Johann Wolfgang Goethe-Universität
Frankfurt am Main

Erkennung kritischer Zustände von Patienten mit der Diagnose „Septischer Schock“ mit einem RBF-Netz

**Fred H. Hamker, Jürgen Paetz, Sven Thöne,
Rüdiger Brause, Ernst Hanisch**

INTERNER BERICHT 4/00

Fachbereich Informatik
Robert-Mayer-Straße 11-15
60054 Frankfurt am Main

ISSN 1432-9611

erschienen im September 2000

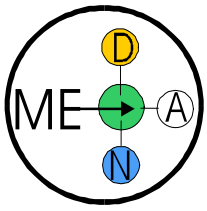
Autoreninformation:

Dr.-Ing. Fred H. Hamker	Hamker@Informatik.Uni-Frankfurt.de oder fred@klab.caltech.edu
Dipl.-Math. Jürgen Paetz	Paetz@Informatik.Uni-Frankfurt.de
Sven Thöne, AiP	st@nerd.de
PD Dr.rer.nat. Rüdiger Brause	Brause@Informatik.Uni-Frankfurt.de
Prof. Dr.med. Dr.med.dent. E. Hanisch	E.Hanisch@Em.Uni-Frankfurt.de

Danksagung:

Im Voraus möchten wir uns bei allen Medizinern bedanken, die uns mit ihrem Expertenwissen bei der Analyse zur Seite standen, insbesondere bei Frau Dr. Holzer. Ein Dankeschön geht auch an Herrn cand. inf. Marcus Pauen, der im Rahmen seiner Diplomarbeit unterstützend bei der Clusteranalyse mitgewirkt hat.

Näheres zum Projekt MEDAN findet sich unter:



http://medan.de

Dieses Projekt wird von der Deutschen Forschungsgemeinschaft unter dem AZ HA1456/7-1 gefördert.

Inhaltsverzeichnis

INHALTSVERZEICHNIS 3

1. EINLEITUNG 5

2. DATEN 8

2.1. Frühere Untersuchungen 8

2.2. Datenvorverarbeitung 9
2.2.1. Eliminierung von Mustern mit zu vielen fehlenden Werten 9
2.2.2. Normalisierung 9

2.3. Verwendete Variablen 9
2.3.1. Studie A1 10
2.3.2. Studie A2 11
2.3.3. Studie B 12
2.3.4. Studie C 14

3. DAS VERWENDETE NEURONALE NETZ 16

3.1. Kurzdarstellung des verwendeten Algorithmus 16
3.1.1. Aufbau 16
3.1.2. Adaption der Repräsentationsschicht 17
3.1.3. Einfügen von Knoten in die Repräsentationsschicht 17
3.1.4. Adaption der Ausgabeschicht 17
3.1.5. Adaption der Zähler und Kanten der Knoten in der Repräsentationsschicht 18

4. VORGEHENSWEISE 19

4.1. Untersuchungen 19

4.2. Kriterien 19
4.2.1. Gesamtprognose, Sensitivität und Spezifität 19
4.2.2. Test und Validierung der Ergebnisse 20
4.2.3. Selektion eines neuronalen Netzes 21
4.2.4. Gewinnung eines Erwartungsintervalls 22

4.3. Durchführung 23
4.3.1. Simulationsparameter und Training 23
4.3.2. Ein Gütemaß der Diagnose 24
4.3.3. Sensitivität und Spezifität 25

5. ERGEBNISSE DES NEURONALEN NETZES 26

5.1. Studie A1 26
5.1.1. Gesamtprognose 26
5.1.2. Gegenüberstellung von Sensitivität und Spezifität 27
5.1.3. Individuelle Warnungen bei verstorbenen Patienten 27

5.1.4. Individuelle Warnungen bei überlebenden Patienten 28
5.1.5. Mittlere Anzahl an Warnungen 29
5.1.6. Tagesanalyse 30
5.1.7. Gütemaß 30

5.2. Studie A2 31
5.2.1. Gesamtprognose 31
5.2.2. Gegenüberstellung von Sensitivität und Spezifität 32
5.2.3. Individuelle Warnungen bei verstorbenen Patienten 33
5.2.4. Individuelle Warnungen bei überlebenden Patienten 34
5.2.5. Mittlere Anzahl an Warnungen 35
5.2.6. Tagesanalyse 35
5.2.7. Gütemaß 35

