts E_{th} der Annotation bei den Klassifikationsmodellen.

Betrachtet man die Coverage unter bestimmten Fehlern für $E_{\text{th}} = 5\,\text{FE}$ und $E_{\text{th}} = 20\,\text{FE}$ im Vergleich, wird deutlich, dass durch eine Erhöhung von E_{th} in der Annotation auch Segmente mit niedrigem Fehler deutlich zuverlässiger miterfasst werden. Der direkte Vergleich ist in Tabelle 6.5 für den XGB-Klassifikator gezeigt.

	insgesamt	$E_{\text{th}} = 5\,\text{FE}$	$E_{\text{th}} = 20\,\text{FE}$
$E_{\text{HR}} < 5\,\text{FE}$	32,61 %	18,85 %	26,06 %
$E_{\text{HR}} < 10\,\text{FE}$	43,21 %	21,70 %	33,22 %
$E_{\text{HR}} < 15\,\text{FE}$	51,64 %	23,33 %	38,65 %
$E_{\text{HR}} < 20\,\text{FE}$	59,38 %	24,65 %	43,35 %

Tabelle 6.5: Coverage unter bestimmten Fehlern E_{HR} nach Klassifikation mittels XGB-Klassifikator für verschiedene Schwellwerte der Annotation.

Bei den Regressionsmodellen ist der Einfluss der Variation von E_{th} deutlich stärker, wie in Tabelle 6.6 zu sehen ist. Hier variiert der MAE um ca. 5,5 FE für das RF-Modell und um ca. 4,5 FE für den XGB-Regressor. Die Coverage zeigt eine stärkere Variation von knapp 60 % für das RF- bzw. knapp 50 % für das XGB-Modell. Auch sind die erreichten durchschnittlichen Fehler für $E_{\text{th}} = 20\,\text{FE}$ um 1 bis 2 FE höher als bei den Klassifikationsmodellen und für $E_{\text{th}} = 5\,\text{FE}$ jeweils 1 bis 2 FE niedriger. Die AUC dagegen verhält sich ähnlich zu den Klassifikationsmodellen.

	$E_{\text{th}} = 5\,\text{FE}$	$E_{\text{th}} = 10\,\text{FE}$	$E_{\text{th}} = 15\,\text{FE}$	$E_{\text{th}} = 20\,\text{FE}$
Coverage annotiert [%]	32,61	43,21	51,64	59,45
Coverage klassifiziert [%]	20,19	46,36	71,16	79,47
MAE [FE]	9,83	12,57	14,61	15,39
F1-Score	0,52	0,64	0,72	0,78
AUC	0,77	0,75	0,74	0,73

(a) RF-Regressor

	$E_{\text{th}} = 5\,\text{FE}$	$E_{\text{th}} = 10\,\text{FE}$	$E_{\text{th}} = 15\,\text{FE}$	$E_{\text{th}} = 20\,\text{FE}$
Coverage annotiert [%]	32,61	43,21	51,64	59,45
Coverage klassifiziert [%]	26,76	47,59	63,74	74,18
MAE [FE]	10,52	12,66	14,04	14,94
F1-Score	0,55	0,64	0,71	0,77
AUC	0,75	0,74	0,73	0,73

(b) XGB-Regressor

Tabelle 6.6: Variation des Schwellwerts E_{th} der Annotation bei den Regressionsmodellen.

Auch hier wird die Coverage unter bestimmten Fehlern E_{HR} untersucht. Die Ergebnisse der Regressionsmodelle für niedrige E_{th} weisen eine deutlich geringere Coverage auf, aber für $E_{\text{th}} = 20$ wird nahezu alles informative Signal erkannt. Vor allem fällt auf, dass bei $E_{\text{th}} = 5\,\text{FE}$ sehr viel Signal mit nur geringem Fehler nicht erkannt wird. Der Vergleich zwischen $E_{\text{th}} = 5\,\text{FE}$ und $E_{\text{th}} = 20\,\text{FE}$ beim RF-Regressor ist in Tabelle 6.7 gezeigt.

	insgesamt	$E_{\text{th}} = 5\,\text{FE}$	$E_{\text{th}} = 20\,\text{FE}$
$E_{\text{HR}} < 5\,\text{FE}$	32,61 %	13,75 %	31,14 %
$E_{\text{HR}} < 10\,\text{FE}$	43,21 %	15,05 %	40,48 %
$E_{\text{HR}} < 15\,\text{FE}$	51,64 %	15,75 %	47,78 %
$E_{\text{HR}} < 20\,\text{FE}$	59,38 %	16,36 %	54,37 %

Tabelle 6.7: Coverage unter bestimmten Fehlern E_{HR} nach Klassifikation mittels RF-Regressor für verschiedene Schwellwerte der Annotation.

