

Umsetzung und Bewertung eines Eventdetektors für Beinbewegungen im Schlaf

Aaron Troll

Studienarbeit

Aufgabenstellung für die Anfertigung einer Studienarbeit

Studiengang: Elektrotechnik
Name: Aaron Troll
Matrikelnummer: 4758601
Immatrikulationsjahr: 2018
Titel: Umsetzung und Bewertung eines Eventdetektors für Beinbewegungen im Schlaf

Ziele der Arbeit

Bei der automatisierten Bewertung von Biosignalen ist ein häufiger Zwischenschritt die Eventdetektion im Biosignal. Ein Beispiel dafür ist die Detektion von Beinbewegungen (LM) im EMG des Polysomnogramms während des Schlafs. Diese LMs werden später als periodische Beinbewegungen (PLM) weiter ausgewertet und als Index (Events/h) abstrahiert.

Häufig wird die Eventdetektion mithilfe von maschinellem Lernen umgesetzt und gegen den medizinischen Goldstandard der manuellen Annotation verglichen. Für diesen Vergleich gibt es verschiedene Ansätze und Metriken, deren klinische Aussagekraft jedoch nicht validiert ist. In der klinischen Praxis ist häufig bereits das Vorliegen eines Events mit einem zugehörigen Zeitpunkt eine ausreichende Detektionsgenauigkeit, um daraus den Index zu berechnen. Die üblichen technischen Metriken zur Bewertung von Eventdetektionen basieren jedoch darauf, die exakte zeitliche Übereinstimmung des Zeitfensters der Detektion mit der manuellen Annotation am höchsten zu belohnen (Sample-weiser Vergleich). Deshalb ist bereits der Vergleich von maschineller und manueller Annotation und die Auswahl einer Metrik nicht trivial.

Diese Arbeit hat zwei Schwerpunkte. Einerseits sollen Metriken zur Bewertung der Eventdetektion recherchiert und angewendet werden. Andrereits soll die automatisierte Detektion von LMs recherchiert und ein Detektor umgesetzt werden. Die erarbeiteten Metriken sollen anschließend zur quantitativen Bewertung des Detektors verwendet und ihre Eignung diskutiert werden.

Schwerpunkte der Arbeit

- Recherche zur Bewertung von Eventdetektionen
- Umsetzung verschiedener Bewertungsmaßnahmen
- Recherche zur automatisierten Detektion von LMs
- Umsetzung eines automatisierten LM-Detektors
- Quantitative Bewertung der Detektionsgüte
- Diskussion der Eignung verschiedener Metriken zur Bewertung des LM-Detektors

Betreuer: Dipl.-Ing. Miriam Goldammer
Dipl.-Inf. Franz Ehrlich

Ausgehändigt am: 11. Oktober 2022
Einzureichen am: 11. April 2023

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

Selbstständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht. Bei der Auswahl und Auswertung des Materials sowie bei der Herstellung des Manuskripts habe ich Unterstützungsleistungen von folgenden Personen erhalten:

1. Dipl.-Ing. Miriam Goldammer
2. Dipl.-Inf. Franz Ehrlich

Weitere Personen waren an der geistigen Herstellung der vorliegenden Arbeit nicht beteiligt. Mir ist bekannt, dass die Nichteinhaltung dieser Erklärung zum nachträglichen Entzug des Diplomabschlusses ((Masterabschlusses)) führen kann.

Dresden, den

Unterschrift

0.1 Abstrakt

Periodische Beinbewegungen im Schlaf sind Symptome verschiedener Krankheitsbilder, welche in einem Polysomnogramm aufgezeichnet werden können. Die Anzahl der periodischen Beinbewegungen (PLM) pro Stunde (PLM-Index) kann einen Hinweis darauf geben, wie sehr der Schlaf gestört wird. Um den PLM-Index automatisch zu bestimmen, werden Eventdetektoren eingesetzt.

Diese Arbeit untersucht zunächst theoretisch die Vergleichbarkeit der Eventdetektoren und zeigt, dass die bisher verwendeten Metriken nur begrenzt zur Bewertung verwendet werden können. Das hier vorgeschlagene Kostenfunktional basiert darauf, die Fehler aufzusummen, die beim Bestimmen des PLM-Indexes entstanden sind. Zusätzlich werden Möglichkeiten vorgeschlagen, wie die Aussagekraft des Kostenfunktional über den Detektor erhöht und weitere Information über den Detektor aus den Annotationssignalen extrahiert werden kann. Die gefundenen Ansätze werden an einem Detektor aus der Literatur [5] unter Verwendung eines Datensatzes des Uniklinikums Dresden (6.1) überprüft und ausgewertet.

0.2 Abstract

Periodic leg movements during sleep are symptoms of many medical conditions, which can be recorded on a polysomnogram. The number of PLM per hour (PLM-Index) can indicate how much the sleep is disturbed. In order to automatically determine the PLM-Index, event detectors are used.

This work investigates the comparability of event detectors and shows that the previously used metrics are limited in their evaluation capabilities. The cost-functional proposed in this work is described by the sum of the errors made by the detector in determining the PLM-Index. In addition, ideas are proposed to increase the descriptiveness of the cost-functional and to extract further information about the detector from the annotation signals. The developed metrics are evaluated on a detector from the literature [5] using a dataset of the University Hospital Dresden 6.1.

Inhaltsverzeichnis

0.1 Abstrakt	iii
0.2 Abstract	iv
Abbildungsverzeichnis	vii
Tabellenverzeichnis	ix
Abkürzungsverzeichnis	x
1 Einleitung	1
2 Medizinische Grundlagen	3
2.1 Symptom	3
2.2 Ursachen	5
2.3 Therapie	5
2.4 Biophysikalischer Signalursprung	6
2.5 Störgrößen	6
3 Stand der Technik	8
3.1 Datenverarbeitungskette	8
3.2 Detektoren	9
3.3 Klassische Metriken	10
4 Präzisierung der Aufgabenstellung	14

5 Wahl der Metrik	16
5.1 Klassische Metriken	16
5.2 Kostenfunktional	17
5.3 Verbesserung der Einordnung des Detektors	22
6 Anwendung der Metriken	24
6.1 Datensatz	24
6.2 Wahl des Detektors	25
6.3 Funktion des Detektors	26
7 Ergebnisse	29
7.1 Klassische Metriken	29
7.2 Kostenfunktional	30
7.3 Verbesserung der Einordnung des Detektors	31
8 Diskussion	35
8.1 Datensatz	35
8.2 Kostenfunktional	37
8.3 Verbesserung der Einordnung des Detektors	38
8.4 Optimierung des Detektors	41
9 Fazit und Ausblick	42
Quellenverzeichnis	x
Anhang	xiii
A Anhang	xiv

Abbildungsverzeichnis

2.1	Graphische Veranschaulichung der AASM-Kriterien.	4
3.1	Veranschaulichung der Datenverarbeitungskette. Das EMG-Signal wird von medizinischem Personal mithilfe von Annotationsunterstützung verarbeitet. Rechts im Bild sind die Kennwerte dargestellt, die von einem Algorithmus berechnet werden und das Ergebnis der Kette beschreiben.	9
3.2	Klassifizierung von richtig-positiv (TP), richtig-negativ (TN), falsch-positiv (FP) und falsch-negativ (FN) nach segment- und eventweiser Berechnung bei segmentierter manueller und quasizeitkontinuierlicher automatischer Annotation.	11
4.1	Erweiterte Datenverarbeitungskette aus 3.1 zur Veranschaulichung der Vergleichsmöglichkeiten	14
5.1	Beispielannotation bei der die automatische Annotation nur positiv ist. Es entstehen Kosten von vier obwohl klassische Metriken Eins sind.	19
5.2	Beispielannotation bei der die automatische Annotation nur negativ ist. Es entstehen Kosten von vier obwohl klassische Metriken Eins sind.	19
5.3	Beispielannotation bei der die automatische Annotation identisch zur manuellen Annotation ist. Die Güte des Detektors wäre in diesem Beispiel Eins.	20
5.4	Beispielannotation bei der die automatische Annotation einen Fehler durch Xto1-Matching aufweist.	20
5.5	Beispielannotation bei der die automatische Annotation einen Fehler aufgrund eines FP aufweist.	20
5.6	Beispielannotation bei der die automatische Annotation einen Fehler aufgrund eines ungenauen Startwertes aufweist.	20
5.7	Beispielannotation bei der die automatische Annotation einen Fehler aufgrund eines FN aufweist. Es entstehen relative Kosten von 0.2.	21
5.8	Beispielannotation bei denen Kosten entstehen, obwohl die PLM Anzahl gleich ist. Es entstehen absolute Kosten von Acht.	21
5.9	Darstellung der Metriken, die zum Informationsgewinn oder zur Bewertung des Detektors verwendet werden können.	23

6.1	Histogramm der Demographie des Datensatzes.	25
6.2	Veranschaulichung der Funktionsweise des implementierten Detektors: vorverarbeitetes EMG-Signal mit oberem Schwellwert (oben), Zwischenergebnis des Annotationssignals vor der Nachbearbeitung (mittig), finale automatische und manuelle Annotation zum Vergleich (unten). Die Abtastfrequenz beträgt 200 Hz.	27
6.3	Ausschnitt eines unregelmäßigen EMG-Signals zur Veranschaulichung der dynamischen Anpassung des Schwellwertes (oben) und die daraus resultierende automatischer Annotation (unten). Die Abtastfrequenz beträgt 200 Hz.	27
6.4	EMG-Signal mit EKG Einkopplung (oben), bei der die manuell annotierten Beinbewegungen von dem Detektor nicht erkannt wurden.	28
7.1	Veranschaulichung der Korrelation von Differenz aus ergebniserhöhenden Fehlern und ergebnisvermindernden Fehlern (K_{diff}) und Differenz aus automatisch annotierten und manuell annotierten PLM.	32
7.2	Darstellung der Kostenverteilung mit einer automatisch erstellten Balkenweite .	32
7.3	Histogramm über die Differenz der Schwerpunkte aus automatischer und manueller Annotation. Die Breite der Balken wurde für die gesamte Arbeit automatisch erstellt.	33
7.4	Histogramm über die falsch-positiv Rate. Die Breite der Balken wurde automatisch erstellt.	34
8.1	Beispiel EMG-Signal mit Schwellwert (oben) und Annotationspaar (unten), bei der augenscheinlich keine Beinbewegungen stattgefunden haben. Die manuelle Annotation weist trotzdem viele Events auf.	36
8.2	Ausschnitt eines EMG-Signals mit zeitweise schlechter Signalqualität (oben); laut manueller Annotation wurden in diesem Bereich trotzdem LMs gefunden (unten)	36
8.3	Ausschnitt eines EMG-Signals mit sehr hoher Eventdichte(oben); das selbe EMG-Signal über die ganze Nacht dargestellt (mittig) mit zugehörigen Annotationen (unten); welche eine hohe Schwerpunktdifferenz aufweisen.	39
1.1	Histogramm über die Mittlere Abweichung der Startzeitpunkte von TP bei einer Abtastfrequenz von 200 Hertz	xiv
1.2	Histogramm über die Mittlere Abweichung der Endzeitpunkte von TP bei einer Abtastfrequenz von 200 Hertz	xv
1.3	Histogramm über die Standardabweichung der Startzeitpunkte von TP bei einer Abtastfrequenz von 200 Hertz	xv
1.4	Histogramm über die Standardabweichung der Endzeitpunkte von TP bei einer Abtastfrequenz von 200 Hertz	xvi
1.5	Histogramm über das Verhältnis aus automatischer zu manueller LM Anzahl .	xvi
1.6	Histogramm über das Verhältnis aus automatischer zu manueller PLM Anzahl .	xvii

Tabellenverzeichnis

3.1	Stand der Technik, Teil 1; Der Zusatz 'r' bezeichnet einen rekonstruierten Wert	12
3.2	Stand der Technik, Teil 2	13
7.1	Stand der Technik für eventweise Auswertung zur Einordnung des implementierten Detektors, Teil 1; Der Zusatz 'r' bezeichnet einen rekonstruierten Wert	29
7.2	Stand der Technik für eventweise Auswertung zur Einordnung des implementierten Detektors, Teil 2	30
7.3	Stand der Technik für segmentweise Auswertung zur Einordnung des implementierten Detektors	30
7.4	Korrelation zwischen Kostenfunktional und klassischen Metriken	31
7.5	Beiträge der Fehlerarten an den Gesamtkosten. (+) hinter der Fehlerart beschreibt, dass dieser Fehler zu einer Erhöhung des PLM-Indexes geführt hat. (-) steht analog für eine Verminderung.	31
7.6	Metriken aus dem Kapitel 5.3: "Verbesserung der Einordnung des Detektors"	33

Abkürzungsverzeichnis

AASM	American Academy of Sleep Medicine
Acc	Korrektklassifikationsrate (englisch: accuracy)
DNN	tiefes neuronales Netz (englisch: Deep Neural Network)
FN	falsch-negativ (englisch: false negative)
FP	falsch-positiv (englisch: false positive)
FPrate	Falsch positiv Rate (englisch: false positive rate)
HMC	Sleep Centre, Medisch Centrum Haaglanden and Bronovo-Nebo
Korr	Pearsons Korrelationskoeffizient (englisch: Pearson correlation coefficient)
LM	Beinbewegung (englisch: limb movement)
MrOS	MrOS Sleep Study
NPV	Falschausslassungsrate (englisch: negative predictive value)
PLM	periodische Beinbewegungen (englisch: periodic limb movement)
Prec	Genauigkeit (englisch: Precision)
RLS	Wittmaack-Ekbom-Syndrom (englisch: Restless-Legs-Syndrom)
Sens	Sensitivität (englisch: sensitivity)
Spez	Spezifität (englisch: specificity)
SSC	Stanford Sleep Cohort
TN	richtig-negativ (englisch: true negative)
TP	richtig-positiv (englisch: true positive)
WSC	Wisconsin Sleep Cohort

1 Einleitung

Periodische Beinbewegungen, die im Schlaf auftreten, sind Symptome verschiedener Krankheitsbilder. Sie können den Schlaf stören und werden deshalb zu Therapie- und Diagnosezwecken im Schlaflabor quantifiziert. Die Auswertung der aufgenommenen Polysomnogramme wird manuell von medizinischem Personal durchgeführt. Dies dauert bei Experten pro ausgewertete Nacht circa zwei bis vier Stunden [20]. Insbesondere bei monotonen Aufgaben treten dabei vermehrt Fehler auf. Es werden Beinbewegungen (LM) vor allem dann übersehen, wenn viele Events auftreten [2]. Besonders im späteren Verlauf der Nacht und in Schlafstadien, in denen keine LM erwartet werden [20], wird weniger zuverlässig gearbeitet. Eine Variabilität im Datensatz entsteht auch durch Subjektivität und den allgemeinen Gemütszustand des medizinischen Personals [2]. Eine ganz oder teilweise Automation würde die Annotationen vereinheitlichen und beschleunigen. Das Arbeiten mit halbautomatischen Annotationsunterstützungen ist 2.41 [4] bis 2.8 [25] mal schneller.

Einige Algorithmen zum Erkennen von Beinbewegungen (Detektoren) wurden in der Literatur bereits vorgeschlagen [20, 5, 10]. Im Idealfall arbeitet ein Detektor gut genug, sodass die verantwortlichen Entscheidungsträger dem System vertrauen und die Ergebnisse nicht einzeln überprüft werden müssen. Die Kaufentscheidung wird in manchen Fällen von Verwaltungskräften getroffen und nicht von Schlafspezialisten [19]. Diese Verwaltungskräfte brauchen also eine einfach verstehbare Möglichkeit die Detektoren zu vergleichen, um entscheiden zu können, welches System gekauft werden soll.

Derzeit werden Detektoren entwickelt und veröffentlicht, in denen die Autoren verschiedene Metriken angeben. Da immer mehrere der klassischen Metriken angegeben werden, widersprechen diese sich teilweise untereinander in der Aussage zur Güte des Detektors. Auch die Berechnungsweise der Metriken unterscheidet sich in den Veröffentlichungen. Für neu entwickelte Detektoren lässt sich also nicht immer eindeutig entscheiden, ob der Stand der Technik verbessert werden konnte. Ziel dieser Arbeit ist es, Metriken zu identifizieren, welche eine fundiertere Vergleichbarkeit der Detektoren ermöglichen.

Zuerst wird der medizinische Kontext der periodischen Beinbewegungen erläutert, um in die Thematik einzuführen und medizinisch relevante Kenngrößen zu identifizieren. Anschließend wird erklärt, wie die Messwerte aus dem Schlaflabor zu den medizinischen Kenngrößen weiterverarbeitet werden und wie Detektoren aus der Literatur derzeit verglichen werden. Hieraus

ergibt sich das Ziel und die Notwendigkeit dieser Arbeit im Kapitel 4. Im Hauptteil dieser Arbeit sollen zuerst die Lösungsansätze theoretisch hergeleitet werden und anschließend praktisch getestet werden. Für die Anwendung der neuen Metriken soll ein Detektor aus der Literatur ausgewählt werden und dieser mithilfe eines Datensatzes bewertet werden. Der Datensatz stammt aus dem Uniklinikum Dresden und beinhaltet 6216 polysomnographische Aufzeichnungen.

2 Medizinische Grundlagen

2.1 Symptom

Periodische Gliedmaßenbewegungen () sind wiederholte Bein- oder seltener Armbewegungen, die ein- oder beidseitig auftreten und symmetrisch oder alternierend sein können. Sie treten meist nach einer typischen Beugung der großen Gelenke (Hüfte, Knie, Sprunggelenk) und Streckung der großen Zehe auf. Zu beobachten ist dieses Symptom episodisch meist beim Übergang zwischen dem Wachzustand und dem oberflächlichen Schlaf (Schlafphase N1) und im stabilen Leichtschlaf (Schlafphase N2), aber auch in Ruhephasen im Wachzustand. Die Auftretenswahrscheinlichkeit nimmt mit dem Alter zu und versechsfacht sich ab dem 50. Lebensjahr auf circa 30 Prozent. [29]

Die relevantesten Definitionen über periodische Beinbewegungen (PLM) im Schlaf kommen von der American Academy of Sleep Medicine (AASM) und der World Association of Sleep Medicine, welche Beinbewegungen und deren periodisches Auftreten quantitativ beschreiben. Die Kriterien und Zahlenwerte sind sich weitestgehend ähnlich [2]. Diese Arbeit stützt sich auf die Kriterien der AASM (2016) [24]:

1. Das Signal der Elektromyographie (EMG) sollte bei entspanntem Tibialis anterior-Muskel bestimmt werden
2. Der Beginn des LM wird festgelegt, wenn die Zunahme des EMGs acht Mikrovolt über dem Ruhesignal liegt. Das Ende des LM wird bei einer Reduktion des EMG-Signals auf weniger als zwei Mikrovolt über dem Ruhesignal für mindestens 0.5 Sekunden festgelegt.
3. Die zeitliche Dauer eines LM liegt zwischen minimal 0.5 Sekunden und maximal zehn Sekunden.
4. Beinbewegungen an zwei unterschiedlichen Beinen werden dann als eine zusammenhängende LM klassifiziert, wenn weniger als fünf Sekunden zwischen den Anfängen beider Bewegungen liegen.
5. Als periodisch gelten mindestens vier LM, die jeweils innerhalb eines Zeitraumes zwischen fünf und 90 Sekunden aufeinander folgen.