5.3. Studie B 36
5.3.1. Gesamtprognose 36
5.3.2. Gegenüberstellung von Sensitivität und Spezifität 37
5.3.3. Individuelle Warnungen bei verstorbenen Patienten 38
5.3.4. Individuelle Warnungen bei überlebenden Patienten 39
5.3.5. Mittlere Anzahl an Warnungen 39
5.3.6. Tagesanalyse 40
5.3.7. Gütemaß 40

5.4. Studie C 41
5.4.1. Gesamtprognose 41
5.4.2. Gegenüberstellung von Sensitivität und Spezifität 42
5.4.3. Individuelle Warnungen bei verstorbenen Patienten 43
5.4.4. Individuelle Warnungen bei überlebenden Patienten 44
5.4.5. Mittlere Anzahl an Warnungen 45
5.4.6. Tagesanalyse 45
5.4.7. Gütemaß 45

5.5. Fazit 46

6. ERGEBNISSE DER CLUSTERANALYSE 46

6.1. Zustände der Patienten 47

6.2. Korrelierte Variablen der Patienten 49

7. ZUSAMMENFASSUNG 50

LITERATUR 52

1. Einleitung

Das von der abdominalen Sepsis verursachte Multiorganversagen prägt wesentlich das Bild der Intensivmedizin und ist mit einer hohen Letalität assoziiert. Die folgende Untersuchung ist ein Teilergebnis der Forschung der MEDAN-Arbeitsgruppe, mit dem Ziel, die Letalität auf der Intensivstation zu senken, und basiert auf Daten von zwei Vorstudien.

Erste Betrachtungen des Datenmaterials zeigen, dass kaum ein einzelner Parameter allein einen Hinweis über die Überlebenswahrscheinlichkeit eines Patienten liefert (Abbildung 1.1)[1]. Die Untersuchung geht daher der Frage nach, ob kritische Patientenzustände[2], die letztendlich zum Tode des Patienten führen, frühzeitig mit einem neuronalen Netz[3] erkannt werden können. In Ergänzung zu diesen umfassenden Zuständen ist die Betrachtung hinsichtlich des Ausfalls spezifischer Subsysteme wie Herz/Kreislauf, Leber, Lunge, Niere o.ä. ebenfalls möglich. Somit lässt sich das Ziel dieses Teilgebiets im MEDAN-Projekt folgendermaßen beschreiben:

> Das längerfristige Ziel dieses Teilgebiets des MEDAN-Projektes ist die Entwicklung eines Frühwarnsystems, welches dem behandelnden Arzt eine Bewertung des aktuellen Patientenzustands erleichtern soll. Eine mögliche Aussage könnte beispielsweise sein: Vorsicht, die Mehrheit der Patienten, die sich in einem ähnlichen Zustand wie der aktuelle Patient befanden, sind verstorben.

Es ist unrealistisch anzunehmen, dass ein System zu jedem Zeitpunkt das tatsächliche Ergebnis (verstorben/nicht verstorben) vorhersagen kann, da Voruntersuchungen von Patienten mit SIRS, Sepsis und septischem Schock auf starke Überlappungen der Datenbereiche zwischen überlebenden und verstorbenen Patienten hinweisen. Folglich sollte der Arzt durch ein System unterstützt werden, welches den Zustand des Patienten überprüft und eine Warnung bei kritischen Zuständen ausgibt. Eine Entwarnung soll aber bewusst nicht gegeben werden: Der Arzt wird durch das System nur zusätzlich informiert. Er darf sich nicht absolut auf die Angaben verlassen, sondern wird lediglich gewarnt. Die Umkehrung des Warnmechanismus (wenn nicht gewarnt wird, ist der Patient unkritisch) darf dem Arzt nicht nahegelegt werden.

Eine Warnung ist allerdings nur dann sinnvoll, wenn nicht zu häufig ein Fehlalarm auftritt und wenn tatsächlich kritische Zustände oder Trends aufgefunden werden können. Was ist aber ein kritischer Zustand, und wie kann man ihn auffinden?

