

TECHNISCHE UNIVERSITÄT MÜNCHEN

Fakultät für Medizin

Klinik für Anästhesiologie

# Einsatz von Maschinellen Lernverfahren zur Erkennung intraoperativer Wachheit

Moritz Tacke

Vollständiger Abdruck der von der Fakultät für Medizin der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Medizin

genehmigten Dissertation.

Vorsitzender: Univ.-Prof. Dr. E. J. Rummeny

Prüfer der Dissertation:

1. Univ.-Prof. Dr. G. E. Schneider
2. Univ.-Prof. Dr. E. Kochs
3. Univ.-Prof. Dr. K. A. Kuhn

Die Dissertation wurde am 04.07.2012 bei der Technischen Universität München eingereicht und durch die Fakultät für Medizin am 17.07.2013 angenommen.



## Danksagung

Für die Unterstützung bei dieser Dissertation möchte ich vor allem Dr. Denis Jordan, Prof. Gerhard Schneider, und Prof. Stefan Kramer danken. Die Zusammenarbeit war immer produktiv, und immer sehr angenehm.



# Inhaltsverzeichnis

1. Einleitung	1
2. Fragestellung	5
3. Material und Methodik	7
3.1. Elektroenzephalogramm und evozierte Potentiale . . . . .	7
3.2. Protokolldesign und Datenerhebung . . . . .	10
3.3. Vorverarbeitung der Daten . . . . .	12
3.4. Statistische Auswertung . . . . .	18
3.5. Klassifikationsverfahren . . . . .	25
3.6. Verfahren zur Parameterauswahl . . . . .	32
3.7. Parameterloser Ansatz . . . . .	35
3.8. Vorgehen . . . . .	36
3.9. Quellen der verwendeten Software . . . . .	37
4. Ergebnisse	39
4.1. Ergebnisse der einzelnen Parameter . . . . .	39
4.2. Parameterauswahl . . . . .	40
4.3. Klassifikatoren . . . . .	45
4.4. Klassifikation roher EEG-Daten . . . . .	47
4.5. Evaluation der Ergebnismetrik . . . . .	48
4.6. Endauswertung . . . . .	50

5. Diskussion	53
5.1. Interpretation der Befunde . . . . .	54
5.2. Vergleichbarkeit der Ergebnisse . . . . .	58
5.3. Verwandte Arbeiten . . . . .	60
5.4. Methodenkritik . . . . .	63
5.5. Schlussfolgerungen . . . . .	67
6. Zusammenfassung	69
A. Die Berechnung des $P_K$ -Werts	73
B. Informationsentropie und Information Gain	75

# Symbolverzeichnis

AEP	Auditorisch evozierte Potentiale
ALS	Amyotrophe Lateralsklerose
ApEn	Approximierte Entropie
BCI	Brain-Computer Interface
BIS	Bispectral Index
CV	Cross-Validation
EEG	Elektroenzephalogramm
EMG	Elektromyogramm
EOG	Elektrookulogramm
EP	Evozierte Potentiale
HEx	Hurst Exponent
J48	Entscheidungsbaumlernverfahren

LOC Loss of Consciousness

LZc Lempel-Ziv-Komplexität

MLAEP AEPs mittlerer Latenz

PCA Principal Component Analysis

PeEn Permutationsentropie

RBF Radial Base Function

ROC Return of Consciousness

SNR Signal/Rauschen-Verhältnis

SVM Support-Vektor-Maschine

WSMF Weighted Spectral Median Frequency

ZNS Zentrales Nervensystem

# 1. Einleitung

Eine seltene Komplikation der Allgemeinanästhesie ist die sogenannte „intraoperative Wachheit“. Es handelt sich hierbei um das bewusste, unbeabsichtigte Miterleben von Vorgängen während einer Operation durch den Patienten. Ihre Häufigkeit (mit nachfolgender Erinnerung an intraoperative Vorkommnisse) liegt Bereich von 0,1 - 0,2 % aller Allgemeinanästhesien [52, 60]. Episoden, an die sich die Patienten postoperativ nicht bewusst erinnern können, sind wahrscheinlich häufiger[53]. Den Inhalt von Erinnerungen an intraoperative Ereignisse bilden akustische Wahrnehmungen, das Gefühl von Muskelschwäche und Lähmung, und Schmerzen [39, 58]. Dies kann eine äußerst belastende Erfahrung sein. Mögliche Folge sind die Symptome einer posttraumatischen Belastungsstörung [31, 52]. Hierzu gehören Schlafstörungen, Alpträume, plötzliches Wiedererleben der Situation (sogenannte Flashbacks) und Ängstlichkeit.

Daher ist es wünschenswert, das Vorliegen intraoperativer Wachheit schnell und zuverlässig zu entdecken, um durch eine Veränderung der Medikamentendosierung reagieren zu können. Diese Arbeit soll einen Beitrag hierzu leisten.

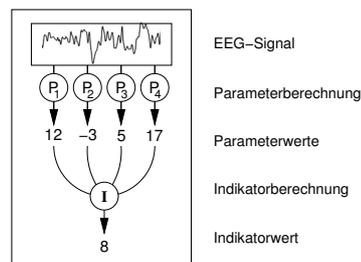
Die Erkennung intraoperativer Wachheit durch Beobachtung des Patienten oder über die Interpretation der Vitalzeichen ist schwierig. Deshalb besteht hier ein Bedarf an weiteren Messverfahren. Da diese Verfahren bei relaxierten Patienten funktionieren müssen, sind Varianten, die auf Bewegungen oder Äusserungen des Patienten beruhen, ungeeignet. Die Überwachung des Zielorgans der Nar-

kose, des Gehirns, liegt daher nahe. Eine Methode, die intrazerebrale Vorgänge reflektiert, ist das Elektroenzephalogramm (EEG). Hierbei werden an der Kopfhaut elektrische Spannungen gemessen, die entstehen, wenn sich das Membranpotential kortikaler Nervenzellen verändert. Das EEG kann um die Verwendung evozierter Potentiale (EP) erweitert werden [15, 22]. Hierbei werden wiederholt Reize wie z.B. Klicks oder Lichtblitze appliziert. Die Reaktion des Gehirns auf einen einzelnen Reiz wird vom spontanen EEG überlagert. Durch synchronisierte Überlagerung der Messungen lassen sich aber typische Reaktionsverläufe darstellen. Sowohl das EEG als auch die auditorisch evozierten Potentiale („AEPs“) ändern ihre Eigenschaften abhängig vom Bewusstseinszustand [8, 43].

Der Einsatz eines EEG-Monitors zur visuellen Kontrolle der Hirnströme ist nicht praktikabel, da das EEG komplexe, hochdimensionale Daten liefert, die nur durch entsprechend ausgebildetes Personal sicher interpretiert werden können. Es werden Methoden vorgezogen, die aus dem EEG einzelne, quantitative Werte extrahieren. Als Wertebereich können beispielsweise alle Zahlen zwischen 0 (=tiefe Narkose) und 100 (wach) gewählt werden. Solche Verfahren befinden sich bereits im klinischen Einsatz, z.B. der Bispektral Index (BIS) und der Narcotrend-Index [48]. Sie können zwar einen Beitrag zur Erkennung intraoperativer Wachheit leisten, es gibt aber auch Kritik und Hinweise, dass sie hierfür nur eingeschränkt geeignet sind [11, 36, 40, 51]. In einer Studie mit 2000 Patienten konnte durch die Verwendung des BIS auch bei Hochrisikopatienten keine Senkung der Inzidenz intraoperativer Wachheit erreicht werden [52]. Eine andere Studie ergab Hinweise darauf, dass der Bispektrale Index gerade beim Aufwachvorgang keine zuverlässige Erkennung des Bewusstseinszustands erlaubt [64]. Bei Hochrisikopatienten scheint die Verwendung der endtidalen Narkosegaskonzentration zur Narkosesteuerung dem BIS überlegen zu sein [2]. Die dem Bispektralen und dem Narcotrend-Index zugrunde liegenden Berechnungsmethoden werden geheimgehalten. Einschätzungen über die Eignung und Probleme dieser Geräte können also nur auf empirischen Beobach-

tungen und nicht auf unabhängigen, systematischen Analysen der Algorithmen beruhen. Die Ursachen dokumentierter Fehlfunktionen können nur gemutmaßt werden[47]. Es wird als unsicher angesehen, ob die derzeit verfügbaren EEG-Monitore zu verbesserten klinischen Ergebnissen führen[44]. Diese Ausgangsbedingungen motivieren dazu, die Forschung nach wirksamen EEG-Analysemethoden weiterhin zu betreiben.

Mathematische Verfahren, die das EEG auf einzelne Zahlenwerte reduzieren, werden als Parameter bezeichnet. Es wurden bereits mehrere solcher Parameter entwickelt. Da jeder einzelne Parameter gegenüber anderen Vor- und Nachteile aufweist, liegt es nahe, mehrere Parameter zu einem eindimensionalen Indikator zu kombinieren. Abbildung 1.1 stellt dies schematisch dar. Entscheidend sind hierfür die Auswahl der einfließenden Parameter und die technische Realisierung der Zusammenfassung.



**Abbildung 1.1.** – Illustration der Berechnung der Parameter aus dem EEG-Signal sowie ihrer Zusammenfassung zu einem Indikator

Für beides stellt das „maschinelle Lernen“ als Teildisziplin der Informatik Verfahren zur Verfügung. Hierzu zählt ein großes Repertoire an Klassifikationsalgorithmen mit verschiedenen Grundprinzipien, von rein statistischen Verfahren bis zur Nachahmung biologischer neuronaler Netze. Diese „lernenden“ Algorithmen müssen vor ihrem Einsatz anhand von Beispieldaten trainiert werden. Sie sind für die Verarbeitung hochdimensionaler Daten ausgelegt. Dies bedeutet, dass mehrere Parameter die Eingabemenge bilden können, aber auch, dass die Verfahren direkt auf die unverarbeiteten EEG-Daten anwendbar sind.

Die vorliegende Arbeit beschäftigt sich durch den Einsatz solcher Methoden mit dem Ziel, intraoperative Wachheit besser zu erkennen als bisher. Als Datenquelle dienen EEG- und AEP-Aufzeichnungen, die direkt vor oder nach dem Bewusstseinsverlust bzw. dem Wiedererwachen von Patienten im Rahmen einer Allgemeinanästhesie aufgezeichnet wurden. Es handelt sich also nicht um Daten, die aus Vorkommen tatsächlicher, ungeplanter intraoperativer Wachheit resultieren, sondern um Material, das in diesem Zusammenhang aus einer definierten Versuchsanordnung gewonnen wurde. Im Rahmen der vorliegenden Untersuchung wird erstmals in diesem Feld die Eignung automatischer Parameterauswahlmethoden überprüft. Eine große Anzahl von Klassifikationsverfahren wird systematisch verglichen. Als neuer Ansatz für die Erkennung intraoperativer Wachheit wird die Anwendung der Klassifikationsverfahren auf den „rohen“ EEG-Daten selbst untersucht.

Die Arbeit ist folgendermaßen aufgebaut: Zunächst wird in Kapitel 2 die Fragestellung präzisiert. In Kapitel 3 werden die verwendeten Daten und Verfahren vorgestellt. In den Kapitel 4 und 5 werden die Ergebnisse präsentiert und ihre Bedeutung beurteilt, bevor in Kapitel 6 ein Resümee gezogen wird.

## 2. Fragestellung

Ziel dieser Arbeit ist, zu beurteilen, inwieweit sich Verfahren des Maschinellen Lernens zur Erkennung intraoperativer Wachheit eignen.

Es werden zwei Ansätze untersucht, die Kombination quantitativer Parameter und die direkte Verarbeitung der EEG-Daten. In beiden Fällen kommen Klassifikationsverfahren zum Einsatz.

Im ersten Fall muss geklärt werden, wie aus einem Repertoire von über hundert zur Verfügung stehenden Parametern die am besten geeigneten ausgewählt werden können. Dies kann durch einen menschlichen Experten geschehen; eine Alternative, und der in dieser Arbeit gewählte Weg, sind automatische Parameterauswahlverfahren. Es werden verschiedene solcher Verfahren untersucht. Ebenso werden mehrere Klassifikationsverfahren systematisch miteinander verglichen. Weiterhin wird überprüft, wie sich Einschränkung und Ausweitung des verwendeten Datenmaterials auf die Klassifikationsqualität auswirken. Die Menge der verwendeten Parameter lässt sich hierbei sowohl durch die Wahl des EEG-Bandes als auch durch die Beschränkung auf entweder EEG- oder AEP-basierte Parameter variieren.

Bei der Verwendung der rohen EEG-Daten ist es von Interesse, inwieweit Vorverarbeitungen (Normalisierung der Daten, Einsatz der Hauptachsentransformation, Berechnung der Leistungsdichte) die Vorhersageleistung beeinflussen.



## 3. Material und Methodik

Dieses Kapitel gliedert sich in folgende Teile: In den Abschnitten 3.1 bis 3.3 werden die verwendeten Daten beschrieben. Es werden die Messgrößen - EEG und AEP - vorgestellt, das zur Erhebung verwendete Studiendesign erläutert und die durchgeführten Vorverarbeitungsschritte dargestellt. Abschnitt 3.4 beschreibt die Grundsätze, die für die Bewertung aller untersuchten Verfahren gelten. Diese sind Gegenstand der Abschnitte 3.5 bis 3.7. Abschnitt 3.8 skizziert die Reihenfolge, in der die Untersuchungen vorgenommen wurden. Das Kapitel schließt in Abschnitt 3.9 mit einer Angabe der verwendeten Software und ihrer Bezugsquellen.

### 3.1. Elektroenzephalogramm und evozierte Potentiale

Die grundlegenden verwendeten Datenquellen bilden das Elektroenzephalogramm (EEG) und die akustischen evozierten Potentiale (AEP). Beides sind Methoden, die den Zustand der neuronalen Dynamik widerspiegeln.

Das Gehirn ist ein aus verschiedenen Zelltypen aufgebautes Organ. Die hauptsächlich für die Verarbeitung von Informationen zuständigen Zellen sind die Nervenzellen, die in der Lage sind, in Form von Änderungen ihres Oberflächenpotentials Informationen weiterzuleiten. Die aufsummierten Potentialän-

derungen der oberflächlichen Nervenzellschichten erreichen eine Größenordnung, die es ermöglicht, sie mittels Elektroden von der Kopfhaut abzuleiten. Diese Ableitung wird als Elektroenzephalogramm (EEG) bezeichnet. Das so gewonnene Signal ist von verschiedenen Einflussgrößen abhängig, z.B. dem Ort der Elektroden und dem Bewusstseinszustand des Patienten. Insbesondere sind die verschiedenen Schlafphasen des Menschen über charakteristische EEG-Muster definiert [43], und das EEG zeigt unter Narkose spezifische Veränderungen [55]. Es gibt darüber hinaus viele Faktoren, durch die das EEG-Signal beeinflusst werden kann, neben Artefakten gehören hierzu z.B. ungenaue Elektrodenpositionen, der Blutzucker, das Geschlecht des Patienten, sowie Sauerstoff- und Kohlendioxidkonzentrationen im Blut [28]. Um eine Vergleichbarkeit zwischen verschiedenen Aufzeichnungen zu erhalten, wurden die Elektrodenpositionen im sogenannten 10-20-System international standardisiert [28].

Das durch das EEG aufgezeichnete Signal hat eine unregelmäßige Struktur. Typische Eigenschaften des EEGs lassen sich auch visuell erkennen; hierzu zählen die Vorherrschaft verschiedener Frequenzbereiche oder „Bänder“ in bestimmten Situationen (z.B. Wachheit, Müdigkeit, und Schlaf)[69]. Es werden folgende Bänder unterschieden:

- $\delta$ : 0.5 - 4 Hz
- $\theta$ : 4 - 8 Hz
- $\alpha$ : 8-12 Hz
- $\beta$ : 12 - 30 Hz
- $\gamma$ : über 30 Hz

Das „klassische“ EEG umfasst die Bänder bis 30 Hz. Vor allem in den darüber liegenden Frequenzbereichen kann es zu Einflüssen durch Muskel- oder Augenbewegungen kommen; diese Artefakte werden als Elektromyogramm (EMG) oder Elektrookulogramm (EOG) bezeichnet. Dies hat direkte Konsequenzen

für die Erkennung intraoperativer Wachheit. Das Vorliegen von EOGs und EMGs ist ein Hinweis darauf, dass der Patient wach ist [54]. Allerdings sind beide Artefakttypen durch eine Relaxierung des Patienten aufhebbar. Es besteht also die Gefahr, dass eine auf ihnen beruhende Methode bei wachen relaxierten Patienten fälschlicherweise Hinweise auf Bewusstlosigkeit liefert. Auf EMG und EOG beruhende Verfahren sind also für die Hauptrisikogruppe intraoperativer Wachheit nicht einsetzbar. Sie könnten das Risiko, eine solche Episode zu übersehen, erhöhen, da bei hier ohnehin spärlichen klinischen Hinweisen auf Wachheit einem apparativen Verfahren, das unbedenkliche Werte liefert, mehr Gewicht zugeordnet würde als bei klinisch suspekten Patienten.

Auditorisch evozierte Potentiale (AEPs) sind im EEG aufnehmbare stereotype Reaktionen der Hörbahn auf akustische Reize[67]. Man unterscheidet, entsprechend dem zeitlichen Ablauf, zwischen frühen oder „Hirnstamm“-AEPs (bis 12.5 ms nach dem Reiz), AEPs mittlerer (12.5 - 50 ms) und langer (50-250 ms) Latenz (diese entsprechen der bewussten Verarbeitung des Tonsignals)[70]. Für diese Untersuchung wurden AEPs mittlerer Latenz (MLAEP) eingesetzt. AEPs eignen sich für die Bestimmung des Bewusstseinszustands, da die Wahrnehmung akustischer Reize bei fortschreitendem Bewusstseinsverlust am längsten aufrecht erhalten und bei Rückkehr des Bewusstseins als erste wieder aktiviert wird[8, 57].

Die messbare Amplitude von AEPs liegt im Bereich von  $1\mu V$ ; sie werden vom Spontan-EEG, das Amplituden in der Größenordnung von  $200\mu V$  erreicht, überlagert. Um sie zu analysieren werden Reize repetitiv dargeboten (im vorliegenden Fall mit einer Frequenz von 8.33 Hz, also alle 120 ms, mit einer randomisierten Ungenauigkeit zur Vermeidung von Gewöhnungseffekten) und die entstehenden Durchläufe triggersynchron überlagert. Das Signal/Rauschen-Verhältnis (SNR, „Signal-to-noise-ratio“) lässt sich durch die Anwendung eines digitalen Hochpassfilters vor der Überlagerung weiter verbessern.

Der Aufwand zur Parametergewinnung liegt bei AEPs über dem, der für die

reinen EEG-Parameter notwendig ist. Die akustischen Stimuli erfordern den Einsatz zusätzlicher Geräte; die zeitliche Auflösung ist wegen der Mittlung geringer als bei EEG-Parametern [54].

### 3.2. Protokolldesign und Datenerhebung

Die dieser Arbeit zugrunde liegenden Daten wurden in einer klinischen Studie erhoben[54], die von der Ethikkommission der Technischen Universität München genehmigt wurde. Das Patientenkollektiv umfasste 40 erwachsene Patienten, die über die Studie aufgeklärt wurden und danach ihre Zustimmung zur Teilnahme schriftlich dokumentierten. Keiner der Patienten erfüllte die Ausschlusskriterien der Studie; diese umfassten Kontraindikationen gegenüber den verwendeten Medikamenten, eine Anamnese mit neurologischen oder psychiatrischen Erkrankungen, der Gebrauch oder Missbrauch von Substanzen, die auf das zentrale Nervensystem wirken, Schwangerschaft, und die Indikation zur Ileuseinleitung. Bei allen Patienten standen operative Eingriffe an; die für die Datenerhebung notwendigen Schritte wurden im Operationsaal unmittelbar prä- und postoperativ durchgeführt.

Die Patienten wurden randomisiert in zwei balancierte Gruppen aufgeteilt, die sich hinsichtlich der eingesetzten Hypnotika unterschieden: Zusätzlich zu Remifentanyl, das in beiden Gruppen mit einer minimalen Infusionsgeschwindigkeit von  $0,2 \frac{\mu\text{g}}{\text{kg}\cdot\text{min}}$  verabreicht wurde, erhielten die Patienten entweder Sevofluran oder Propofol.