6. Sie werden in allen Schlafstadien und im Wachzustand bestimmt.
7. Leg Movements am Ende einer Apnoe oder Hypopnoe, die gemeinsam mit dem Hyperventilationsbeginn auftraten, werden nicht bewertet, wenn sie in einem Zeitfenster auftreten, das 0.5 Sekunden vor dem Anfang einer Apnoe oder Hypopnoe beginnt und 0.5 Sekunden nach dem Ende einer Apnoe oder Hypopnoe endet

Die Kriterien wurden evidenzbasiert getroffen, basierten auf Literatur-Reviews oder auf Konsensverfahren [29]. Zur Übersicht wurden die Abbildung 1.6 erstellt. Die Muskelaktivität spielt

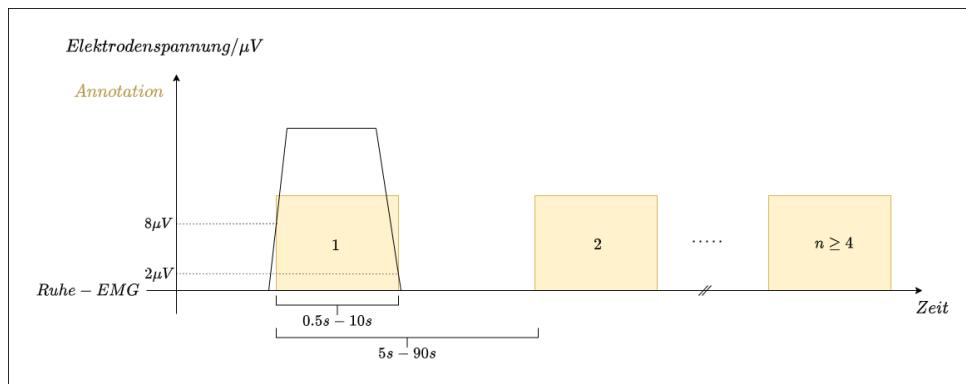


Abb. 2.1: Graphische Veranschaulichung der AASM-Kriterien.

auch bei folgenden Symptomen eine Rolle, muss aber von diesen abgegrenzt werden: hypnagogem Fußzittern (Hypnagogic Foot Tremor), alternierender Beinmuskelaktivität (Alternating Leg Muscle Activation, ALMA), exzessivem fraktionierten Myoklonus im NonREM-Schlaf, phasischen REM-twitches und Wadenkrämpfen [29]. Die Kriterien zur Unterscheidung von anderen Bewegungsstörungen sind in [24] beschrieben.

Die Beinbewegungen können mit Arousals einher gehen, welche zu einer partiellen, temporären oder vollständigen Weckreaktion führen. Diese äußern sich durch eine abrupte Frequenzänderung im Elektroenzephalogramm und sollen, laut der AASM, den Beinbewegungen zugeordnet werden. Da die Beinbewegungen überwiegend in den Schlafstadien N1 und N2 auftreten und oft mit Weckreaktionen einhergehen, sind REM- und Tiefschlaf häufig vermindert. Darunter leidet die Schlafeffizienz und die gesunde Periodizität der Schlafphasen wird gestört. Bei länger anhaltenden Beschwerden entwickeln viele Patienten zusätzlich psychische Fehlhaltungen und Verhaltensweisen, die die Schlafqualität weiter verschlechtern. Der Schlaf ist dadurch weniger erholsam. Es kommt zu Symptomen einer Hypersomnie, wie zum Beispiel eine ausgeprägte Tagesschläfrigkeit und Monotonieintoleranz sowie sekundäre depressive Symptome, imperative Einschlafattacken, Gedächtnis- und Aufmerksamkeitsstörungen. Diese können den Alltag stark beeinträchtigen und beispielsweise beim Autofahren sogar tödlich sein. Darüber hinaus werden die PLM mit Herzkrankheiten, hohem Blutdruck und Herzversagen assoziiert [20]. [29]

In der Praxis wird der PLMS-Index, also die Anzahl der pro Stunde auftretenden periodischen Beinbewegungen im Schlaf, genutzt, um das Ausmaß einer möglichen Erkrankung festzustellen. Dabei gelten bis zu fünf pro Stunde als unauffällig, zwischen fünf pro Stunde und 20 pro

Stunde als leichte Störung, zwischen 20 pro Stunde und 60 pro Stunde als moderate Störung und bei einem PLMS-Index über 60 pro Stunde als schwere Erkrankung. [29]

2.2 Ursachen

Periodische Beinbewegungen sind Symptome vieler unterschiedlicher Krankheiten. Diese treten beim Wittmaack-Ekbom-Syndrom (RLS) in 80 Prozent der Fälle, bei der REM-Verhaltensstörung in 70 Prozent und bei Narkolepsie in 45 Prozent bis 60 Prozent der Fälle auf. Bei der Krankheit Insomnie sind in 1 Prozent bis 15 Prozent der Fälle periodische Beinbewegungen symptomatisch und bei einer obstruktiven Schlafapnoe sind auch oft PLM vorhanden. Außerdem werden diese häufig in Assoziation mit psychiatrischen Störungen und bei neurologischen Erkrankungen wie Multisystematrophie oder Rückenmarksverletzungen gefunden. Auch Nebenwirkungen durch eingenommene Medikamente sind möglich. Dies ist allerdings nur eine Auswahl aller möglichen Ursachen. Falls das Symptom nicht durch eine der obigen Krankheiten erklärt werden kann, wird die Diagnose "Periodic-Limb-Movement-Disorder" gestellt, auch wenn die Krankheiten simultan auftreten können [14]. [29]

Da die Krankheitsbilder sehr unterschiedlich sind, sind auch die Ursachen für periodische Beinbewegungen vielfältig. Bei dem RLS und PLMD ist die Erkrankung ähnlich, aber noch weitgehend unverstanden. Diese entsteht wahrscheinlich im Zentralnervensystem durch eine Störung des dopaminergen oder opioidergen Systems [14], welche durch eine Störung des Eisenstoffwechsels auftreten könnte. Diese These wird gestützt durch das vermehrte Auftreten von PLM im höheren Alter bei einem gleichzeitigen Verlust an Dopamin und Dopaminrezeptoren. Die mit RLS korrelierenden Gene könnten auf eine frühe Entwicklungsstörung des zentralen Nervensystems hindeuten. [29]

2.3 Therapie

Periodische Beinbewegungen sollen nur dann behandelt werden, wenn sie eine Insomnie oder Hypersomnie auslösen, welche nicht durch andere Erkrankungen erklärt werden kann. Grundsätzlich gilt, dass nach Möglichkeit zuerst die Grunderkrankung behandelt werden sollte, falls diese in Verbindung mit PLM auftritt. [29]

Das RLS und PLMD werden pharmakologisch behandelt, indem die oben beschriebenen Mängel ausgeglichen werden. Hier soll zuerst dopaminerig, danach antiepileptisch und als letzte Maßnahme opioiderg behandelt werden. Hier eignen sich die Wirkstoffe L-Dopa, Rotigotin und Oxykodon je nach Schweregrad. Zusätzlich ist eine Eisensubstitution empfehlenswert. Insomnien sollten generell mit nichtmedikamentösen Verfahren kombiniert werden, indem über Grundlagenwissen der Schlafhygiene (z.B. Zubettgeh-Ritual, Matratze, Raumtemperatur, Geräuschisolierung und Abdunkelungsmöglichkeiten) aufgeklärt wird. Die Verwendung von schlaffördernden Mitteln wirkt lediglich symptomatisch und sollte nur über einen begrenzten Zeitraum erfolgen. [14, 29]

Die meisten Krankheitsbilder werden in erster Linie durch Fragebögen festgestellt. Die periodischen Beinbewegungen werden jedoch in der Regel nicht durch den Patienten in der

Nacht wahrgenommen. Eine mögliche schlafstörende Wirkung kann durch ein Elektroenzephalogramm bestätigt werden. Dies ist nötig, da sonst keine Therapie verschrieben werden darf. Bei diagnostisch unklaren Fällen, sowie bei Kindern und Jugendlichen kann es auch gerechtfertigt sein, den Schlaf im Schlaflabor zu überwachen. [14]

2.4 Biophysikalischer Signalursprung

In der folgenden Arbeit werden Daten aus Elektromyogrammen ausgewertet. Deswegen ist es an dieser Stelle wichtig, zu verstehen, welchen biophysikalischen Ursprung die ausgewerteten Signale haben. Da Beinbewegungen im PSG nicht direkt gemessen werden können, werden im Schlaflabor ersatzweise die Muskelkontraktionen des Beines gemessen, welche die Bewegungen verursachen. Um in diesem Kontext eine Muskelkontraktion zu verursachen, muss zunächst ein Signal in Form eines Aktionspotentials einer Nervenzelle zu dem Muskel weitergeleitet werden. Diese Nervenzelle bewirkt eine Depolarisation in der anliegenden Muskelzelle, in der Calciumionen freigesetzt werden [16]. Die Calciumionen ermöglichen dann die Kontraktion des Muskels [26]. Dieser Vorgang verändert das Potentialfeld, welches an der Hautoberfläche mit einer Elektrode gemessen werden kann. Die Änderung ergibt sich aus der Überlagerung der Depolarisation von Nerven- und Muskelzellen, welche anschließend durch aktives Zurückpumpen von Calciumionen repolarisiert werden [26]. Da Muskeln aus vielen Fasern bestehen, welche jeweils nur kurz und mehrfach an einer Kontraktion des gesamten Muskels beteiligt sind, ist die Potentialänderung ein stochastisches Signal in einem Frequenzbereich von zwischen zehn Hertz und 500 Hertz [22]. Die ein bis drei Quadratmillimeter große Depolarisationszone wird mit einer Geschwindigkeit von zwei bis sechs Meter pro Sekunde entlang der Muskelfaser weitergeleitet [16]. Laut AASM sollen die Kontraktion beider Beine anhand jeweils zweier Elektroden gemessen werden, welche entlang des Tibialis Anterior Muskels (Tibialis anterior muscle) platziert sind [24].

2.5 Störgrößen

Der Vorteil in der Nutzung zweier Elektroden ergibt sich aus der verbesserten Unterdrückung von Störgrößen. Das Signal ist dadurch im Idealfall unabhängig von Einflussgrößen, welche auf beide Elektroden wirken und daher nicht von dem gewünschten Signal verursacht wurden. Diese Störgrößen beinhalten das statische Potentialfeld, Haut-Elektrode Übergänge, kapazitive und induktive Einkopplung von anderen elektrischen Geräten und der Netzspannung. Im Idealfall heben sich die Halbzellspannungen auf, welche durch den Elektrode-Haut-Übergang entstehen. Die oben genannten Gleichtaktstörgrößen beeinflussen das Signal jedoch wenig, falls die Bioimpedanzen klein sind im Vergleich zu der Innenimpedanz des Messinstruments. Die AASM empfiehlt Impedanzen von 5000 Ohm [24]. [22]

Problematisch sind auch zeitliche Änderung der Haut-Elektrode Übergänge, welche durch die Bewegung oder Schweiß entstehen können [22, 12]. Zwischen dem Muskel und der Elektrode befindet sich weiteres Gewebe, welches das Signal nichtlinear verzerrn kann und wie ein Tiefpass Filter wirkt, dessen Grenzfrequenz mit der Gewebedicke abnimmt. Eine weitere

Störgröße ist die Überlagerung von Potentialfeldern, welche durch das Herz oder die Atmung verursacht wurden [5, 14]. [23]

Das Schlafverhalten des Patienten ist möglicherweise aufgrund der Verkabelung und Überwachung im Schlaflabor gestört [14]. Zudem kann es selbst bei einer schweren Erkrankung Nächte geben, in denen wenig periodische Beinbewegungen auftreten [29].

3 Stand der Technik

3.1 Datenverarbeitungskette

Die Datenerhebung findet über eine Polysomnographie im Schlaflabor statt. Richtlinien dazu finden sich in der [24]. In einer Standarduntersuchung werden ein Elektroenzephalogramm (EEG), ein Elektrookulogramm und Elektromyogramme (EMG) erstellt [14]. Das EMG besteht aus den Elektroden, einer Verstärkerschaltung, einem Analog-Digital-Umsetzer und Filtern [16] [12]. Der Analog-Digital-Umsetzer soll mit mindestens 200 Hertz (möglichst mit 500 Hertz) und einer 12 Bit Quantisierung arbeiten [24]. Üblich sind heutzutage 16 bis 22 Bit [19].

Außerdem können bei Bedarf 50 Hz beziehungsweise 60 Hz (in Amerika) Notchfilter eingesetzt werden. Die AASM rät davon allerdings ab, da die Aufzeichnung der Muskelaktivität beeinträchtigt werden könnte [12]. Der Start der eigentlichen Messung beginnt mit dem Löschen des Lichtes (Licht aus) und endet mit dem Anschalten des Lichtes (Licht an) [14]. Die Aufzeichnung wird auf den Monitoren mit einer Geschwindigkeit von zehn Millimeter pro Sekunde dargestellt, sodass auf einen Bildschirm 30 Sekunden passen [12]. Die Einteilung ist historisch aus den Aufzeichnungen auf Endlospapier entstanden, welches nach 30 Sekunden umklappte. Die gesammelten Daten werden anschließend verwendet, um unter anderem motorische, respiratorische und EEG-bezogene Ereignisse zu finden und zu klassifizieren. [29]

Die Software der Aufzeichnungsgeräte verfügt meistens direkt über eine Annotationsunterstützung, welche Vorschläge für die Start- und Endzeitpunkte der Beinbewegungen (LM) macht. Diese werden vom Assistenzpersonal überarbeitet. Vorschläge, die mit atembezogenen Events (zum Beispiel einer Schlafapnoe) zusammenhängen, sollen laut AASM gelöscht werden. Abschließend werden Regeln zur Bestimmung von periodischen Beinbewegungen angewandt (siehe Kapitel: „medizinische Grundlagen“), um die PLMS-Kennwerte zu bestimmen (Anzahl von PLMS, Anzahl von PLMS mit Arousals, PLMS-Index, PLMS-Arousal-Index) [10]. Die Annotationen werden später von einem Somnologen oder von einem Arzt für Schlafmedizin überprüft [29]. [12]

In dem Prozessbild 3.1 ist die Datenverarbeitungskette dargestellt. Als Eingangssignal auf der linken Seite ist das EMG Signal aus dem Schlaflabor zu sehen, welches durch das medizinische Personal -meist mithilfe der Annotationsunterstützung- auf Beinbewegungen untersucht

wird. Durch die Verarbeitung entsteht ein binäres Annotationssignal, welches jedem Zeitpunkt aus dem ursprünglichen EMG-Signal zuweist, ob an diesem Zeitpunkt eine Beinbewegung vorliegt oder nicht. Anhand dieses Annotationssignals wendet ein Algorithmus die Kriterien der AASM an und berechnet Kennwerte, die repräsentativ für das Annotationssignal stehen.

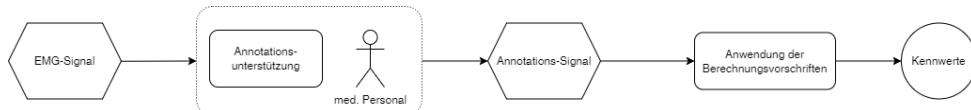


Abb. 3.1: Veranschaulichung der Datenverarbeitungskette. Das EMG-Signal wird von medizinischem Personal mithilfe von Annotationsunterstützung verarbeitet. Rechts im Bild sind die Kennwerte dargestellt, die von einem Algorithmus berechnet werden und das Ergebnis der Kette beschreiben.

3.2 Detektoren

Die Detektoren können nach der Art der Vorverarbeitung der Eingangsdaten, der Merkmalsextraktion und der Klassifizierung unterschieden werden. Es gibt sehr unterschiedliche Möglichkeiten die Merkmalsextraktion umzusetzen. Sie kann beispielsweise im Frequenzbereich (spektrale Kantenfrequenz¹, fraktaler Exponent²) [13], im Zeit-Frequenzbereich (Waveletkoeffizienten) [27] oder im Zeitbereich (Spikewiederholrate, durchschnittliche Amplitude, Länge des Zeitfensters in dem Spikes auftreten) [9] stattfinden. Die Auswertung kann zum Beispiel über statistische Klassifikatoren [11, 13] oder Neuronale Netze [10, 27] erfolgen.

Am meisten verbreitet sind die Detektoren, die im Zeitbereich eine Schwellwertklassifikation mit einem absoluten [9, 7, 20, 28, 30] oder dynamischen [3, 5] Schwellwert vornehmen. Hierbei wird meist die Amplitude des vorverarbeiteten Signals als Merkmal genutzt. Der Detektor von Carvelli et al. nutzt ein tiefes neuronales Netz (englisch: Deep Neural Network) (DNN) für die Klassifikation.

Für ein besseres Verständnis der Funktionsweise wird im Folgenden der Algorithmus von Moore et al. zusammengefasst. Dieser wurde gewählt, da der Ansatz weit verbreitet ist und Ideen von Tauchmann, Ferri et al. und Wetter et al. aufgegriffen und weiterentwickelt wurden.

1. Einlesen des Datensatzes

Die beiden EMG-Signale der Beine werden zu einem Signal zusammengeführt und die Elektrokardiogrammstörung mithilfe eines adaptiven Filters vermindert. Das gesäuberte Signal wird gleichgerichtet zu $x(n)$. Für die Annotation wird das RMS $y(n)$ von $x(n)$ mit einem beidseitigem 0.15-sekündigen Fenster gebildet.

2. Berechnung des Grundrauschens

Aus dem $x(n)$ wird mithilfe eines 20-sekündigen gleitenden Mittelwertes das vorläufige Grundrauschen $\eta(n)$ ermittelt. Das Grundrauschensignal ist nur vorläufig, da alle Beinbe-

¹Frequenz unter der ein bestimmter Prozentsatz der Gesamtleistung liegt.