Eine Möglichkeit einen kritischen Zustand zu beschreiben, sind die Zustände, die von nicht überlebenden Patienten angenommen werden. Wie könnte man die Zustände aus den Beispielen früherer Patienten bestimmen, und innerhalb welcher Bereiche sind sich die Zustände bezüglich ihrer medizinischen Konsequenz für den Patienten ähnlich? Man kann so vorgehen, dass zunächst jedes Datentupel aus Messungen für einen Zeitpunkt eines Patienten, ein sog. Muster, mit dem Label *nicht Exitus* versehen wird, wenn er letztendlich überlebt und jedes Muster eines Patienten mit dem Label *Exitus* versehen wird, wenn er auf der Intensivstation stirbt. Der durch die Variablen der Messungen aufgespannte Raum, der Musterraum, wird nun mittels sog. radialen Basisfunktionen unterteilt und jeder Punkt (jeder Patientenzustand) im Raum unterschiedlich stark auf die Klassen *Exitus* und *nicht Exitus* abgebildet. Hierzu kann ein wachsendes RBF-Netz verwendet werden, welches ähnliche Zustände hinsichtlich ihrer medizinischen Konsequenz interpoliert und damit die Beispieldaten früherer Patienten generalisiert (Abbildung 1.2). Ein kritischer Zustand zeichnet sich bei dieser Methode dadurch aus,

---

1 Für weitere Histogramme siehe http://www.medan.de/Histogramme/Vorstudie/Exitus/histogramme.html

2 Ein Patientenzustand soll durch die gemessenen Parameter, die Medikation und die weiteren Behandlungsparameter definiert sein.

3 Für eine Übersicht des Einsatzes neuronaler Netze in der Medizin siehe [BRAUSE, 1999].

dass er durch den Lernvorgang auf die Klasse *Exitus* abgebildet wird, allerdings unabhängig vom Zeitpunkt der Erhebung des Messwertes. Dieser Fall tritt ein, wenn die Mehrzahl der Patientenzustände aus der Nähe[4] dieses punktuellen Zustands von Patienten stammen, die verstorben sind. In Bereichen mit überlappenden Klassen, geht die Auftrittswahrscheinlichkeit der Klassen mit in die Zuordnung des Zustands ein. Durch die Veränderung der Auftrittswahrscheinlichkeiten im Trainingsdatensatz lässt sich Einfluss auf die Sensitivität (die Empfindlichkeit dafür, eine Warnung an den Arzt zu geben) nehmen. Über die Angabe eines Sicherheitsmaßes, welches aus dem neuronalen Netz gewonnen wird, kann eine weitere Gewichtung des kritischen Zustands angegeben werden.

Das Datenmaterial besteht aus den wichtigsten klinischen Parametern auf der Intensivstation, wobei deren Signifikanz für einen kritischen Zustand noch nicht eindeutig geklärt ist. Die oben beschriebene Hauptlinie wird durch folgende Hypothesen und Überlegungen ergänzt:

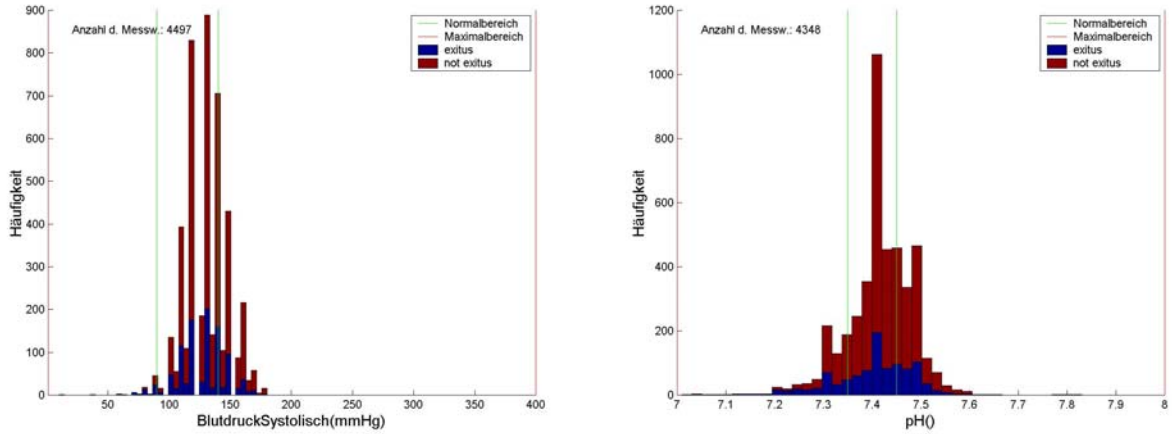


Abbildung 1.1. Histogramme des systolischen Blutdrucks und des pH Wertes.