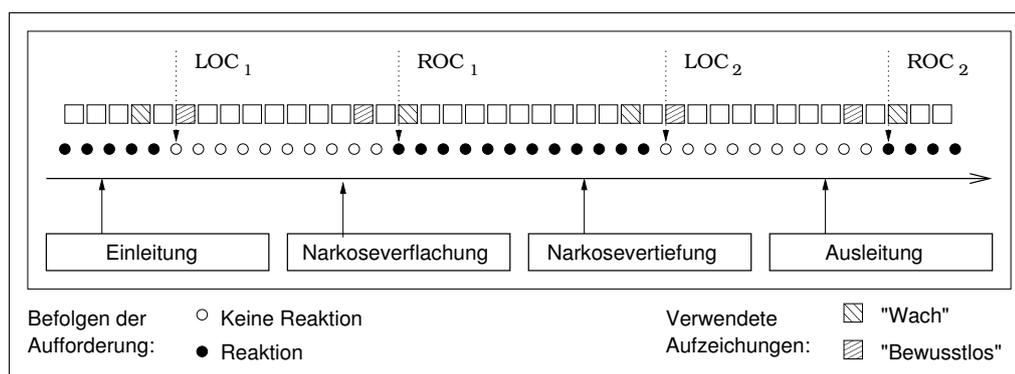
Die Remifentanilinfusion wurde ohne Prämedikation gestartet. Alle 30 Sekunden wurden die Patienten dazu aufgefordert, durch das Drücken der Hand des Untersuchers zu zeigen, dass sie noch bei Bewusstsein sind. Sobald dies geschehen war, wurde die Aufforderung zum Händedruck sofort wiederholt, um eine Fehlinterpretation unwillkürlicher Bewegungen auszuschließen.

Während der nun folgenden langsamen Narkoseeinleitung wurde der Zeitpunkt, zu dem der Patient der Aufforderung zum Händedruck das erste mal nicht mehr nachfolgen konnte, als erster Bewusstseinsverlust ( $LOC_1$ , nach „Loss of Consciousness“) dokumentiert. Bei gesteigerter Verabreichung des Hypnotikums wurde der Unterarm des Patienten für fünf Minuten mit einer Blutdruckmanschette vom Blutkreislauf abgeschnitten, um ihn von der folgenden Gabe von Succinylcholin ( $1.0 \frac{mg}{kg}$ ) abzuschirmen[68]. Der Patient wurde intubiert, anschließend wurde das Hypnotikum langsam ausgeschlichen, bis der Patient wieder den Aufforderungen Folge leistete. Dieser Zeitpunkt wurde als erstes Aufwachen (bzw.  $ROC_1$  für „Return of Consciousness“) in den Aufzeichnungen markiert.

Die Konzentration des Hypnotikums wurde bis zu einem zweiten Bewusstseinsverlust des Patienten erhöht. Dieser Zeitpunkt wurde als  $LOC_2$  dokumentiert. Die Allgemeinanästhesie wurde daraufhin richtliniengerecht fortgesetzt, und die Operation durchgeführt. Mit dem Abschluss der Operation wurde die Gabe von Remifentanil und Propofol bzw. Sevofluran eingestellt und die Aufforderungen, die Hand zu drücken, wieder aufgenommen. Sobald der Patient einer dieser Aufforderungen das erste Mal wieder nachkam, wurde dieser Zeitpunkt als zweites Aufwachen bzw.  $ROC_2$  markiert.

Während des gesamten Zeitraums wurden EEG- und AEP-Daten durch ein explizit für die intraoperative Überwachung entworfenes Gerät [22] aufgezeichnet. Hierzu wurden Zwei-Kanal-EEG-Signale an den Elektrodenpositionen AT1 (zwischen lateralem Augenwinkel und Oberrand Ohr), M2 (rechtes Mastoid - für die AEPs), Fpz (Referenz - Stirnmitte) und F7 (Erdung) abgeleitet. Für die Standardüberwachung diente ein Datex<sup>®</sup> AS/3<sup>™</sup>- Monitor (Datex Ohmeda, Helsinki, Finnland).

Die aufgezeichneten Daten wurden in Zehn-Sekunden-Blöcke unterteilt, wobei diese Blöcke entsprechend den Markierungen für den Bewusstseinsverlust bzw. -wiedergewinn als „bewusstlos“ oder „wach“ markiert wurde. Diejenigen



**Abbildung 3.1.** – Schematische Darstellung des Untersuchungsablaufs, der Reaktionen des Patienten während der Untersuchung, sowie der Herkunft der für die Auswertung verwendeten Daten. LOC und ROC: „Loss of consciousness“ bzw. „Return of consciousness“. Tatsächliche Länge der Datenblöcke: 10 Sekunden (für EEG-Parameter, bei den AEPs: variabel), tatsächliche Intervalldauer zwischen zwei Aufforderungen, den Händedruck zu erwidern: 30 Sekunden

Blöcke, die am nächsten am Wechsel des Bewusstseinszustands lagen, dabei aber noch sicher einem der beiden Zustände zugeordnet werden können, bilden die Grundlage der weiteren Untersuchungen. Bei  $LOC_X$  handelt es sich um den Datenblock direkt vor der letzten befolgten Aufforderung sowie um den ersten Block, nachdem der Aufforderung das erste Mal nicht mehr nachgekommen wurde; das Vorgehen war bei  $ROC_X$  analog. Pro Patient wurden dementsprechend acht Datenblöcke verwendet. Abbildung 3.1 skizziert den Studienablauf.

### 3.3. Vorverarbeitung der Daten

Vor dem Einsatz der maschinellen Lernverfahren wurden die Daten auf verschiedene Art bearbeitet. Hierzu gehören Filterung, Artefakterkennung[54], und die Berechnung der EEG- und AEP-Parameter, deren Kombination durch die Parameterauswahlverfahren und Klassifikatoren untersucht werden sollte. Für den parameterlosen Ansatz wurde das Frequenzspektrum der aufgezeich-

neten Daten berechnet.

Alle Messungen, bei denen entweder zumindest abschnittsweise ein konstantes Signal (also eine Nulllinie), Amplitudendifferenzen von mehr als  $\frac{140\mu V}{ms}$  oder eine Spannung von mehr als  $250\mu V$  aufgezeichnet wurde, wurden als Artefakte markiert und von den weiteren Analysen ausgeschlossen.

### 3.3.1. Beschreibung der EEG- und AEP-Parameter

Zur Erstellung der Indikatoren wurden mehrere unterschiedliche EEG- und AEP-Parameter verwendet, die alle zehn Sekunden neu berechnet wurden. Ein digitaler Tiefpassfilter mit wählbarer Grenzfrequenz wurde eingesetzt. Sowohl das klassische EEG-Band mit einer Grenzfrequenz von 30 Herz als auch erweiterte Frequenzspektren bis 49 bzw. 90 Herz konnten ausgewählt werden.

#### EEG-Parameter

Die EEG-Parameter quantifizieren verschiedene Eigenschaften des EEG-Signals [48, 66, 71]. Grundlegend ist hierbei die Unterscheidung in „lineare“ und „nicht-lineare“ Parameter. Die klassische Interpretation des EEGs beruht auf der Beobachtung, dass unter verschiedenen Bedingungen - z.B. Schlaf, Wachheit - unterschiedliche Frequenzbereiche, die sogenannten „Bänder“, das Signal beherrschen. Aus dieser Sicht abgeleitete Parameter, deren Grundprinzip es ist, aus den Frequenzmustern des EEGs quantitative Werte zu errechnen, werden als „lineare“ Parameter bezeichnet. Im Gegensatz hierzu quantifizieren „nicht-lineare“ Parameter Charakteristika des EEG, die einer Frequenzanalyse entgehen, wie z.B. die Signalkomplexität [13]. Für alle Parameter gilt, dass es Möglichkeiten gibt, das Verhalten des Parameters zu modifizieren. Dies kann auf verschiedene Weisen geschehen. Neben allgemeinen Maßnahmen wie z.B. der Auswahl der Frequenzbereiche, die der Berechnung zugrunde liegen, kann die Pa-

parameterberechnung durch parameterspezifische Variablen beeinflusst werden. Da die einzelnen Parameter nicht Gegenstand dieser Untersuchung waren, wird im folgenden darauf verzichtet, diese Einstellmöglichkeiten vorzustellen. Häufig wurden mehrere Parameter eines Typs, die sich nur durch unterschiedliche Einstellungen unterschieden, als individuelle Parameter behandelt. Es spielte in der Untersuchung keine Rolle, ob zwei Parameter sich grundlegend unterschieden, oder ob ihre Differenzen nur auf verschiedenen Einstellungen beruhten.

Zu den linearen Parametern zählte die gewichtete spektrale Medianfrequenz (WSMF, nach „Weighted Spectral Median Frequency“) [26]. Es handelt sich hierbei um eine Verallgemeinerung der vorbeschriebenen spektralen Parameter MF und SEF95 [59, 48]. Sie errechnet den Mittelpunkt eines gewichteten EEG-Amplitudenspektrums zwischen 8 Hertz und der jeweiligen Grenzfrequenz des Tiefpassfilters. Dieser Parameter wurde explizit für die Unterscheidung zwischen Bewusstlosigkeit und Wachheit entworfen. Hiervon abgeleitet wird der Parameter  $q\text{WSMF}$ , der durch das Verhältnis  $\frac{\text{WSMF}_1}{\text{WSMF}_2}$  definiert ist.  $\text{WSMF}_1$  und  $\text{WSMF}_2$  unterscheiden sich durch die berücksichtigten Frequenzbereiche, und zwar wird für  $\text{WSMF}_1$  das Spektrum zwischen 8 und 30 Hertz verwendet, während in die Berechnung von  $\text{WSMF}_2$  nur der Bereich zwischen 4 und 7 Hertz einfließt [17].

In die Kategorie der nichtlinearen Parameter fallen beispielsweise die Approximierte Entropie (ApEn) [6, 7], die Lempel-Ziv-Komplexität (LZc) [30] und der Hurst Exponent (HEX) [32]. Sie können EEG-Eigenschaften quantifizieren, die einer Spektralanalyse entgehen.

Die Permutationsentropie (PeEn) [4, 25] ist ein weiteres Maß für die nichtlineare Dynamik und kommt dabei ohne einschränkende Annahmen über das zu analysierende Signal aus. Da EEG-Signale hochdimensional vorliegen, und außerdem mit einer Überlagerung durch Rauschen zu rechnen ist, können die verschiedenen Bewusstseinszustände mittels der robusten nicht-parametrischen PeEn besser voneinander unterschieden werden, als dies mit anderen nichtlinearen

Parametern der Fall ist [24, 25]. Insbesondere ist  $PeEn$  unabhängig von den absoluten Amplituden des Signals, was dazu führt, dass der Parameter von Schwankungen des Amplitudenbereichs (z.B. durch die Impedanz der Elektroden) unbeeinflusst bleibt.

#### AEP-Parameter

Bei den vorliegenden Daten wurden die AEP-Werte als Durchschnitt aus 300 artefaktfreien Durchläufen ermittelt. Diese müssen insgesamt aus einem Intervall von 90 Sekunden Länge stammen; falls dieses Intervall überschritten wird, führt dies zu einem Ausschluss des entsprechenden AEP-Wertes von der weiteren Bearbeitung. Hierdurch soll es möglichst vermieden werden, dass sich die verschiedenen Bewusstseinszustände in einem AEP-Wert überschneiden.

Die AEP-Parameter wurden mittels Wavelet-Analyse bestimmt; hierdurch konnte eine Zeit- und Frequenzauflösung erzielt werden, außerdem war es möglich, die typischen AEP-Wellenformen mit in die Klassifikation einzubeziehen [33].

Die Zeit-Frequenz-Charakteristik wurde durch eine „à trous“- Wavelet-Transformation [33] bis zum Detailniveau 5 (D1-D5) auf den AEP-Daten ermittelt. Hierzu kamen ein Daubechies-3-Wavelet und das Auffüllen mit Nullen zum Einsatz. Auf der Grundlage der Wavelet-Koeffizienten wurden 56 verschiedene Wavelet-Parameter bestimmt. Da die Durchlaufänge 120 ms beträgt, liegen diesen Parametern die AEP-Antworten mittlerer Latenz zugrunde. Die Parameter lassen sich in fünf verschiedene Kategorien einteilen [33]:

1. Aus 11 Koeffizienten des Niveaus D3 und 11 Koeffizienten aus D4 errechnete Indexwerte
2. Amplituden und Verzögerungen der ersten drei Minima und Maxima unter den Waveletkoeffizienten aus D4 bzw. der ersten vier Extremwerte aus D5

3. Zwei Signalenergien der Waveletkoeffizienten aus D3 und D4
4. Die maximale Amplitude eines rekonstruierten AEPs aus D3
5. Die Varianz der zweiten Ableitung der Koeffizienten aus D4

### 3.3.2. Modifikationen für den parameterlosen Ansatz

Die Interpretation von EEG-Aufzeichnungen erfolgt üblicherweise nicht über die Messwerte zu einem bestimmten Zeitpunkt, sondern über das Auftreten charakteristischer Frequenzmuster im längeren Verlauf. Dieses Vorgehen erlaubt beispielsweise die Identifikation verschiedener Schlafphasen oder die Diagnose zerebraler Anfälle[43, 62].

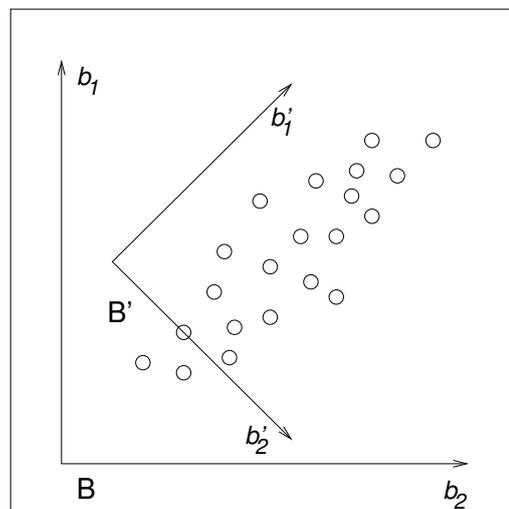
Deshalb wurde davon ausgegangen, dass auch für die Erkennung intraoperativer Wachheit die Frequenzinformationen richtungweisend sind. Die Fouriertransformation ist ein mathematisches Verfahren, durch das bei einem kontinuierlichen Signal die Gewichtung einzelner Sinusschwingungen, aus denen das Signal zusammengesetzt ist, bestimmt werden kann[48]. Sie eignet sich also dafür, das zunächst unregelmäßige EEG-Signal in einzelne Schwingungskomponenten zu zerlegen, und dabei den Einfluss dieser Komponenten auf das Gesamtsignal zu quantifizieren. Die EEG-Aufzeichnungen wurden deshalb für den parameterlosen Ansatz mittels einer Fourier-Transformation in ihr Frequenzspektrum überführt. Als Resultat lagen für jeden Messpunkt bei einer Auflösung von 0.1 Hz die Amplitudenwerte  $A(f_i)$  für die Frequenzen  $f_i$  zwischen 0 und 500 Hz vor.

Diese Werte wurden auf verschiedene Art und Weise vorbehandelt: Es wurde untersucht, wie sich die Vorhersageleistung bei einer Veränderung des Frequenzbereichs und der Auflösung verändert. Hierzu wurden Frequenzblöcke definiert, deren Grenzen bei 30, 49, 90, 300 und 500 Hz lagen. Die untersuch-

ten Auflösungen waren 0.1, 0.2, 0.5, und 1.0 Hz. Um Fehlklassifikationen durch möglicherweise vorliegende Amplitudenunterschiede bei ansonsten äquivalenten Frequenzspektren zu begegnen, wurde die Leistungsdichte der einzelnen Frequenzen nach folgender Formel errechnet:

$$PSD_0^{f_{max}}(f_i) = \frac{A^2(f_i)}{\sum_{f=0}^{f_{max}} A^2(f)}$$

Für die dimensionale Reduktion der fouriertransformierten Daten kam die Hauptachsentransformation (Principal Component Analysis, PCA) [29] zum Einsatz. Dies ist ein Verfahren, das bei mehrdimensionalen Daten eine neue Basis des entsprechenden Vektorraums ausrechnet. Die Basisvektoren werden hierbei so bestimmt, dass der erste Basisvektor die Richtung angibt, in der die Varianz der gegebenen Stichprobe maximal ist; auch die weiteren Basisvektoren sind hinsichtlich des Anteils der Varianz der entsprechenden Richtung an der Gesamtvarianz sortiert. Abbildung 3.2 illustriert dies. Falls die Dimensionali-



**Abbildung 3.2.** – Ein Beispiel für die Neuberechnung einer Vektorbasis durch die PCA. Die neue Basis  $B'$  ist so rotiert, dass der erste Basisvektor  $b'_1$  in die Richtung der größten Varianz zeigt.

tät der Daten verringert werden soll, kann der Informationsverlust reduziert werden, indem die Dimensionen weggelassen werden, deren Basisvektoren den höchsten Index haben; in der Abbildung wäre dies also  $b'_2$ . Häufig wird die Anzahl von Vektoren, die für die weitere Bearbeitung verwendet werden, anhand des Anteils der Varianz, der sich durch diese Vektoren erklären lässt, bestimmt. Es werden also so viele Vektoren berechnet, wie notwendig sind, um z.B. 95% der beobachteten Gesamtvarianz zu erklären.

Weiterhin wurde der Einfluss einer Normalisierung, bei der der Wertebereich für jede Frequenz in das Intervall  $[0, 1]$  projiziert wurde, untersucht.

### 3.4. Statistische Auswertung

Die verwendeten Daten wurden vor und nach den klinisch festgestellten Zeitpunkten  $LOC_1$ ,  $ROC_1$ ,  $LOC_2$ , und  $ROC_2$  ausgewertet. Es sind also die Messungen, die einem Bewusstseinsverlust oder -wiedergewinn vorausgingen bzw. nachfolgten. Nach der Artefakterkennung standen von den ursprünglich 320 Messungen noch 279 zur Verfügung (darunter 133 von bewusstlosen, die übrigen 146 von wachen Patienten). Die Daten eines Patienten wurden völlig aus dem Datensatz entfernt, so dass sich im tatsächlich verwendeten Datensatz nur Messungen von 39 Patienten befinden.

Für den Hauptteil der Arbeit wurden 103 EEG- und AEP-Parameter verwendet. In der Teiluntersuchung, wie sich die Breite des EEG-Spektrums auf die Vorhersagequalität auswirkt, waren es bei stärker eingeschränktem Spektrum (Grenzfrequenz von 30 Hz anstelle den ansonsten verwendeten 49 Hz) 96 Parameter, bei einem weiteren Spektrum bis 90 Hz insgesamt 115 Parameter.

Insgesamt lagen die Daten von 39 Patienten vor. Diese wurden in einen Arbeitsdatensatz (29 Patienten) und einen Validationsdatensatz (zehn Patienten)

aufgespalten. Auf dem Arbeitsdatensatz wurde ein Screening durchgeführt, in dem die unterschiedlichen Parameterauswahlverfahren und Klassifikatoren evaluiert wurden. Die Verfahren, die hier die höchste Vorhersagewahrscheinlichkeit erbrachten, wurden anhand des Validationsdatensatzes erneut untereinander verglichen.

#### 3.4.1. Grundlagen des Einsatzes maschineller Lernverfahren

Maschinelle Lernverfahren funktionieren nach dem Prinzip, dass einem Algorithmus zunächst ein Satz von Trainingsdaten, die sogenannte „Trainingsmenge“, präsentiert wird. Sie besteht aus den gemessenen Daten und ihren Klassifikationen; ein Datenpunkt zwischen  $ROC_1$  und  $LOC_2$  würde also aus den zu diesem Zeitpunkt berechneten Parametern und der Klassifikation „wach“ bestehen.

Ein Lernverfahren sucht nach Regelmäßigkeiten in der Trainingsmenge und erstellt eine Hypothese über den Zusammenhang zwischen Daten und Klassifikationen. Ein einfaches Beispiel für eine solche Hypothese wäre bei zwei Parametern  $A$  und  $B$  „wenn  $A$  größer als 10 und  $B$  kleiner als 5 ist, ist der Patient wach“.

Hat der Lernalgorithmus seine Hypothese erarbeitet, wird eine zweite Menge, die „Testmenge“, verwendet, um die Qualität dieser Hypothese zu beurteilen. Die Testmenge besteht ausschließlich aus Messwerten, die Klassifikation dieser Werte fehlt also. Aufgabe des Lernalgorithmus ist es nun, unter Verwendung der Hypothese vorherzusagen, welche Klasse den einzelnen Datenpunkte aus der Testmenge zuzuordnen ist. Der Vergleich dieser Vorhersagen mit den tatsächlichen Klassifikationen ermöglicht es, Aussagen über die Qualität des Lernverfahrens zu treffen.