²Maß für Signalkomplexität: Steigung der Geraden in der doppellogarithmischen Darstellung spektrale Leistungsdichte gegen Frequenz [13].

wegungen mit in den Mittelwert eingerechnet werden und somit fälschlicherweise das Grundrauschen erhöhen. Aus dem Grundrauschen wird auch der vorläufige erste und zweite Schwellwert berechnet (α und β)

$$\alpha(n) = \begin{cases} \eta(n) \log(\eta(n) + 1) + U, & \eta \leq 50 \\ \infty, & \eta > 50 \end{cases}$$

$$\beta(n) = \frac{L}{U} \alpha(n)$$

3. Berechnung der Annotation

Diese Schwellwerte weichen bewusst von den definierten Werten der AASM Kriterien ab, orientieren sich jedoch an diesen durch die Startwerte L und U. Der Startzeitpunkt für einen Beinbewegungskandidaten ist der Zeitpunkt, bei dem $y(n)$ das erste Mal den oberen Schwellwert überschreitet. Die Beinbewegung endet, wenn das RMS-Signal für 0.05 Sekunden unter dem unteren Schwellwert bleibt. Kandidaten, die weniger als 0.1 Sekunden voneinander trennen, werden zusammengefasst. Um den Verlauf des Grundrauschens besser zu approximieren, wird an den Stellen, an denen Beinbewegungskandidaten erkannt wurden, der Wert des absoluten EMG-Signals auf die Hälfte des zweiten Schwellwertes $\beta(n)$ gesetzt. Dieses Signal ist das finale Grundrauschen.

3.3 Klassische Metriken

Die Metriken können segmentweise oder eventweise berechnet werden. Ein Segment wird meist als positiv gezählt, wenn das binäre quasizeitkontinuierliche Signal bei mehr als 50% des Segmentzeitraumes positiv ist. Wenn die Segmente die Länge eines Abtastzeitraumes haben, geht die segmentweise Klassifikation in eine sampleweise Klassifikation über. Bei der eventweisen Klassifikation erfolgt die Zuordnung analog. Hier wird jedoch meistens (mit Ausnahme von [28]) ein TP gezählt, sobald sich die beiden Annotationszeiträume mit mindestens einem Abtastwert überlappen. Dabei kann es passieren, dass mehrere Events in einem Signal auftreten, während in der anderen Annotation nur ein Event gefunden wurde. Hier findet eine multiple Zuordnung statt, bei der für alle dieser Events ein TP gezählt wird. Die Notation ist analog zu Wetter et al. mit Xto1-Matching bei mehreren manuellen Annotationen und 1toX-Matching bei mehreren automatischen Annotationen.

Die Abbildung 3.2 gibt einen Überblick über die segmentweise und eventweise Klassifikation zusammen. Zu beachten ist, dass die Events in der Auswertung nicht mit ihrer Zeit gewichtet werden, sondern nur gezählt werden.

Die wichtigsten klassischen Metriken aus der Literatur werden im Folgenden vorgestellt.

Pearsons Korrelationskoeffizient (Korr) (PLM/h)

$$\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (3.1)$$

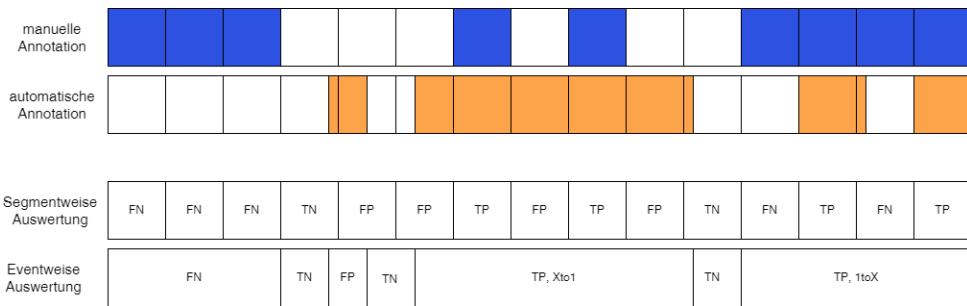


Abb. 3.2: Klassifizierung von TP, TN, FP und FN nach segment- und eventweiser Berechnung bei segmetierter manueller und quasizeitkontinuierlicher automatischer Annotation.

Cohens κ

$$P_e = \frac{(TP + FN)(TP + FP) + (TN + FP)(TN + FN)}{(TP + TN + FP + FN)^2} \quad (3.2)$$

$$P_0 = \frac{TP + TN}{2} \quad (3.3)$$

$$\kappa = \frac{P_0 - P_e}{1 - P_e} \quad (3.4)$$

F1-Maß

$$\frac{2TP}{2TP + FP + FN} \quad (3.5)$$

AUROC

Integral unter der Grenzwertoptimierungskurve beim Verändern des Entscheidungsschwellwertes

AUPRC

Integral unter der Genauigkeits-Sensitivity-Kurve beim Verändern des Entscheidungsschwellwertes

Genauigkeit (Prec)

$$\frac{TP}{TP + FP} \quad (3.6)$$

Falschauslassungsrate (NPV)

$$\frac{TN}{TN + FN} \quad (3.7)$$

Korrektklassifikationsrate (Acc)

$$\frac{TP + TN}{TP + TN + FP + FN} \quad (3.8)$$

Sensitivität (Sens)

$$\frac{TP}{TP + FN} \quad (3.9)$$

Spezifität (Spez)

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

Falsch positiv rate (FPrate)

$$\frac{FP}{FP + TP} \quad (3.11)$$

Die Tabellen 3.1 und 3.2 beschreiben den Stand der Technik und geben einen Überblick welche Detektoren aus der Literatur bereits welche Ergebnisse erzielen konnten.

Tab. 3.1: Stand der Technik, Teil 1; Der Zusatz 'r' bezeichnet einen rekonstruierten Wert

	[20]	[5]	[5]	[3]	[7]
Datensatz	15 RLS (24)	WSC (1073)	SSC (760)	HMC (70)	WSC (60)
Art des Klassifikators	statisch	dynamisch	dynamisch	dynamisch	statisch
Berechnungsart	Event	Event [21]	Event [21]	Segment (1s)	Event
Sens	0.95	0.6	0.75	0.82	0.85
Prec	–	0.88	0.82	0.71	0.62
Spez	0.92	1	1	1	0.99
NPV	–	0.99	0.99	–	1
Acc	–	0.98	0.99	1	0.99
Cohens κ	–	0.71	0.87	0.73	0.72
$F1$ – Maß	0.93r	–	–	0.73	–
Korrealtion PLM/h	0.97	0.94	0.94	–	–
relative # PLM	1.05	0.63	0.92	0.99	1.33

Bei der eventweisen Berechnungsart wurde ein TP bei jeglicher Überlappung gewertet. Die Werte für den Detektor von Wetter et al. sind aus [5] übernommen. Der Zahlenzusatz 'r' steht dafür, dass dieser Wert nicht veröffentlicht wurde, aber aus den anderen Metriken rekonstruiert werden konnte. Die Zahl in Klammern am Ende der Datensatzbeschreibung bezeichnet die Gesamtanzahl der Patienten im verwendeten Datensatz.

Tab. 3.2: Stand der Technik, Teil 2

	[9]	[10]	[10]	[10]
Datensatz	WSC (60)	WSC (275)	SSC (177)	MrOS Sleep Study (348)
Art des Klassifikators	statisch	DNN	DNN	DNN
Berechnungsart	Event	Event	Event	Event
Sens	0.96	0.9	–	–
Prec	0.47	0.81	–	–
Spez	0.98	–	–	–
NPV	1	–	–	–
Acc	0.98	–	–	–
Cohens κ	0.62	–	–	–
F1 – Maß	–	0.83	0.71	0.77
Korrealtion PLM/h relative # PLM	–	–	–	–
	2.05	–	–	–

4 Präzisierung der Aufgabenstellung

Um herauszufinden, bei welchem Detektor die Güte am höchsten ist, lohnt es sich zunächst das Prozessbild aus dem Stand der Technik zu 4.1 zu erweitern . Die in der Literatur vorgeschlagenen Detektoren sind hier auf der Höhe der Annotationsunterstützungssoftware und des medizinischen Personals dargestellt, da Sie deren Aufgaben ganz oder teilweise übernehmen sollen.

Da die Detektoren eine sehr unterschiedliche Funktionsweise haben können, müssen die Detektoren anhand ihrer Ergebnisse untereinander verglichen werden. Dies kann jedoch nicht zu qualitativen Aussagen führen, da nicht entschieden werden kann, welche Beinbewegungen als richtig annotiert gelten. Aus diesem Grund werden die Ergebnisse paarweise mit den Ergebnissen den medizinischen Fachpersonals verglichen. Da aus dem Annotationssignal die Kennwerte berechnet werden, bieten sich diese zusätzlich als Vergleichsmöglichkeit in der Datenverarbeitungskette an. Dieses Vorgehen hat den zusätzlichen Vorteil, dass auch Detektoren untereinander verglichen werden können, welche andere Eingangssignale nutzen. Hierzu zählen: Videoanalyse [18], Aktigramm [29], Kraftsensoren unter der Matratze [1], Bioimpedanzänderung bei Bewegung [22], Elektrocardiogramm [8], Flugzeitsensoren [17].

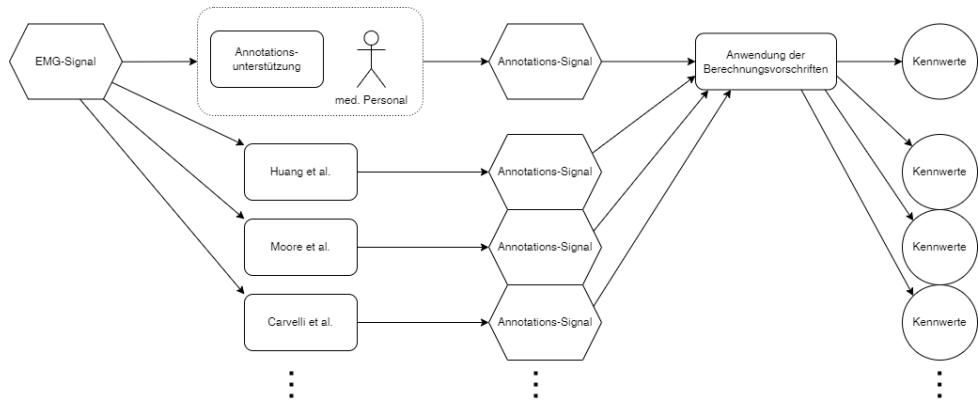


Abb. 4.1: Erweiterte Datenverarbeitungskette aus 3.1 zur Veranschaulichung der Vergleichsmöglichkeiten

Als Metrik wird hier ein skalarer Wert definiert, mit dem Detektoren verglichen werden können. Aus den oben genannten Problemen ergibt sich die dringende Notwendigkeit, eine Metrik zu finden, welche

1. Einfache, einheitliche und eindeutige Vergleichbarkeit zwischen Detektoren ermöglichen,
2. die Güte in dem medizinischen Kontext bewerten und
3. benutzt werden können, um Detektoren zu verbessern.

In der folgenden Arbeit sollen diese Probleme gelöst werden. Die AASM ist sich dieser Problematik bewusst und arbeitet daran Metriken zu untersuchen [31].

5 Wahl der Metrik

5.1 Klassische Metriken

Bei der Berechnung jeder Metrik gehen Informationen verloren, da zwei zeitliche Signale, welche jeweils komplizierte Zusammenhänge zwischen den einzelnen Events haben, auf eine reelle Zahl abgebildet werden. Also sollte man genau betrachten, welche Informationen bei dem Berechnen der klassischen Metriken verloren gehen.

Das Ziel der polysomnographischen Untersuchung und Auswertung ist es, die klinischen Kennzahlen der AASM [24] aus dem Annotationssignal zu bestimmen, um herauszufinden, wie stark ein Krankheitsverlauf den Schlaf stört. Daher wäre es sinnvoll den Detektor an dem Vergleich dieser Kennzahlen zu bemessen. Der Vergleich der Kennzahlen liefert allerdings kein vollständiges Bild über die korrekte Arbeitsweise des Detektors. Die Kennzahlen werden bei dem Detektor aus TP- und falsch-positiv (FP) Werten berechnet. Letztere sind allerdings nicht durch eine korrekte Arbeitsweise entstanden, sondern durch einen Fehler des Detektors.

Für die Bewertung des Annotationssignales kann die Berechnung von TP, FP, TN und FN Werten segmentweise oder eventweise geschehen.

Selbst bei einer schweren Erkrankung mit einem PLM-Index von 60 pro Stunde [29] und maximaler Beinbewegungslänge von zehn Sekunden liegt das Verhältnis von Beinbewegungen zu Ruhephasen bei circa eins zu fünf. Bei einer segmentweisen Berechnung hätte also eine starke Ungleichverteilung in beiden Klassen. Da die Anzahl der TN sehr hoch sein wird, bieten Sensitivität, Spezifität, und Korrektklassifikationsrate keinen informativen Gehalt [20]. Die Länge der jeweiligen LM ist für medizinische Entscheidungen nicht relevant. Da nur die Anzahl (und Abstände) der Events zählen, ist es sinnvoller die Metriken eventbasiert zu berechnen. Bei dieserzählweise sind die Klassen wesentlich gleicher verteilt, da gilt: $TP + FP + 1 = TN + FN$ (falls die Ränder negativ sind).

Die Metriken, die sich anhand des Annotationssignales berechnen lassen, sind eher ungeeignet, um einem Detektor eine Güte zuzuweisen. Bei den Metriken Genauigkeit, Spezifität und negativem Vorhersagewert fehlt jeweils eine der vier Klassen, weswegen die Werte einzeln betrachtet den Detektor nicht gut repräsentieren. Beispielsweise wäre die Sensitivität perfekt auf dem Wert 1 wenn der Detektor jeden Abtastwert als Beinbewegung klassifiziert. Es kann

außerdem passieren, dass die klassischen Metriken unendlich große Werte annehmen beziehungsweise neu definiert werden müssten. Dies kommt zum Beispiel bei der Genauigkeit zustande, wenn weder TP noch FP gefunden wurden.

Die klassischen Metriken lassen auch nicht immer eine eindeutige Aussage zu. Beispielsweise ist die Sensitivität bei dem Detektor von Huang et al. größer als die bei dem Detektor von Moore et al. Bei der Spezifität verhält es sich umgekehrt. Zudem lassen sich die Metriken sehr schlecht untereinander vergleichen, wenn sie auf unterschiedlichen Datensätzen berechnet werden. [6]

Andere Metriken wie Cohens κ beinhalten zwar alle der vier Klassen und gelten als aussagekräftig, jedoch geht durch die komplexe, abstrakte Berechnungsweise die Nachvollziehbarkeit und Erklärbarkeit verloren. Das Integral unter der ROC-Kurve sowie das Integral unter der Genauigkeits-Sensitivity-Kurve bieten für viele Arten von Detektoren eine gute Aussagekraft, sind jedoch für den Vergleich von Detektoren, welche auf Schwellwerten basieren, nicht anzuwenden. Ein Effekt der eventweisen Berechnung ist außerdem, dass im Normalfall (isierte Events) durch ein FP immer ein weiteres TN entsteht (im Vergleich zu dem Fall, dass es das FP nicht gegeben hätte). Genauso geht durch ein FN ein TN verloren. Dies führt dazu, dass die klassischen Metriken nicht die Aussage treffen, nach der es beim Betrachten der Formeln aussieht. Hier geht die Information verloren, wodurch (hier) das TN entstanden ist.

Die klassischen Metriken als Gesamtes betrachtet haben eine gewisse Aussagekraft, wenn alle der klassischen Metriken ausreichend hoch sind, da beispielsweise bei einem sehr hohen F1-Maß die errechneten Kennwerte fast ausschließlich aus den richtig erkannten Beinbewegungen stammen müssen. Komplizierter wird es allerdings, wenn die klassischen Metriken zweier Detektoren ähnlich sind oder nicht ausreichend gut. In diesen Fällen ist die Informationsreduktion in den klassischen Metriken zu groß, um fundierte Aussagen über die jeweiligen Detektoren treffen zu können. Der Detailgrad der Information beschränkt sich jedoch auf vage Aussagen wie „Sensitivität ist höher als Genauigkeit, was signalisiert, dass das Modell sehr inklusiv ist“ [10]. Außerdem muss für die Definition einer Güte ein skalarer Wert definiert werden. Ein weiteres Problem ist, dass zeitliche Zusammenhänge zwischen den Events verloren gehen. Die Berechnung der medizinisch wichtigen Kennwerte ist jedoch teilweise durch die Zeiten zwischen den LM-Startzeiten relevant. Deswegen lassen sich nicht alle Informationen aus der Anzahl der TP, TN, FP und FN-Werte ableiten.

5.2 Kostenfunktional

Es wird also eine neue Metrik benötigt, welche die Kriterien im Ziel der Arbeit besser erreicht. Das Kostenfunktional ist hier definiert als eine Funktion, welche zwei binäre Annotationssignale auf einen Skalar abbildet. Dieser Skalar beschreibt die Fehler (Kosten), die von einem Detektor gemacht wurden beim Berechnen des automatischen Annotationssignals im Vergleich zu dem manuellen Annotationssignal. Die Güte (G) um den Detektor zu beschreiben kann reziprok zu den Kosten (K) definiert werden:

$$G = \frac{1}{1 + K} \quad (5.1)$$

. Diese Definition hat den Vorteil, dass die Güte in einem Wertebereich zwischen Null und Eins liegt.

Für die Beschreibung des Kostenfunktionalen in dieser Arbeit wird sich auf den PLMS/h als medizinisch relevanten Wert beschränkt. Die Berechnungsweise lässt sich leicht erweitern, um andere Indices zu beschreiben. Das Kostenfunktional soll die aufgezeigten Probleme mit den herkömmlichen Metriken lösen, indem die Fehler gezählt werden, die gemacht wurden beim Bestimmen des PLMS-Indexes.

Für die Bestimmung der Fehler ist es zunächst notwendig, die Beinbewegungen zu finden, welche von dem Detektor richtig erkannt wurden. Für den Menschen ist es intuitiv zu wissen welche Beinbewegungen von dem Detektor „gemeint“ sind. Diese Intuition ist vermutlich von der Umgebung der Beinbewegungen abhängig. So würden LM einander großzügiger zugeordnet werden, wenn sich in der Umgebung keine weiteren LM befinden. Ist jedoch die Eventdichte hoch, würden strengere Regeln gelten. Für die Anwendung auf die Detektoralgorithmen ist dieses Vorgehen eher unbrauchbar. In der Literatur wird meistens ein TP gezählt, wenn ein Abtastwert von beiden Annotationen als positiv eingestuft wird. Dieser Ansatz wird hier übernommen. Dabei kann es passieren, dass die manuelle Annotation für einen Zeitraum nur ein LM vorsieht, der Detektor aber mehrere findet (1toX-Matching). Umgekehrt kann werden auch alle manuellen Annotationen, die mit einer automatischen Annotation überlappen dieser zugeordnet (Xto1-Matching).