- Da die Normalwerte der physiologischen Parameter von Männern und Frauen gelegentlich verschieden voneinander sind, ist es eventuell ratsam, zwischen diesen Kategorien zu unterscheiden. Eine Möglichkeit ist die Verwendung eines binären Eingabeparameters, mit dem Nachteil, dem neuronalen Netz die Lernaufgabe zu überlassen. Gerade bei RBF-Netzen ist der Ansatz, nach Mann und Frau getrennte neuronale Modelle zu trainieren, vielversprechender, da nominale Variablen aufgrund der zur Aktivierungsberechnung verwendeten euklidischen Metrik schlecht verarbeitet werden können. Da eine vorherige Histogrammanalyse[5] allerdings keine grundsätzlichen Unterschiede in der Verteilung jeder Variable aufzeigte, wird zunächst auf eine Unterscheidung verzichtet.
- Es wird vermutet, dass mit dem fortschreitenden Aufenthalt auf der Intensivstation eine bessere Prognose hinsichtlich der Gefahr zu sterben geleistet werden kann. Eventuell sind besonders kritische Zustände kurz vor dem Tod vorhanden[6]. Allerdings ist dann zu klären, wie in diesem Fall der Arzt möglichst frühzeitig gewarnt werden kann. Mit dem oben beschriebenen Ansatz sollten auch diese Zustände erkennbar sein. Indem die Prognose jeweils an verschiedenen Tagen des Aufenthalts, relativ bezogen zur Entlassung oder Todesfall, durchgeführt wird, soll das Potential einer möglichst frühzeitigen Diagnose unter Verwendung wesentlicher physiologischer Parameter ermittelt werden. Weitere interessante Phasen sind die Entwicklung im Zeitraum ab der Aufnahme auf der Intensivstation und um das erstmalige Auftreten des septischen Schocks herum.
- Durchläuft jeder Patient andere Zustände oder weisen die Patienten Gemeinsamkeiten auf, die sich generalisieren lassen? Ein Ansatz dieses zu untersuchen, besteht darin, den Unter-

4 Die Ähnlichkeit ist bei diesem Verfahren ein Maß, welches vom Zustand abhängig ist.

5 Siehe http://medan.de/Histogramme/Vorstudie/Geschlecht/histogramme.html

6 Leider werden die Messungen oft seitens der Medizin kurz vor dem Tod des Patienten eingestellt.

schied zu messen, wenn vollkommen unbekannte Patienten in der Diagnosegruppe (Testdatensatz) sind, gegenüber dem Fall, dass lediglich die Wahl unbekannter Zustände die Diagnosegruppe kennzeichnet. Im letzten Fall kann bereits bekannte Information über einen Patienten verwendet werden, im ersten Fall muss vollständig von anderen Patienten auf den neuen Patienten geschlossen werden. Falls der Unterschied zwischen beiden Diagnosegruppen groß ausfällt, besitzt jeder Patient individuell ausgeprägte Zustände auf der Intensivstation. Weitere Hinweise gibt eine Clusteranalyse. Indem mehr Patienten dokumentiert und in den Trainingsdatensatz aufgenommen werden, lässt sich dem Unterschied teilweise entgegenwirken.