Viele Lernalgorithmen sind in der Lage, statt einer reinen Klassifikation einen

Zahlenwert auszugeben, der angibt, wie sicher sich der Algorithmus bei der Klassifikation ist, z.B.  $1.0 \hat{=}$  sicher wach,  $0.0 \hat{=}$  sicher nicht wach,  $0.75 \hat{=}$  vermutlich wach. Hierdurch können Schwellwerte mit verschiedener Sensitivität und Spezifität ausgewählt werden. Außerdem ermöglicht dies den Einsatz über Sensitivität und Spezifität hinausgehender Qualitätsmaßstäbe, wie des in dieser Untersuchung verwendeten  $P_K$ -Werts (siehe Abschnitt 3.4.2).

Trainings- und Testmengen können unter den zur Verfügung stehenden Daten frei ausgewählt werden, wobei die Bedingung zu beachten ist, dass keine Datenpunkte in beiden Mengen gleichzeitig auftauchen. Dies würde dazu führen, dass die Leistung des Lernalgorithmus überschätzt wird, da er zufällige, tatsächlich irrelevante Eigenschaften der Daten mit Erfolg zur Klassifikation verwenden könnte. Bei überschneidungsfreien Mengen ist das Risiko, zufälligen Gesetzmäßigkeiten zu Opfer zu fallen, deutlich geringer [38].

In der vorliegenden Untersuchung wurde diese Bedingung noch etwas gestrafft: Es wurde nicht nur gefordert, dass kein Datenpunkt in Trainings- und Testmenge gleichzeitig vorkommen darf, sondern auch, dass alle Datenpunkte, die von ein und demselben Patienten stammen (maximal acht), in dieselbe Menge gelangen. Dies war notwendig, um zu verhindern, dass beispielsweise die zweite Phase der Bewusstlosigkeit eines Patienten unter Verwendung von Daten aus der ersten Phase erkannt wird.

Die Aufteilung in Test- und Trainingsmenge verringert die Menge der zur Verfügung stehenden Daten. Deshalb wird, um die Datenmenge bestmöglich auszunutzen, häufig eine Kreuzvalidation (auch im deutschen Sprachgebrauch eher als Cross-Validation oder CV bezeichnet) eingesetzt.

Das Vorgehen bei der Cross-Validation ist, die gesamte Datenmenge zufällig in eine vorgegebene Anzahl überschneidungsfreier Untermengen aufzuteilen; bei beispielsweise zehn dieser Untermengen spricht man von zehnfacher Cross-Validation. Es wird dann zyklisch je eine dieser Untermengen als Testmenge

ausgewählt und die anderen zur Trainingsmenge zusammengefasst. Sobald jede Teilmenge einmal als Testmenge verwendet wurde, existiert für jeden Datenpunkt genau eine Klassifikation durch den Lernalgorithmus.

Je kleiner die Teilmengen sind, desto größer wird in diesem Verfahren die Trainingsmenge. Dies erhöht die Wahrscheinlichkeit, dass ein Lernalgorithmus korrekte Hypothesen erzeugt. Je größer die einzelnen Teilmengen, desto schneller terminiert der Algorithmus. Es ist also eine Abwägung zwischen Vorhersagequalität und Laufzeit, durch die die Anzahl der Teilmengen gewählt wird.

Im vorliegenden Fall kamen im Arbeitsdatensatz (also alle Daten ohne die zurückgehaltenen) die Daten von 29 Patienten vor. Diese konnten nur patientenweise aufgeteilt werden (s.o.), so dass die feinstmögliche Aufteilung, das sogenannte Leave-One-Out-Verfahren gewählt wurde: Jede Teilmenge wurde von den Daten eines einzigen Patienten gebildet, so dass am Ende 29 Teilmengen zur Verfügung standen.

Es ist notwendig, die Trennung zwischen Test- und Trainingsmenge streng einzuhalten. Dies betrifft auch Modifikationen der Daten, die vor der eigentlichen Klassifikation vorgenommen werden. Im Falle einer Normalisierung, bei der der Wertebereich jedes Parameters im Bereich zwischen Null und Eins zu liegen kommen soll, dürfen die Grundlagen dieser Verschiebung, also das beobachtete Maximum und Minimum, nur in der Trainingsmenge bestimmt werden. Diese Extremwerte werden dann auch zur Normalisierung der Testmenge herangezogen; dies kann dazu führen, dass innerhalb dieser auch nach der Normalisierung noch Werte vorliegen, die außerhalb des Intervalls zwischen Null und Eins liegen.

### 3.4.2. Evaluation der Vorhersagequalität

Als Maß für die Qualität einer Kombination von Lernverfahren und Parameterauswahl wurde der sogenannte  $P_K$ -Wert verwendet [65].

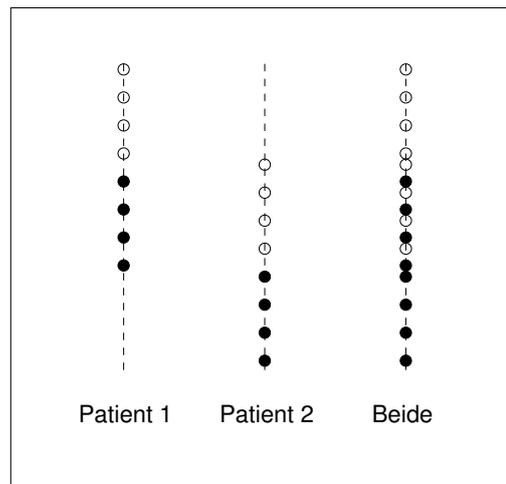
Der  $P_K$ -Wert ist ein quantitatives Maß dafür, wie gut sich eine bestimmte Variable  $x$  dafür eignet, als Grundlage für eine Klassifikation mittels eines Schwellwerts eingesetzt zu werden. Man kann damit also Verfahren, die ihre Klassifikation in Zahlenwerten ausdrücken, miteinander vergleichen.

Bei zwei möglichen Klassen „0“ und „1“ wird der  $P_K$ -Wert genau dann eins, wenn  $x$  für alle Beispiele der Klasse „1“ einen höheren Wert hat als für alle Beispiele der Klasse „0“. Ein  $P_K$ -Wert von Null besagt, dass  $x$  für alle Datenpunkte der niedrigeren Klasse „0“ höher ist als für die der Klasse „1“; bei einem  $P_K$ -Wert von 0.5 lassen sich aus dem Betrag von  $x$  keine Rückschlüsse auf die Klassenzugehörigkeit ziehen. Die Berechnung von  $P_K$  wird in Anhang A dargestellt.

Da für jeden Patienten acht verschiedene Datenpunkte zur Verfügung standen, und die Aufteilung der Gesamtdaten nach dem Leave-One-Out-Verfahren erfolgte, wurden die  $P_k$ -Werte ebenfalls für die einzelnen Patienten bestimmt und dann zu  $\bar{P}_K$  gemittelt, um Informationen über die Streuung der Werte zu erhalten. Dieses Vorgehen beinhaltet die Möglichkeit, dass bei von Patient zu Patient unterschiedlichen Schwellwerten der verschiedenen Indikatoren  $\bar{P}_K$ -Werte erzielt werden, die höher sind, als die bei Gruppen von Patienten erzielbaren. Dies wird in Abbildung 3.3 dargestellt.

Um diesen Effekt abzuschätzen erfolgte ein Durchlauf mit einer siebenfachen Cross-Validation, deren Ergebnisse mit denen aus dem Leave-One-Out-Verfahren verglichen wurden. Bei einer siebenfachen CV bestand die Testmenge immer aus mindestens vier Patienten (da  $\frac{29}{7} > 4$ ).

Wenn einzelne Patienten unterschiedliche Schwellwerte aufweisen, ist dies ein



**Abbildung 3.3.** – Darstellung der Unterschiede bei individuell errechneten  $P_K$  im Gegensatz zu denen bei gruppierten Ergebnissen. Für jeden der beiden Patienten lässt sich ein perfekter Schwellwert finden; für die zusammengefassten Daten trifft dies nicht mehr zu.

Hinweis darauf, dass sich durch eine Kalibrierung der Indikatoren auf den einzelnen Patienten unter Umständen eine Verbesserung der Erkennungsrate im praktischen Einsatz erreichen ließe.

### 3.4.3. Endauswertung

Ziel der Endauswertung sind statistisch aussagekräftige Ergebnisse. Hierzu wurde eine geringe Anzahl aussichtsreicher Verfahren auf „frischen“ Daten, nämlich den Messungen von zehn Patienten, die zunächst zurückgehalten wurden, überprüft. Sie bildeten die Testmenge, während der bisherige Arbeitsdatensatz als Trainingsmenge verwendet wurde.

Zur Bestimmung von Konfidenzintervallen wurde die Validierung der erzielten Ergebnisse auf der zurückgehaltenen Datenmenge, die aus den Messungen von zehn zufällig ausgewählten Patienten bestand, mittels eines Bootstrap-

Verfahrens durchgeführt[9]. Unter Bootstrap-Verfahren versteht man Algorithmen, die es erlauben, bei Daten mit unbekanntem zugrunde liegenden Verteilungen und bei verhältnismäßig kleinen Datensätzen Konfidenzintervalle zu bestimmen. Kernprinzip dieser Verfahren ist das sogenannte „Resampling“, also das wiederholte Ziehen von Stichproben aus einer Stichprobe.

Wenn also z.B. bei einem Datensatz  $D$  von  $n$  Datenpunkten, die jeweils einfache Zahlen sind, nicht nur der Mittelwert, sondern auch eine Annäherung seiner Konfidenzintervalle bestimmt werden soll, wäre das typische Vorgehen bei einem Bootstrap-Ansatz:

1. Das  $k$ -fache Ziehen von Stichproben der Größe  $n$  aus  $D$  (mit Zurücklegen).
2. Die Berechnung der  $k$  Mittelwerte
3. Die Sortierung der Mittelwerte
4. Die Schätzung der Grenzen des Konfidenzintervalls: Bei einem 95%-Konfidenzintervall würden diese durch die Mittelwerte angenähert, die in der sortierten Mittelwertliste auf den Positionen  $k \cdot 0.025$  und  $k \cdot 0.975$  (jeweils gerundet) zu finden wären.

Es handelt sich um ein stochastisches Verfahren vom Monte-Carlo-Typ, also um einen Algorithmus, dessen Ergebnisse nicht sicher richtig sind, bei dem die Ergebnisqualität mit steigender Iterationszahl (hier also: mit steigendem  $k$ ) besser wird.

Hier wurde dieses Verfahren folgendermaßen eingesetzt:

1. Training des zu untersuchenden Lernalgorithmus mit dem gesamten Arbeitsdatensatz von 29 Patienten
2. Klassifikation aller Datenpunkte aus dem Kontrolldatensatz

3. Bootstrap: Resampling der Klassifikationsergebnisse, Bestimmung des  $P_K$  jeder Stichprobe, Bestimmung der Konfidenzintervalle

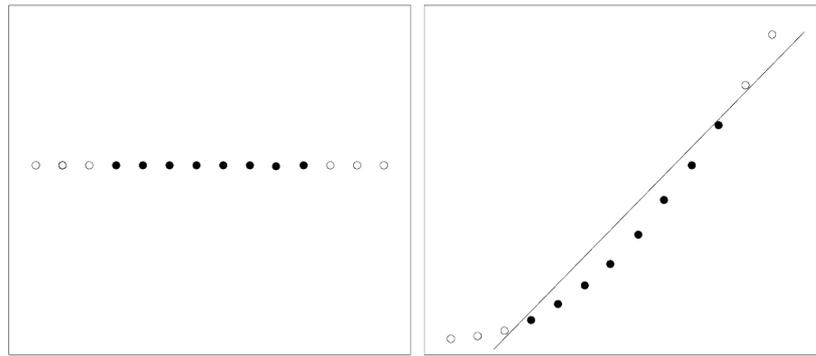
### 3.5. Klassifikationsverfahren

Auf die beschriebenen Daten wurden verschiedene Klassifikationsverfahren angewendet: Support-Vektor-Maschinen [56], künstliche neuronale Netze, Entscheidungsbaum-Lernverfahren, logistische Regression und Bayessche Klassifikatoren.

Alle dieser Klassifikationsverfahren können nicht nur direkte Klassifikationen durchführen, sondern auch quantitativ die Zuverlässigkeit der Klassifikation abschätzen. Durch ein Schwellwertverfahren kann also eine Abwägung zwischen Sensitivität und Spezifität vorgenommen werden. Hierdurch eignen sie sich auch dafür, durch den  $P_K$ -Wert untereinander verglichen zu werden.

#### 3.5.1. Support-Vektor-Maschinen

Support-Vektor-Maschinen [56] oder SVMs sind Algorithmen, die bei der Hypothesenkonstruktion in zwei Schritten vorgehen. Zuerst werden die als Eingabe verwendeten Daten in einen höherdimensionalen, u.U. unendlichdimensionalen Vektorraum abgebildet. In diesem Vektorraum wird dann eine Ebene (die sogenannte „Hyperebene“) gesucht, die die Beispiele mit den verschiedenen Markierungen (also „wach“ bzw. „bewusstlos“) auf bestmögliche Art und Weise trennt. Der Maßstab, mit dem die Qualität einer Hyperebene quantifiziert wird, ist der „Margin“- hierbei handelt es sich um den Abstand zwischen der Hyperebene und den nächsten Punkten. Die Hyperebene mit dem größten Margin wird zur folgenden Klassifikation der Testmenge verwendet: Jeder Datenpunkt aus der Testmenge wird in den höherdimensionalen Vektorraum

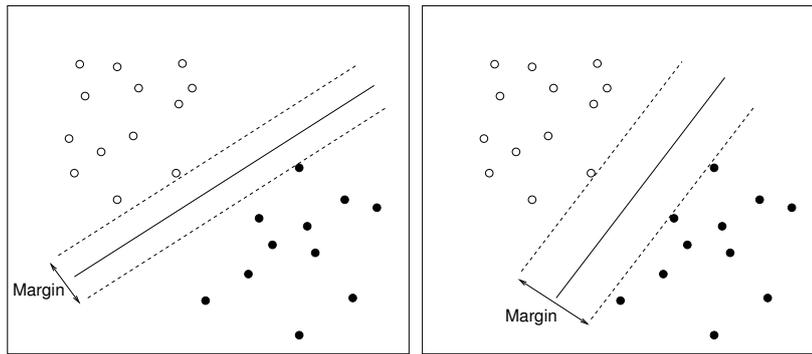


**Abbildung 3.4.** – Wirkung der Abbildung in einen höherdimensionalen Vektorraum. Die zunächst in einem eindimensionalen Raum, d.h. auf einer Gerade, vorliegenden Punkte können nicht durch eine Ebene entsprechend ihrer Färbung getrennt werden. Durch die Abbildung in einen höherdimensionalen Vektorraum - in diesem Fall durch Quadrierung - wird dies ermöglicht

abgebildet, und je nachdem, auf welcher Seite der Hyperebene er zu liegen kommt, entsprechend klassifiziert.

Abbildung 3.4 illustriert die Wirkung der Abbildung in einen höherdimensionalen Vektorraum auf einen Datensatz, der erst durch eine Projektion entsprechend der vorkommenden Klassen aufteilbar wird. Abbildung 3.5 zeigt das Prinzip des „Margins“: Zwei Hyperebenen, die hinsichtlich der Aufteilung der gelieferten Datenpunkte dieselbe Qualität haben - sie ist in beiden Fällen perfekt - unterscheiden sich im Margin, und lassen dadurch die Entscheidung für eine der beiden Ebenen zu.

Die Besonderheit bei den Support-Vektor-Maschinen liegt darin, dass durch den Einsatz bestimmter Funktionen, der sogenannten Kernelfunktionen (oder „Kernel“), die Projektion in den höherdimensionalen Raum umgangen werden kann, die Ergebnisse der Klassifikation aber genau denjenigen entsprechen, die unter der Verwendung der Projektion resultieren würden. Dies ermöglicht auch erst das Verwenden unendlichdimensionaler Vektorräume.



**Abbildung 3.5.** – Darstellung des „Margins“. Beide Hyperebenen trennen die Beispiele perfekt hinsichtlich ihrer Färbung auf; der „Margin“, d.h. der Abstand der nächsten Punkte zur Hyperebene, ist im zweiten Bild größer. Eine Support-Vektor-Maschine würde die zweite Ebene als Hypothese verwenden.

Eine der Möglichkeiten, das Verhalten einer SVM zu beeinflussen, ist die Auswahl der verwendeten Kernelfunktion. In dieser Arbeit kamen verschiedene Kernel zum Einsatz: Lineare, quadratische, sigmoidale, sowie RBF-Kernel.

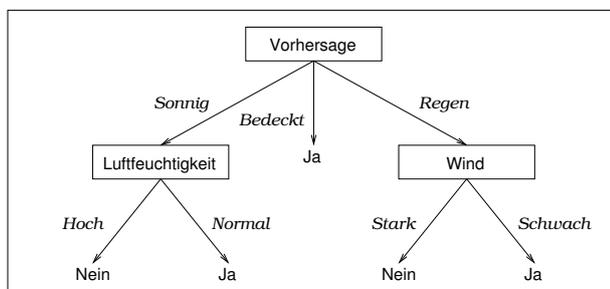
### 3.5.2. Entscheidungsbaum-Lernverfahren

Entscheidungsbaum-Lernverfahren [46] verfolgen einen klassischen Ansatz zur Klassifikation: Die aus der Trainingsmenge abgeleiteten Hypothesen haben die Form von Entscheidungsbäumen, also Graphen, die z.B. den aus medizinischen Lehrbüchern bekannten Flussdiagrammen ähneln.

Abbildung 3.6 zeigt ein Beispiel für einen Entscheidungsbaum, wie er von diesen Verfahren erzeugt wird.

Jeder „Knoten“ des Baums entspricht einem Test auf einen bestimmten Parameter; alle Datenpunkte, die den Test erfüllen, werden in einen folgenden Ast des Baums weiter überprüft, während die Datenpunkte, für die das nicht zutrifft, in den anderen Ast gelangen. Die Blätter des Baums entsprechen den Klassi-

fikationen. Ein großer Vorteil dieser Art der Hypothesendarstellung ist eine, verglichen mit anderen Ansätzen (z.B. SVMs), hohe Lesbarkeit der gewonnenen Hypothesen, die ermöglicht, neue Einsichten in das untersuchte Feld zu gewinnen.



**Abbildung 3.6.** – Beispiel eines Entscheidungsbaumes, der unter Verwendung dreier Parameter entscheidet, ob sich ein gegebener Tag zum Tennisspielen eignet.

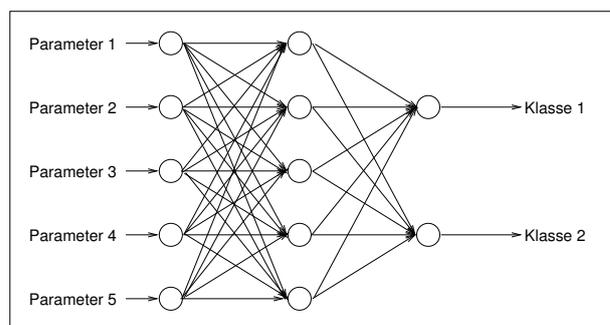
In dieser Arbeit wurde der J48-Algorithmus verwendet. Er ist eine Variante des klassischen C4.5-Verfahrens [46], das auf der Grundlage informationstheoretischer Überlegungen arbeitet. Der Entscheidungsbaum wird hierbei von der Spitze her aufgebaut. Bei jedem neu in den Baum einzufügenden Knoten wird überprüft, mit welchem Parameter und, gegebenenfalls, Schwellwert sich die in der Trainingsmenge vorliegenden Datenpunkte bestmöglich aufteilen lassen. Eine Aufteilung durch einen solchen Parametertest wird als gut angesehen, wenn sie zu möglichst reinen Teilmengen führt. Bei einer Datenmenge mit zwei verschiedenen Klassen „0“ und „1“ wird also der Parameter ausgewählt, der es erlaubt, die gegebenen Daten in zwei Teilmengen aufzuteilen, von denen in der einen die Klasse „0“ überwiegt, in der anderen die Klasse „1“.

Zur Messung dieser Qualität wird der sogenannte „Information Gain“ verwendet, der in Anhang B erklärt wird.

Der so entstandene Baum wird am Ende noch „beschnitten“, um die Gefahr des „Overfittings“, also der Erstellung eines Baums, der auf Kosten der Leistung bei den Testdaten zu sehr auf die Trainingsdaten ausgelegt ist, zu mindern.

## 3.5.3. Künstliche Neuronale Netze

Künstliche neuronale Netze [50] sind ein in vielen Disziplinen beliebtes Lern- und Klassifikationsverfahren. Sie beruhen auf einer groben Simulation natürlicher neuronaler Netze, wie sie im zentralen Nervensystem vieler Organismen vorkommen. Ein künstliches neuronales Netz besteht aus zwei Komponenten, den sogenannten Knoten oder Neuronen, sowie den Verbindungen zwischen ihnen. Zu jedem Knoten gehören eine Menge eingehender und ausgehender Verbindungen. Die Knoten fassen die an ihren Eingängen anliegenden Daten - es handelt sich um Zahlenwerte - zusammen, wobei den verschiedenen Eingängen unterschiedliche Gewichte beigemessen werden. Dieser zusammengefasste Wert wird dann über die Ausgänge an andere Knoten weitergegeben.