Für eine zu hohe PLM Anzahl kann es folgende Gründe geben:

1. Vorhandensein eines FP innerhalb einer PLMS-Serie in [5-90] Sekunden Abstand zu einem zugehörigen LM
2. Dazuzählen eines TP, da es fälschlicherweise in das [5-90] Sekunden Intervall zählt
3. Vorhandensein von 1toX-Matching
4. Gründung einer PLM-Serie durch oben genannte Gründe

Für eine zu kleine PLM-Anzahl kann es folgende Gründe geben:

1. Vorhandensein eines FN in [5-90] Sekunden Abstand zu einem PLM
2. Nichtinkludieren eines TP, da es sich nicht in dem [5-90] Sekunden Intervall befindet
3. Vorhandensein von Xto1-Matching
4. PLM-Serie kommt nicht zu Stande aus oben genannten Gründen

Die Mengen FP und TP sind disjunkt und es kann deswegen nicht zu doppelten Zählungen kommen. Da die 1toX-Matches ein Teil der TP sind, gibt es beispielsweise einen Grenzfall, bei dem der erste oder letzte LM eines 1toX-Matches sich innerhalb des [5-90] Sekunden Intervalls befindet, das LM der manuellen Annotation jedoch noch nicht. Die Umsetzung des Kostenfunktionalen speichert die Beinbewegungen, die bereits zu einer Veränderung des Indexes beigetragen haben, sodass diese nicht doppelt gezählt werden.

Die Punkte eins bis drei tragen direkt zu einer Veränderung des PLM Indexes bei. Die Fehler unter Punkt vier tragen nichtlinear zum Ergebnis bei, da eine PLM-Serie wird erst als solche gezählt, wenn mindestens vier LM zu dieser Serie gezählt werden. Falls durch einen Fehler beispielsweise sich nur drei anstatt vier LM in der Serie befinden, wird das Ergebnis durch nur einen Fehler um vier verringert anstatt nur um eins. Das Kostenfunktional soll in diesem Fall auch die Veränderung der PLM Anzahl widerspiegeln.

Die Arbeitsweise des Kostenfunktionalen lässt sich anhand eines hypothetischen Annotationssignal gut veranschaulichen. In den folgenden Beispielen wurde ein Ausschnitt eines manuellen Annotationssignals aus dem Datensatz verwendet, welches eine PLM-Serie mit insgesamt vier PLM aufweist. Das erste automatische Beispieldaten ist in Bild 5.1 dargestellt und besteht nur aus einem LM, welches über die ganze Nacht andauert. Für einen Vergleich wurde in diesem Abschnitt zusätzlich ein manuell annotiertes Beispieldaten aus dem Datensatz 6.1 verwendet.

Die klassischen Metriken kommen hier bei der eventweisen Berechnung auf einen Wert von eins und treffen damit die Aussage, dass der Detektor perfekt funktioniert. Das Kostenfunktional hingegen berechnet absolute Kosten von vier, da die manuelle Annotation aus vier PLM besteht und keine Serie richtig erkannt wurde. Die Fehler wurden alle richtigerweise Xto1-Matches klassifiziert.

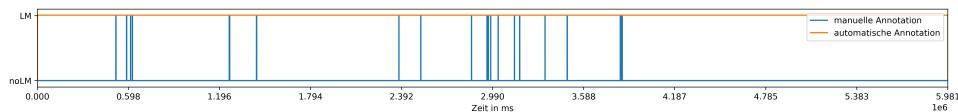


Abb. 5.1: Beispielannotation bei der die automatische Annotation nur positiv ist. Es entstehen Kosten von vier obwohl klassische Metriken Eins sind.

In dem Beispiel 5.2 wurde die automatische Annotation auf null gesetzt für die gesamte Nacht. Das Kostenfunktional kommt hier wieder auf einen Wert von 4 und auf relative Kosten von Eins. Die Genauigkeit und falsch positiv Rate sind in diesem Fall nicht definiert, da Zähler und Nenner Null sind. In diesem Fall werden die Metriken teilweise auch als Eins also als perfekte Genauigkeit definiert [15]. Laut Spezifität ist ein Detektor, der trivialerweise immer Null ausgibt also auch perfekt.

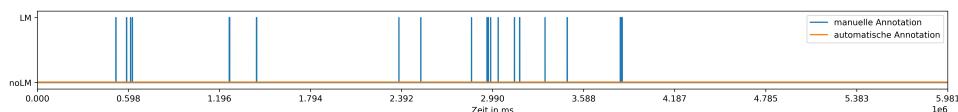


Abb. 5.2: Beispielannotation bei der die automatische Annotation nur negativ ist. Es entstehen Kosten von vier obwohl klassische Metriken Eins sind.

Für einen Detektor, der die manuelle Annotation exakt nachbildet, wie in Bild 5.3 dargestellt, wären beide Annotationssignale identisch. Da hier keine Fehler gemacht werden sind die absoluten und relativen Kosten Null. Die Güte hat somit den perfekten Wert von Eins.

Für die nachfolgenden Ergebnisse wurden Annotationssignale selbst erstellt, damit die einzelnen Kostenbeiträge besser veranschaulicht werden können. Die Funktionsweise der mul-

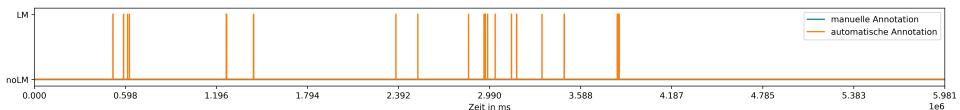


Abb. 5.3: Beispielannotation bei der die automatische Annotation identisch zur manuellen Annotation ist. Die Güte des Detektors wäre in diesem Beispiel Eins.

tiplen Zuordnung lässt sich in Abbildung 5.4 veranschaulichen. In dem Beispiel wurden der einen automatischen Annotation vier manuelle zugeordnet. Die absoluten Kosten sind wieder vier, da keins der vier manuell gefundenen PLM entdeckt wurde. Alle klassischen Metriken würden anhand dieses Signals einen perfekten Wert ausgeben. Das 1toX-Matching lässt sich analog darstellen.

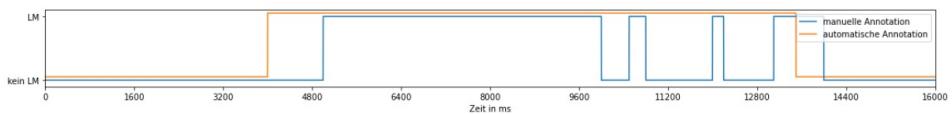


Abb. 5.4: Beispielannotation bei der die automatische Annotation einen Fehler durch Xto1-Matching aufweist.

In Abbildung 5.5 wird der Fehler aufgrund eines FP dargestellt, welcher somit eine Serie feststellt, obwohl die manuelle Annotation keine PLM erkennt. Die Kosten sind hier ebenfalls vier.



Abb. 5.5: Beispielannotation bei der die automatische Annotation einen Fehler aufgrund eines FP aufweist.

Die Abbildung 5.6 zeigt, dass eine zeitlich ungenaue Annotation in dem letzten LM ebenfalls zu einem Fehler führen kann. Hier erfüllt das LM der manuellen Annotation nicht die von der AASM geforderten Zeitbedingungen um als PLM zu zählen.

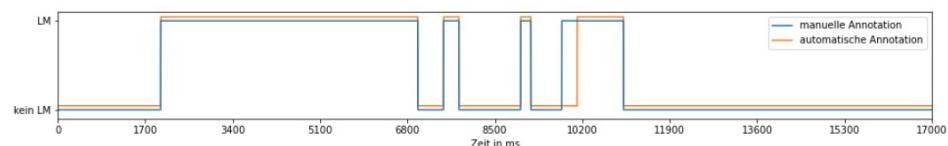


Abb. 5.6: Beispielannotation bei der die automatische Annotation einen Fehler aufgrund eines ungenauen Startwertes aufweist.

In dem Beispiel 5.7 wurde das letzte LM nicht erkannt und der Fehler aufgrund FN beträgt Eins. Die relativen Kosten sind hier nur 0.2, da trotzdem vier der fünf LM in der PLM-Serie richtig erkannt wurden.

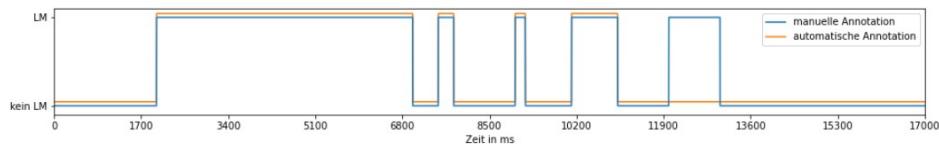


Abb. 5.7: Beispielannotation bei der die automatische Annotation einen Fehler aufgrund eines FN aufweist. Es entstehen relative Kosten von 0.2.

Im Bild 5.8 ist dargestellt, wie bei dem ersten automatisch erkannten LM ein Xto1-Matching vorliegt. Dieses wird aber gleichzeitig in eine weitere PLM-Serie aus FP gezählt. Obwohl die PLM Anzahl für beide Annotationen gleich ist, entstehen hier absolute Kosten von acht und relative Kosten von zwei.

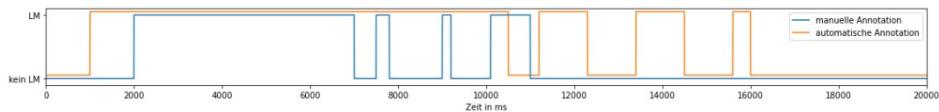


Abb. 5.8: Beispielannotation bei denen Kosten entstehen, obwohl die PLM Anzahl gleich ist. Es entstehen absolute Kosten von Acht.

Falls es keine anderen außer die oben genannten Fehler gibt, müsste

$$\frac{\text{ergebniserhöhende Fehler} - \text{ergebnisverkleinernde Fehler}}{\text{automatisch annotierte PLM} - \text{manuell annotierte PLM}} = \text{(5.2)}$$

gelten. Diese Gleichung beschreibt alle Fehler, die durch das Kostenfunktional gefunden werden können und setzt diese in Beziehung zu den tatsächlich auftretenden Fehlern. Mit dieser Gleichung kann also überprüft werden, ob alle einzeln gemachten Fehler auch gefunden wurden.

Es gibt jedoch auch Detektorfehler, die das Ergebnis nicht verändern aber trotzdem als Kosten zählen sollen. Zum Beispiel kann es passieren, dass eine manuell gefundene PLM Serie durch einen Zeitfehler (siehe Punkt 2) in zwar aus der einen Serie exkludiert wird, aber in einer anderen Serie damit inkludiert wird. Da sowohl Verkleinerung als auch Vergrößerung des Ergebnisses als falsch zu beurteilen ist, werden beide Fehlerarten addiert und bilden damit das Kostenfunktional:

$$K_{abs} = \text{ergebniserhöhende Fehler} + \text{ergebnisverkleinernde Fehler} \quad (5.3)$$

Die absolute Anzahl an Fehlern kann bei unterschiedlichen Datensätzen irreführend sein, da zwar die Wahrscheinlichkeit einen Fehler zu machen gleich sein sollte, aber bei mehr Events auch mehr Fehler entstehen. Daher sollte das Kostenfunktional auf die Anzahl der manuell annotierten PLM in der jeweiligen Serie normiert werden. Somit sind die Kosten auch unabhängig von der durchschnittlichen Länge der PLM-Serien. Das finale Kostenfunktional stellt also dar wie viele Fehler pro richtig erkannte PLM gemacht werden.

5.3 Verbesserung der Einordnung des Detektors

Dieses Kapitel bezieht sich darauf weitere nützliche Information aus den Annotationssignalen zu extrahieren, um mit diesen den Detektor besser einschätzen zu können. Diese Metriken gehen nicht in die Berechnung der Güte ein, können aber informativ genutzt werden.

Da sich die anderen Metriken hauptsächlich auf einzelne LM beziehen, könnte eine Metrik nützlich sein, welche große Veränderungen der Annotationen im Laufe der Nacht, darstellt. Diese könnte durch den Schwerpunkt der Startzeitpunkte der LM berechnet werden. Da das Gewicht jedes LM eins ist, und der Abstand gleich des Startzeitpunktes ist, geht die Berechnung des Schwerpunktes in die des Mittelwertes über. Dieser Wert könnte dafür genutzt werden, große Veränderungen, bei denen sich die Annotation im Laufe der Nacht ändert, zu erkennen. Interessant ist auch, ob diese Veränderung von beiden Annotationen gleichermaßen detektiert wurde. Also berechnet sich die Metrik aus der Differenz der beiden Schwerpunkte.

Eine weitere Information steckt in der zeitlichen Genauigkeit der Start- und Endzeitpunkte der Beinbewegungen. Diese lassen sich nur bei der eventweisen Zuordnung von Beinbewegungen zu TP analysieren. Falls eines der beiden Annotationssignale segmentweise ausgewertet wurde, geht genauere Information zu den Startzeitpunkten verloren. Diese Genauigkeit könnte in dem anderen Annotationssignal höher sein. Betrachtet man nun die Differenz der jeweiligen Startzeitpunkte von TP entsteht eine Verteilung. Die nächsten beiden Metriken beschreiben also den Mittelwert und die Standardabweichung dieser Verteilung. Da die Wahrscheinlichkeit des zeitlich höher aufgelösten Annotationssignal ein LM-Start zu haben unabhängig von den Segmentgrenzen des andren Annotationssignals ist, liegt der Erwartungswert bei null. Analog lässt sich auch eine Verteilung für die Endzeitpunkte der LM berechnen.

Eine hohe Anzahl an FP in einzelnen Nächten könnte darauf hindeuten, dass Events manuell entweder nicht betrachtet wurden oder nach einer Annotation im Nachhinein wieder gelöscht wurden. Die klassische Metrik der falsch positiv Rate könnte ein Indiz über diese Art von Uneinigkeit geben.

Des Weiteren lassen sich Information gewinnen, welche genutzt werden kann, um die Arbeitsweise des Detektors besser zu verstehen, um den Detektor gegebenenfalls zu optimieren. So kann beispielsweise das Verhältnis aus der Anzahl automatisch gefundener LM zu der Anzahl manuell gefundener LM darauf hindeuten, wie schnell der Detektor auf positive Änderungen im EMG reagiert.

Eine Verwandte Information liefert die Anzahl der multiplen Zuordnung (Xto1- und 1toX-Matching). Diese Werte könnten beschreiben, wie schnell der Detektor auf eine negative Änderung im EMG reagiert.

Da Beinbewegungen laut AASM eine bestimmte Mindest- und Maximaldauer haben, könnte die Zahl der Verstöße gegen diese Zeiten aufschlussreich über die Art der gefundenen Muskelkontraktionen.

Diese Informationen sind allerdings nur im Kontext der Funktionsweise des Detektors nützlich und sollten nicht zur Bewertung verwendet werden.

In der Grafik 5.9 sind die nützlichen Informationen zum Detektor dargestellt. Der obere Kreis bezieht sich auf den Vergleich zu bisherigen Detektoren. Hierfür sind die angegebenen klassischen Metriken am nützlichsten, da diese von den bisher entwickelten Detektoren angegeben

wurden. Diese können außerdem einen guten Überblick darüber geben, wie viele von den Beinbewegungen circa richtig erkannt werden. Besonders Cohens κ und das F1-Maß treffen eine vergleichsweise genaue Aussage. Links im Bild sind die Informationen zu dem Detektor dargestellt, welche Aussagen über die Funktionsweise des Detektors liefern. Der rechte Unterpunkt beschreibt das Kostenfunktional und seine Einflussfaktoren.

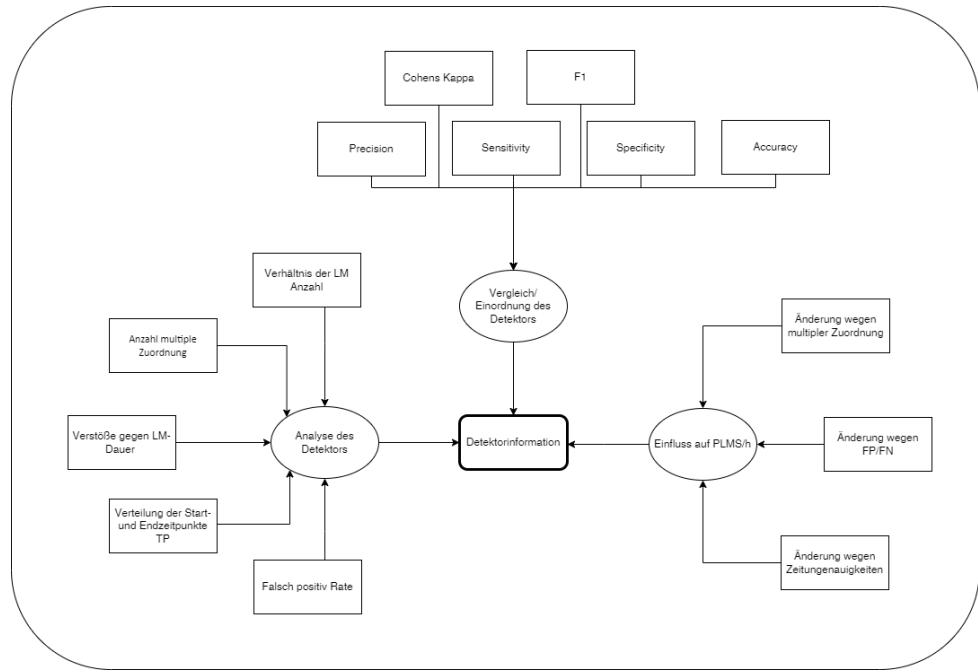


Abb. 5.9: Darstellung der Metriken, die zum Informationsgewinn oder zur Bewertung des Detektors verwendet werden können.

6 Anwendung der Metriken

6.1 Datensatz

Der Datensatz, auf dem der Detektor ausgewertet werden soll, wurde im Uniklinikum Dresden unter Verwendung des Alice System der Version 5 von Phillips erhoben. Für die Auswertung der EMG-Signale wurden die Kriterien der AASM angewendet und die Annotationsunterstützung des Alice Systems genutzt.

Es wurde ein Hochpassfilter mit einer Grenzfrequenz von 10 Hz und ein Tiefpassfilter mit einer Knickfrequenz von 93.6 Hz verwendet. Zusätzlich kam ein 50 Hz Notchfilter zur Anwendung. Die Abtastfrequenz beträgt 200 Hz.

Beim Bearbeiten des Datensatzes wurden 296 Dateien ohne manuelle Annotation gefunden und ausgeschlossen. Bei weiteren 12 Dateien wurde der gleitende Mittelwert nicht richtig berechnet. Dieser Fehler entsteht, wenn das vom Detektor erkannte Hintergrundrauschen $\eta(n)$ $50 \mu V$ überschreitet. Der Algorithmus setzt in diesem Fall den oberen Schwellwert richtigerweise auf unendlich. Bei Benutzung der Bibliothek scilab entstehen bei der Faltung jedoch ungültige Werte, welche nicht für die weitere Berechnung verwendet werden können. Der Fehler konnte anhand einiger Beispiele nachvollzogen werden und es wird davon ausgegangen, dass bei allen 12 Dateien das gleiche Problem vorlag. Die Ergebnisse basieren somit auf 5908 von ursprünglich 6216 Dateien im EDF-Format.