Bei der Betrachtung von Coverage und MAE aller vier Modelle mit den verschiedenen Schwellwerten E_{th} im Vergleich zeigt sich, dass der Tradeoff zwischen Coverage und MAE nahezu linear ist (Abbildung 6.5).

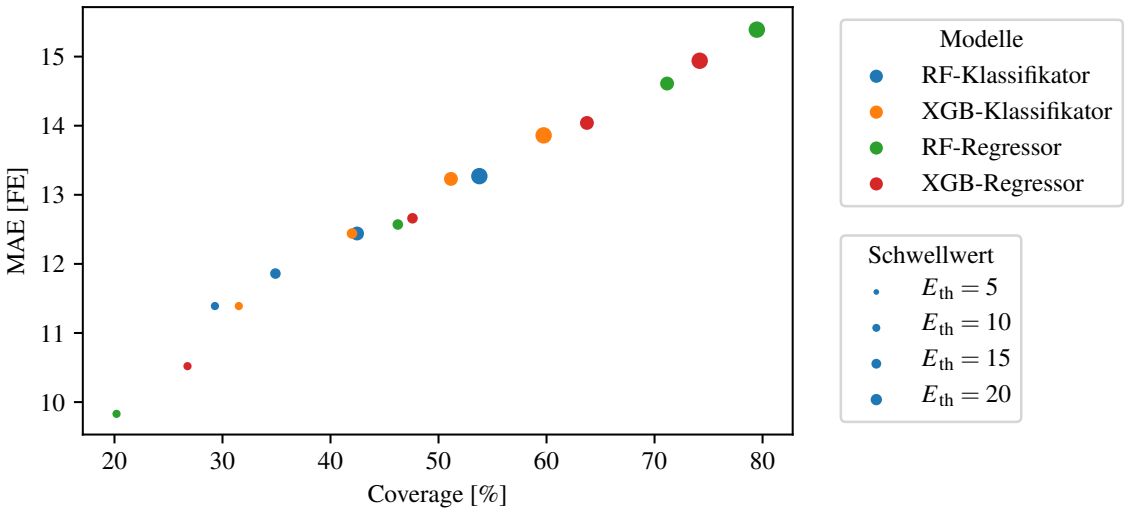


Abbildung 6.5: Graphische Darstellung des Zusammenhangs von MAE und Coverage abhängig von dem gewählten Schwellwert E_{th}.

Die Untersuchungen zeigen, dass die Wahl von E_{th} einen großen Einfluss vor allem auf die erreichte Coverage der Klassifikation hat. Auch hier muss abgewogen werden, ob eine hohe Genauigkeit oder eine hohe Coverage wichtiger ist, allerdings ist der Gewinn der Coverage durch die Wahl eines höheren Schwellwertes E_{th} in den meisten Fällen vermutlich größer.

6.3 Einfluss der Segmentlänge

Neben dem Einfluss des Schwellwertes der Annotation E_{th} wird auch der Einfluss der verwendeten Segmentlänge s untersucht. Für diese Untersuchung wird aufgrund der Ergebnisse der vorherigen Untersuchungen $E_{\text{th}} = 20$ gewählt. Untersuchte Segmentlängen sind 5, 10, 20 und 30 Sekunden.

Bei den Klassifikationsmodellen zeigt sich deutlich, dass mit steigender Segmentlänge die Coverage höher und der MAE niedriger wird, siehe Tabelle 6.8. Es muss zusätzlich beachtet werden, dass der MAE über die Segmente insgesamt mit steigender Segmentlänge ebenfalls sinkt, die Verbesserung also eventuell darin begründet ist. Auch der F1-Score steigt mit der Segmentlänge; die AUC dagegen wird etwas niedriger. Bei dem RF-Modell sind die Veränderungen in MAE und Coverage etwas größer als bei dem XGB-Klassifikator. Alles in allem führen größere Segmente bei den Klassifikationsmodellen zu besseren Ergebnissen bezüglich Coverage und MAE, allerdings lassen sich die beiden Klassen nach AUC etwas schlechter voneinander trennen.