**Abbildung 3.7.** – Beispielhafte Darstellung eines neuronalen Netzes mit fünf Ein- und zwei Ausgängen

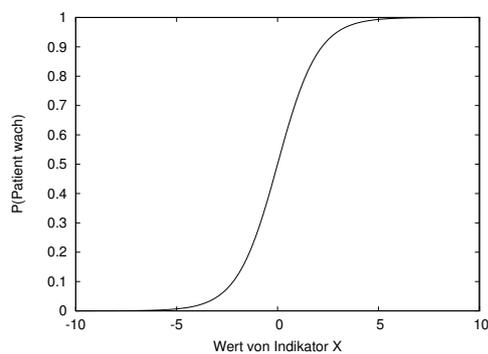
Die Eingänge zur ersten Ebene von Knoten bilden die Schnittstelle zur Dateneingabe, während die Ausgänge der letzten Ebene von Knoten die Ausgabe der Klassifikationen durchführen. Eine Darstellung der Struktur eines neuronalen Netzes zeigt Abbildung 3.7.

Die Struktur des Netzes ist im Allgemeinen vorgegeben, das „Lernen“ besteht in der Suche nach einer Gewichtung der Eingänge aller Knoten.

### 3.5.4. Logistische Regression

Zur Regression auf Daten, bei denen die abhängige Variable eine Dichotomie ist (d.h. nur zwei verschiedene Werte annehmen kann), kann die logistische Regression angewendet werden [20].

Grundannahme bei diesem Verfahren ist, dass sich die Wahrscheinlichkeit, mit der ein bestimmtes Beispiel der einen oder anderen Klasse zugewiesen wird, durch eine logistische Kurve darstellen lässt. Dies ist eine Funktion, die ihre Eingabewerte auf einen Wert zwischen Null und Eins abbildet; ein einfaches Beispiel (mit eindimensionalen Eingabewerten) ist in Abbildung 3.8 dargestellt. Die Gewichte, mit denen die einzelnen Parameter in diese Funktion einfließen, werden im Trainingsschritt bestimmt. Im Vorhersageschritt wird zunächst der Funktionswert für jedes Beispiel berechnet und dann die Klassifikation durch ein Schwellwertverfahren durchgeführt.



**Abbildung 3.8.** – Beispiel für eine logistische Kurve, die für einen fiktiven Parameter  $X$  angibt, wie wahrscheinlich bei verschiedenen Werten für  $X$  ein Patient wach ist. Bei einem  $X$ -Wert von 0 sind in diesem Beispiel 50% der Patienten wach.

### 3.5.5. Bayessche Lernverfahren

Allgemein gesehen ist die Eingabemenge für ein Lernverfahren immer eine Stichprobe aus einer unbekanntem Wahrscheinlichkeitsverteilung. Bayessche Lernverfahren [38] erstellen Hypothesen über diese Verteilung. Zu jedem gegebenen Datenpunkt wird im Klassifikationsschritt die Wahrscheinlichkeit der verschiedenen Klassen geschätzt.

Bei der den Daten tatsächlich zugrunde liegenden Wahrscheinlichkeitsverteilung sind die einzelnen Parameter im Normalfall nicht völlig unabhängig voneinander. Da die Abhängigkeitsverhältnisse unter den Parametern nicht bekannt sind, müssen sie bei der Hypothesenerstellung ebenfalls modelliert werden. Dies geschieht über verschiedene Annahmen, durch die sich die Bayesschen Lernverfahren unterscheiden.

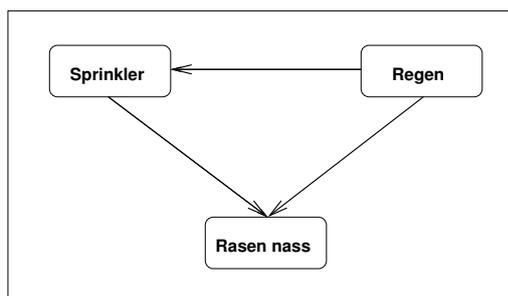
Der „naive Bayessche Klassifikator“ basiert auf der Annahme, dass alle Parameter paarweise unabhängig sind. Hierdurch wird die Schätzung der Wahrscheinlichkeitsverteilung stark vereinfacht, da bei völliger Unabhängigkeit der einzelnen Parameter - und nur dann - die Gesamtwahrscheinlichkeit, mit der ein Datenpunkt einer Klasse zugewiesen wird, das Produkt der Wahrscheinlichkeiten ist, mit denen diese Zuweisung auf der Basis der einzelnen Parameter geschehen würde.

Allerdings ist diese vereinfachende Annahme in dieser Arbeit, wie auch in den meisten anderen Fällen, in denen das Verfahren zum Einsatz kommt, nicht zutreffend. Insbesondere ist es nicht in der Lage, zu berücksichtigen, wenn einige Parameter stark miteinander korreliert (oder sogar identisch) sind. Trotz dieses prinzipiellen Problems hat sich das naive Bayessche Lernverfahren aber als ein in vielen Domänen erfolgreicher und robuster Ansatz etabliert [38].

Im Gegensatz zur groben Vereinfachung des naiven Ansatzes bildet ein „Bayesisches Netz“ seine Abhängigkeitsannahmen als einen Graphen ab; dementspre-

chend können komplexe Abhängigkeitsschemata dargestellt werden. Abbildung 3.9 zeigt ein sehr einfaches Beispiel für die Struktur eines Bayesschen Netzes.

Im Gegensatz zu den künstlichen neuronalen Netzen muss hier die Struktur des Netzes ebenfalls gelernt werden; insgesamt sind Bayessche Netze ein Verfahren, das durch hohe algorithmische Komplexität und lange Laufzeiten gekennzeichnet ist[45].



**Abbildung 3.9.** – Ein Bayessches Netz, das die Zusammenhänge darstellt, von denen es abhängt, ob ein Rasen nass ist. Hier ist z.B. der Zustand der Sprinkleranlage abhängig davon, ob es regnet, während der Regen von keinem der anderen Sachverhalte abhängt. Jedem Knoten  $k$  ist eine Wahrscheinlichkeitsverteilung zugeordnet, die von allen Knoten, von denen aus Pfeile auf  $k$  deuten, abhängig ist.

### 3.6. Verfahren zur Parameterauswahl

Um die Dimensionalität der Eingabedaten zu verringern, wurden Parameterauswahlverfahren eingesetzt. Es gibt hierfür zwei verschiedene Grundprinzipien: Entweder werden die Klassifikationsverfahren selbst zur Auswahl der Parameter eingesetzt, oder es wird, unabhängig von den später verwendeten Lernverfahren, eine Rangliste der Parameter erstellt, anhand derer die Parametermengen bestimmt wurden.

Die Bewertung eines Auswahlverfahrens erfolgte analog zu der eines Klassifi-

kationsalgorithmus: Es wurde ein Referenzklassifikator ausgewählt, dessen Ergebnisse bei Verwendung der verschiedenen Parametermengen verglichen wurden. Zur Parameterauswahl stand hierbei immer nur die Trainingsmenge zur Verfügung, und die Testmenge wurde zur Beurteilung dieser Auswahl verwendet.

### 3.6.1. „Wrapper“-basierte Parameterauswahl

Eines der Auswahlverfahren war die sogenannte Wrapper-basierte „greedy forward feature selection“ [27].

Unter einem „Wrapper“ versteht man ein Verfahren, das durch Aufrufe eines zweiten Algorithmus zu seinen Ergebnissen kommt. Das „eingepackte“ Verfahren ist im vorliegenden Fall ein Lernalgorithmus. Seine Ergebnisse unter Verwendung verschiedener Parametermengen sind das Maß, anhand dessen die Parameter ausgewählt werden. Die Parameterauswahl ist ein „greedy“-Verfahren - dies bedeutet, dass ein einmal ausgewählter Parameter in späteren Schritten nicht mehr aus der Parametermenge entfernt wird. Das „forward“ im Namen des Algorithmus steht dafür, dass eine zunächst leere Parametermenge nach und nach gefüllt wird. Im Falle einer Menge, die zunächst alle verfügbaren Parameter beinhaltet und Schritt für Schritt verkleinert wird, spricht man von einer „backward feature selection“. In der vorliegenden Arbeit wurde die „forward“-Variante gewählt, da davon auszugehen war, dass bei der großen Zahl der Parameter die Unterschiede, die sich nach dem Entfernen eines Parameters aus der vollständigen Parametermenge in der Vorhersageleistung niederschlagen, nicht die tatsächliche Eignung dieses Parameters widerspiegeln, sondern durch zufällige Einflüsse zustande kommen.

Das Vorgehen zur Auswahl der Parameter verlief folgendermaßen: Bei einer leeren Startmenge wurde zunächst derjenige Parameter ausgewählt, der alleine die besten Ergebnisse bei der Klassifikation erzielte. In jedem folgenden Schritt

wurde die Menge der ausgewählten Parameter um den Parameter erweitert, dessen Hinzunahme zur größten Steigerung der Klassifikationsleistung führte. Dies wurde solange wiederholt, bis die Erweiterung der Parametermenge keine Steigerung der Vorhersageleistung mehr erbrachte.

### 3.6.2. Informationsbasierte Parameterauswahl

Im Gegensatz zur Wrapper-basierten Parameterauswahl wird bei der informationsbasierten Variante für jeden Parameter auf Grundlage der vorliegenden Daten ein Wert bestimmt, sein „Nutzen“, der, unabhängig vom später verwendeten Klassifikator, die Eignung dieses Parameters für die Klassifikation einschätzt.

Kernpunkt dieses Verfahrens ist die ausgewählte Nutzenfunktion. Im vorliegenden Fall wurde der auf den Grundlagen der Informationstheorie beruhende „information gain“ oder Informationsgewinn herangezogen[46]. Er gibt an, wie die Informationsentropie der Daten durch eine Aufteilung anhand des untersuchten Parameters abnimmt. Die Informationsentropie ist ein Maß für die Reinheit einer Datenmenge. Eine völlig einheitliche Menge, also eine Menge, in der beispielsweise nur Messpunkte von wachen Patienten vorliegen, hat eine Informationsentropie von 0. Bei einem Zwei-Klassen-Problem (wie es hier vorliegt) ist die maximale Informationsentropie 1 und wird genau dann erreicht, wenn eine gleiche Anzahl von Datenpunkten beider Klassen in der Menge vorhanden ist. Die Berechnung von Entropie und Informationsgewinn werden in Anhang B dargestellt.

Die Bestimmung des Informationsgewinns der Parameter ermöglicht es, sie zu sortieren, und dadurch eine Rangfolge der besten Parameter zu erstellen. Hierbei können die Parameter unabhängig voneinander betrachtet werden, es besteht aber auch die Möglichkeit, die Parameter nacheinander auszuwählen und die Eigenschaften bereits selektierter Parameter zu berücksichtigen.

Dies geschieht, indem die Datenmengen sukzessive aufgespalten werden. Jeder Parameterauswahl liegt eine gewisse Anzahl von Datenmengen zugrunde - bei der Auswahl des ersten Parameters zunächst nur eine. Sobald der Parameter mit dem höchsten Informationsgewinn bestimmt wurde, ist dadurch auch die Aufteilung der Datenmenge, durch die dieser Informationsgewinn realisiert wird, bekannt. Entsprechend dieser Aufteilung wird die ursprüngliche Datenmenge in zwei Untermengen (und in späteren Schritten - solange alle Aufspaltungen Mengen ergeben, die noch nicht perfekt sortiert sind - vier, acht,...) aufgespalten. Nachfolgende Parameter werden anhand des akkumulierten Informationsgewinn, der sich durch die Aufspaltung aller diese Mengen erzielen lässt, ausgewählt. Nach vollständiger Aufteilung der Datenmenge terminiert das Verfahren.

### 3.6.3. Randomisierte Parameterauswahl

Um zu beurteilen, ob die gezielte Parameterauswahl im Gegensatz zu einem zufälligen Ziehen der Parameter überhaupt von Vorteil ist, wurde eine randomisierte Methode angewendet. Sie bestand daraus, dass in jedem Schritt fünfzig zufällige Parametermengen erzeugt, die Klassifikation anhand dieser Parameter durchgeführt wurde, und die Ergebnisse gemittelt wurden.

## 3.7. Parameterloser Ansatz

Die oben beschriebene Berechnung der Parameter ist ein für den Einsatz der Klassifikationsverfahren nicht prinzipiell erforderlicher Schritt. Diese sind in der Lage, die anfallenden EEG-Rohdaten zu verarbeiten. Mithilfe einer Support-Vektor-Maschine (mit RBF-Kernel) wurde untersucht, inwieweit dieses Vorgehen zur erfolgreichen Klassifikation verwendbar ist.

Der Begriff „Rohdaten“ bedeutet hier, dass keine Parameter erstellt wurden. Die Messwerte wurden allerdings vorverarbeitet, siehe Abschnitt 3.3.2.

### 3.8. Vorgehen

Die in dieser Arbeit untersuchten Verfahren beeinflussen sich gegenseitig. Die Leistungsfähigkeit eines Klassifikators ist von den ausgewählten Parametern abhängig, während die Eignung eines Parameterauswahlverfahrens nur indirekt, also durch die Anwendung eines Klassifikationsverfahrens auf den ausgewählten Parametern, beurteilt werden kann. Dementsprechend mussten ein Startpunkt und eine geeignete Reihenfolge der Untersuchungen bestimmt werden.

Das angewendete Vorgehen war folgendes: Da die einzelnen Parameter selbst als eigenständige Indikatoren betrachtet werden können, wurde zunächst für jeden Parameter der auf den Daten erreichte  $\bar{P}_K$ -Wert ermittelt. Diese bildeten dann die Vergleichswerte für die zusammengesetzten Indikatoren. Im nächsten Schritt wurden die verschiedenen Parameterauswahlverfahren untersucht. Um sie vergleichen zu können, muss ein Klassifikationsverfahren eingesetzt werden. Aufgrund vorheriger Erfahrungen und Ergebnissen eines Vergleichstests wurden die Support-Vektor-Maschinen als insgesamt aussichtsreichstes Verfahren angesehen[37]. Den Empfehlungen der Softwaredokumentation folgend wurde daher eine Support-Vektor-Maschine mit einem RBF-Kernel als Referenzklassifikator ausgewählt, und die Parameterauswahlverfahren so miteinander verglichen. Nach der Auswahl des besten Selektionsverfahrens wurden die hierdurch bestimmten Parametermengen als Eingabe für die verschiedenen Lernverfahren verwendet. Zuletzt wurde die Klassifikation der rohen EEG-Daten durch die Referenz-Support-Vektor-Maschine evaluiert.

In einem abschließenden Schritt wurden die Verfahren, die sich auf dem Arbeitsdatensatz als die vielversprechendsten herauskristallisiert hatten, mit den

Daten aus dem Validationsdatensatz überprüft.

### 3.9. Quellen der verwendeten Software

Zur Klassifikation wurden zwei verschiedene Softwarepakete verwendet: Die Support-Vektor-Maschinen entstammen der libSVM-Softwarebibliothek<sup>1</sup>, während für alle anderen Klassifikatoren das WEKA-Paket<sup>2</sup> eingesetzt wurde. Bei beiden handelt es sich um frei verfügbare Software.

Alles weitere (Parameterauswahlverfahren, Auswertung, etc) wurde in der Programmiersprache „Python“ implementiert und kann auf Anfrage vom Autor dieser Arbeit bezogen werden.

---

<sup>1</sup><http://www.csie.ntu.edu.tw/~cjlin/libsvm>

<sup>2</sup>[www.cs.waikato.ac.nz/ml/weka/](http://www.cs.waikato.ac.nz/ml/weka/)



## 4. Ergebnisse

Der Ergebnisteil ist folgendermaßen aufgebaut: Zunächst werden die Resultate der Klassifikation mit einzelnen Parametern dargestellt. In den Abschnitten 4.2 und 4.3 werden Parameterauswahlverfahren beziehungsweise Klassifikatoren verglichen. Die Ergebnisse der Verwendung von rohen EEG-Daten finden sich in Abschnitt 4.4.

Abschnitt 4.5 stellt dar, inwieweit die gemittelten  $\bar{P}_K$ -Werte, die auf den  $P_K$ -Werten der einzelnen Patienten beruhen, mit denen, bei denen die Klassifikationsergebnisse mehrerer Patienten in die Berechnung eines  $P_K$ -Wertes einfließen, vergleichbar sind.

Im Abschnitt 4.6 finden sich die Resultate einer Bootstrap-Validation der Verfahren aus den Abschnitten 4.1 bis 4.4, die die höchsten  $\bar{P}_K$ -Werte erzielten.

### 4.1. Ergebnisse der einzelnen Parameter

Die zehn höchsten  $\bar{P}_K$ -Werte, die sich mit einzelnen Parametern erreichen lassen, sind in Tabelle 4.1 zusammengefasst.

Drei dieser Parameter werden im Folgenden zum Vergleich mit den in dieser Arbeit vorgestellten Verfahren verwendet. Es sind die beiden Parameter, durch

Parameter	$\bar{P}_K$ -Wert
PeEn0_49_5_200	$0.913 \pm 0.114$
WSMF8_49_a	$0.905 \pm 0.111$
PeEn0_40_5_200	$0.905 \pm 0.124$
PeEn0_39_6_200	$0.896 \pm 0.136$
WSMF0_49_a	$0.891 \pm 0.130$
ATrous_db3_5zpdD3_WKs_31-131_Energie	$0.891 \pm 0.121$
ATrous_db3_5zpdD3_sig_24-120_MaxAbsAmp	$0.881 \pm 0.138$
WSMF8_30_a	$0.869 \pm 0.139$
PeEn0_30_5_100	$0.869 \pm 0.138$
ATrous_db3_5zpdD4_Koeffs_Deriv2Var	$0.861 \pm 0.140$

**Tabelle 4.1.** – Ergebnisse der verschiedenen Parameter bei Errechnung und Mittlung des  $\bar{P}_K$ -Werts der einzelnen Patienten

die die größten  $\bar{P}_K$ -Werte erreicht wurden, also „PeEn0\_49\_5\_200“ - ein Parameter aus der Familie der Permutationsentropien, mit Quellfrequenzen zwischen 0 und 49 Hertz, und „WSMF8\_49\_a“, aus der Gruppe der gewichteten medianen Frequenzen, sowie, da beides EEG-Parameter sind, der AEP-Parameter mit der höchsten Vorhersagewahrscheinlichkeit.

## 4.2. Parameterauswahl

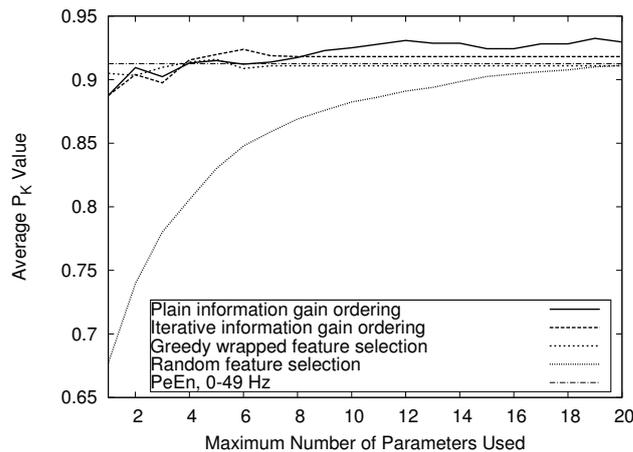
Einen Überblick über die Leistungen der einzelnen Parameterauswahlverfahren bietet Abbildung 4.1. Zum Vergleich enthält das Schaubild als Hilfslinie den höchsten durch einen einzelnen EEG-Parameter erreichbaren  $\bar{P}_K$ -Wert.

Die hier verglichenen Verfahren sind:

1. Die Wrapper-basierte Parameterauswahl
2. Informationsbasierte Auswahlmechanismen, und zwar

- a) Einfache informationsbasierte Auswahl: Unabhängige Bewertung der einzelnen Parameter
- b) Die iterative Variante: Hinzunahme der besten Parameter unter Berücksichtigung der bereits verwendeten.

### 3. Das zufällige Ziehen der Parameter



**Abbildung 4.1.** – Parameterauswahlverfahren im Überblick. Die Vorhersageleistung einer Support-Vektor-Maschine mit RBF-Kernel in Abhängigkeit der verwendeten Parameterauswahl und der maximal erlaubten Größe der Parametermengen. Als Hilfslinie eingezeichnet ist der höchste Wert, der sich durch einen einzelnen Parameter erreichen ließ.