- Statt des Zustands eines Patienten könnten auch seine Zustandsänderungen oder Trends Hinweise über seine zukünftige Entwicklung geben. Möglicherweise zeigen gerade Zustandsänderungen den Mangel eines Patienten auf, die Drift von bestimmten Parametern nicht ausreichend kompensieren zu können. Um dieser Hypothese nachzugehen, soll jeweils die Änderung des Zustands von einem auf den anderen Tag als Parameter verwendet werden. Sinnvoll erscheint die Betrachtung der Änderung im Zeitraum eines Tages und im Zeitraum von zwei Tagen.
- Da bei der ausschließlichen Betrachtung von Zustandsänderungen der Kontext des tatsächlichen Zustands nicht eingeht, erscheint zusätzlich auch die Kombination von Zustand und Zustandsänderung als untersuchenswert.
- Zur Einschätzung der Prognose des neuronalen Netzes für einen konkreten Patienten durch den behandelnden Arzt sollte zusätzlich zum Klassifikationsergebnis ein Gütemaß dafür angegeben werden. Es ist daher zu untersuchen, ob das neuronale Netz auch zur Angabe eines Gütemaßes geeignet ist.
- Kritische Zustände werden im oben beschriebenen Verfahren aus dem Vergleich der Anzahl zwischen Zuständen von Patienten, die überleben und die sterben, ermittelt. Dabei wird derzeit aufgrund fehlender Information jeder einzelne Zustand nicht weiter gewichtet. Auch ein Patient, der überlebt, kann Messungen haben, die als kritisch zu bezeichnen sind; ebenso könnte ein Patient, der stirbt, Messungen aufweisen, die weniger kritisch sind. Die Einteilung in "kritisch" und "weniger kritisch" wird erst durch den lokalen Mehrheitsentscheid getroffen. Es wäre daher wünschenswert zu prüfen, ob die Messungen nicht vorher in relevante (typische) sowie unrelevante Beispiele unterteilt werden können.
- Da wie erwähnt, nicht davon ausgegangen werden kann, dass jede Messung einen eindeutigen Hinweis zur Überlebenswahrscheinlichkeit gibt, erlaubt die Prognosegüte bezüglich jedes Zustandes keine eindeutige Bewertung. Sicherlich ist eine hohe Prognosegüte besser als eine niedrige. Aber eine niedrige Prognosegüte kann auch die Folge von stark überlappenden Klassenbereichen sein, so dass viele Patientenzustände uneindeutig sind. Für ein in der Praxis taugliches Verfahren sollte allerdings der Arzt bei möglichst vielen verstorbenen Patienten zumindest eine Warnung bekommen, und die Anzahl der Warnungen sollte bei verstorbenen Patienten höher ausfallen als bei überlebenden.

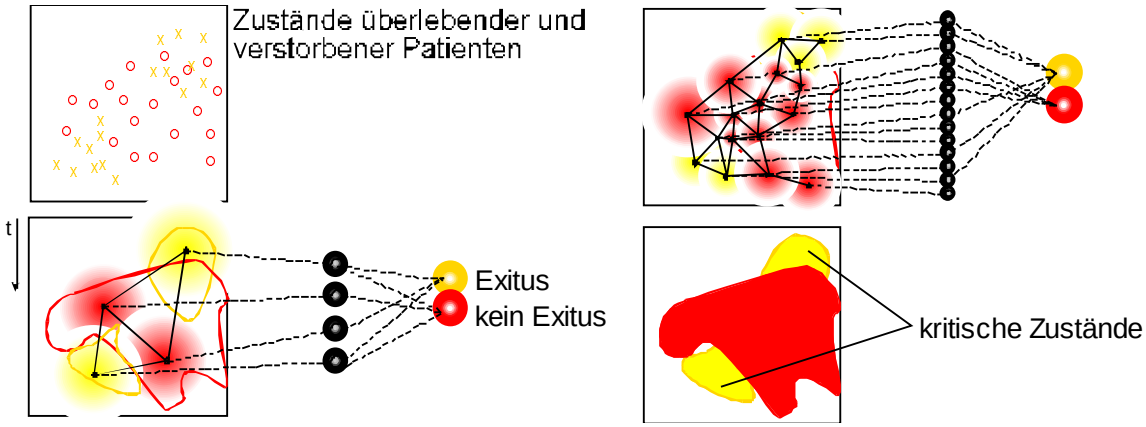


Abbildung 1.2. Vorgehensweise zur Detektion von kritischen Zuständen. *Links oben*: Viele Zustände von überlebenden und versterbenden Patienten sind einander ähnlich. Es sollten allerdings Bereiche vorhanden sein, in denen vorwiegend Zustände von sterbenden Patienten zu finden sind. Andernfalls besteht keine Chance ein Frühwarnsystem auf Basis der Zustände zu entwickeln. *Links unten*: Ein wachsendes RBF-Netz deckt den Zustandsraum zunächst mit nur wenigen radialen Basisfunktionen ab. Eine weitere Schicht bildet die Aktivierungen der radialen Basisfunktionen auf zwei Klassen ab. *Rechts oben*: Durch ein weiteres Einfügen von Knoten und der Modifikation von Verbindungsgewichten versucht das Netz die Zustände der überlebenden und versterbenden Patienten besser den Klassen zuzuordnen. *Rechts unten*: Indem Zustände aus gemischten Bereichen und aus Bereichen von überlebenden Patienten vorwiegend den Ausgabeknoten für kein Exitus aktivieren, führt die Präsentation von Mustern aus den Bereichen, die mehrheitlich Zustände verstorbener Patienten aufweisen, zu der Aktivierung Exitus. In diesem Fall sollte dem Arzt eine Warnung ausgegeben werden. Die Größe der Bereiche für kritische Zustände kann problemlos ausgedehnt oder verkleinert werden.