Es wurden von 3025 Personen das Alter und das Geschlecht aufgenommen. Das Geschlechtsverhältnis (M/F) beträgt 1.48 und die Alterszusammensetzung ist in Abbildung 6.1 dargestellt.

Die manuelle Annotation ist in 0.5 Sekunden Segmente unterteilt. Die Analyse der manuellen Annotationen ergibt eine durchschnittliche Anzahl von LM im Schlaf von 266. Am häufigsten aufgetreten sind Aufzeichnungen mit 17 annotierten LM. Der durchschnittliche PLMS/h-Wert liegt bei $9.95 * 10^{-3}$. Dabei haben 875 Dateien einen Wert von Null.

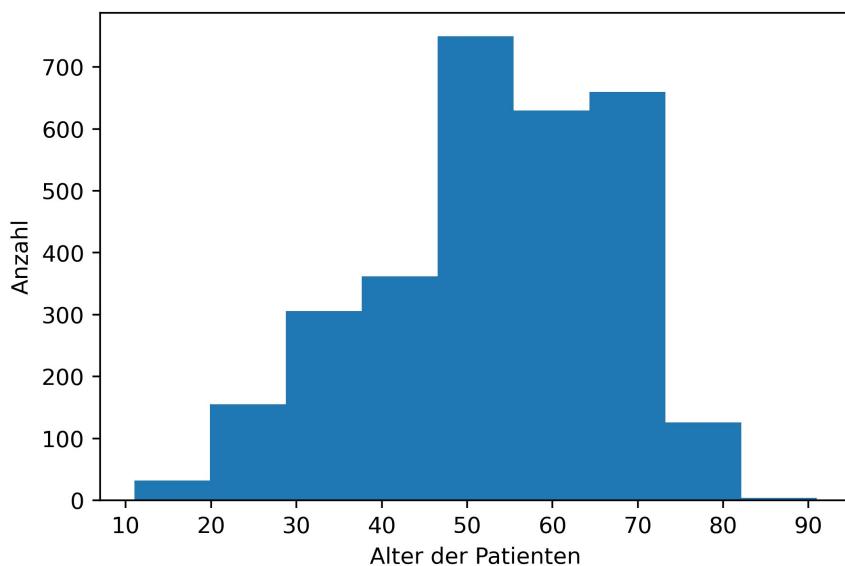


Abb. 6.1: Histogramm der Demographie des Datensatzes.

6.2 Wahl des Detektors

Um die gefundenen Metriken anzuwenden, soll für diese Arbeit ein Detektor umgesetzt werden. Augenscheinlich sind die vorgeschlagenen Detektoren in der Tabelle im Stand der Technik ausreichend performant, sodass hier auf die Entwicklung eines neuen Detektors verzichtet werden kann. Die am weitesten verbreitete Methode – und somit auch am weitesten entwickelte Methode – nutzt eine EMG-Signalvorverarbeitung und einen doppelten Schwellwertvergleich mit Nachbearbeitung des Annotationssignals. Der Vorteil eines Schwellwertvergleiches liegt außerdem in einer geringen Rechenlaufzeit, da keine Modelle trainiert werden müssen und die Berechnungen leicht zu parallelisieren sind. Der Algorithmus von Huang et al. wurde nur auf einem sehr kleinen Datensatz getestet und benötigt eine hohe EMG-Qualität, da ein absoluter Schwellwert implementiert ist. Die Variante von Moore baut auf den Erkenntnissen von Ferri et al., Tauchmann und Wetter et al. auf.

Also wurde für den Rahmen dieser Arbeit der Detektor von Moore et al. implementiert. Es gibt zwar eine weiterentwickelte Variante von Alvarez-Estevez, welche bessere Ergebnisse liefert, diese ist jedoch wesentlich komplizierter aufgebaut und schwieriger vergleichbar mit den anderen Detektoren, da die Metriken anhand von einsekündigen Segmenten erstellt wurden. Da es in dieser Arbeit hauptsächlich um die Untersuchung der Metriken geht, ist die von Moore et al. erreichte Klassifikationsgüte ausreichend.

Die Funktionsstruktur des Programms wurde im Stand der Technik beschrieben und ist hier in Python implementiert. Deswegen werden hier nur die Anpassungen beschrieben: Beim Einlesen der Dateien werden die leeren Annotationssignale übersprungen, da diese keinen Informationsgehalt bieten. Die beiden EMG-Signale der Beine werden laut [5] zu einem Signal zusammengeführt. Die Signale für beide Beine einzeln zu berechnen wie in [3] funktioniert hier nicht, da für die Nachbearbeitung der Annotationssignale ein Grundrauschen des kombinierten Signals gebraucht wird. Das RMS-Signal und der gleitende Mittelwert werden mit einer

Randeffektanpassung errechnet, um das Signal nicht an den Rändern zu verfälschen. Hierbei werden die erkannten LM gelöscht, welche vor dem „Licht aus“-Zeitpunkt und nach „Licht An“-Zeitpunkt stattgefunden haben, da diese auch nicht in dem manuellen Annotationssignal vorhanden sind. Es werden auch Beinbewegungen entfernt, welche laut AASM die Maximallänge von 10 Sekunden überschreiten. Durch das Zusammenführen der beiden Beine dürfen laut AASM alternierende Bewegungen zusammengefasst werden. Mit zusätzlicher Filterung sind Beinbewegungen im Signal von mehr als 15 Sekunden möglich. Für bessere Vergleichbarkeit mit der Literatur wurde der Wert aus der Arbeit von Moore et al. implementiert.

Bei der Umsetzung des Detektors wurde die adaptive Filterung der Elektrokardiogrammstörung und die Löschung der Beinbewegungen in der Nähe von atembezogenen Events verzichtet, da die benötigten Signale nicht zur Verfügung gestellt wurden.

Die automatischen und manuellen Annotationen werden mit den jeweiligen Start- und Endzeitpunkten in einer CSV-Datei gespeichert. Ein separates Programm kann die Annotationssignale aus den CSV-Dateien auslesen und daraus die Metriken bestimmen. Hier wird auch das Signal zu den Schlafstadien geladen, um die LM zu entfernen, die nicht im Schlaf stattgefunden haben. Des Weiteren gibt es ein Programm, welches alle pro Nacht berechneten Metriken einliest und zwischen den Nächten vergleicht.

6.3 Funktion des Detektors

An folgendem Beispiel lässt sich gut die Funktionsweise des Detektors anhand des Schwellwertes und die folgende Nachbearbeitung nachvollziehen. In Abbildung 6.2 ist oben das vorbearbeitete EMG-Signals (blau) und der davon abhängige dynamische obere Schwellwert zu erkennen. In dem Annotationssignal darunter ist ein Zwischenergebnis des Detektors zu sehen, bei dem alle Zeiträume annotiert wurden, bei denen das EMG-Signal über dem Schwellwert liegt. Das unterste Bild zeigt das finale automatische (orange) Annotationssignal nach der Nachbearbeitung. Dazu ist die manuelle Annotation zum Vergleich in blau auf der gleichen Achse dargestellt.

Damit bei unregelmäßigem EMG-Signal keine LM erkannt werden, passt sich der dynamische Schwellwert an die veränderte Qualität an. In Abbildung 6.3 lässt sich der Vorteil eines dynamischen Schwellwertes veranschaulichen. Für den dargestellten Zeitraum wurden manuell keine Beinbewegungen annotiert.

Das folgende Signal (Abb. 6.4) zeigt einen Ausschnitt einer EKG Störung mit einer Herzfrequenz von ungefähr einem Hertz und den daraus resultierenden erhöhten oberen Schwellwert. Durch diese Einkopplungen sind Beinbewegungen nicht mehr klar erkennbar.

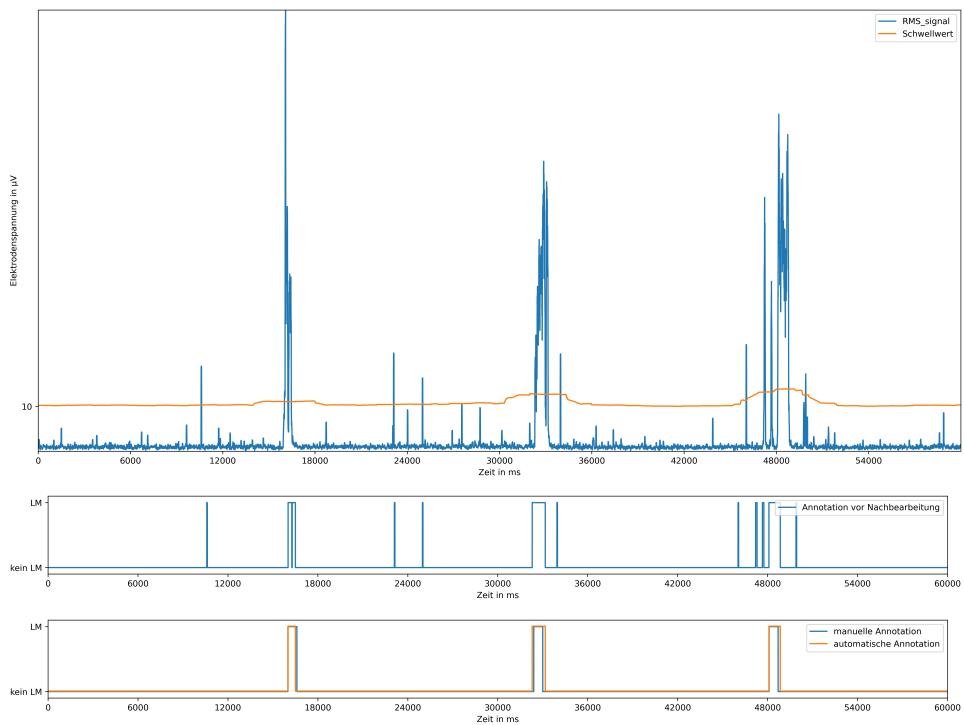


Abb. 6.2: Veranschaulichung der Funktionsweise des implementierten Detektors: vorverarbeitetes EMG-Signal mit oberem Schwellwert (oben), Zwischenergebnis des Annotationssignals vor der Nachbearbeitung (mittig), finale automatische und manuelle Annotation zum Vergleich (unten). Die Abtastfrequenz beträgt 200 Hz.

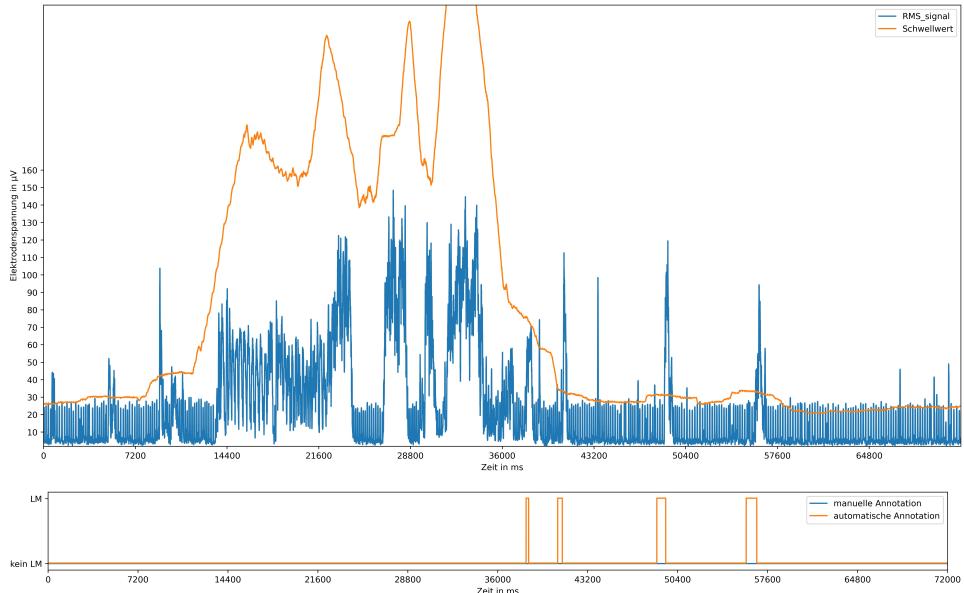


Abb. 6.3: Ausschnitt eines unregelmäßigen EMG-Signals zur Veranschaulichung der dynamischen Anpassung des Schwellwertes (oben) und die daraus resultierende automatischer Annotation (unten). Die Abtastfrequenz beträgt 200 Hz.

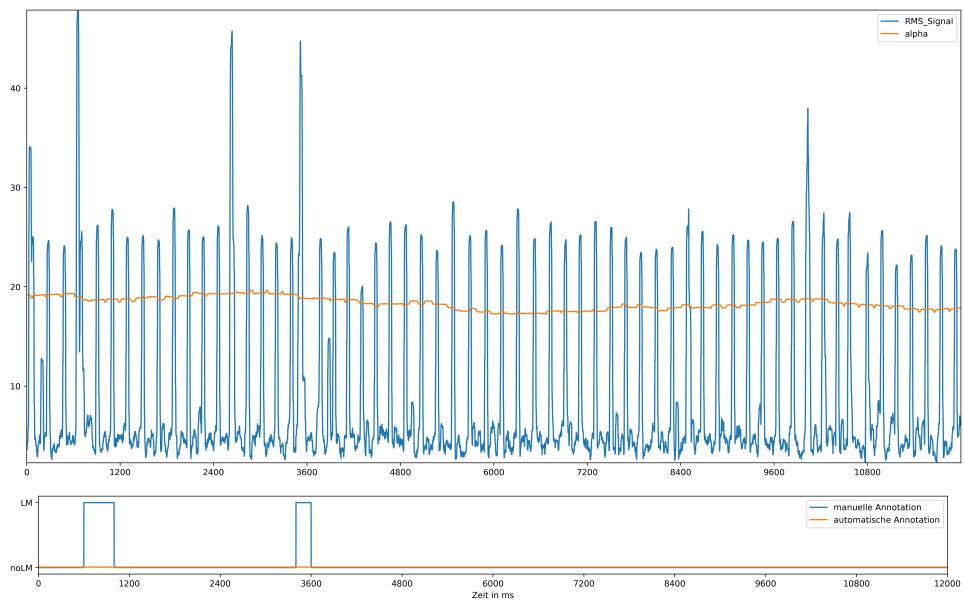


Abb. 6.4: EMG-Signal mit EKG Einkopplung (oben), bei der die manuell annotierten Beinbewegungen von dem Detektor nicht erkannt wurden.

7 Ergebnisse

7.1 Klassische Metriken

In den Tabellen 7.1 und 7.2 wird der in dieser Arbeit implementierte Detektor mit den Detektoren aus der Literatur anhand eventweise berechneter klassischen Metriken verglichen. Zu beachten ist, dass der Detektor größtenteils aus der Veröffentlichung von Moore et al. [5] übernommen wurde. In der Statistik wurden Annotationen ausgeschlossen, bei denen die jeweilige Metrik unendlich wurde. Die Mittelwerte werden stark durch die große Abweichung weniger Extremfälle beeinflusst. Um die Verteilung besser zu repräsentieren, wurde zusätzlich der Median angegeben.

Tab. 7.1: Stand der Technik für eventweise Auswertung zur Einordnung des implementierten Detektors, Teil 1; Der Zusatz 'r' bezeichnet einen rekonstruierten Wert

	[20]	[5]	[5]	[3]	[7]
Datensatz	15 RLS (24)	WSC (1073)	SSC (760)	HMC (70)	WSC (60)
Art des Klassifikators	statisch	dynamisch	dynamisch	dynamisch	statisch.
Berechnungsart	Event	Event [21]	Event [21]	Segment (1s)	Event
Sens	0.95	0.6	0.75	0.82	0.85
Prec	-	0.88	0.82	0.71	0.62
Spez	0.92	1	1	1	0.99
NPV	-	0.99	0.99	-	1
Acc	-	0.98	0.99	1	0.99
Cohens κ	-	0.71	0.87	0.73	0.72
$F1 - \text{Maß}$	0.93r	-	-	0.73	-
Korrealtion PLM/h	0.97	0.94	0.94	-	-
relative # PLM	1	0.63	0.92	0.99	1.3

In Tabelle 7.3 ist gesondert der Vergleich der segmentweisen Auswertung zu sehen. Zu beachten ist, dass in dem implementierten Detektor besonders die Metriken einen hohen Wert

Tab. 7.2: Stand der Technik für eventweise Auswertung zur Einordnung des implementierten Detektors, Teil 2

	[9]	[10]	[10]	[10]	implement. Detektor
Datensatz	WSC (60)	WSC (275)	SSC (177)	MrOS (348)	Uniklinik DD (5908)
Klassifikatorart	statisch	DNN	DNN	DNN	dynamisch
Berechnungsart	Event	Event	Event	Event	Event
Sens	0.96	0.9	–	–	0.6
Prec	0.47	0.81	–	–	0.68
Spez	0.98	–	–	–	0.84
NPV	1	–	–	–	0.78
Acc	0.98	–	–	–	0.72
Cohens κ	0.62	–	–	–	0.41
F1 – Maß	–	0.83	0.71	0.77	0.57
Korrealtion PLM/h	–	–	–	–	0.48
relative # PLM	2	–	–	–	1.4

aufweisen, die ein TN im Zähler haben.

Tab. 7.3: Stand der Technik für segmentweise Auswertung zur Einordnung des implementierten Detektors

	[3]	implement. Detektor
Datensatz	HMC (70)	Uniklinik DD (5908)
Klassifikatorart	dyn.	dyn.
Berechnungsart	Segment (1s)	Segment (1s)
Sens	0.82	0.49
Prec	0.71	0.6
Spez	1	0.99
NPV	–	0.98
Acc	1	0.97
Cohens κ	0.73	0.46
F1 – Maß	0.73	0.47
Korrealtion PLM/h	–	0.48
relative # PLM	0.99	1.4

7.2 Kostenfunktional

Das Kostenfunktional korreliert nur schwach mit den klassischen Metriken (siehe Tabelle 7.4).

Tab. 7.4: Korrelation zwischen Kostenfunktional und klassischen Metriken

	F1-Maß	Cohens κ	Spez	Prec	Sens	Acc	NPV
<i>Kostenfunktional</i>	-0.184	-0.125	0.251	0.175	-0.262	-0.159	-0.283

Die Tabelle 7.5 schlüsselt die Beiträge der verschiedenen Fehler zu der Gesamtkostenzahl auf. Zur besseren Einordnung der Mittelwerte sind zusätzlich noch die Anzahl der Annotationsen aufgelistet, die keine Kosten verursacht haben. Die relativen Kosten sind definiert durch die absoluten Kosten bezogen auf die manuelle PLM Anzahl. An dieser Aufschlüsselung der Kosten ist interessant, dass wenig Fehler durch Xto1- und 1toX- Matches entstanden sind. Die Hauptfehlerquelle ist eine Verminderung der PLM Anzahl aufgrund von FN (41 Prozent).