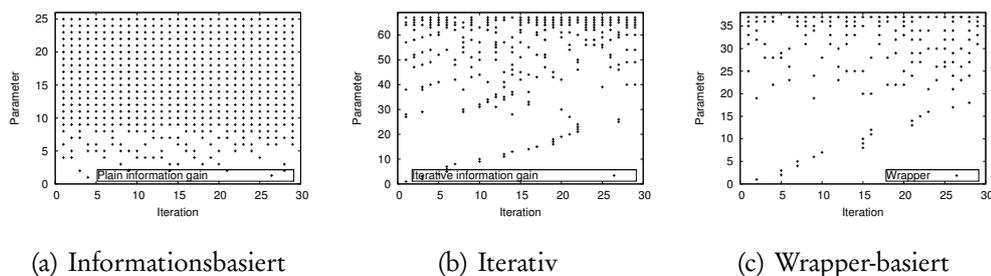
Auf der Ordinate finden sich die erreichten  $\bar{P}_K$ -Werte, auf der Abszisse die Anzahl von Parametern, die von der jeweiligen Auswahlmethode jeweils maximal ausgewählt werden durfte.

Es zeigt sich, dass die einfache informationsbasierte Parameterauswahl bei kleineren Parametermengen geringerer Vorhersagewahrscheinlichkeiten erreicht als die iterativen Verfahren. Allerdings können diese nur eine begrenzte Anzahl von Parametern auswählen (sie terminieren, sobald eines weiteren Parameters die Vorhersageleistung auf der Trainingsmenge nicht mehr steigert, oder die

Teilmengen, in die die Daten aufgesplittet wurden, allesamt nur noch Datenpunkte einer Klasse enthalten - siehe Abschnitt 3.6). Die einfache Parameterauswahl, die diese Einschränkung nicht hat, kann die Vorhersageleistung weiterhin steigern.

Die Ergebnisse der zufälligen Parameterauswahl liegen deutlich unterhalb denen der gezielten Vorgehensweisen; sie nähern sich erst bei den größeren Parametermengen an.

Bei der einfachen informationsbasierte Parameterauswahl wurden in den verschiedenen Durchläufen der Kreuzvalidation ähnliche Parametergruppen ausgewählt. Bei insgesamt 29 Zyklen, in denen jeweils 20 Parameter zu bestimmen waren, fanden sich insgesamt 25 Parameter in mindestens einer Gruppe wieder.



**Abbildung 4.2.** – Durch die verschiedenen Parameterauswahlverfahren selektierte Parametermuster. Bei der einfachen informationsbasierten Variante (a) werden meistens dieselben Parameter ausgewählt, bei der iterativen Modifikation dieses Verfahrens (b) und der Wrapper-basierten Version (c) unterscheiden sich die Parametergruppen deutlich

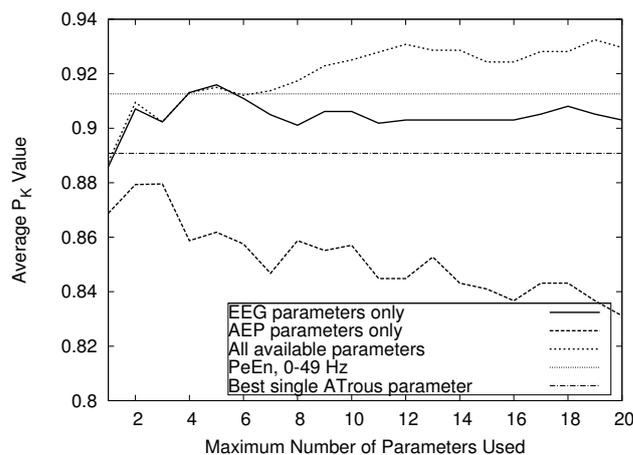
Im Gegensatz hierzu werden bei den iterativen Verfahren je nach Durchlauf unterschiedliche Parametermengen ausgewählt. Dies zeigt sich in einer deutlich höheren Gesamtzahl ausgewählter Parameter, bei der Wrapper-basierten Parameterauswahl insgesamt 48, im Fall der iterativen informationsbasierten Auswahl 67. Dies kontrastiert damit, dass beide dieser Verfahren in keinem einzigen Zyklus 20 Parameter auswählten, da die jeweiligen Algorithmen vorher

terminierten.

Abbildung 4.2 stellt dies graphisch dar. Die Teilabbildungen (a), (b) und (c) zeigen die Auswahlmuster der einfachen und der iterativen informationsbasierten Parameterauswahl sowie die des Wrapper-basierten Verfahrens. Die ins Auge springende ansteigende Linie bei den Teilabbildungen (b) und (c) ist ein Artefakt der Darstellung.

Zur Parameterauswahl wurde im weiteren immer die einfache informationsbasierte Variante verwendet.

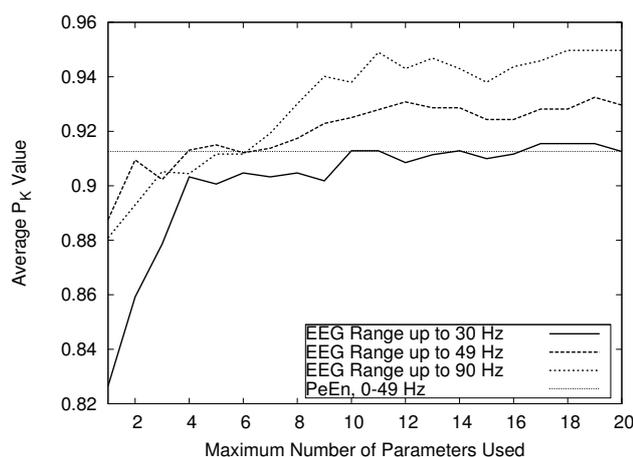
Durch eine Einschränkung der zur Verfügung stehenden Parametermenge kann die Auswahl beeinflusst werden. Es wurden zwei verschiedene Vorauswahlmethoden untersucht: Die Beschränkung auf entweder nur EEG- oder nur AEP-basierte Parameter sowie die Veränderung der Grenzfrequenz des Tiefpassfilters.



**Abbildung 4.3.** – Vergleich der Klassifikationsleistung mit ausschließlich EEG-basierten bzw. AEP-basierten Parametern sowie mit vollständigem Datensatz. Zum Vergleich: Die Ergebnisse der EEG- und AEP-Referenzparameter.

Abbildung 4.3 zeigt den Effekt der Kombination von AEP- und EEG-basierten Parametern gegenüber Datensätzen, die nur Daten aus einer dieser beiden Ka-

tegorien enthalten. Der maximal erreichbare  $\bar{P}_K$ -Wert von  $0.932 \pm 0.11$  ist bei der freien Parameterauswahl höher als bei beiden eingeschränkten Varianten. EEG-Werte alleine zeigen mit einem  $\bar{P}_K$  von  $0.916 \pm 0.11$  eine höhere Klassifikationsleistung, als es durch AEP-Parameter möglich ist ( $\bar{P}_K = 0.880 \pm 0.14$ ), und tolerieren eine Steigerung der Größe der erlaubten Parametermengen, während die Vergrößerung der Parametermenge bei ausschließlich auf AEPs beruhenden Messwerten die Gesamtleistung verringert.



**Abbildung 4.4.** – Einfluss der EEG-Filterung auf die Vorhersageleistung. Als Hilfslinie: Das Ergebnis eines Referenz-EEG-Parameters

Zur Bestimmung des Einflusses der Tiefpassfiltereinstellung auf die Klassifikationsleistung wurden drei verschiedene Grenzfrequenzen verglichen: Neben der im übrigen Teil dieser Arbeit verwendeten Grenzfrequenz von 49 Hz wurde das Standard-EEG-Band mit einer Grenzfrequenz von 30 Hz und ein erweitertes Band bis 90 Hz untersucht. Die Ergebnisse sind in Abbildung 4.4 dargestellt. Je höher die Grenzfrequenz eingestellt ist, je größer also die Menge der zur Verfügung stehenden Daten, desto höher ist im Endeffekt auch die Klassifikationsleistung. Bei einer maximalen Größe der Parametermenge von 20 beträgt der höchste erreichte  $\bar{P}_K$ -Wert im Standard-EEG-Band  $0.915 \pm 0.12$ , während Grenzfrequenzen von 49 beziehungsweise 90 Hertz zu maximalen  $\bar{P}_K$ -Werten von  $0.932 \pm 0.11$  respektive  $0.950 \pm 0.10$  führen.

### 4.3. Klassifikatoren

Um die Klassifikatoren untereinander zu vergleichen, wurden die Parameter anhand des Wertes der durch die Parameterauswahl vorgegebene Nutzenfunktion sortiert. Es wurden 20 Parametermengen erstellt. Hierbei entsprach die  $n$ -te Menge den  $n$  Parametern mit den höchsten Funktionswerten. Für jeden Klassifikator wurde die Vorhersageleistung bei jeder der Parametergruppen bestimmt.

Name	$\bar{P}_K$ -Wert	Parameterzahl
SVM, Polynomial	$0.935 \pm 0.11$	12
SVM, Sigmoid	$0.932 \pm 0.11$	19
SVM, RBF	$0.932 \pm 0.11$	19
SVM, Linear	$0.931 \pm 0.10$	12
NaiveBayes -K	$0.922 \pm 0.11$	10
NaiveBayes	$0.919 \pm 0.11$	5
Logistic Regression	$0.912 \pm 0.13$	12
Multilayer Perceptron	$0.894 \pm 0.13$	2
Bayes Net	$0.885 \pm 0.12$	11
J48	$0.802 \pm 0.19$	13
Zum Vergleich: PeEn0_49_5_200	$0.913 \pm 0.114$	1

**Tabelle 4.2.** – Die höchsten bei der Verwendung von maximal 20 Parameter erzielten  $\bar{P}_K$ -Werte für die verschiedenen Lernverfahren. Außerdem der Einzelparameter, durch den der größte  $\bar{P}_K$ -Wert erreicht wurde

Tabelle 4.2 zeigt die von den einzelnen Lernverfahren erreichten  $\bar{P}_K$ -Werte. In der ersten Spalte steht der Namen des verwendeten Klassifikationsverfahrens, in der zweiten der höchste  $\bar{P}_K$ -Wert, der unter den geschilderten Bedingungen erreicht werden konnte, und in der letzten Spalte die Anzahl der Parameter, die hierfür verwendet wurde.

Es zeigt sich, dass die Support-Vektor-Maschinen die höchsten  $\bar{P}_K$ -Werte errei-

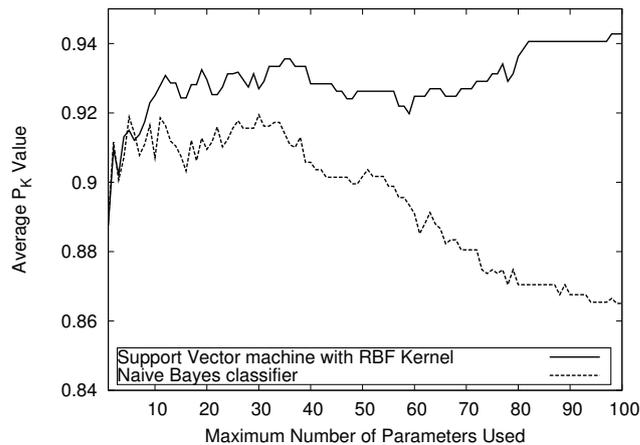
chen konnten; weiterhin fällt auf, dass zwei der Support-Vektor-Maschinen ihr Optimum erst bei höheren Parameterzahlen erreichen, während dies bei den anderen Verfahren nicht der Fall ist.

Tabelle 4.3 zeigt die  $\bar{P}_K$ -Werte, die sich bei vorgegebenen Parametermengen der Größe 10 erreichen ließen. Auch hier liefern die Support-Vektor-Maschinen höhere Vorhersagegenauigkeiten als die anderen Verfahren.

Name	$\bar{P}_K$
SVM, Polynomial	$0.928 \pm 0.11$
SVM, Sigmoid	$0.926 \pm 0.11$
SVM, RBF	$0.925 \pm 0.11$
NaiveBayes -K	$0.922 \pm 0.11$
SVM, Linear	$0.921 \pm 0.11$
NaiveBayes	$0.907 \pm 0.12$
Logistic Regression	$0.893 \pm 0.13$
Multilayer Perceptron	$0.864 \pm 0.19$
Bayes Net	$0.852 \pm 0.16$
J48	$0.770 \pm 0.19$

**Tabelle 4.3.** – Ergebnisse der Lernverfahren bei einer vorgegebenen Parametermenge der Größe 10

Die Verfahren unterschieden sich auch in der Fähigkeit, größere Parametermengen zu verarbeiten. Dies ist durch Abbildung 4.5 dargestellt, in der zwei Algorithmen verglichen werden: Die Referenz-Support-Vektor-Maschine mit einem RBF-Kernel und das beste Klassifikationsverfahren, das nicht auf dem Support-Vektor-Maschinen-Ansatz beruht, der naive Bayes-Algorithmus mit Kernelschätzung der Wahrscheinlichkeitsdichten, „Naive-Bayes -K“. Die durch den naiven Bayes-Algorithmus erreichten  $\bar{P}_K$ -Werte sinken mit zunehmender Parameterzahl; die Support-Vektor-Maschine zeigt weitere kleine Zuwächse bei den  $\bar{P}_K$ -Werten.

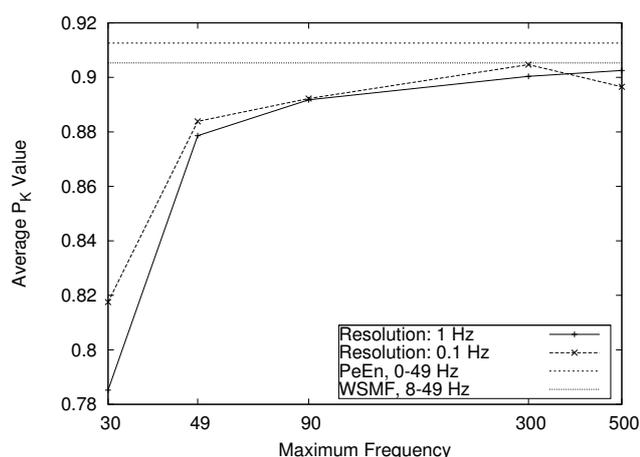


**Abbildung 4.5.** – Vergleich der Vorhersagequalität einer Support-Vektor-Maschine und eines Naiven Bayes-Verfahren bei großen Parametermengen

#### 4.4. Klassifikation roher EEG-Daten

Bei Anwendung einer Support-Vektor-Maschine auf das (bis auf die Fourier-Transformation) unbearbeitete Rohsignal wurden  $\bar{P}_K$ -Werte bis  $0.905 \pm 0.122$  erreicht. Dieser Wert kam bei Verwendung des Frequenzbereichs von 0 bis 300 Hz und einer Auflösung von 0.1 Hz zustande. Den Zusammenhang zwischen der Grenzfrequenz, der Auflösung und der Vorhersagequalität ist in Abbildung 4.6 dargestellt. Bei Auflösungen von 0.1 und 1.0 Hz sind hier die  $P_K$ -Werte dargestellt, die sich bei oberen Grenzfrequenzen von 30, 49, 90, 300, und 500 Hertz erreichen ließen. Die untere Grenzfrequenz lag jeweils bei 0 Hertz.

Abbildung 4.7(a) zeigt den Einfluss verschiedener Vorbehandlungen der Daten. Es handelt sich in dieser Darstellung um die Verwendung der Leistungsdichte einerseits und um die Normalisierung der einzelnen Frequenzen andererseits. Zum Vergleich ist wiederum die Klassifikationsleistung unter Verwendung nativer Daten angegeben. Durch die verschiedenen Vorbehandlungen ließen sich keinerlei Ergebnissteigerungen erreichen. Die Berechnung der Leistungsdichte führte zu einem deutlichen Abfall der  $P_K$ -Werte.



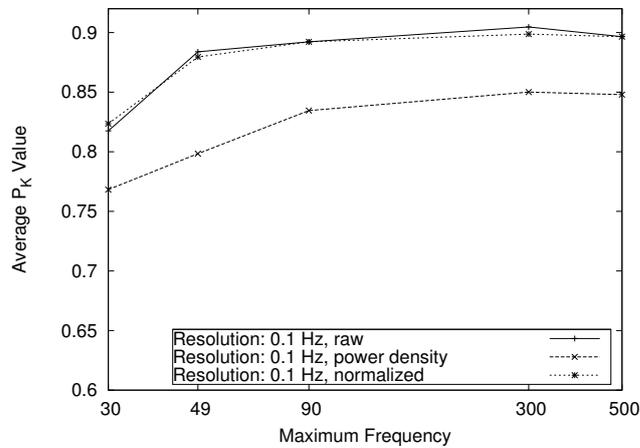
**Abbildung 4.6.** – Einfluss von oberer Grenzfrequenz und Auflösung auf die Vorhersagequalität unter Verwendung roher EEG-Daten. Auf der X-Achse abgetragen ist die obere Grenzfrequenz (die untere war immer bei 0 Hz), auf der y-Achse findet sich der erreichte  $P_K$ -Wert. Zum Vergleich sind die Resultate von zwei Referenzparametern aufgeführt.

In der Abbildung 4.7(b) sind die Effekte der Anwendung einer Hauptachsentransformation dargestellt. Hierbei wurde die Dimensionalität der Eingabedaten so gewählt, dass 90, 95 und 99 % der auf der Trainingsmenge beobachteten Varianz erklärt wurden. Außerdem ist ein Durchlauf aufgeführt, bei der vor der Anwendung der Hauptachsentransformation eine Normalisierung durchgeführt wurde. Auch hier zeigte sich eine Senkung der Vorhersagequalität durch die Vorbehandlung der Daten.

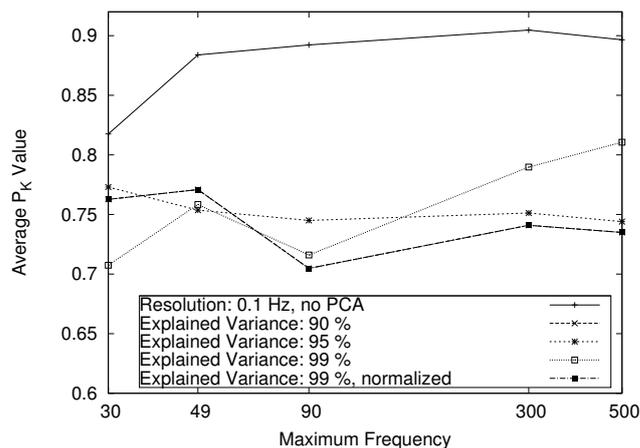
#### 4.5. Evaluation der Ergebnismetrik

Die bisherigen Ergebnisse entstanden, indem der  $P_K$ -Wert bei jedem Patienten errechnet wurde; diese Werte wurden dann gemittelt.

Um zu überprüfen, ob sich diese Mittelwerte mit  $P_K$ -Werten über größere Kol-



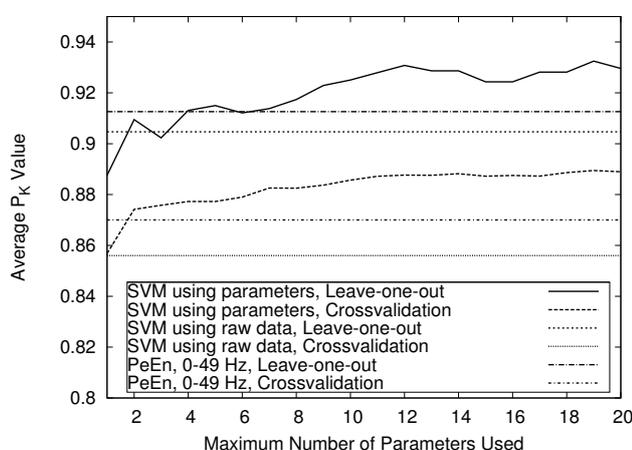
(a)



(b)

**Abbildung 4.7.** – Vorverarbeitung von Daten beim direkten Einsatz von Support-Vektor-Maschinen auf die fouriertransformierten Rohdaten. Abbildung (a): Modifikationen ohne Reduktion der Dimensionalität (Normalisierung durch Leistungsdichte sowie lineare Normalisierung). Abbildung (b): Einsatz der Hauptachsentransformation, mit und ohne vorhergehender Normalisierung.

lektive vergleichen lassen, wurde eine siebenfache Kreuzvalidation durchgeführt. Hierbei entstehen Patientengruppen der Größe 4 (sowie eine der Größe 5) als Testmengen, mit dementsprechend kleineren Trainingsmengen. Die Ergebnisse sind in Abbildung 4.8 dargestellt. Es zeigt sich, dass die  $\bar{P}_K$ -Werte, zu deren Berechnung die Klassifikationsergebnisse mehrerer Patienten Eingang gefunden haben, systematisch geringer als die individuell bestimmten sind - unabhängig davon, ob die Klassifikation mittels zusammengesetzter Indikatoren, einzelner Parameter, oder der direkten Verarbeitung der Rohdaten geschehen ist.