	$s = 5$	$s = 10$	$s = 20$	$s = 30$
Coverage annotiert [%]	39,30	43,21	45,47	45,87
MAE insgesamt [FE]	25,96	21,85	19,99	19,31
Coverage klassifiziert [%]	45,94	53,78	60,93	62,01
MAE klassifiziert [FE]	14,20	13,27	13,03	12,68
F1-Score	0,62	0,65	0,66	0,67
AUC	0,75	0,74	0,73	0,73

(a) RF-Klassifikator

	$s = 5$	$s = 10$	$s = 20$	$s = 30$
Coverage annotiert [%]	39,30	43,21	45,47	45,87
MAE insgesamt [FE]	25,96	21,85	19,99	19,31
Coverage klassifiziert [%]	54,18	59,73	64,81	67,00
MAE klassifiziert [FE]	14,20	13,86	13,23	13,12
F1-Score	0,61	0,65	0,66	0,67
AUC	0,73	0,73	0,72	0,72

(b) XGB-Klassifikator

Tabelle 6.8: Variation der Segmentlänge s bei den Klassifikationsmodellen.

Betrachtet man die Coverage aufgeschlüsselt nach bestimmten Fehlern E_{HR}, zeigt sich, dass der Fehler wie schon beschrieben bei längeren Segmenten niedriger ist, aber auch, dass die Klassifikation bei längeren Segmenten näher an der Annotation ist. Für den RF-Klassifikator ist die Coverage für die Segmentlängen $s = 5$ und $s = 30$ in Tabelle 6.9 aufgeschlüsselt.

	$s = 5$		$s = 30$	
	insgesamt	klassifiziert	insgesamt	klassifiziert
$E_{\mathrm{HR}} < 5\,\mathrm{FE}$	29,88 %	21,72 %	33,33 %	27,45 %
$E_{\mathrm{HR}} < 10\,\mathrm{FE}$	39,29 %	26,36 %	45,86 %	36,16 %
$E_{\mathrm{HR}} < 15\,\mathrm{FE}$	47,38 %	30,07 %	54,97 %	42,23 %
$E_{\mathrm{HR}} < 20\,\mathrm{FE}$	55,17 %	33,48 %	62,85 %	47,11 %

Tabelle 6.9: Coverage unter bestimmten Fehlern E_{HR} nach Klassifikation mittels RF-Klassifikator für verschiedene Segmentlängen s.

Bei den Regressionsmodellen ist, wie in Tabelle 6.10 gezeigt, ähnliches zu beobachten. Hier ist der Effekt auf den MAE etwas stärker als bei den Klassifikationsmodellen; bei beiden Modellen beträgt die Veränderung über 2 FE. Auch hier ist er beim RF-Modell etwas stärker. Die Veränderung der Coverage ist beim RF-Regressor dagegen minimal; zwischen $s = 5$ und $s = 30$ beträgt der Unterschied nur knapp 1,5 Prozentpunkte. Auch beim XGB-Modell ist die Veränderung der Coverage weniger stark als bei den Klassifikationsmodellen, aber zumindest bei ca. 8 Prozentpunkten.

	$s = 5$	$s = 10$	$s = 20$	$s = 30$
Coverage annotiert [%]	39,30	43,21	45,47	45,87
MAE insgesamt [FE]	25,96	21,85	19,99	19,31
Coverage klassifiziert [%]	78,93	79,47	80,15	80,31
MAE klassifiziert [FE]	17,62	15,39	14,35	14,01
F1-Score	0,62	0,66	0,68	0,68
AUC	0,76	0,75	0,74	0,73

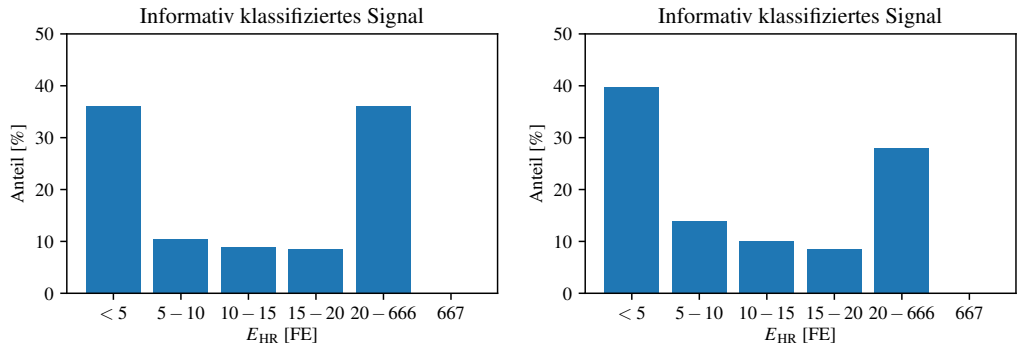
(a) RF-Regressor

	$s = 5$	$s = 10$	$s = 20$	$s = 30$
Coverage annotiert [%]	39,30	43,21	45,47	45,87
MAE insgesamt [FE]	25,96	21,85	19,99	19,31
Coverage klassifiziert [%]	69,40	74,18	76,69	77,14
MAE klassifiziert [FE]	16,34	14,94	14,17	13,96
F1-Score	0,63	0,66	0,68	0,68
AUC	0,74	0,74	0,72	0,71