Tab. 7.5: Beiträge der Fehlerarten an den Gesamtkosten. (+) hinter der Fehlerart beschreibt, dass dieser Fehler zu einer Erhöhung des PLM-Indexes geführt hat. (-) steht analog für eine Verminderung.

Fehlerart	Mittelwert	# Dateien ohne Kosten
<i>FP</i> (+)	42	1341
<i>1toX</i> (+)	1.25	4718
<i>Zeitungenaugkeiten</i> (+)	72	1108
<i>FN</i> (-)	117	873
<i>Xto1</i> (-)	2.05	4274
<i>Zeitungenaugkeiten</i> (-)	72.9	988
<i>absolute Kosten</i>	285	488
<i>relative Kosten</i>	2.26	488

Die Differenz aus ergebniserhöhenden Fehlern und ergebnisvermindernden Fehlern (K_{diff}) korreliert gut mit der Differenz aus automatisch annotierten PLM und manuell annotierten PLM (siehe Bild 7.1). Der Korrelationskoeffizient liegt bei 0.987. Dieser Wert beschreibt den Zusammenhang zwischen den erklärten Fehlern aus dem Kostenfunktional und dem medizinisch wichtigen PLM-Index.

Die Verteilungen der absoluten und relative Kosten (absolute Kosten bezogen auf die manuelle PLM-Anzahl) sind in Abb. 7.2 dargestellt. Es ist zu erkennen, dass Annotationen an denen höhere Kosten entstanden sind, seltener sind.

7.3 Verbesserung der Einordnung des Detektors

Die Metriken zu Verbesserung der Einordnung des Detektors sind in Tabelle 7.6 zu sehen. Der Anhang "mean" beschreibt den Mittelwert der Verteilung über die Differenz der Start- oder Endzeitpunkte von TP. Die Anhand β_{std} beschreibt analog die Standardabweichung der Verteilung.

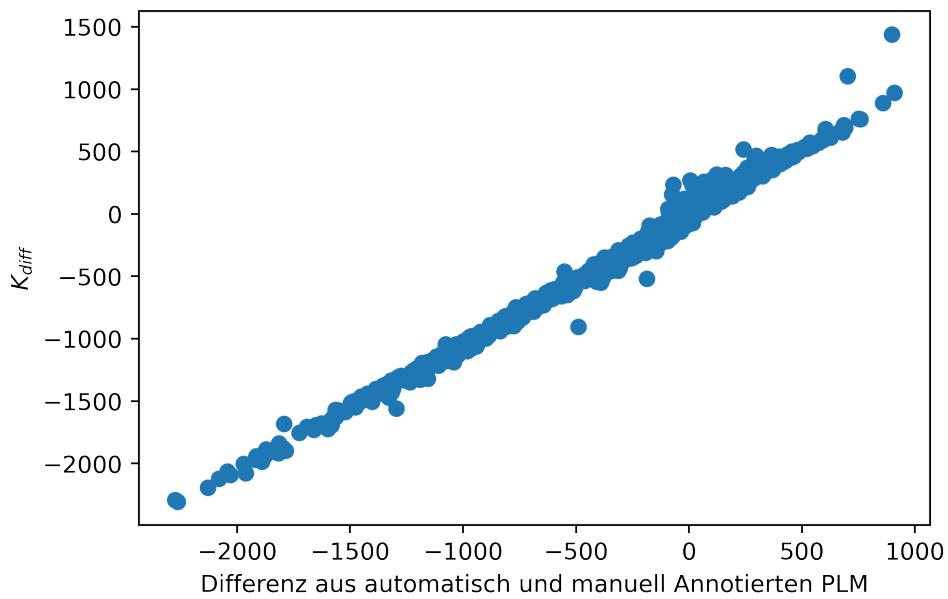


Abb. 7.1: Veranschaulichung der Korrelation von Differenz aus ergebniserhöhenden Fehlern und ergebnisvermindernden Fehlern (K_{diff}) und Differenz aus automatisch annotierten und manuell annotierten PLM.

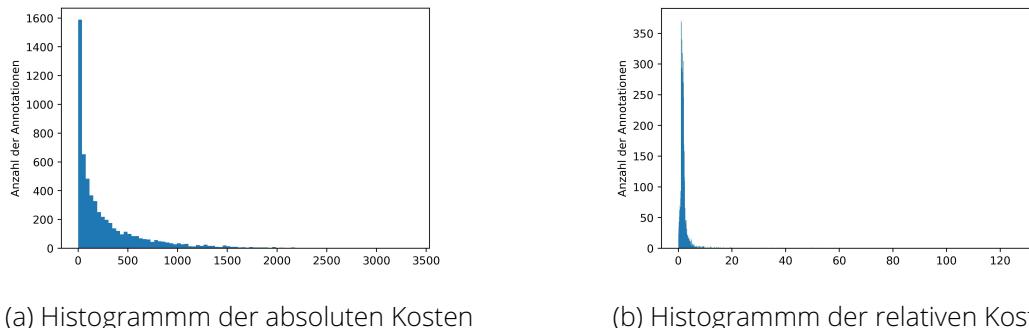


Abb. 7.2: Darstellung der Kostenverteilung mit einer automatisch erstellten Balkenweite

Zur besseren Einordnung der LM Verhältnisse wurde die Differenz der Anzahl zwischen automatisch und manuell annotierten LM berechnet. Der Mittelwert beträgt -87.1. Da das Vorzeichen negativ ist, wurden im Mittel mehr manuelle LM annotiert.

Das Histogramm über alle Schwerpunkttdifferenzen ist in Grafik 7.3 dargestellt. Es sind keine Teilmengen im Histogramm zu erkennen, die die Festsetzung eines Schwellwertes erlauben würden.

Dieses Histogramm zeigt die falsch positiv Rate. Die Häufung um den Wert 0.5 deutet darauf hin, dass FP und TN ähnlich häufig vorkommen. In dem Histogramm sind nur die Dateien mit ohne FP auffällig als eigenständige Teilmenge.

Tab. 7.6: Metriken aus dem Kapitel 5.3: "Verbesserung der Einordnung des Detektors"

	Mittelwert	Median
<i>Falsch positiv Rate</i>	0.55	0.56
<i>Verstöße gegen die LM – Zeitkriterien</i>	0	0
<i>Anzahl der Xto1 – Matches</i>	2.9	0
<i>Anzahl der 1toX – Matches</i>	1.6	0
<i>Schwerpunkt</i>	136 Sekunden	50 Sekunden
<i>LM – Anzahl (automatisch / manuell)</i>	1.6	0.89
<i>Startzeitpunkte_{mean}</i>	-0.38 Sekunden	-0.26 Sekunden
<i>Startzeitpunkte_{std}</i>	0.83 Sekunden	0.67 Sekunden
<i>Endzeitpunkte_{mean}</i>	0.76 Sekunden	0.64 Sekunden
<i>Endzeitpunkte_{std}</i>	1.2 Sekunden	1 Sekunde

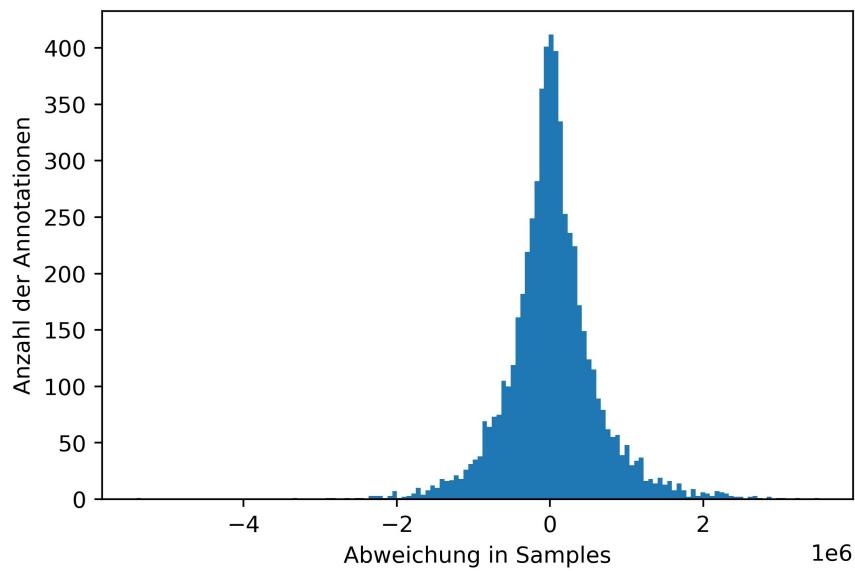


Abb. 7.3: Histogramm über die Differenz der Schwerpunkte aus automatischer und manueller Annotation. Die Breite der Balken wurde für die gesamte Arbeit automatisch erstellt.

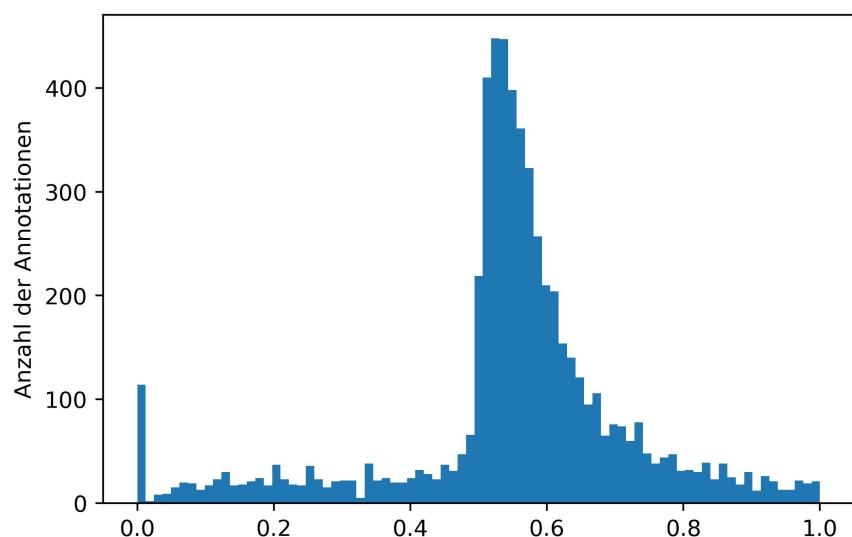


Abb. 7.4: Histogramm über die falsch-positiv Rate. Die Breite der Balken wurde automatisch erstellt.

8 Diskussion

8.1 Datensatz

Die Analyse der Metadaten zu dem Datensatz zeigt, dass die Probanden hauptsächlich einer älteren Demographie angehören und überwiegend männlich sind. Die verfügbaren Informationen beziehen sich allerdings nur auf circa die Hälfte des gesamten Datensatzes. Es kann daher nicht exakt bestimmt werden für welche Patienten die gefundenen Ergebnisse gelten. Es wurden außerdem nur die Beinbewegungen betrachtet, die im Schlaf aufgetreten sind.

An dem gegebenen Datensatz schneidet der Detektor im Vergleich zu den Detektoren aus der Literatur relativ schlecht ab. Dies ist besonders an den vergleichsweise kleinen Korrelationskoeffizienten des PLM-Indexes (0.48) sowie Cohens κ (0.41) und F1-Maß (0.57) zu erkennen. Insbesondere sind diese Ergebnisse wesentlich schlechter als in der Veröffentlichung von Moore et al. [5], obwohl diese als Vorlage für den Detektor gedient hat.

Dies könnte durch die Abweichung in der Implementierung zu dem Detektor von Moore et al. sein (siehe 6.2). Unter der Annahme, dass nur selten LM aufgrund von atembezogenen Events gelöscht werden mussten, sollte die größte Abweichung durch die fehlende Filterung des EKG zustande kommen.

In den Abbildungen 6.4 und 6.3 erhöhen die Einkopplungen das Grundrauschen auf 20-30 Millivolt. Da die Ausschläge im EMG der Beinbewegungen wesentlich höher sind, sollte diese Art von Störung nur einen geringen Einfluss haben und kann alleine die starke Abweichung nicht erklären.

In der Abbildung 8.1 ist ein EMG-Signal zu sehen, welches keine nennenswerten Ausschläge aufgezeichnet hat. Die manuelle Annotation erkennt in diesem Signal sehr viele Events und es ist zu vermuten, dass das Signal von einer Annotationsunterstützung erstellt und nicht von Schlafspezialisten begutachtet wurde, bevor es in den Datensatz aufgenommen wurde. Das Beispiel in Abbildung 8.2 zeigt eine zeitlich begrenzte Verschlechterung der Signalqualität, welche von dem dynamischen Schwellwert erkannt wird. Zu beachten ist, dass die manuelle Annotation in diesem Zeitraum Events aufweist. Wahrscheinlich war nicht für alle Patienten eine explizite Untersuchung auf Beinbewegungen indiziert, sodass die Annotation möglicherweise nicht immer von Schlafspezialisten überprüft wurde. Die manuelle Annotation wurde vermut-

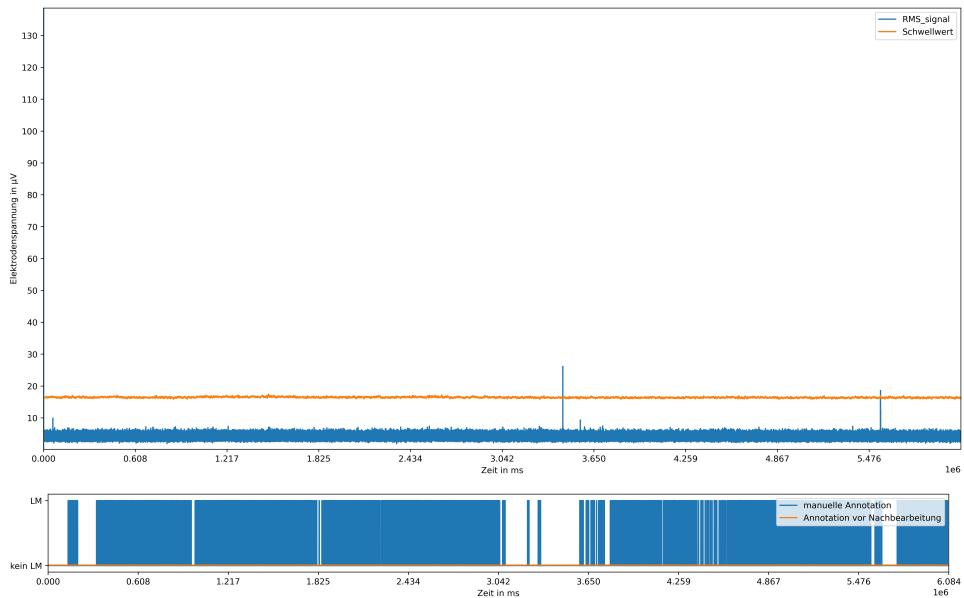


Abb. 8.1: Beispiel EMG-Signal mit Schwellwert (oben) und Annotationspaar (unten), bei der augenscheinlich keine Beinbewegungen stattgefunden haben. Die manuelle Annotation weist trotzdem viele Events auf.

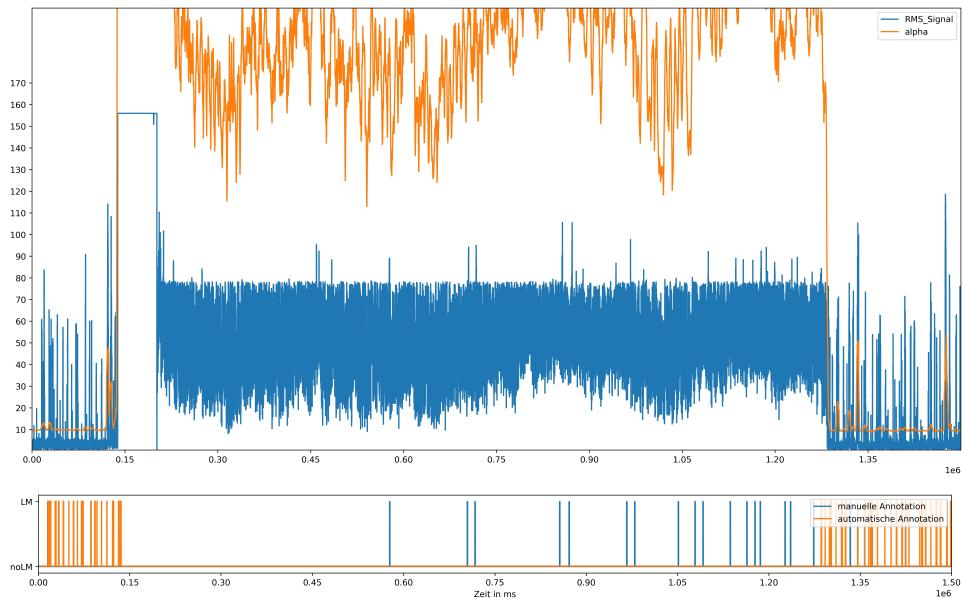


Abb. 8.2: Ausschnitt eines EMG-Signals mit zeitweise schlechter Signalqualität (oben); laut manueller Annotation wurden in diesem Bereich trotzdem LMs gefunden (unten)

lich in diesen Fällen von einer Annotationsunterstützung erstellt. Der Detektor wird in diesen Fällen auch an der manuellen Annotation bewertet und erzielt damit wesentlich schlechtere Ergebnisse. Die Grenzfälle mit einer sehr großen Abweichung zwischen manueller und automatischer Annotation beeinflussen den Mittelwert stark. Die hier erzeugten Ergebnisse erlauben somit keinen Vergleich mit Detektoren, die auf anderen Datensätzen angewendet wurden.

Um die Qualität des Datensatzes zu verbessern, empfiehlt die AASM für die Schwellwert-

Klassifikation mit $8\mu V$ und $2\mu V$ Schwellwerten ein Ruhe-EMG-Signal dessen Amplitude kleiner als $\pm 10\mu V_{pp}$. Da hier in dem gewählten Detektor ein dynamischer Schwellwert genutzt wird, kann für diese Arbeit nicht definiert werden ab welcher Grundrauschamplitude Signale ausgeschlossen werden können.