**Abbildung 4.8.** – Vergleich der Ergebnisse einer siebenfachen Kreuzvalidation (bei denen die Testmengen jeweils mindestens 4 Patienten enthalten) mit dem Leave-One-Out-Verfahren (also ein Patient je Testmenge). Größere Testmengen führen zu niedrigeren Vorhersagewahrscheinlichkeiten.

#### 4.6. Endauswertung

Diese Arbeit soll die Frage beantworten, ob sich durch die vorgestellten Verfahren die Vorhersagequalität gegenüber der Verwendung einfacher Parameter steigern lässt.

Für eine statistisch valide Beantwortung dieser Frage waren die Daten von 10 Patienten zurückgehalten worden (siehe Kapitel 3.4). Diese wurden als Testmenge verwendet, während alle anderen Daten die Trainingsmenge bildeten.

Durch die Parameterauswahl wurden 20 Parameter bestimmt. Das eingesetzte Klassifikationsverfahren war eine Support-Vektor-Maschine mit einem RBF-Kernel. Mittels eines Bootstrap-Verfahrens wurden der durchschnittliche  $P_K$ -Wert und die Konfidenzintervalle geschätzt.

Als Einzelparameter wurden die im vorherigen Verlauf als Maßstab gewählten Parameter verwendet (Siehe Abschnitt 4.1).

Der mit der Support-Vektor-Maschine erreichte durchschnittliche  $P_K$ -Wert betrug 0.921, bei einem von 0.835 bis 0.986 reichenden Konfidenzintervall. Im Gegensatz hierzu erreichten die einzelnen Parameter teilweise höhere Werte, siehe Tabelle 4.4. Aus der Anwendung der Support-Vektor-Maschine auf das Roh-EEG resultierte hier ein  $P_K$ -Wert von 0.914.

Modus	Mittelw.	Konf.intervall
SVM RBF, 20 Parameter	0.921	[0.835, 0.986]
Roh-EEG, 0-300 Hz, Aufl. 0.1 Hz	0.914	[0.838, 0.972]
PeEn0_49_5_200	0.900	[0.806, 0.977]
WSMF8_49_a	0.933	[0.857, 0.986]
Bestes ATrous	0.860	[0.768, 0.938]

**Tabelle 4.4.** – Vergleich einzelner Parameter mit zusammengesetzten Indikatoren, mit dem regulären Held-Back-Kollektiv

Diese Ergebnisse sind statistisch nicht signifikant. Eine andere, zufällig ausgewählte Testmenge von zehn Patienten führte zu völlig anderen Ergebnissen (Tabelle 4.5).

Modus	Mittelw.	Konf.intervall
SVM RBF, 20 Parameter	0.851	[0.797,0.901]
Roh-EEG, 0-300 Hz, Aufl. 0.1 Hz	0.829	[0.774,0.881]
PeEn0_49_5_200	0.836	[0.777,0.889]
WSMF8_49_a	0.842	[0.785,0.893]
Bestes ATrous	0.802	[0.742,0.858]

**Tabelle 4.5.** – Vergleich einzelner Parameter mit zusammengesetzten Indikatoren, bei zufällig ausgewählter zurückgehaltener Patientengruppe

## 5. Diskussion

Ziel dieser Arbeit war, die Eignung von Verfahren des maschinellen Lernens für die Erkennung intraoperativer Wachheit zu überprüfen. Hierbei sollte ein Überblick über die verschiedenen Methoden gewonnen und aussichtsreiche Kandidaten identifiziert werden und untersucht werden, ob diese statistisch signifikante Erhöhungen der Vorhersagewahrscheinlichkeit ermöglichen. Zwischen diesen Zielen besteht ein Konflikt: Eine möglichst große Anzahl von Verfahren sollte untersucht werden. Mit ihrer Anzahl wächst die Wahrscheinlichkeit des Auftretens von Methoden, die zufälligerweise auf den verwendeten Daten eine hohe Vorhersageleistung zeigen, ohne, dass dies für die Grundgesamtheit aller Patienten gälte. Deshalb muss dieses „multiple testing“ bei der Bewertung der Ergebnisse berücksichtigt werden. Dies kann z.B. durch die Anwendung des Bonferroni-Kriteriums geschehen[61]. Bei der vorliegenden hohen Anzahl von Hypothesen führt die Anwendung des Bonferroni-Kriteriums dazu, dass die Unterschiede zwischen den Vorhersageleistungen unrealistisch hoch sein müssen, um als statistisch signifikant anerkannt zu werden. Der in dieser Arbeit gewählte Weg war daher, die Auswahl der Verfahren und die statistische Validation in zwei unabhängigen Schritten mit unterschiedlichen Datensätzen durchzuführen.

In einem ersten Schritt wurden die Daten von 29 Patienten verwendet, um Verfahren aus dem Bereich des maschinellen Lernens zu identifizieren, die für die

Erkennung intraoperativer Wachheit geeignet erscheinen. Zur statistischen Absicherung dieser Ergebnisse wurde ein Validationsdatensatz, der die Daten von zehn Patienten beinhaltet, zurückgehalten. Anhand dieses Datensatzes wurde überprüft, ob mithilfe der ausgewählten Methoden eine Unterscheidung zwischen Wachheit und Bewusstlosigkeit mit höherer Wahrscheinlichkeit möglich ist als mit den vorbeschriebenen EEG- und AEP-Parametern. Hierbei zeigten sich keine statistisch signifikanten Unterschiede zwischen den verschiedenen Verfahren.

Es ließen sich allerdings im ersten Abschnitt der Untersuchungen Tendenzen feststellen, die über den gesamten Verlauf der Arbeiten konstant blieben. Sie legen es nahe, dass man die eingangs gestellte Frage, ob sich Techniken aus dem Bereich des maschinellen Lernens eignen, die Erkennung intraoperativer Wachheit zu verbessern, in weiteren Untersuchungen mit zusätzlichem Datenmaterial positiv beantworten wird.

Im folgenden werden zunächst (Abschnitt 5.1) die verschiedenen Ergebnisse des ersten Abschnitts der Untersuchungen interpretiert. Abschnitt 5.2 erörtert, inwieweit sich die quantitativen Resultate mit denen verwandter Arbeiten vergleichen lassen, bevor in Abschnitt 5.3 ein Überblick über diese Veröffentlichungen gegeben wird. In Abschnitt 5.4 wird diskutiert, wie das Verfehlen statistisch signifikanter Ergebnisse in dieser Untersuchung zu beurteilen ist. Ein Resümee findet sich im letzten Abschnitt.

## 5.1. Interpretation der Befunde

### 5.1.1. Datenquellen

Erster Schritt bei der Erstellung eines Indikators ist die Auswahl der Daten, die hierfür aufgezeichnet werden sollen. Die EEG-Daten können um AEPs ergänzt

werden, und die Weite des EEG-Spektrums kann unterschiedlich gewählt werden. In der vorliegenden Arbeit zeigte sich, dass sowohl die Hinzunahme von AEP-Daten als auch die größtmögliche Ausweitung des EEG-Spektrums zu einer Steigerung der Vorhersageleistung führten. Reine EEG- und AEP-Indikatoren scheinen keinen Vorteil gegenüber den besten Einzelparametern zu haben. Dieses Ergebnis ist allerdings nicht gleichbedeutend damit, dass die Verwendung der größtmöglichen Datenmenge automatisch zu einem im klinischen Einsatz verbesserten Indikator führt. Eine Erweiterung des EEG-Spektrums über das klassische EEG-Band hinaus erhöht den Einfluss des EMGs; an dieser Stelle muss eine Abwägung zwischen der erreichbaren Qualität des Indikators und seiner Störanfälligkeit getroffen werden. Es muss berücksichtigt werden, dass aufgrund seiner hohen Amplitude Einflüsse eines vorhandenen EMGs auch bei eingeschränkten EEG-Bändern nicht völlig aus der Berechnung der Parameter ausgeschlossen werden können. Auch bei der Hinzunahme von AEP-Daten muss eine Abwägung zwischen verbesserter Vorhersageleistung und größerem apparativen Aufwand sowie einer Empfindlichkeit gegenüber Störgeräuschen erfolgen.

### 5.1.2. Parameterauswahlverfahren und Parameterzahl

Die Dimension der Eingabedaten von Lernverfahren ist im Prinzip nicht begrenzt, es könnte also immer der volle Parametersatz aus EEG- und AEP-Parametern verwendet werden. Einige Argumente sprechen allerdings für die Begrenzung der Menge verwendeter Parameter:

1. Die Berechnung jedes Parameter ist mit einem Ressourcenbedarf verbunden. Je weniger Parameter berechnet werden müssen, desto kürzer ist die Laufzeit des Verfahrens, und desto geringer die Hardwareanforderungen.
2. Mit der Anzahl der verwendeten Parameter steigt das Risiko, dass einzelne Werte aufgrund von Artefakten nicht zu verwenden sind

3. Die Vorhersageleistung einiger Klassifikationsverfahren sinkt ab einer gewissen Parametermenge.
4. Geringere Parametermengen reduzieren das Risiko des „Overfittings“, also der Bevorzugung von Parameterkombinationen, die auf den Trainingsdaten, nicht aber in der Realität hohe Vorhersagewahrscheinlichkeiten erbringen[12].
5. Die Ergebnisse eines Auswahlverfahrens, also die Identität der berücksichtigten sowie der abgelehnten Parameter, kann als Maßstab für die Qualität dieser Parameter gesehen werden, und, ob und wie gut sie sich miteinander kombinieren lassen.

Deshalb wurden Auswahlverfahren, sogenannte „Feature-Selection“-Algorithmen, in der vorliegenden Untersuchung berücksichtigt. Insgesamt wurden drei verschiedenen Methoden eingesetzt, eine Wrapper-basierte Variante sowie zwei informationsbasierte Verfahren. Diese unterscheiden sich in ihren Vor- und Nachteilen.

Für die Wrapper-basierte Parameterauswahl spricht die Plausibilität des Ansatzes. Ein Lernverfahren wird dazu verwendet, die Parameter auszuwählen, auf deren Basis das selbe Verfahren später arbeiten soll. Demgegenüber stehen die informationsbasierten Methoden, bei denen die Parameterauswahl und die spätere Klassifikation voneinander unabhängig sind, und unklar ist, ob die selektierten Parameter eine geeignete Grundlage für die Prädiktion darstellen.

Ein pragmatischer Nachteil des Wrapper-basierten Verfahrens ist seine hohe Laufzeit: In jedem einzelnen Schritt muss für jeden der noch auswählbaren Parameter errechnet werden, wie sich seine Hereinnahme auf die Gesamtleistung auswirkt. Dies geschieht durch eine Kreuzvalidation. Da das Gesamtverfahren von Auswahl, Training, und Klassifikation ebenfalls im Rahmen einer Kreuzvalidation bewertet wird, laufen bei der Wrapper-basierten Parameterauswahl

zwei ineinander verschachtelte Kreuzvalidationen ab. Hierdurch wird die Anzahl notwendiger Lern- und Vorhersagevorgänge größenordnungsmäßig quadriert. Auch das zweite iterative Parameterauswahlverfahren, die iterative informationsbasierte Variante, hat im Gegensatz zur einfachen informationsbasierten Methode eine längere Laufzeit.

Beide iterativen Verfahren terminieren nach der Auswahl einer bestimmten Menge von Parametern. In den einzelnen Schritten der Kreuzvalidation selektierten sie deutlich unterschiedliche Parametermengen. Dies deutet darauf hin, dass beide Verfahren eine Tendenz haben, Parameter aufgrund zufälliger Eigenschaften auszuwählen. Die einfache informationsbasierte Parameterauswahl erstellt demgegenüber nur eine Reihenfolge über die verschiedenen Parameter, die in den Validationsschritten relativ konstant blieb, und die Auswahl beliebig großer Parametermengen erlaubt. In Abschnitt 4.2 zeigte sich, dass hierdurch eine höhere Vorhersageleistung möglich ist (Abbildung 4.1).

Zusammenfassend erscheint die informationsbasierte Parameterauswahl als ein geeigneter Weg, aus der großen Gesamtzahl zur Verfügung stehender Parameter eine Teilmenge zu selektieren. Die optimale Größe dieser Teilmenge hängt vom verwendeten Klassifikationsverfahren ab.

### 5.1.3. Verwendung der Rohdaten

Das Heranziehen roher (d.H. ausschließlich Fourier-transformierter) Daten als Grundlage der Klassifikation hat sich als mögliche Alternative oder Erweiterung zu den vorbeschriebenen Parametern herausgestellt. Die Vorhersageleistung reicht zwar nicht an die unter Verwendung der Parameter im besten Fall zu Erreichende heran; dennoch ließen sich Werte erreichen, die mit denen bei der Verwendung einzelner Parameter erreichten vergleichbar waren. Auffallend war, dass dieser - seinem Wesen nach lineare - Ansatz seine besten Ergebnisse erst bei Einbeziehung höhere Frequenzbereiche erreicht. Dies könnte entwe-

der auf eine Ausnutzung höherfrequenter EEG-Anteile, aber auch auf die Abhängigkeit vom Vorliegen von EMGs hindeuten. Es ergaben sich keine Hinweise darauf, dass die Vorverarbeitung der Daten per Normalisierung oder Hauptachsentransformation zu einer Steigerung der Vorhersagewahrscheinlichkeit führt.

#### 5.1.4. Klassifikatoren

In der vorliegenden Arbeit wurde eine umfangreiche, aber nicht vollständige Auswahl von Lernalgorithmen untersucht. Unter den überprüften Algorithmen scheinen die Support-Vektor-Maschinen für die gegebene Aufgabe am besten geeignet zu sein. Sie zeigen schon bei kleinen Parametermengen eine höhere Vorhersageleistung als die anderen Klassifikatoren und sind robuster bei der Verarbeitung höherdimensionaler Eingabedaten.

## 5.2. Vergleichbarkeit der Ergebnisse

Die vorliegende Arbeit erlaubt in dem Rahmen, in dem dies bei fehlender Signifikanz möglich ist, die Formulierung qualitativer Aussagen, wie z.B. dass die Kombination von AEP- und EEG-Parametern zu einer Steigerung der Vorhersagequalität führt. Die Verwendbarkeit der hier erreichten  $P_K$ -Werte zum quantitativen Vergleich mit den Ergebnissen anderer Arbeiten ist dagegen eingeschränkt. Hierfür gibt es mehrere Gründe:

1. Die zu geringe Größe des Validationsdatensatzes
2. Die Verwendung des „leave-one-out“-Verfahrens
3. Verschiedene Zielrichtungen: Erkennung von Wachheit bzw. Monitoring der Narkosetiefe

Auf die Größe des Validationsdatensatzes und die Konsequenzen wird weiter unten eingegangen.

In der vorliegenden Arbeit wurden das „leave-one-out-Verfahren“ zur Bestimmung der  $P_K$ -Werte verwendet. Dies bedeutet, dass die Testmenge in den einzelnen Validationsschritten durch die Datenpunkte eines einzigen Patienten gebildet wurde. Dies kann zu der in Abbildung 3.3 skizzierten Situation führen, dass die Qualität eines Indikators für Patientenkollektive überschätzt wird. Dieser Effekt konnte in einer Teiluntersuchung festgestellt werden (siehe Abbildung 4.8). Es zeigte sich bei der Klassifikation der zusammengefassten Messwerte mehrerer Patienten eine Senkung des  $P_K$ -Wertes sowohl der zusammengesetzten Indikatoren, als auch der einzelnen Parameter und der mit den Rohdaten operierenden SVMs. Für dieses Ergebnis sind verschiedene Ursachen denkbar: Aus der Schlafmedizin ist bekannt, dass die Fähigkeit von Probanden, auf äußere Reize zu reagieren, und ihr somnographisch festgestelltes Schlafstadium nicht eindeutig korrelieren[63]. Falls dies im Falle der Narkose auch gilt, würde es dazu führen, dass Indikatoren, die äquivalente Messwerte „korrekt“ gleich klassifizieren, dennoch den klinisch festgestellten Vigilanzzustand nicht in jedem Fall feststellen können. Dies ließe sich unter Umständen durch eine Eichung der Indikatoren auf die einzelnen Patienten ausgleichen. Die Notwendigkeit, ein Verfahren zu eichen, würde die Eignung für den klinischen Einsatz deutlich senken. Abgesehen von neurophysiologischen Unterschieden zwischen den Patienten können Faktoren, die nichts mit dem Aufwachvorgang per se zu tun haben, ebenfalls dazu führen, dass sich für verschiedene Patienten keine einheitliche Schwelle zwischen Wachheit und Bewusstlosigkeit feststellen lässt. Hierzu gehören z.B. akustische Unterschiede (Schwerhörigkeit des Patienten, Hintergrundgeräusche, in unterschiedlicher Lautstärke geäußerte Aufforderungen, die Hand zu drücken) sowie, durch die die klinische „Erkennung“ der Wachheit - durch Händedruck - trotz identischer Vigilanz und bei äquivalentem EEG und AEP früher oder später geschieht.

Die vorliegende Arbeit unterscheidet zwischen zwei Vigilanzzuständen: „Wach“ und „bewusstlos“. Dies unterscheidet sie von anderen Publikationen, in denen ein quantitativer Wert für die Narkosetiefe bestimmt wird. Die  $P_K$ -Werte beider Fragestellungen sind nicht direkt vergleichbar.

Quantitative Vergleiche der Ergebnisse dieser und anderer Arbeiten sind zusammenfassend nur eingeschränkt möglich. Das in dieser Arbeit ausgewählte Verfahren, für jeden Patienten einen  $P_K$ -Wert zu errechnen, und dann den Mittelwert zu bilden, steigert die Vorhersagewahrscheinlichkeit im Gegensatz zu der Variante, in der die Messungen mehrere Patienten zusammengefasst zur Berechnung der Vorhersagewahrscheinlichkeit verwendet werden.

### 5.3. Verwandte Arbeiten

Die Aufgabenstellung und die Instrumente, die in dieser Arbeit behandelt wurden, finden sich in etlichen Publikationen wieder, die sich mit diesem und ähnlichen Forschungsgebieten befassen.

Von der Grundlagen- bis zur klinischen Forschung ist die Erkennung intraoperativer Wachheit ein aktives Feld. Es ist nicht bekannt, wie genau die in der Anästhesiologie eingesetzten Hypnotika zu einem Bewusstseinsverlust führen [1]; auch die Gemeinsamkeiten und Unterschiede zwischen Schlaf, Koma, und Narkose sind nicht restlos geklärt [5] und werden weiter untersucht [10]. Die Definition der Bewusstlosigkeit ist nicht präzise gefasst, und alle Methoden, Bewusstlosigkeit oder Wachheit nachzuweisen, funktionieren indirekt. Die Erkennung von Wachheit gelingt nur, wenn eine Kette von Systemen funktionsfähig ist [21]. Ein Bruch dieser Kette an beliebiger Stelle führt zu der Diagnose „Bewusstlosigkeit“. Die in den bisherigen Studien bestimmte Häufigkeit intraoperativer Wachheit [60] unterschätzt also vermutlich die tatsächliche Inzidenz [53]. Die in der vorliegenden Arbeit verwendete Überwachung des EEGs als Mittel

zur Narkosesteuerung ist ein klassischer Ansatz[15]. Inwieweit die hierzu zur Verfügung stehenden Methoden zur Erkennung intraoperativer Wachheit zu einem tatsächlichen Nutzen für die Patienten führen ist unklar[44]. Eine größere Multicenterstudie konnte in Bezug auf Verhinderung intraoperativer Wachheit mit Erinnerung keine Vorteile für die Verwendung des „Bispektralen Index“ (BIS), eines der kommerziell erhältlichen EEG-Monitoringverfahren, gegenüber der Überwachung der endtidalen Narkosegaskonzentration feststellen[2]. Dies kann allerdings nicht als endgültiges Urteil über die Verwendbarkeit aller EEG-gestützten Methoden angesehen werden, zumal es Hinweise gibt, dass die Leistungsfähigkeit anderer EEG-Parameter die des BIS übertrifft[54].

Es gibt eine große Anzahl an Parametern, die zur Erkennung intraoperativer Wachheit vorgeschlagen wurden. Sie beruhen auf EEG- und AEP-Daten und bildeten eine Grundlage dieser Arbeit.

Auch die Kombination dieser Parameter zu zusammengesetzten Indikatoren ist bereits auf verschiedene Weise erfolgt. Durch uneinheitliche Ergebnismetriken sind die Ergebnisse dieser Arbeiten allerdings nicht direkt mit denen der vorliegenden Arbeit zu vergleichen. Die Verfahren, mittels derer die Zusammenfassung zu den Indikatoren durchgeführt wurde, beinhalteten z.B. die logistische Regression sowie eine Anzahl verschiedener Vorgehensweisen, die dem Bereich der „Fuzzy Logic“ entstammen [16, 18, 23, 54]. Auch die in dieser Arbeit im Fokus stehenden Support-Vektor-Maschinen wurden bereits für diese Aufgabe eingesetzt [34].