(b) XGB-Regressor

Tabelle 6.10: Variation der Segmentlänge s bei den Regressionsmodellen.

Der direkte Vergleich der Verteilung von E_{HR} auf den als informativ klassifizierten Segmenten zeigt deutlich, dass bei $s = 30$ ein deutlich größerer Teil des als informativ klassifizierten Signals einen kleinen Fehler E_{HR} aufweist. Bei $s = 5$ dagegen ist der Anteil des Signals mit $E_{\text{HR}} > 20$ ähnlich groß wie der mit $E_{\text{HR}} < 5$.



(a) $s = 5$ (b) $s = 30$

Abbildung 6.6: Verteilung von E_{HR} bei den vom RF-Regressor als informativ klassifizierten Segmenten mit verschiedenen Segmentlängen s.

Betrachtet man alle Modelle im Vergleich, indem man den Zusammenhang von MAE und Coverage graphisch darstellt, siehe Abbildung 6.7, zeigt sich, dass die Unterschiede der Modelle mit variierender Segmentlänge sichtbar sind. Das Modell mit geringstem MAE und geringster Coverage ist der RF-Klassifikator, das mit höchstem MAE und Coverage

ist der RF-Regressor. Auch zeigt sich erneut, dass jeweils Regressions- und Klassifikationsmodelle eine hohe Ähnlichkeit der Ergebnisse zeigen. Auch wird sehr deutlich, dass bei gleichem Fehler die Regressionsmodelle eine höhere Coverage erreichen. So variiert bei einem MAE von knapp über 14 FE die Coverage zwischen ca. 45 % und ca. 80 %. Das Modell mit der höchsten Coverage ist dabei der RF-Regressor.

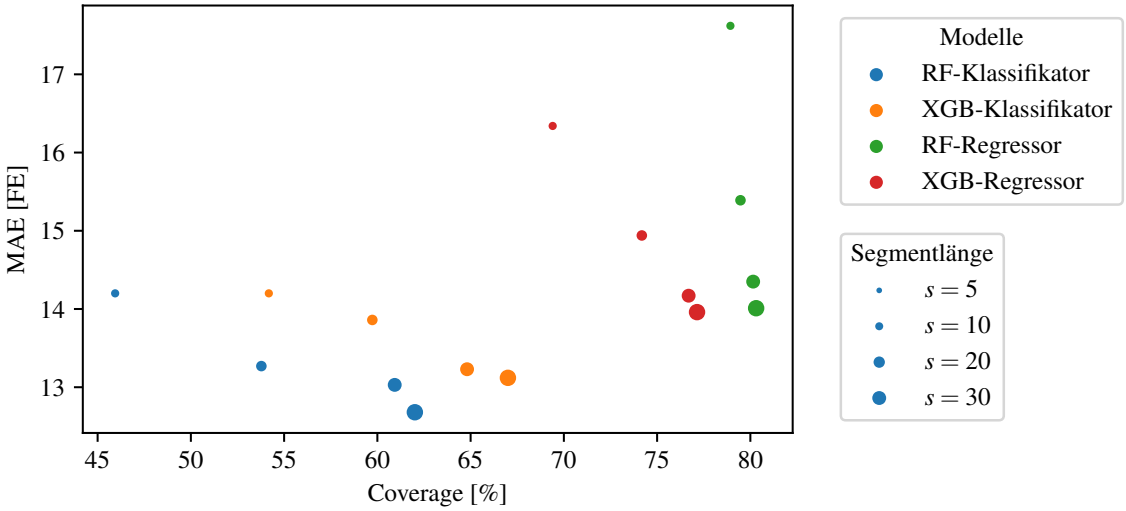


Abbildung 6.7: Graphische Darstellung des Zusammenhangs von MAE und Coverage abhängig von der Segmentlänge s.