Selbst bei einem hochqualitativen Datensatz kann die manuelle Annotation von der Wirklichkeit abweichen. In der Veröffentlichung von Wetter et al. wird die Sensitivität zwischen zwei menschlichen Experten mit 97% und die Präzision mit nur 92% angegeben. Diese Abweichung ist für die Bewertung eines Detektors kritisch, da selbst ein hypothetisch perfekt funktionierender Detektor keine perfekten Ergebnisse hervorbringen würde. In Huang et al. wurde der Datensatz zusätzlich von Schlafexperten mit jahrelanger Erfahrung ausgewertet und hat somit zu einer besseren Bewertung des Detektors geführt (87.7% im Vergleich zu 94.4% Übereinstimmung). Bei Carvelli et al. wurde ein Konsens aus mehreren Schlafspezialisten durchgeführt, um die Wirklichkeit zu approximieren. Diese Optionen waren für den gegebenen Datensatz nicht vorhanden. Da die Detektoren nur untereinander verglichen werden müssen reicht es aus die gleichen Ausgangsbedingungen für die Detektoren zu schaffen. Um die gleichen Bedingungen herzustellen, sollten Detektoren nur verglichen werden, wenn sie auf dem gleichen Datensatz getestet wurden. Dies wird deutlich in dem Vergleich der Ergebnisse des Detektors von Moore et al., welcher auf den Daten von Alvarez-Estevez wesentlich schlechtere Ergebnisse liefert obwohl der gleiche Algorithmus verwendet wurde.

8.2 Kostenfunktional

Dass das Kostenfunktional nur sehr schwach mit den klassischen Metriken korreliert, ist damit zu begründen, dass die Ergebnisse des hier implementierten Detektors stark von den manuellen Annotationen abweichen. Dies ist insbesondere an dem kleinen Korrelationskoeffizienten der PLM-Indices zu erkennen.

Ein Problem bei der eventweisen Klassifikation ist, dass ein Abtastwert den Unterschied in der Anzahl der LM bedeuten kann. Da die Entscheidung, welche der LM einander zugeordnet werden, intuitiv von der Umgebung abhängt, ist es nicht einfach Regeln zu definieren, wann ein TP gezählt werden sollte.

Beim Bewerten des Detektors wurden Fehler ausschließlich anhand des binären Annotationssignals ohne vertiefende Fachkenntnis eines Schlafspezialisten definiert. Diese hätten möglicherweise bestimmte Fehler anders beurteilt, sodass sie in dem medizinischen Kontext ein aussagekräftigeres Ergebnis lieferten.

Aufgrund der Definition der Güte 5.1 der relativen Kosten kommt es außerdem zu einer Verzerrung der Interpretation bei gut funktionierenden Detektoren. Falls die relativen Kosten sehr klein sind und die Differenz der relativen Kosten zweier Detektoren relativ klein ist, wird es aufgrund der Nichtlinearität der Güte zu einem unerwartet großen Unterschied kommen. Die Güte ist also für kleine Kosten schlechter konditioniert. Das hat aber zusätzlich den Vorteil, dass sehr ähnliche Detektoren gut voneinander getrennt werden können.

Aufgrund dessen, dass für das Kostenfunktional Fehler gezählt werden, gibt es keine Obergrenze die der Wert annehmen kann. Die untere Grenze liegt bei Null. Das bedeutet auch,

dass die Güte (zumindest bei unendlicher Signallänge) unendlich sein kann, wenn keine Fehler gemacht werden. Falls manuell keine PLM-Serien erkannt werden, sind die relativen Kosten wenig aussagekräftig, und gehen gegen unendlich, auch wenn nur ein Fehler gemacht wird. Wenn zusätzlich keine PLM-Serien vom Detektor erkannt werden, sind Kosten und Güte nicht definiert. Signale ohne manuelle Annotation sollten aus diesem Grund nicht zur Bewertung verwendet werden.

Ein weiteres Problem ist, dass es bei einer PLM-Serie mehrere Gründe geben kann, warum diese fehlerhaft sind. So kann es beispielsweise dazu kommen, dass der erste TP eines 1toX-Matchings (welches die Serie verändert) gleichzeitig fälschlicherweise die [5-90] Sekunden Intervallgrenze überschreitet. Diese Fehler werden in der Aufschlüsselung der Kosten bei beiden Gründen sichtbar. Im Kostenfunktional wird in solchen Fällen nur ein Fehler gezählt. Das Problem mit der Berechnung des Kostenfunktional ist, dass es möglicherweise noch wesentlich mehr Grenzfälle gibt, als in dem in dieser Arbeit in Betracht gezogen wurden. Es eignet sich hierfür eine Implementierung, bei der für jedes LM die Gründe für die Fehler analysiert werden, anstatt für jeden Fehlergrund die LM zu suchen, die dagegen verstößen.

Die Formel 5.2

$$\frac{\text{ergebniserhöhende Fehler} - \text{ergebnisverkleinernde Fehler}}{\text{automatisch annotierte PLM} - \text{manuell annotierte PLM}}$$

beschreibt inwieweit die Fehler, die von dem Detektor verursacht wurden, (PLM Differenz) von dem Kostenfunktional (Differenz aus Zuvielzählen und Zuwenigzählen) erklärt werden können. Da diese beiden Differenzen stark korreliert sind ($r^2 = 0.99$), kann geschlussfolgert werden, dass Kostenfunktional eine zuverlässige Beschreibung der Güte des Detektors ist. Die weitere Untersuchung auf unbekannte Grenzfälle vernachlässigt werden, da diese das Ergebnis nur wenig beeinflussen.

Das Kostenfunktional ist auf die Kriterien der AASM zur Erkennung von PLM spezialisiert und kann deswegen auch nur in diesem Rahmen angewandt werden. Sollten sich diese Regeln ändern oder wird ein anderes Regelwerk bevorzugt, müssen die hier vorgeschlagenen Berechnungen angepasst werden und es lassen sich somit auch keine Vergleiche mehr zwischen den Bewertungen vor dieser Regeländerung herstellen. Die Prinzipien, für die Erstellung des Kostenfunktional können auf ähnliche regelbasierte medizinische Fragestellungen angewandt werden.

8.3 Verbesserung der Einordnung des Detektors

Da einige Gründe bekannt sind, aus denen die manuelle Annotation fehlerhaft sein könnte, bietet es sich an, die erhobenen Metriken aus dem Kapitel Verbesserung der Einordnung des Detektors (Kapitel 5.3) zu nutzen, um die Qualität des Datensatzes zu verbessern. Im Folgenden sind einige dieser Gründe aufgezählt und interpretiert. Zuerst kann die Annahme getroffen werden, dass der Detektor seine Annotationsentscheidungen nur anhand der vorliegenden Daten trifft. Bei einer manuellen Annotation können andere Faktoren wie Monotoniemüdigkeit (siehe 1) eine Rolle spielen. Sind beispielsweise Annotationspaare vorhanden, bei denen die manuelle Annotation ausschließlich im vorderen Teil der Nacht befinden, während automatische Annotationen über die ganze Nacht verteilt sind, könnte das bedeuten, dass die manuelle

Annotation aus unbekannten Gründen abgebrochen wurde.

Die Teilmenge an Daten die vorzeitig abgebrochen wurde, weist tendenziell eine größere Abweichung der Schwerpunktdifferenz auf. Ein Extrembeispiel dafür könnte in Abbildung 8.3 zu sehen sein. Hier könnte die hohe Eventdichte für einen frühzeitigen Abbruch der manuellen Annotation geführt haben. Die Schwerpunktdifferenz liegt in dem folgenden Beispiel bei 3.03 Stunden.

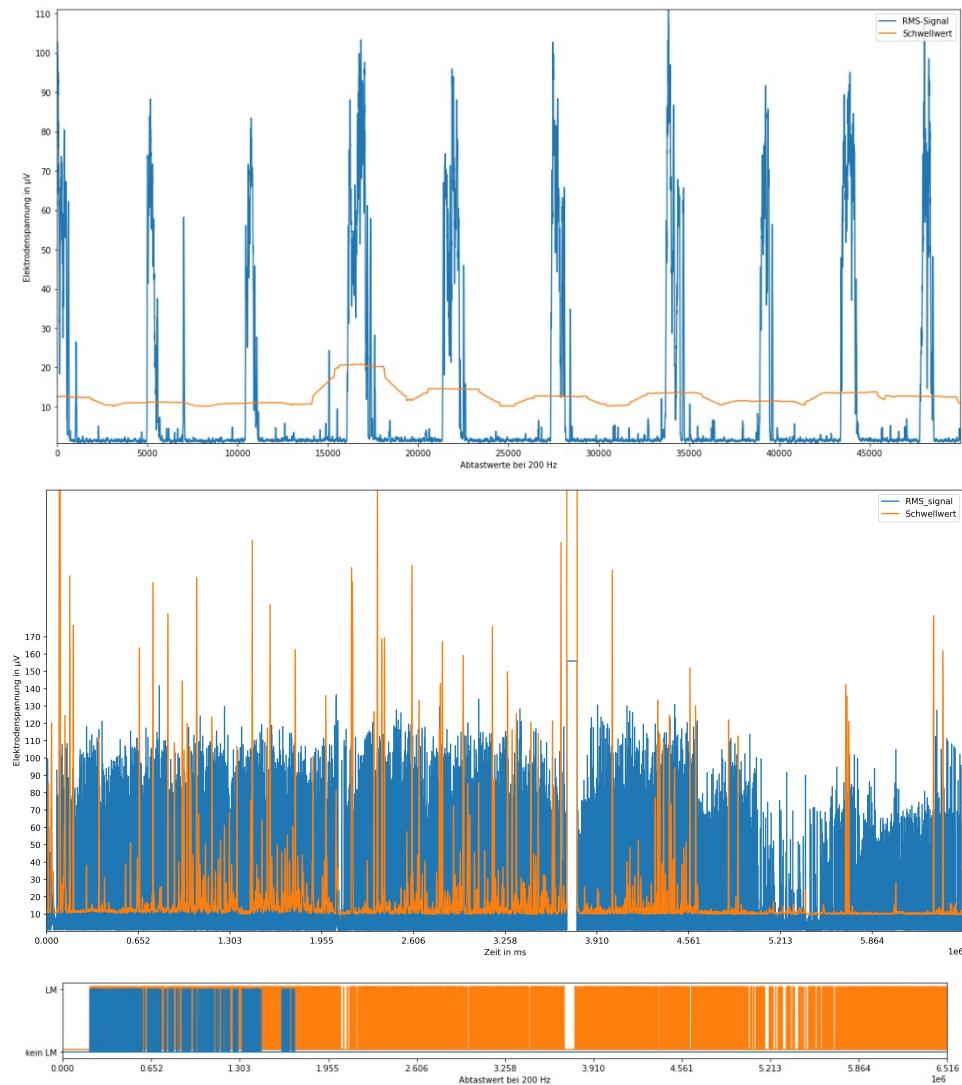


Abb. 8.3: Ausschnitt eines EMG-Signals mit sehr hoher Eventdichte(oben); das selbe EMG-Signal über die ganze Nacht dargestellt (mittig) mit zugehörigen Annotationen (unten); welche eine hohe Schwerpunktdifferenz aufweisen.

Im Histogramm 7.3 ist eine solche Teilmenge weit entfernt von dem Mittelwert. Es kann jedoch ohne medizinisches Fachwissen kein Schwellwert definiert werden, unter dem die Daten aussortiert werden könnten.

Bei einer segmentweisen manuellen Annotation wird der tatsächliche Startzeitpunkt in die Segmente diskretisiert. Da die meisten Detektoren quasizzeitkontinuierlich auswerten, kann dieser bei ausreichend guter Detektionsqualität die Wirklichkeit zeitlich genauer darstellen. Gäbe es beispielsweise einen ungenau annotierenden Angestellten, der jede echte Beinbe-

wegung eine Sekunde vorher (gerundet auf die Segmentgrenzen) schon als positiv markiert, würde der daraus resultierende Mittelwert der Startzeitpunkte über diese Nacht gesehen einen Ausreißer im Histogramm produzieren. Diese Metrik könnte also einen Hinweis darauf geben, welche Dateien genauer untersucht werden sollten. Es könnte argumentiert werden, dass Annotationen, welche im Mittel ungenauer sind als die Segmentgrenzen es zulassen, eher hinderlich für die Bewertung des Detektors sind, da die Startzeitpunkte den PLM-Index direkt beeinflussen.

Die Löschung von LM, die in der Nähe von atembezogenen Events stattgefunden haben, ist beispielsweise für den Detektor, ohne die benötigten Atem-Signale nicht zu erkennen. Unter der Annahme, dass diese Art von Events nicht bei allen Patienten auftritt, könnte es sinnvoll sein, die Nächte genauer zu untersuchen, bei der die falsch positiv Rate stark von dem Mittelwert abweicht. Anhand des Histogramms 7.4 aus den Ergebnissen, ist nicht zu erkennen, ab welchem Wert eine starke Abweichung zu vermuten ist. Diese Abweichung wird vermutlich erst bei sehr guten Ergebnissen sichtbar.

Ein sehr hohes LM Verhältnis bedeutet, dass der Detektor wesentlich öfter auf einen Anstieg im EMG-Signal reagiert als das medizinische Personal es tun würde. Diese Metrik könnte also einen Hinweis darauf geben, wie sinnvoll es ist, den Erkennungsschwellwert anzuheben. Die Aussage des Mittelwertes und des Medians ist für den implementierten Detektor widersprüchlich. Betrachtet man das die Differenz der gefundenen LM (automatisch gefundene LM – manuell gefundene LM) wird ersichtlich, dass der Detektor in den meisten Fällen weniger LM als das medizinische Personal erkennt. Im Durchschnitt erkennt der Detektor 87.1 von 226 Beinbewegungen zu wenig. In diesem Fall könnte also die Formel für den Schwellwert angepasst werden, sodass mehr LM erkannt werden.

Eine hohe Anzahl an 1toX-Matches impliziert, dass der Detektor das Ende eines LM zu früh erkennt. In diesen Fällen könnte beispielsweise eine Änderung in den Zeiten der Nachbearbeitung des Annotationssignals oder die Definition des unteren Schwellwertes angepasst werden. Die Anzahl der Xto1-Matches trifft entgegengesetzte Aussagen. In dem hier implementierten Detektor sind beide Anzahlen nicht besonders hoch. Falls es Verstöße gegen die definierte Länge der LM gibt, könnte das bedeuten, dass der Detektor entweder Ausschläge im EMG als LM klassifiziert, die nicht aus Beinbewegungen stammen oder die Dauer eines richtig gefundenen LM wird falsch eingeschätzt. Diese Metrik ist für den Detektor per Definition bei Null, da in der Nachbearbeitung des Annotationssignals die Länge der gefundenen LM überprüft wird.

Die Mittelwerte der Verteilung (über alle Nächte berechnet) kann genutzt werden, um den Detektor auf systematische Fehler zu untersuchen. Liegen beide Mittelwerte bezogen auf die manuelle Annotation zeitlich gesehen im positiven Bereich, kann das darauf hindeuten, dass die Vorverarbeitung des EMG-Signals einen zu starken Tiefpasscharakter aufweist und somit die Flanken der LM weniger schnell ansteigen und abfallen. Dies würde dazu führen, dass alle automatisch erkannten LM leicht zeitlich versetzt zu den korrespondierenden manuell erkannten LM sind. Es ist zu vermuten, dass in der manuellen Annotation auch systematische Fehler gemacht wurden. Da die Mittelwerte der Startzeitpunkte negativ und die Mittelwerte der Endzeitpunkte positiv sind, lässt sich vermuten, dass das medizinische Personal, anstatt exakt ab- und aufzurunden, lieber das ganze Segment als positiv definiert hat, in dem ein LM stattgefunden hat. Aus diesem Grund können systematische Fehler nur zwischen den Detektoren

untereinander verglichen werden.

Maschinelle Lernalgorithmen sind aufgrund der hohen Anzahl der Parameter in der Lage sich an den Trainingsdatensatz stark anzupassen. Es kann also passieren, dass der Algorithmus die Segmentierung der Annotation mitlernt, obwohl diese in dem Eingangssignal nicht vorhanden sind. Die Verteilung der LM-Startzeitpunkte hat aufgrund der Segmentierung der manuellen Annotation eine bestimmte Standardabweichung. Wird von einem Detektor, der mithilfe eines solchen Lernalgorithmus arbeitet diese Standardabweichung unterschritten, könnte eine Überanpassung an die Segmentgrenzen stattgefunden haben. Diese Überanpassung lässt sich anhand dieser Metrik feststellen. Ein Neuronales Netz welches die Segmentgrenzen auswendig gelernt hat, würde auf einem anderen Datensatz mit quasikontinuierlicher manueller Annotation eine unerwartet hohe Standardabweichung aufweisen und damit auch zu schlechteren Ergebnissen führen. Die Struktur eines Schwellwertdetektor hat zu wenig wählbare Parameter, um Segmentgrenzen zu erlernen. Daher kann die Metrik hier nicht ausgewertet werden.

8.4 Optimierung des Detektors

Es sollte darauf geachtet werden, dass die Bewertung des Detektors nicht mit der Optimierung des Detektors vermischt wird. Da die Optimierung anhand der Metriken im Kapitel Verbesserung der Einordnung des Detektors (Kap. 5.3) unabhängig von dem Kostenfunktional ist, können diese genutzt werden. Es wird trotzdem empfohlen die Bewertung des Detektors anhand von Daten durchzuführen, die nicht zu einer Veränderung des Detektors geführt haben. Das Kostenfunktional lässt sich durch die Zählweise nicht mathematisch ableiten und kann somit nicht durch Optimierungsverfahren genutzt werden, die auf Gradienten beruhen. Trotzdem lassen sich Detektoren aufgrund der Definition einer Güte, welche als Inverse der Kosten definiert ist, durch Verfahren wie genetischen Algorithmen verbessern.

9 Fazit und Ausblick

Aus der vorliegenden Arbeit ging hervor, dass die klassischen Metriken ungeeignet für die Bewertung der Eventdetektoren im Kontext der periodischen Beinbewegungen sind. Diese sind insbesondere ungeeignet, da wichtige Informationen über zeitliche Zusammenhänge verloren gehen, Metriken sich untereinander qualitativ widersprechen können und auch mit trivialen Detektoren perfekte Ergebnisse erzielt werden können.

Es wurde daher ein Kostenfunktional vorgestellt, welches auf dem medizinisch relevanten PLM-Index basiert und die Fehler, die vom Detektor verursacht wurden, quantifiziert darstellt. Die aufgetretenen Fehler konnten sehr gut ($r^2 = 0.99$) durch das Kostenfunktional erklärt werden. Durch dieses Vorgehen wird eine einfache und eindeutige Vergleichbarkeit ermöglicht, indem jedem Detektor genau ein Gütwert zugewiesen wird.