Eine sowohl in Hinblick auf die verwendeten Daten als auch auf die eingesetzten Methoden verwandte Arbeit [35] untersucht die Eignung verschiedener Kern-Methoden, darunter auch die SVMs, für das Narkosetiefenmonitoring. Auch hier stellt sich heraus, dass die Vorhersagewahrscheinlichkeit eines durch eine SVM erstellter Indikators oberhalb der durch die einzelnen Parameter erreichbaren liegt. Trotz der eingeschränkten Vergleichbarkeit scheinen die  $P_K$ -Werte allerdings systematisch geringer zu sein als in der vorliegen-

den Untersuchung. Eine mögliche Ursache ist die Auswahl der Eingangsdaten: Anstelle einer automatischen, hinsichtlich der auszuwählenden Parameter freien Parameterauswahl wurden in der erwähnten Arbeit manuell Parametergruppen selektiert und untereinander kombiniert. Diese Gruppen wurden aufgrund bestimmter Eigenschaften der Parameter wie ihren Ausgangsdaten - EEG oder AEP - oder ihres Grundprinzips, also z.B. lineare oder nichtlineare EEG-Parameter, zusammengesetzt. Es scheint, als ob eine automatische Parameterauswahl hier zu wesentlichen Zuwächsen in der Vorhersagewahrscheinlichkeit geführt hätte. Diese Auswahlmethoden sind einer der wesentlichen Neuansätze, die durch die vorliegende Arbeit in den Aufgabenbereich der Erkennung intraoperativer Wachheit eingeführt wurden. Eine andere Untersuchung zusammengesetzter Indikatoren erreicht unter Verwendung eines Fuzzy-Logik-Systems, das, im Gegensatz zu den Support-Vektor-Maschinen, das artefaktbedingte Ausfallen einzelner Parameter tolerieren kann, bei einem kleineren Patientenkollektiv für die Erkennung von Bewusstlosigkeit einen  $P_K$ -Wert von 0.94[19]. Eine Bestätigung dieser Ergebnisse steht noch aus, zumal bei dieser Arbeit Test- und Trainingsmenge identisch waren; hierdurch steigt die Gefahr einer Überschätzung der Leistungsfähigkeit.

Die Klassifikation mittels des direkten Einsatzes von Support-Vektor-Maschinen auf Fourier-transformierten EEG-Daten wurde in dieser Domäne unseres Wissens ebenfalls noch nicht untersucht. Es gibt allerdings andere Disziplinen, in denen dieses Vorgehen zu den etablierten Vorgehensweisen gehört, z.B. die Forschung an sogenannten Hirn-Maschinen-Schnittstellen („brain-computer interfaces“, BCI) . Hier geht es darum, dass der Untersuchte mittels der Vorstellung von Bewegungsmustern (z.B. dem Strecken und Beugen eines Fingers) über das EEG Informationen mit der Umgebung austauscht. Hierdurch kann z.B. Patienten mit fortgeschrittener Amyotropher Lateralsklerose die Interaktion mit der Umgebung ermöglicht werden [42]. Bei der Entwicklung von BCIs wurde eine große Bandbreite an Klassifikationsverfahren untersucht, wobei sich die Support-Vektor-Maschinen als einer der populärsten Ansätze her-

ausgestellt haben [14, 41].

## 5.4. Methodenkritik

### 5.4.1. Endpunkte

In dieser Studie werden die Patienten in die zwei Stadien „wach“ und „bewusstlos“ eingeteilt. Dies entspricht dem Kern der Aufgabenstellung, der Vermeidung intraoperativer Wachheit. Es gibt allerdings Argumente für eine feinere Unterscheidung, und zwar sowohl in der Unterteilung der Bewusstlosigkeit in verschiedene Narkosetiefen als auch für die Berücksichtigung der „Richtung“, d.H. der Unterscheidung zwischen Ein- und Ausleitung der Narkose.

Die Bestimmung einer quantitativen Narkosetiefe ist einerseits eine im Prinzip anspruchsvollere Aufgabe, da ein Indikator hier bei heterogeneren Eingabedaten über einen weiteren Wertebereich hinaus konsistente Einschätzungen erbringen muss. Demgegenüber stehen mehrere Eigenschaften dieser Variante, die zu methodischen und klinischen Vorteilen führen: Die in dieser Arbeit verwendeten Datenpunkte „wach“ und „bewusstlos“ liegen zeitlich nah aneinander und die Konzentrationen der Hypnotika sind aufgrund des Protokolls (langsame Einleitung) ähnlich. Hierdurch besteht die Gefahr, dass Unschärfen auftreten. Bei der Beurteilung der Narkosetiefe können die einzelnen Narkosetiefen unter visueller EEG-Kontrolle angesteuert und gehalten werden, so dass eine höhere Reinheit der Daten bei deutlich unterschiedlichen Funktionszuständen des ZNS, und dadurch klarer unterscheidbaren EEG-Signalen, erreicht wird.

Von klinischer Seite erbringt ein Narkosetiefenmonitor den Vorteil, dass auch eine Überdosierung der Hypnotika vermieden werden kann. Eine quantitative Darstellung ermöglicht es weiterhin, die Dynamik des Aufwachvorgangs zu

erkennen, und daher intraoperative Wachheit zu antizipieren und zu vermeiden, anstelle, wie bei dem hier verwendeten Modell, erst das Auftreten des unerwünschten Ereignisses feststellen zu können. Der Zustand, in dem sich die in dieser Arbeit als „bewusstlos“ eingeteilten Patienten befinden, entspricht zudem keinem intraoperativ anzustrebenden Status, da es sich um eine sehr flache Narkose handelt - das Risiko intraoperativer Wachheit ist also gerade in diesem Zustand besonders hoch.

Dichotome EEG/AEP-Indikatoren werden dementsprechend vermutlich nicht in den klinischen Einsatz gelangen. In der Zusammenschau sprechen diese Argumente dafür, bei zukünftigen Untersuchungen die Narkosetiefe als Zielgröße zu verwenden.

In der Klassifikation in „wach“ oder „bewusstlos“ wird nicht unterschieden, ob ein Patient gerade erwacht ist, oder wenige Sekunden vor dem Einschlafen steht (analog gilt dies für die Bewusstlosigkeit). Falls hierdurch elektrophysiologisch deutlich unterschiedliche Zustände zusammengefasst werden, kann dies die Ergebnisse in jede Richtung verzerren. Im Falle des natürlichen Schlafs ohne Medikamentengabe liegt nahe, dass ein Gehirn, das seit zehn Sekunden schläft, und dies über Stunden hinweg weiter tun wird, in einem anderen Zustand ist als eines, das nach stundenlangem Schlaf spontan erwacht und innerhalb weniger Sekunden die Wachheit erreichen wird[3]. Im Einschlafvorgang werden die Schlafstadien in aufsteigender Reihenfolge durchlaufen (1, 2, 3, 4; nach neuer Nomenklatur N1, N2, N3[63]), das Erwachen geschieht vorwiegend aus den Stadien 2 und REM[49]. Es gibt auch Hinweise darauf, dass im Schlafstadium 2 zwei verschiedene Stadien zusammengefasst sind, die sich u.a. im EEG unterscheiden[3]. Bei der medikamenteninduzierten Hypnose ist die Vigilanz im Wesentlichen ein Resultat der Konzentration der entsprechenden Medikamente im zentralen Nervensystem. Je nach verwendetem Wirkstoff unterscheiden sich die zugrunde liegenden Mechanismen[1]. Die hier im EEG vorliegenden (und von den Befunden im natürlichen Schlaf verschiedenen) Sta-

dien werden beim Erwachen in der umgekehrten Reihenfolge wie bei der Einleitung durchlaufen[5].

Eine Gleichbehandlung der verschiedenen Zustände direkt nach ihrem Eintreten bzw. kurz vor ihrem Ende erscheint also vertretbar, solange sichergestellt ist, dass die An- und Abflutung der Hypnotika in ähnlicher Geschwindigkeit geschieht.

#### 5.4.2. Aufspaltung der Daten

Dass in dieser Arbeit das Ziel statistisch signifikanter Aussagen verfehlt wurde, lag möglicherweise an einer unzureichenden Größe des Validationsdatensatzes, der nur einen Teil der zur Verfügung stehenden Daten umfasste. Die beobachtete Differenz der Ergebnisse, die mit einem alternativen, gesamplen Validationsdatensatz erzielt wurden, ist ein deutlicher Hinweis hierauf.

Der ursprüngliche Grund für die Aufspaltung der vorliegenden Daten war, dass zunächst ein Überblick über die verschiedenen Verfahren gewonnen werden musste. Dies entspricht dem Untersuchen einer großen Anzahl von Hypothesen: Jede neue Kombination von Auswahlverfahren, zugelassener Größe der Parametermenge, und Klassifikationsalgorithmus entspricht einer neuen Hypothese. Um dieses „multiple Testing“ bei nur einem Datensatz korrekt zu behandeln, müssen die Bedingungen, unter denen Ergebnisse als statistisch signifikant anerkannt werden, durch z.B. das Bonferroni-Kriterium verschärft werden[61]. Dies ist bei einer Anzahl von Hypothesen, die im oberen dreistelligen Bereich liegt, kein gangbarer Weg.

Als Alternative bleibt der gewählte Ansatz: Die Aufspaltung der vorhandenen Daten in zwei Teile, die Auswahl einer Hypothese auf einem Teil, und ihre statistische Untersuchung im zweiten. Es bleibt die Frage nach dem besten Verhältnis, in dem diese Aufteilung zu erfolgen hat. Hierbei müssen zwei An-

forderungen abgewogen werden: Zum Einen sollte der Validationsdatensatz so groß wie möglich sein, um die Chance, signifikante Ergebnisse zu erzielen, zu verbessern. Zum Anderen muss der Arbeitsdatensatz aber so gewählt werden, dass die Auswahl geeigneter Hypothesen überhaupt erst möglich wird. Hierbei muss auch der Tatsache Rechnung getragen werden, dass die Leistungsfähigkeit der Lernverfahren auch von der Präsentation geeignet großer Datenmengen abhängt.

Die Methoden, die in der vorliegenden Untersuchung zum Einsatz kamen, basieren auf einem einheitlichen Grundprinzip: Eine „Trainingsmenge“ von Daten wird verwendet, um einen Lernalgorithmus zu trainieren. Hierauf wird der derart vorbereitete Algorithmus zum Klassifizieren anderer Daten, der „Testmenge“, herangezogen. Prinzipiell gilt: Je größer die Trainingsmenge, desto besser. Zu kleine Trainingsmengen haben zur Folge, dass Algorithmen, die prinzipiell in der Lage wären, vorhandene Regelmäßigkeiten in den Daten zu erkennen, hieran scheitern, und stattdessen ihre Klassifikation auf Kriterien stützen, die nur zufällig in der Trainingsmenge vorliegen, in der Gesamtpopulation, aus der diese gezogen wurde, jedoch nicht. Dieser Vorgang, dass eine Spezialisierung auf die Trainingsmenge zu einer verringerten Leistung auf der Testmenge führt, wird als „Overfitting“ bezeichnet. Das Overfitting-Risiko wird durch die Präsentation einer möglichst großen Trainingsmenge reduziert, da mit steigender Stichprobengröße auch die Ähnlichkeit der Stichprobe zur Grundgesamtheit zunimmt.

In jedem Schritt der Untersuchungen werden also möglichst große Trainingsmengen benötigt. Die Daten aus dem Arbeitsdatensatz können problemlos in der abschließenden Untersuchung als Trainingsmenge für den Validationsdatensatz verwendet werden. Die Gegenrichtung ist nicht möglich - würde man Informationen über die zurückgehaltenen Daten zur Auswahl geeigneter Indikatoren verwenden, ist davon auszugehen, dass in der Endauswertung die Qualität der Indikatoren überschätzt wird.

Der Arbeitsdatensatz musste also relativ groß gewählt werden - auf Kosten der Menge an Daten, die für eine abschließende Validation zur Verfügung stand.

## 5.5. Schlussfolgerungen

Einige der vorgestellten Verfahren haben das Potential, die Erkennung intraoperativer Wachheit zu verbessern. Automatische Parameterauswahlverfahren stellen eine gute Alternative zum manuellen Zusammenstellen von Parametergruppen dar. Diese Gruppen können von Support-Vektor-Maschinen wirkungsvoll zu einem Indikator zusammengesetzt werden. Support-Vektor-Maschinen sind auch zur Verarbeitung der EEG-Rohdaten geeignet. Es ergaben sich Hinweise darauf, dass die Eichung der Indikatoren zu einer erhöhten Vorhersagewahrscheinlichkeit führen könnte. Für eine statistisch signifikante Absicherung dieser Ergebnisse ist die Verwendung zusätzlicher Daten notwendig.



## 6. Zusammenfassung

Ziel dieser Arbeit war es, die Eignung verschiedener Verfahren, die aus dem Bereich des „Maschinellen Lernens“ stammen, für die Erkennung von Wachheit und Bewusstlosigkeit zu untersuchen. Das Fernziel ist hierbei eine Methode, die es erlaubt, intraoperative Wachheit zuverlässig zu entdecken. Es wurden erstmals in diesem Feld die Eignung einer automatischen Parameterselektion sowie die Verwendung der EEG-Rohdaten als Eingabe für die Lernverfahren untersucht.

Es standen Daten von 39 Patienten zur Verfügung. Diese Patienten wurden unter kontinuierlicher EEG- und AEP-Ableitung zweimal narkotisiert. Von Interesse waren kurz vor oder nach klinisch bestimmtem Aufwachen bzw. Bewusstseinsverlust aufgezeichnete Daten. Hieraus wurden über hundert verschiedene quantitative EEG- und AEP-Parameter errechnet. Um viele verschiedene Verfahren untersuchen zu können, ohne die Möglichkeit, statistisch aussagekräftige Ergebnisse ungefährdet von diesem „multiple Testing“ zu erhalten, zu verlieren, wurden - randomisiert - zwei Datensätze erstellt, ein „Arbeitsdatensatz“ mit den Daten von 29 Patienten und ein „Validationsdatensatz“, der zehn Patienten umfasste. Letzterer wurde verwendet, um die auf dem Arbeitsdatensatz erzielten Ergebnisse statistisch zu bestätigen. Es gelang nicht, unter Verwendung der zurückgehaltenen Daten statistisch signifikante Ergebnisse zu erreichen. Jedoch zeigten sich nicht-signifikante Tendenzen, die sich über die ver-

schiedenen Untersuchungen konstant beobachten ließen.

Die untersuchten Verfahren umfassten zwei verschiedene Arten von Algorithmen: Auswahlverfahren, die die Aufgabe haben, aus einer Parametermenge die vielversprechendsten Kandidaten auszusuchen. Diese bildeten dann die Eingabedaten für Klassifikationsverfahren, die nach einer Trainingsphase bei vorgelegten Datenpunkte abschätzen, ob sie von einem bewussten oder einem wachen Patienten abgeleitet wurden.

Alle eingesetzten Parameterauswahlverfahren ermöglichten bei den gewünschten kleinen Parametermengen die Erzielung höherer Vorhersageleistungen als eine zufällige Parameterauswahl. Es zeigte sich, dass eine einfache Metrik, der „Information Gain“, bei der alle Parameter unabhängig voneinander beurteilt werden, gegenüber Verfahren, bei denen schrittweise neue Parameter abhängig von den bereits selektierten beurteilt wurden, von Vorteil war. Diese Metrik wurde für die weiteren Schritte ausgewählt.

Unter den verglichenen Klassifikationsverfahren zeigten Support-Vektor-Maschinen die höchsten Vorhersageleistungen, insbesondere bei der Verarbeitung hochdimensionaler Eingabedaten.

Die Eingabedaten konnten nach bestimmten Kriterien gefiltert werden. Es war möglich, ausschließlich EEG- oder AEP-Parameter zu verwenden, oder beide Arten parallel zuzulassen. Außerdem konnte die Grenzfrequenz des Tiefpassfilters, durch den das EEG geleitet wurde, Werte von 30 Hz (dies entspricht dem klassischen EEG-Band), 49 Hz und 90 Hz annehmen. Die Auswahl der größtmöglichen Datenmenge führte in beiden Fällen zu einer Steigerung der Klassifikationsqualität.

Es wurde außerdem untersucht, ob die Klassifikationsalgorithmen es ermöglichen, den Zwischenschritt der Parameterberechnung zu überspringen, und direkt die rohen (also nur Fourier-transformierten) EEG-Daten zu verwenden.

Die Leistungsfähigkeit dieses Ansatzes ist mit der Verwendung der vorbeschriebenen Parameter vergleichbar.

Zusammenfassend hat die Untersuchung Hinweise darauf geliefert, dass die verwendeten Verfahren sinnvoll zur Erkennung intraoperativer Wachheit eingesetzt werden können. Die bereits in vorangegangenen Arbeiten verwendeten Support-Vektor-Maschinen haben sich auch hier als das am besten geeignete Klassifikationsverfahren herausgestellt. Durch sie erstellte Indikatoren übertreffen in ihrer Vorhersageleistung einzelne Parameter, während ihr Einsatz auf den Fourier-transformierten Rohdaten zu Vorhersagewahrscheinlichkeiten führt, die im Bereich der Einzelparameter liegt. Automatische Parameterauswahlverfahren scheinen ein für die Auswahl geeigneter Parametergruppen geeignetes Werkzeug zu sein.

Zukünftige Untersuchungen müssen zeigen, ob sich die Ergebnisse dieser Arbeit bei größeren Datensätzen, die auch Daten aus tieferen Narkosephasen enthalten, bestätigen lassen. In diesem Fall empfiehlt sich, bei der Konstruktion eines neuen Narkose-Monitors das hier skizzierte Verfahren einzusetzen. Die Berechnung eines möglichst umfassenden Parametersatzes, wobei hier auch die Ergebnisse der Support-Vektor-Maschinen auf den Rohdaten als Parameter einfließen können, die Auswahl geeigneter Parameter anhand objektiver Kriterien, und ihre Zusammenfassung durch Support-Vektor-Maschinen sind hierbei die entscheidenden Schritte.



## A. Die Berechnung des $P_K$ -Werts

Der  $P_K$ -Wert ist ein Maßstab, der angibt, wie sehr die Reihenfolgen von durch einen Indikator, der für  $n$  Datenpunkte reellwertige Zahlen  $x_1, x_2, \dots, x_n$  liefert, ausgegebenen Werte mit der der tatsächlich vorliegenden Klassen  $y_1, y_2, \dots, y_n$  übereinstimmt.

Hier muss dementsprechend eine „Reihenfolge“ der verschiedenen Klassen definiert werden; dies geschieht bei dichotomen Klassen im Allgemeinen dadurch, dass der einen Klasse der Wert „0“ zugewiesen wird, der anderen der Wert „1“.

Ein  $P_K$ -Wert von 1 bedeutet intuitiv, dass es einen perfekten Schwellwert  $b$  gibt, so dass bei allen  $x_i, y_i$ -Paare, für die  $x_i > b$  gilt, die Klasse  $y_i = 1$  vorliegt, und für alle anderen  $(x_j, y_j)$  mit  $x_j \leq b$  immer  $y_j = 0$  gilt.

Bei einem  $P_K$ -Wert von 0 liegt ebenfalls eine Grenze vor, anhand derer sich die Datenpunkte entsprechend ihrer Klassen perfekt aufteilen lassen, mit dem Unterschied, dass hier alle Datenpunkte mit der Klasse „1“  $x$ -Werte haben, die unterhalb des Grenzwertes liegen.

Bei einem  $P_K$ -Wert von 0.5 lässt sich überhaupt kein Zusammenhang zwischen dem  $x$ -Wert und der beobachteten Klasse  $y$  feststellen; je näher der  $P_K$ -Wert an der 1 liegt, desto besser gelingt die Aufteilung.

Formal geht die Berechnung des  $P_K$ -Werts folgendermaßen vor sich: Es werden

zunächst drei Wahrscheinlichkeiten  $P_c$ ,  $P_d$ , und  $P_t$  berechnet. Hierbei ist  $P_c$  die Wahrscheinlichkeit, dass bei zwei zufällig gezogenen Paaren  $(x_i, y_i)$  und  $(x_j, y_j)$  mit  $i \neq j$  Konkordanz vorliegt, also  $y_i > y_j \wedge x_i > x_j$  (oder umgekehrt).  $P_d$  ist die Wahrscheinlichkeit einer diskordanten Beziehung, also  $y_i > y_j \wedge x_i < x_j$ . Der letzte Wert,  $P_t$ , gibt an, mit welcher Wahrscheinlichkeit sich Paare finden lassen, bei denen die  $x$ -Werte identisch, die beobachteten Klassen aber verschieden sind, also  $y_i \neq y_j \wedge x_i = x_j$ .