Bei der Bewertung muss jedoch auch beachtet werden, dass die Herzrate über die Berechnung des Medians der geschätzten Intervalllängen bestimmt wird, also dementsprechend mit steigender Segmentlänge auch robuster wird. Auch hier gilt aus diesem Grund, dass die Wahl der Segmentlänge von dem Anwendungsfall abhängt, tendenziell aber längere Segmente zu besseren Ergebnissen führen.

6.4 Test auf Daten von gesunden Personen

Neben den bisher betrachteten Daten der Patient*innen liegen außerdem die im Schlaf aufgenommenen Daten von 8 gesunden Personen vor. Da die Beurteilung der Signalqualität bei schlafenden, gesunden Personen bedeutend einfacher ist, wurden diese Daten bis jetzt nicht betrachtet. Abschließend werden diese genutzt, um die Performance der Modelle für Daten mit anderen Aufnahmebedingungen zu testen. Getestet werden 10 Sekunden lange Segmente mit dem Schwellwert $E_{\text{th}} = 20$.

Schon bei der Betrachtung der Daten fällt der Unterschied der Signalqualität insgesamt auf: Während bei den zuvor betrachteten Daten mit $E_{\text{th}} = 20$ eine Coverage von 59,45 %

durch die Annotation erreicht werden, liegt sie bei den Daten der gesunden Proband*innen bei 83,96 %. Auch liegt der durchschnittliche MAE aller Daten bei den gesunden Personen bei 13,76 FE im Vergleich zu 21,85 FE bei tatsächlichen Patient*innen.

Die Daten werden auf den trainierten Modellen ohne weitere Anpassungen getestet. Es zeigt sich, dass die Klassifikation auch auf unbekannten Daten mit anderen Aufnahmebedingungen funktioniert. Die Ergebnisse für die vier Modelle sind in Tabelle 6.11 abgebildet. Es zeigt sich erneut, dass die Regressionsmodelle eine höhere Coverage bei einem ebenfalls höheren MAE erreichen. Während die Performance der Klassifikationsmodelle auch bei den unbekannten Daten sehr ähnlich ist, zeigt sich bei den Regressionsmodellen, dass das RF-Modell bei der AUC, dem F1-Score und der Accuracy etwas besser abschneidet. Allerdings ist auch der MAE leicht höher - bei einer ebenfalls leicht höheren Coverage.

	Modell	MAE [FE]	Coverage [%]	F1-Score	AUC	Accuracy
insgesamt		13,76	-	-	-	
annotiert		4,34	83,96	-	-	
Klassifikation	RF	6,60	81,58	0,91	0,80	0,85
	XGB	6,60	81,58	0,90	0,80	0,85
Regression	RF	7,58	91,33	0,93	0,81	0,88
	XGB	7,34	88,95	0,92	0,77	0,87

Tabelle 6.11: Resultate der 4 Modelle auf im Schlaf aufgenommenen Daten gesunder Proband*innen.

Die Modelle sind also in der Lage, auch die Signalqualität von Daten mit anderen Aufnahmebedingungen ohne weiteres Training zu beurteilen.

7 Zusammenfassung und Ausblick

Ziel dieser Arbeit war die Untersuchung von Möglichkeiten, die Signalqualität von ballistokardiographischen Signalen mittels Methoden maschinellen Lernens zu beurteilen. Besonderer Fokus lag dabei auf Messdaten von bettlägerigen Patient*innen. Hierzu wurden zunächst die Grundprinzipien des maschinellen Lernens vorgestellt und die Eigenschaften von ballistokardiographischen Signalen untersucht. Auf dieser Basis wurden Besonderheiten der Signalverarbeitung von BKG-Signalen präsentiert und existierende Verfahren zur Beurteilung der Signalqualität vorgestellt.

Die vorliegenden Messdaten wurden vorbereitet und ein geeignetes Verfahren zur Annotation entwickelt. Diese vorbereiteten Daten wurden anschließend genutzt, um die nachimplementierten, existierenden Verfahren zur Beurteilung der Signalqualität zu testen und zu evaluieren. Es zeigte sich, dass diese sich nur bedingt eignen, wenn ein großer Teil des Signals von schlechter Qualität ist, wie es bei in Betten aufgenommenen BKG-Signalen häufig der Fall ist, besonders wenn diese aus Messungen tagsüber stammen. Daher wurden Merkmale entwickelt, anhand derer die Signalqualität beurteilt werden kann.