Das vorgestellte Kostenfunktional ist besonders dann nützlich, wenn Vergleichswerte existieren, um neu entwickelte Detektoren einordnen zu können. Dafür sollte das Kostenfunktional in zukünftigen Arbeiten auf die bereits existierenden Detektoren angewendet werden und ein neuer Stand der Technik festgelegt werden. Für qualitative Aussagen sollte ein möglichst öffentlich zugänglicher Datensatz definiert werden, an dem alle Detektoren und Metriken angewendet werden können. Im Optimalfall hat dieser Datensatz eine gute manuelle Annotation. Primär ist jedoch wichtig, dass die Qualität des Datensatzes für alle Detektoren gleich ist, damit die Unterschiede in den Datensätzen nicht fälschlicherweise mit in die Bewertung des Detektors einfließen.

Für eine Weiterentwicklung des hier vorgestellten Kostenfunktionalen sollten insbesondere Grenzfälle betrachtet werden, um herauszufinden, ob es weitere relevante Fehlerquellen gibt, bei denen der Algorithmus nicht erwartungsgemäß funktioniert.

Die im Kapitel 5.3 präsentierten Ideen können genutzt werden, um einerseits einen Datensatz zu bereinigen, indem Ausreißer gefunden werden können und andererseits fundierte Aussagen über die Funktionsweise des Detektors zuzulassen. Diese Aussagen sind jedoch nur als Hinweise zu verwenden, an welchen Stellen genauere Untersuchungen lohnenswert sind. Für den verwendeten Detektor konnten keine definitiven Aussagen getätigter werden. Die Wirksamkeit der Ideen und die Festlegung von Schwellwerten könnte sich jedoch bei anderen Detektoren als nützlich erweisen und sollte in zukünftigen Arbeiten mithilfe von Experten durchgeführt und überprüft werden.

Quellenverzeichnis

- [1] Adriana M. Adami u. a. „A system for assessment of limb movements in sleep“. In: 2013, S. 419–423. ISBN: 9781467358019. DOI: 10.1109/HealthCom.2013.6720712.
- [2] Richard Allen u. a. *Course 16: Sleep related movements: Standards for scoring, interpreting, reporting, and publishing*.
- [3] Diego Alvarez-Estevez. „A new automatic method for the detection of limb movements and the analysis of their periodicity“. In: *Biomedical Signal Processing and Control* 26 (Apr. 2016), S. 117–125. ISSN: 17468108. DOI: 10.1016/j.bspc.2016.01.008.
- [4] Diego Alvarez-Estevez und Roselyne M Rijsman. „Computer-assisted analysis of polysomnographic recordings improves inter-scorer associated agreement and scoring times“. In: (). DOI: 10.1101/2022.03.23.22272801. URL: <https://doi.org/10.1101/2022.03.23.22272801>.
- [5] Hyatt Moore IV und andere. „Design and Validation of a Periodic Leg Movement Detector“. In: *PLoS ONE* (2014).
- [6] Matthew Reyna und andere. „Rethinking Algorithm Performance Metrics for Artificial Intelligence in Diagnostic Medicine“. In: (2022).
- [7] Raffaele Ferri und andere. „Computer-Assisted Detection of Nocturnal Leg Motor Activity in Patients with Restless Legs Syndrome and Periodic Leg Movements During Sleep“. In: (2005).
- [8] Sharadha Kolappan und andere. „A Low-cost Approach for Wide-spread Screening of Periodic Leg Movements Related to Sleep Disorders“. In: (2017).
- [9] Thomas C. Wetter und andere. *An Automatic Method for Scoring Leg Movements in Polygraphic Sleep Recordings and Its Validity in Comparison to Visual Scoring*. *Sleep* 27(2):324-8. DOI:10.1093/sleep/27.2.324. 2004.
- [10] Lorenzo Carvelli u. a. „Design of a deep learning model for automatic scoring of periodic and non-periodic leg movements during sleep validated against multiple human experts“. In: *Sleep Medicine* 69 (Mai 2020), S. 109–119. ISSN: 18785506. DOI: 10.1016/j.sleep.2019.12.032.

- [11] Matteo Cesari u. a. *Probabilistic Data-Driven Method for Limb Movement Detection During Sleep*. 2018. ISBN: 9781538636466. DOI: 10.0/Linux-x86_64.
- [12] Sudhansu Chokroverty und Sushanth Bhat. *Atlas of Sleep Medicine*. Elsevier B.V, 2014.
- [13] Navin Cooray u. a. *Automated Movement Detection with Dirichlet Process Mixture Models and Electromyography*. 2022.
- [14] Tatjana Crönlein, Wolfgang Galetke und Peter Young. *Schlafmedizin 1 x 1*. Springer-Verlag, 2017.
- [15] Rachel Draelos. *Measuring Performance: AUPRC and Average Precision*. 2. März 2019. URL: <https://glassboxmedicine.com/2019/03/02/measuring-performance-auprc/> (besucht am 19.02.2023).
- [16] Jonas Ebbecke. *Elektromyographie im Überblick*. 12.Jan. 2021. URL: <https://www.biomechanist.net/de/elektromyographie-im-uberblick/> (besucht am 19.02.2023).
- [17] Markus Gall u. a. „Automated Detection of Movements during Sleep Using a 3D Time-of-Flight Camera: Design and Experimental Evaluation“. In: *IEEE Access* 8 (2020), S. 109144–109155. ISSN: 21693536. DOI: 10.1109/ACCESS.2020.3001343.
- [18] Adrienne Heinrich u. a. „Robust and sensitive video motion detection for sleep analysis“. In: *IEEE Journal of Biomedical and Health Informatics* 18 (3 2014), S. 790–798. ISSN: 21682194. DOI: 10.1109/JBHI.2013.2282829.
- [19] Max Hirshkowitz und Amir Sharafkhaneh. *Sleep Disorders Medicine*. Springer-Verlag, 2017.
- [20] Andy S. Huang u. a. „MATPLM1, A MATLAB script for scoring of periodic limb movements: Preliminary validation with visual scoring“. In: *Sleep Medicine* 16 (12 Dez. 2015), S. 1541–1549. ISSN: 18785506. DOI: 10.1016/j.sleep.2015.03.008.
- [21] Hyatt Moore IV. *roc_dlg.m*. 14. Mai 2013. URL: https://github.com/informaton/sev/blob/%2010efbecb206c262833410867d118963fe4bd81f2/roc_dlg.m (besucht am 19.02.2023).
- [22] Roman Kusche. *Mehrkanal-Bioimpedanz-Instrumentierung*. Springer-Verlag, 2019.
- [23] Carlo De Luca. „Physiology and Mathematics of Myoelectric Signals“. In: (1979).
- [24] Alexandre R. Abreu Richard B. Berry Stuart F. Quan. „AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications, Version 2.6“. In: *American Academy of Sleep Medicine* (2020).
- [25] M. Roessen, M. Thijssen und B. Kemp. „Semi-automatic detection of leg movements: program features and scoring results“. In: (1998).
- [26] Judith Schlupf. *Motorische Endplatte*. Ernst Klett Verlag GmbH, 2010.
- [27] Mehrnaz Shokrollahi u. a. „Nonnegative matrix factorization and sparse representation for the automated detection of periodic limb movements in sleep“. In: *Medical and Biological Engineering and Computing* 54 (11 Nov. 2016), S. 1641–1654. ISSN: 17410444. DOI: 10.1007/s11517-015-1444-y.

- [28] Ambra Stefani u. a. „Validation of a leg movements count and periodic leg movements analysis in a custom polysomnography system“. In: *BMC Neurology* 17 (1 Feb. 2017). ISSN: 14712377. DOI: 10.1186/s12883-017-0821-6.
- [29] Boris A Stuck u. a. *Praxis der Schlafmedizin*. Springer-Verlag, 2018.
- [30] N. Tauchmann und T. Pollmacher. *Automatic detection of periodic leg movements (PLM)* [2]. 1996. DOI: 10.1111/j.1365-2869.1996.00273.x.
- [31] Unbekannt. *AASM AI/Autoscoring Pilot Certification is coming soon*. 26. Aug. 2022. URL: <https://aasm.org/aasm-ai-autoscoring-pilot-certification-is-coming-soon/> (besucht am 19.03.2023).

Anhang

A Anhang

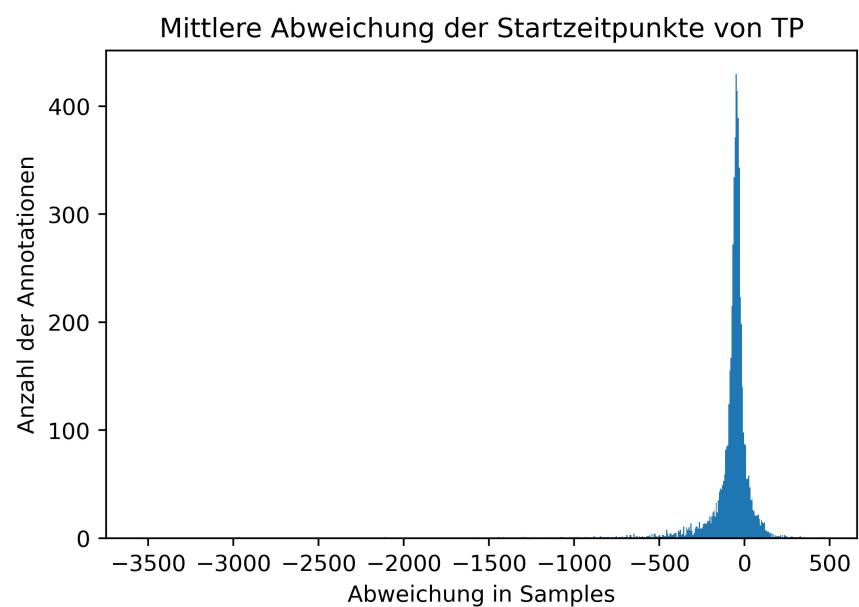


Abb. 1.1: Histogramm über die Mittlere Abweichung der Startzeitpunkte von TP bei einer Abtastfrequenz von 200 Hertz

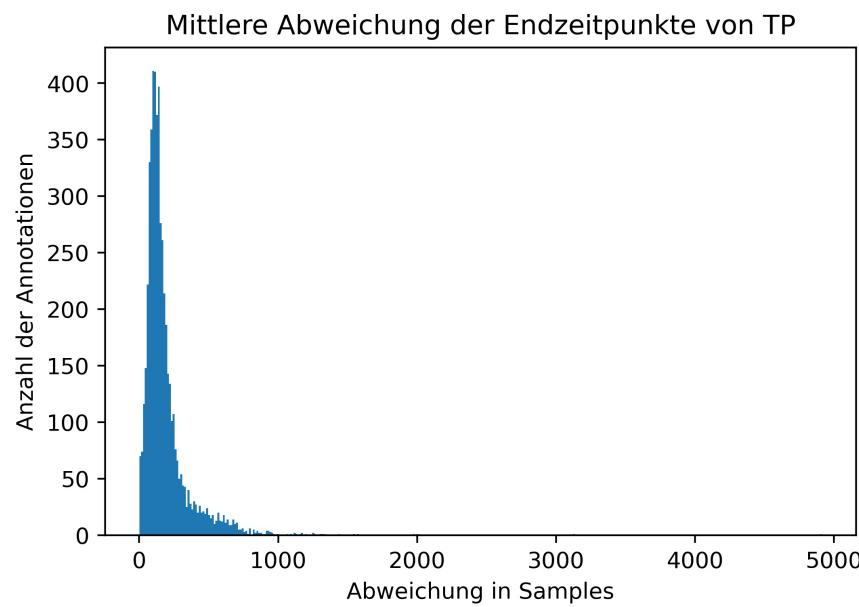


Abb. 1.2: Histogramm über die Mittlere Abweichung der Endzeitpunkte von TP bei einer Abtastfrequenz von 200 Hertz

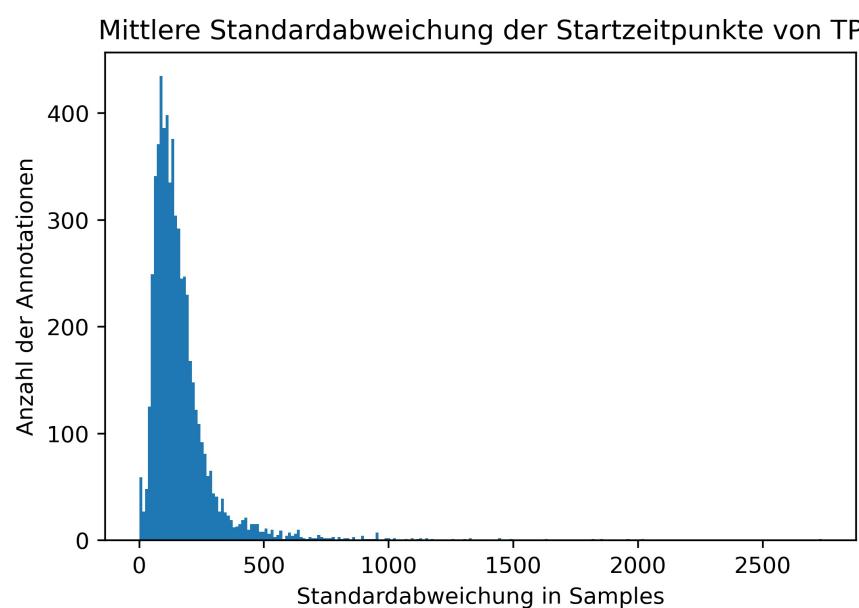


Abb. 1.3: Histogramm über die Standardabweichung der Startzeitpunkte von TP bei einer Abtastfrequenz von 200 Hertz

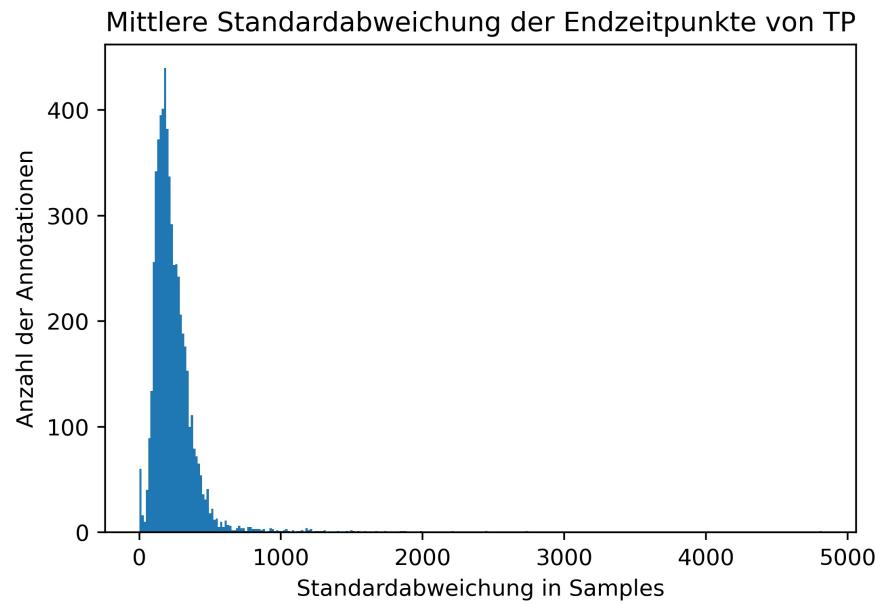


Abb. 1.4: Histogramm über die Standardabweichung der Endzeitpunkte von TP bei einer Abtastfrequenz von 200 Hertz

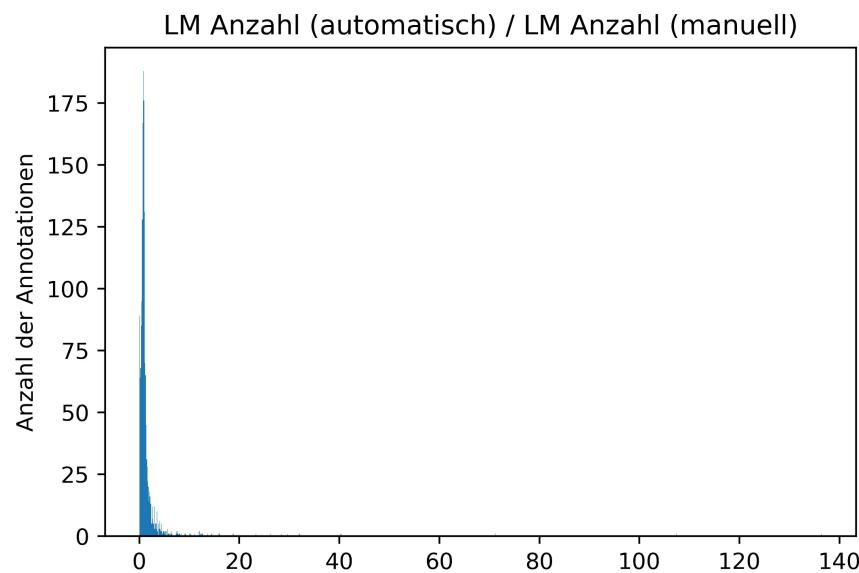


Abb. 1.5: Histogramm über das Verhältnis aus automatischer zu manueller LM Anzahl

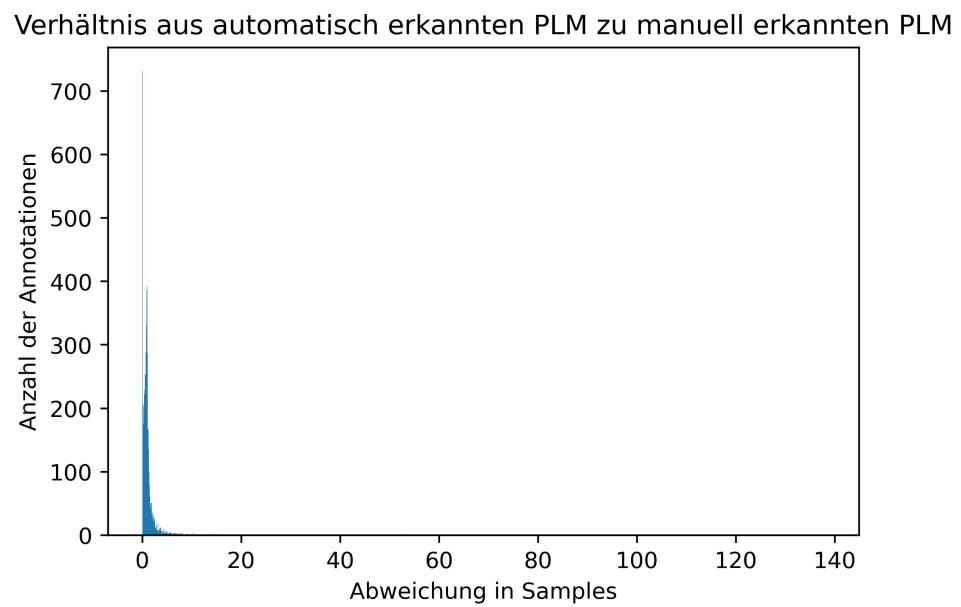


Abb. 1.6: Histogramm über das Verhältnis aus automatischer zu manueller PLM Anzahl