Aus diesen drei Wahrscheinlichkeiten lässt sich dann der  $P_K$ -Wert berechnen.

$$P_K = \left( \frac{P_c - P_d}{P_c + P_d + P_t} + 1 \right) / 2 \quad (\text{A.1})$$

Dieser hat die oben beschriebenen Eigenschaften und stellt ein häufig verwendetes Maß für die Qualität von Narkosetiefenindikatoren dar[65].

## B. Informationsentropie und Information

### Gain

Die Informationsentropie ist ein Maß für die Homogenität von Daten. Sie weist einer Stichprobe  $S$  von Datenpunkten, die zu einer beliebigen Anzahl von Klassen gehören können, einen Zahlenwert zu, der die Reinheit der Stichprobe quantifiziert. Im binären Fall, also beim Vorliegen von nur zwei Klassen 0 und 1, ist die Entropie folgendermaßen definiert:

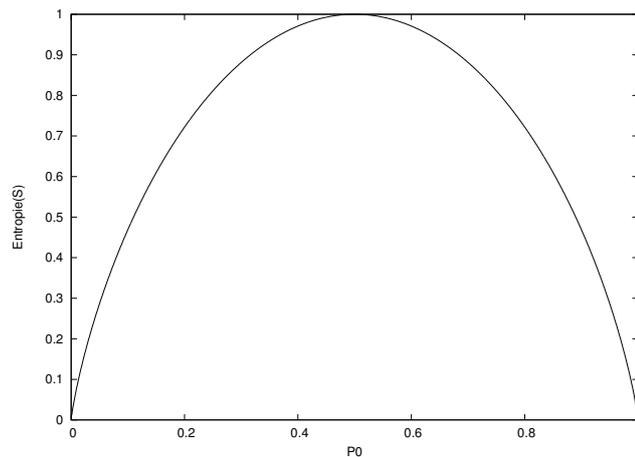
$$Entropie(S) = -p_1 \log_2 p_1 - p_0 \log_2 p_0 \quad (\text{B.1})$$

Hierbei sind  $p_1$  und  $p_0$  der Anteil der Elemente, die zu den Klassen 1 und 0 gehören. Abbildung B.1 zeigt den Entropieverlauf bei dichotomen Mengen.

Die Definition der Informationsentropie lässt sich auf beliebige Anzahlen verschiedener Klassen erweitern. Bei  $c$  verschiedenen Klassen ist die Entropie folgendermaßen definiert:

$$Entropie(S) = \sum_{i=1}^c -p_i \log_2 p_i \quad (\text{B.2})$$

Der Informationsgewinn, oder information gain, gibt für einen Parameter  $A$  und eine Stichprobe  $S$  die Reduktion der Entropie an, die erreichbar ist, wenn



**Abbildung B.1.** – Der Verlauf des Entropiewerts einer Stichprobe, in der zwei verschiedene Klassen, 0 und 1, vorkommen können. Je stärker eine der beiden Klassen überwiegt, desto näher gelangt der Entropiewert an die 0; sind beide Klassen gleichstark vertreten, erreicht die Entropie ihr Maximum von 1

die Stichprobe entsprechend Parameter  $A$  aufgespalten wird. Formal:

$$\text{Gewinn}(S,A) = \text{Entropie}(S) - \sum_{v \in \text{Werte}(A)} \frac{|S_v|}{|S|} \text{Entropie}(S_v) \quad (\text{B.3})$$

Hierbei beschreibt  $\text{Werte}(A)$  die Menge aller möglichen Werte, die  $A$  annehmen kann, und  $S_v$  die Stichprobe, die alle Elemente von  $S$  enthält, in denen  $A$  den Wert  $S$  annimmt.

Im vorliegenden Fall, in dem die verschiedenen Parameter reelle Zahlenwerte annehmen können, lassen sich die obigen Definitionen dennoch sinngemäß verwenden. Allerdings sind hierzu Modifikationen notwendig. Anhand eines Parameters  $R$  kann eine Menge von  $n$  Punkten  $x_1, \dots, x_n$  auf maximal  $n$  Weisen mittels eines Schwellwertverfahrens aufgespalten werden (falls die Punkte entsprechend  $R$  sortiert sind:  $\{(\{x_1\}, \{x_2, \dots, x_n\}), (\{x_1, x_2\}, \{x_3, \dots, x_n\}), \dots\}$ ). Der genaue Wert des Schwellwerts interessiert dementsprechend nicht; jeder Parameter  $R$  kann also durch eine Menge von Parametern  $R_0, \dots, R_n$  ersetzt werden. Jedes  $R_i$  kann für jeden Punkt  $x_j$  genau zwei verschiedene Werte annehmen -

je nachdem, in welcher der beiden Teilgruppen  $x_j$  bei einer Aufteilung entsprechend  $R_j$  eingeordnet wird.

Nach dieser Aufblähung der Parametermenge mit gleichzeitiger Ersetzung der reellwertigen Ursprungsparameter durch Dichotomien können Entropie und Informationsgewinn definitionsgemäß bestimmt werden.



## Literaturverzeichnis

- [1] ALKIRE, M. T., A. G. HUDETZ und G. TONONI: Consciousness and Anesthesia. *Science*, 322(5903):876–880, November 7 2008.
- [2] AVIDAN, M. S., E. JACOBSON, D. GLICK, B. A. BURNSIDE, L. ZHANG, A. VILAFRANCA, L. KARL, S. KAMAL, B. TORRES, M. O’CONNOR, A. S. EVERS, S. GRADWOHL, N. LIN, B. J. PALANCA und G. A. MASHOUR: Prevention of Intraoperative Awareness in a High-Risk Surgical Population. *New Engl J Med*, 365(7):591–600, 2011.
- [3] BRANDENBERGER, G., J. EHRHART und M. BUCHHEIT: Sleep Stage 2: An Electroencephalic, autonomic, and hormonal duality. *SLEEP*, 28(12):1535–1540, 2005.
- [4] BRANDT, C. und B. POMPE: Permutation entropy: a natural complexity measure for time series. *Phys Rev Lett*, 88:174102, 2002.
- [5] BROWN, E. N., R. LYDIC und N. D. SCHIFF: General Anesthesia, Sleep, and Coma. *New Engl J Med*, 363:2638–50, 2010.
- [6] BRUHN, J., H. RÖPCKE und A. HOEFT: Approximate entropy as an electroencephalographic measure of anesthetic drug effect during desflurane anesthesia. *Anesthesiology*, 92:715–26, 2000.

- [7] BRUHN, J., H. RÖPCKE, B. REHBERG, T. BOUILLON und A. HOEFT: Electroencephalogram approximate entropy correctly classifies the occurrence of burst suppression pattern as increasing anesthetic drug effect. *Anesthesiology*, 93:981–5, 2000.
- [8] CHEEK, D. B.: The Anaesthetized Patient can Hear and can Remember. *Am J Proctol*, 13:287–290, 1962.
- [9] CHERNICK, M. R.: Bootstrap Methods, A practitioner's guide. Wiley S Pro. Wiley, 1999.
- [10] CHUNG, F., D. HILLMAN und R. LYDIC: Sleep Medicine and Anesthesia - A New Horizon for Anesthesiologists. *Anesthesiology*, 114:1261–2, 2011.
- [11] EKMAN, A., M. LINDHOLM, C. LENNMARKEN und R. SANDIN: Reduction in the incidence of awareness using BIS monitoring. *Acta Anaesthesiol Scand*, 48:20–26, 2004.
- [12] ESMEIR, S. und S. MARKOVITCH: Occam's Razor Just Got Sharper. In: Proceedings of The 20th International Joint Conference on Artificial Intelligence (IJCAI-2007), Hyderabad, India, 2007.
- [13] FERENETS, R., T. LIPPING, A. ANIER, V. JÄNTTI, S. MELTO und S. HOVILEHTO: Comparison of entropy and complexity measures for the assessment of depth of sedation.. *IEEE Trans Biomed Eng*, 53(6):1067–77, 2006.
- [14] GARCIA, G., T. EBRAHIMI und J. VESIN: Support Vector EEG Classification in the Fourier and Time-Frequency Correlation. In: Proceedings of the IEEE-EMBS First International Conference on Neural Engineering 2003, Commercial and Forensic Applications, S. 591–594. IEEE, 2003.

- [15] GIBBS, F. A., E. L. GIBBS und W. G. LENNOX: Effect on the Electroencephalogram of Certain Drugs which Influence Nervous Activity. Arch Intern Med, 60:154–166, 1937.
- [16] HENSEL, T., G. STOCKMANN, D. JORDAN, E. S. FÜR EEG-AEP-NARKOSEMONITORING, E. F. KOCHS und G. SCHNEIDER: Entwicklung eines Takagi Sugeno Kang Fuzzy Modells für das Narkosemonitoring basierend auf EEG- und AEP-Parametern. Bio-med Tech (Berl), 2006.
- [17] HOCK, A., D. JORDAN, G. SCHNEIDER, S. JOCHUM und E. KOCHS: Improvement of the robustness of EEG spectral parameters against signal artifacts. Anesthesiology, 103(A-53), 2005.
- [18] HOCK, A., G. STOCKMANN, T. HENSEL, MCS-GRUPPE, E. F. KOCHS und G. SCHNEIDER: Kombination von EEG- und AEP-Parametern mit Fuzzy-Logik für das Narkosemonitoring. In: Proceedings der Gemeinsamen Jahrestagung der Deutschen, Österreichischen und Schweizerischen Gesellschaft für Biomedizinische Technik, S. P71, Zürich, Schweiz, 2006.
- [19] HORN, B., S. PILGE, E. F. K. G. STOCKMANN, A. HOCK und G. SCHNEIDER: A Combination of Electroencephalogram and Auditory Evoked Potentials Separates Different Levels of Anesthesia in Volunteers. Anesth Analg, 108(15):12–21, 2009.
- [20] HOSMER, D. W. und S. LEMESHOW: Applied logistic regression. Wiley S Pro. Wiley, Second Edition Aufl., 2000.
- [21] HUDETZ, A. G.: Are We Unconscious During General Anesthesia?. Int Anesthesiol Clin, 46(3):25–42, 2008.
- [22] JORDAN, C., C. WELLER, C. THORNTON und D. NEWTON:

- Monitoring evoked potentials during surgery to assess the level of anaesthesia. J Med Eng Tech, 19(2-3):77–9, Mar-Jun 1995.
- [23] JORDAN, D., G. SCHNEIDER, A. HOCK, T. HENSEL, G. STOCKMANN und E. F. KOCHS: EEG parameters and their combination as indicators of depth of anaesthesia.. Biomed Tech (Berl), 51(2):89–94, 2006.
- [24] JORDAN, D., G. SCHNEIDER und E. KOCHS: EEG permutation entropy separates consciousness from unconsciousness during anesthesia. www.asa-abstracts.com, 2006.
- [25] JORDAN, D., G. STOCKMANN, E. KOCHS und G. SCHNEIDER: Permutation entropy of the EEG indicates increase and decrease of the anesthetic level. www.asa-abstracts.com, A800, 2007.
- [26] JORDAN, D., G. STOCKMANN, E. F. KOCHS und G. SCHNEIDER: Median Frequency Revisited: An Approach to Improve a Classic Spectral Electroencephalographic Parameter for the Separation of Consciousness from Unconsciousness.. Anesthesiology, 107(3):397–405, 2007.
- [27] KOHAVI, R. und G. JOHN: Wrappers for Feature Subset Selection. Artif Intell, 97(1-2):273–324, 1997.
- [28] KRAMME, R.: Medizintechnik. Verfahren - Systeme - Informationsverarbeitung. Springer, Heidelberg, Berlin, 3. Aufl., 2007.
- [29] LAY, D.: Linear Algebra and It's Application, Kap. 7. Addison-Wiley, New York, 2000.
- [30] LEMPEL, A. und J. ZIV: On the Complexity of Finite Sequences. IEEE T Inform Theory, 22(1):75–81, 1976.
- [31] LENNMARKEN, C., K. BILDFORS, G. ENLUND, P. SAMUELSSON und

- R. SANDIN: Victims of awareness. Acta Anaesthesiol Scand, 46:229–231, 2002.
- [32] LIEBOVITCH, L., A. TODOROV, M. ZOCHOWSKI, D. SCHEURLE und L. COLGIN: Nonlinear properties of cardiac rhythm abnormalities. Phys Rev E, 59(3312-3329), 1999.
- [33] LÜCKE, D., G. STOCKMANN, M. GALLINAT, E. KOCHS und G. SCHNEIDER: Determination of significant parameters from auditory evoked potentials and their combination for assessment of depth of anaesthesia. Biomed Tech (Berl), 52:90–95, 2007.
- [34] LÜCKE, D., G. STOCKMANN, M. GALLINAT, E. F. KOCHS und G. SCHNEIDER: Klassifikation akustisch evozierter Potentiale durch Support-Vektor-Maschinen für das Narkosemonitoring. Biomed Tech (Berl), V122, 2006.
- [35] LÜCKE-JANSSEN, D.: Integration von Kern-Methoden in das medizinische Monitoring. Doktorarbeit, Fachbereich Ingenieurwissenschaften der Universität Duisburg-Essen, 2008.
- [36] MCCULLOCH, T. J.: Use of BIS Monitoring Was Not Associated with a Reduced Incidence of Awareness. Anesth Analg, 100:1221, 2005.
- [37] MEYER, D., F. LEISCH und K. HORNIK: The support vector machine under test. Neurocomputing, 55(1-2):169–186, 2003.
- [38] MITCHELL, T. M.: Machine Learning. McGraw-Hill, New York, 1997.
- [39] MOERMANN, N., B. BONKE und J. OOSTING: Awareness and recall during general anesthesia. Facts and Feelings. Anesthesiology, 79:454–464, 1993.
- [40] MYLES, P., K. LESLIE, J. MCNEIL, A. FORBES und M. T. CHAN:

- Bispectral index monitoring to prevent awareness during anaesthesia: the B-Aware randomised controlled trial. *Lancet*, 363:1757–1763, 2004.
- [41] MÜLLER, K.-R., C. W. ANDERSON und G. E. BIRCH: Linear and Nonlinear Methods for Brain-Computer Interfaces. *IEEE T Neur Sys and Reh*, 11(2):165–169, June 2003.
- [42] NEUMANN, N. und A. KÜBLER: Training Locked-in Patients: A Challenge for the Use of Brain-Computer Interfaces. *IEEE T Neur Sys and Reh*, 11(2), June 2003.
- [43] NEWTON HARVEY, E., A. L. LOOMIS und G. A. HOBART: Cerebral States during Sleep as Studied by Human Brain Potentials. *Sci Mon*, 45:191–192, Aug. 1937.
- [44] PALANCA, B. J. A., G. A. MASHOUR und M. S. AVIDAN: Processed electroencephalogram in depth of anesthesia monitoring. *Curr Opin Anesthesio*, 22:553–559, 2009.
- [45] PEARL, J.: Probabilistic reasoning in intelligent systems: Networks of plausible inference. Morgan-Kaufmann, San Mateo, CA, 1988.
- [46] QUINLAN, J. R.: Induction of Decision Trees. In: SHAVLIK, J. W. und T. G. DIETTERICH (Hrsg.): Readings in Machine Learning. Morgan Kaufmann, 1990. Originally published in Machine Learning 1:81–106, 1986.
- [47] RAMPERSAD, S. und M. MULROY: A case of awareness despite an adequate depth of anesthesia as indicated by a Bispectral Index (BIS) Monitor. *Anesth Analg*, 100:1363–1364, 2005.
- [48] RAMPIL, I. J.: A primer for EEG signal processing in anesthesia. *Anesthesiology*, 89(4):980–1002, October 1998.

- [49] RECHTSCHAFFEN, A., P. HAURI und M. ZEITLIN: Auditory awakening thresholds in REM and NREM sleep stages. *Percept Motor Skill*, 22(3):927–942, 1966.
- [50] RUMELHART, D., B. WIDROW und M. LEHR: The basic ideas in neural networks. *Commun ACM*, 37(3):87–92, 1994.
- [51] RUSSELL, I.: The Narcotrend 'depth of anaesthesia' monitor cannot reliably detect consciousness during general anaesthesia: an investigation using the isolated forearm technique. *Brit J Anesth*, 96(3):346–352, 2006.
- [52] SANDIN, R., G. ENLUND, P. SAMUELSSON und C. LENNMARKEN: Awareness during anaesthesia: a prospective case study. *Lancet*, 355:707–711, 2000.
- [53] SCHNEIDER, G.: Intraoperative Wachheit. *Anasth Intensiv Notf*, 38:75–84, 2003.
- [54] SCHNEIDER, G., R. HOLLWECK, M. NINGLER, G. STOCKMANN und E. F. KOCHS: Detection of consciousness by electroencephalogram and auditory evoked potentials.. *Anesthesiology*, 103(5):934–943, November 2005.
- [55] SCHNEIDER, G. und E. KOCHS: The search for structures and mechanics controlling anesthesia-induced unconsciousness. *Anesthesiology*, 107:195–198, 2007.
- [56] SCHOLKOPF, B. und A. J. SMOLA: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, MA, USA, 2001.
- [57] SCHWENDER, D., A. KAISER, S. KLASING, E. FABER-ZUELLIG, W. GOLLING und E. P. ANS K. PETER: Anästhesie mit Flunitrazepam/Fentanyl und Isofluran/Fentanyl. *Unbewusste*

- Wahrnehmung und akustisch evozierte Potentiale mittlerer Latenz..  
Anaesthesist, 43(5):289–297, 1994.
- [58] SCHWENDER, D., H. KUNZE-KRONAWITTER, P. DIETRICH, S. KLASING, H. FORST und C. MADLER: Conscious awareness during general anaesthesia: patient's perceptions, emotions, cognition and reactions. Br J Anaesth, 80:133–139, 1998.
- [59] SCHWILDEN, H. und H. STOECKL: Investigations on several EEG-parameters as indicators of the state of anaesthesia the median: A quantitative measure of the depth of anaesthesia. Anasth Intensivther Notfallmed, 15:279–86, 1980.
- [60] SEBEL, P., T. BOWDLE, M. GHONEIM, I. RAMPIL, R. PADILLA, T. J. GAN und K. DOMINO: The incidence of awareness during anesthesia: A multicenter United States study. Anesth Analg, 99:833–839, 2004.
- [61] SHAFFER, J.: Multiple hypothesis testing. Annu Rev Psycho, 46:561–584, 1995.
- [62] SIEMES, H.: Epilepsien bei Kindern und Jugendlichen, Kap. 4.3. Verlag Hans Huber, 2009.
- [63] SILBER, M. H., S. ANCOLI-ISRAEL, M. H. BONNET, S. CHOKROVERTY, M. M. GRIGG-DAMBERGER, M. HIRSHKOWITZ, S. KAPEN, S. A. KEENAN, M. H. KRYGER, T. PENZEL, M. R. PRESSMAN und C. IBER: The visual scoring of sleep in adults. J Clin Sleep Med, 3(2):121–131, 2007.
- [64] SLEIGH, J. und J. DONOVAN: Comparison of bispectral index, 95% spectral edge frequency and approximate entropy of the EEG, with changes in heart rate variability during induction of general anaesthesia. Brit J Anesth, 82(5):666–671, 1999.
- [65] SMITH, W., R. DUTTON und N. SMITH: Measuring the performance of

- anaesthetic depth indicators. *Anesthesiology*, 84:38–51, 1996.
- [66] THAKOR, N. und S. TONG: Advances in quantitative electroencephalogram analysis methods. *Annu Rev Biomed Eng*, 6:453–95, 2004.
- [67] THORNTON, C. und R. SHAPE: Evoked Responses in Anaesthesia. *Brit J Anesth*, 81:771–781, 1998.
- [68] TUNGSTAN, M. E.: Detecting wakefulness during general anaesthesia for caesarean section. *Brit Med J*, 1(6072):1321, May 21 1977.
- [69] WILHELM, W., J. BRUHN und S. KREUER (Hrsg.): Überwachung der Narkosetiefe. Grundlagen und klinische Praxis, Kap. Entstehung des EEG: Das normale EEG und pathologische Befunde, S. 5–26. Deutscher Ärzte-Verlag, Köln, 2006.
- [70] WILHELM, W., J. BRUHN und S. KREUER (Hrsg.): Überwachung der Narkosetiefe. Grundlagen und klinische Praxis, Kap. Akustisch evozierte Potentiale - Grundlagen, S. 64–80. Deutscher Ärzte-Verlag, Köln, 2006.
- [71] ZHANG, X., R. ROY und E. JENSEN: EEG complexity as a measure of depth of anesthesia for patients. *IEEE Trans Biomed Eng*, 48:1424–1433, 2001.