Zur Untersuchung dieser Merkmale wurden insgesamt vier Modelle des maschinellen Lernens ausgewählt: Regression und Klassifikation jeweils mit Random Forests und Gradient Boosted Trees. Diese Modelle sind robust und schnell lernend und eignen sich somit für eine Evaluation verschiedener Einflüsse wie beispielsweise der Merkmalsauswahl, der Länge der Segmente, die beurteilt werden sollen und des Schwellwertes, der für die Klassifikation verwendet wird. Um die Modelle an Besonderheiten wie Lücken in den Daten anzupassen, wurde ein eigenes Modell entwickelt, das eine bestehende Implementierung eines Modells des maschinellen Lernens erweitert. Um Regressionsverfahren für eine Klassifikation nutzen zu können, wurde auch hierfür ein Modell entwickelt, dass die Vorhersagen des Regressionsmodells in eine binäre Klassifikation umwandelt und zusätzlich Wahrscheinlichkeiten zu den Klassifikationen liefern kann.

Die Evaluation der Modelle zeigte eine deutliche Verbesserung zu den existierenden Methoden zur Beurteilung der Signalqualität. Der Einfluss von Segmentlänge und Schwellwert der binären Annotation wurden gezeigt. Abschließend wurden die Modelle erfolgreich auf Daten einer anderen Aufnahmesituation, von schlafenden gesunden Proband*innen, getestet.

Im Rahmen dieser Arbeit war es nicht möglich, die Daten von Expert*innen annotieren zu lassen. Eine solche Annotation führt zu weniger Fehlern im zum Training verwendeten Datenset und kann so womöglich zu besseren Ergebnissen führen. Auch kann das entwickelte Merkmalsset noch um weitere Merkmale erweitert werden und unwichtigere Merkmale beispielsweise mit rekursiver Merkmalselimination verworfen werden.

Aufgrund von begrenzter Zeit und Rechenleistung wurden die Hyperparameter für diese Arbeit nur eingeschränkt optimiert. Weiteres Hyperparameter-Tuning bietet Potenzial, die Ergebnisse nochmals zu verbessern. Idealerweise wird hierfür eine eigene Metrik entwickelt, die erreichte Coverage und MAE kombiniert bewertet, da diese beiden Werte entscheidend für die Bewertung der Ergebnisse sind.

Eine Betrachtung der Wahrscheinlichkeiten der Klassenzugehörigkeiten kann detaillierter als lediglich durch die Betrachtung der AUC untersucht werden und womöglich weiteren Einblick geben. Eine Variation des Schwellwertes der Wahrscheinlichkeit, ob ein Segment informativ ist, kann eine weitere Anpassung eines Modells an den jeweiligen Anwendungsfall bedeuten.

Auch könnte untersucht werden, ob eine Kombination von Modellen, Segmentlängen und Schwellwerten genutzt werden kann um zunächst eine gröbere Klassifikation vorzunehmen und diese anschließend zu verfeinern.

Insgesamt wurden mit dieser Arbeit Möglichkeiten gezeigt, die Signalqualität von ballistokardiographischen Signalen zu beurteilen. Dies ist eine wichtiger Schritt, um BKG in der Praxis zu verwenden.

A Quelltext der entwickelten Modelle

```
import numpy as np
import pandas as pd
import xgboost as xgb
from sklearn.base import BaseEstimator, ClassifierMixin
from sklearn.metrics import accuracy_score


class OwnClassifier(BaseEstimator, ClassifierMixin):

    def __init__(self, model):
        """Initialize self
        """
        self.model = model

    def fit(self, X, y):
        """Fits the underlying model to a binary label

        :param X: input samples {array-like, sparse matrix} of shape (n_samples, n_features}
        :param y: target label
        """
        if type(X) is not pd.DataFrame:
            X = pd.DataFrame(X)
        if type(y) is not pd.Series:
            y = pd.Series(y, index=X.index)
        mask_nan = X.isna().any(axis=1)
        y_true = y.loc[X[~mask_nan].index]
        if type(self.model) == xgb.XGBClassifier:  # class weight
            scale_pos_weight = len(y[~y].index) / len(y[y].index)
            self.model.set_params(scale_pos_weight=scale_pos_weight)
        self.model.fit(X.loc[~mask_nan], y_true)
        self.classes_ = [False, True]  # order is important for AUC
        return self

    def predict(self, X):
        """Predict class for X

        :param X: input samples {array-like, sparse matrix} of shape
                  (n_samples, n_features}
        :returns: np array of shape (n_samples,) with binary labels
        """
        if type(X) is not pd.DataFrame:
            X = pd.DataFrame(X)
        mask_nan = X.isna().any(axis=1)
        X_not_na = X.loc[~mask_nan]
        y_pred = self.model.predict(X_not_na)
        y = pd.Series(index=X.index, data=np.full((len(X.index),), False), name='pred')
        y.loc[X_not_na.index] = pd.Series(y_pred, X_not_na.index, dtype=bool)
        return y.to_numpy()
```

```
    def get_params(self, deep=True):
        return {'model': self.model}

    def set_params(self, **params):
        self.model.set_params(**params)
        return self

    def predict_proba(self, X):
        """Predicts the probability of both classes for given feature vectors

        :param X: input samples {array-like, sparse matrix} of shape
                  (n_samples, n_features)
        :returns: class probabilites in the order [False, True] in form (n_samples,2)
        """
        if type(X) is not pd.DataFrame:
            X = pd.DataFrame(X)
        mask_nan = X.isna().any(axis=1)
        X_not_na = X.loc[~mask_nan]
        y_proba_not_na = self.model.predict_proba(X_not_na)
        y_proba = pd.DataFrame(index=X.index, columns=self.model.classes_)
        y_proba.loc[mask_nan, True] = np.array([0])
        y_proba.loc[mask_nan, False] = np.array([1])
        y_proba.loc[X_not_na.index, self.model.classes_[0]] = y_proba_not_na[:, 0]
        y_proba.loc[X_not_na.index, self.model.classes_[1]] = y_proba_not_na[:, 1]
        return y_proba.to_numpy()


class RegressionClassifier(BaseEstimator, ClassifierMixin):

    def __init__(self, model, threshold=10, scale_sqrt=4):
        """Initialize self
        """
        self.model = model
        self.threshold = threshold
        self.scale_sqrt = scale_sqrt

    def fit(self, X, y):
        """Fits the underlying model to a continuous target

        :param X: input samples {array-like, sparse matrix} of shape
                  (n_samples, n_features}
        :param y: needs to be continuous target
        """
        y = np.power(y, 1/self.scale_sqrt)
        self.model.fit(X, y)
        self.classes_ = [False, True]  # order is important for AUC
        return self

    def predict(self, X):
        """Predict class for X

        :param X: input samples {array-like, sparse matrix} of shape
                  (n_samples, n_features}
        :returns: np array of shape (n_samples,) with binary labels
        """
        y_continuous = self.model.predict(X)
        y_continuous = np.power(y_continuous, self.scale_sqrt)
        y = [False if curr > self.threshold else True for curr in y_continuous]
```

```
        return y

    def get_params(self, deep=True):
        return {'model': self.model,
                'threshold': self.threshold,
                'scale_sqrt': self.scale_sqrt}

    def set_params(self, **params):
        if "scale_sqrt" in params:
            self.scale_sqrt = params.pop("scale_sqrt")
        self.model.set_params(**params)
        return self

    def predict_proba(self, X):
        """Predicts the probability of both classes for given feature vectors

        :param X: input samples {array-like, sparse matrix} of shape
                  (n_samples, n_features}
        :returns: class probabilites in the order [False, True] in form (n_samples,2)
        """
        y_continuous = self.model.predict(X)
        y_continuous = np.power(y_continuous, self.scale_sqrt)
        # e function with f(th)=0.5
        proba_true = np.array([np.math.exp(np.log(0.5)/self.threshold * y)
                               for y in y_continuous])
        proba_false = 1 - proba_true
        ret = np.ones(shape=(len(y_continuous), 2))
        ret[:, 0] = proba_false
        ret[:, 1] = proba_true
        return ret

    def score(self, X, y):
        """Calculates accuracy score for given input-output pairs

        :return: accuracy score
        """
        y = [False if curr > self.threshold else True for curr in y]
        return accuracy_score(y, self.predict(X))
```

Literatur

Albukhari, Almothana, Frederico Lima und Ulrich Mescheder (2019). „Bed-embedded heart and respiration rates detection by longitudinal ballistocardiography and pattern recognition“. In: *Sensors (Switzerland)* 19.6.

Brüser, Christoph, Kurt Stadlthanner et al. (2011). „Adaptive beat-to-beat heart rate estimation in ballistocardiograms“. In: *IEEE Transactions on Information Technology in Biomedicine* 15.5, S. 778–786.

Brüser, Christoph, S. Winter und S. Leonhardt (2013). „Robust inter-beat interval estimation in cardiac vibration signals“. In: *Physiological Measurement* 34.2, S. 123–138.

de Lalla, V., M. A. Epstein und H. R. Brown (1950). „Analysis of H wave of ballistocardiogram.“ In: *Circulation* 2.5, S. 765–769.

Elgendi, Mohamed, Mirjam Jonkman und Friso De Boer (2010). „Frequency bands effects on QRS detection“. In: *BIOSIGNALS 2010 - Proceedings of the 3rd International Conference on Bio-inpsired Systems and Signal Processing*. Bd. 1, S. 428–431.

Gordon, J W (1877). „Certain Molar Movements of the Human Body produced by the Circulation of the Blood.“ In: *Journal of Anatomy and Physiology* 11.Pt 3, S. 533–6.

Harrison, Matt (2019). *Machine Learning Pocket Reference - Working with Structured Data in Python*. O'Reilly Media, Inc.

Hoog Antink, Christoph et al. (Aug. 2020). „Ballistocardiography can estimate beat-to-beat heart rate accurately at night in patients after vascular intervention“. In: *IEEE Journal of Biomedical and Health Informatics* 24.8, S. 2230–2237.

Howell, Luis und Bernd Porr (Aug. 2019). *Popular ECG R peak detectors written in python*. Version 0.9.6. URL: `https://doi.org/10.5281/zenodo.3357365`.

Inan, Omer T. et al. (2015). „Ballistocardiography and Seismocardiography: A Review of Recent Advances“. In: *IEEE Journal of Biomedical and Health Informatics* 19.4, S. 1414–1427.

Medizinische elektrische Geräte - Teil 2-27: Besondere Festlegungen für die Sicherheit einschließlich der wesentlichen Leistungsmerkmale von Elektrokardiographie-Überwachungsgeräten (Apr. 2015). Norm.

Nizami, Shermeen, James R. Green und Carolyn McGregor (2013). „Implementation of Artifact Detection in Critical Care: A Methodological Review“. In: *IEEE Reviews in Biomedical Engineering* 6, S. 127–142.

Orphanidou, Christina et al. (2015). „Signal-quality indices for the electrocardiogram and photoplethysmogram: Derivation and applications to wireless monitoring“. In: *IEEE Journal of Biomedical and Health Informatics* 19.3, S. 832–838.

Paalasmaa, Joonas, Hannu Toivonen und Markku Partinen (2015). „Adaptive heartbeat modeling for beat-to-beat heart rate measurement in ballistocardiograms“. In: *IEEE Journal of Biomedical and Health Informatics* 19.6, S. 1945–1952.

Pedregosa, F. et al. (2011). *Scikit-learn: Machine Learning in Python*.

Pinheiro, Eduardo, Octavian Postolache und Pedro Girão (2010). „Theory and Developments in an Unobtrusive Cardiovascular System Representation: Ballistocardiography“. In: *The Open Biomedical Engineering Journal* 4.1, S. 201–216.

Pino, Esteban J., Javier A.P. Chavez und Pablo Aqueveque (2015). „Noninvasive ambulatory measurement system of cardiac activity“. In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. Bd. 2015-November, S. 7622–7625.

Porr, Bernd und Luis Howell (2019). „R-peak detector stress test with a new noisy ECG database reveals significant performance differences amongst popular detectors“. In: *bioRxiv*.

Rosales, Licet et al. (2012). „Heartbeat detection from a hydraulic bed sensor using a clustering approach“. In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, S. 2383–2387.

Sadek, Ibrahim, Jit Biswas und Bessam Abdulrazak (2019). „Ballistocardiogram signal processing: a review“. In: *Health Information Science and Systems* 7.1.

Sadek, Ibrahim, Jit Biswas, Zhu Yongwei et al. (2016). „Sensor data quality processing for vital signs with opportunistic ambient sensing“. In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. Bd. 2016-October, S. 2484–2487.

Starr, Isaac et al. (1939). „Studies on the Estimation of Cardiac Ouptut in Man, and of Abnormalities in Cardiac Function, From the Heart's Recoil and the Blood's Impacts; the Ballistocardiogram“. In: *American Journal of Physiology-Legacy Content* 127.1, S. 1–28.

Yu, Xinchi et al. (Aug. 2020). „Noncontact Monitoring of Heart Rate and Heart Rate Variability in Geriatric Patients Using Photoplethysmography Imaging“. In: *IEEE Journal of Biomedical and Health Informatics*, S. 1–1.

Zink, Matthias Daniel et al. (2017). „Unobtrusive Nocturnal Heartbeat Monitoring by a Ballistocardiographic Sensor in Patients with Sleep Disordered Breathing“. In: *Scientific Reports* 7.1.