2. Daten

Die Daten der Untersuchung entstammen einer Datenbank, die Patienten der chirurgischen Intensivstation des Klinikums der J.W. Goethe-Universität Frankfurt am Main aus dem Zeitraum 8.11.1993 bis 25.11.1997 enthält. Diese Datenbank lag in zwei Teilen vor, da mit den Teilen bereits prospektive Studien durchgeführt wurden (vgl. [WADE ET AL., 1998; HANISCH ET AL., 1998]). Eine Auswahl der physiologischen Messwerte und der Medikamente wurden aus den Datenbankteilen in eine einheitliche Datenbank zusammengeführt. Die Textdaten, wie Diagnosen und Operationen waren zwar dokumentiert, waren aber einer automatischen Auswertung nicht zugänglich. In der Regel wurden die Messwerte und die durchgeführten Maßnahmen einmal am Tag in der Datenbank dokumentiert, in Ausnahmefällen öfters am Tag. Diese Ausnahmefälle wurden allerdings in dieser Untersuchung nicht betrachtet. Bei der Dokumentation vor Ort (während den medizinischen Studienphasen) ist aus praktischen Gründen nicht auf die Verwendung von Messungen mit gleicher Uhrzeit geachtet worden. Ein Patientenzustand setzt sich daher in dieser Untersuchung aus der jeweils ersten Messung des Tages zusammen. In dieser Untersuchung werden lediglich Patienten mit der Diagnose "septischer Schock" herangezogen. Die Einteilung in SIRS, Sepsis und septischer Schock wurde aufgrund der Patienten- und Behandlungsparameter in Anlehnung an die „Consensus Conference ACCP/CCM“ [BONE ET AL., 1992] getroffen und für jeden Tag in der Datenbank dokumentiert.

2.1. Frühere Untersuchungen

Basierend auf den oben genannten Daten wird bereits von HANISCH ET AL. [1998] eine Gruppe von 52 Patienten mit der Diagnose „septischer Schock“ zur individuellen Prognose durch ein neuronales Netz herangezogen. Eingangsparameter des neuronalen Netzes sind Patienten-

daten (Geschlecht, Alter, pO_2, pCO_2, pH, Leukocyten, TPZ, Thrombozyten, Laktat, Kreatinin, Puls, systolischer RR, Temperatur) und Behandlungsparameter (Atemfrequenz, FiO_2, maximaler Beatmungsdruck, AT III, Dopamin, Dubotrex, Natriumcarbonat, kontinuierliche venovenöse Hämofiltration).

Bei einem Test auf 23 Patienten kann bei Berücksichtigung obiger Parameter am ersten Tag der Diagnose des septischen Schocks für 93,4% eine richtige Vorhersage für das Überleben getroffen werden. Ein ähnlich gutes Ergebnis wird bei der Berücksichtigung von Änderungen der Daten vom ersten auf den zweiten Tag erzielt. Da sich diese Ergebnisse allerdings nur auf einen einzigen Testdatensatz beziehen, fehlt eine allgemeingültige Aussage, insbesondere weil die untersuchte Patientengruppe sehr klein ist. Eine erneute Untersuchung ergab, dass die problematischen Fälle nur im Trainings- und Validierungsdatensatz (und nicht im Testdatensatz) enthalten waren. Das erzielte Ergebnis ist daher nicht repräsentativ und als allgemeine Aussage hinsichtlich der Prognosefähigkeit von Patienten unzulässig.

2.2. Datenvorverarbeitung

Ein nicht zu unterschätzendes Problem, das einen erheblichen Zeitaufwand verursacht, ist eine adäquate Vorverarbeitung der Daten (vgl. [PAETZ ET AL., 2000]). Die oben genannten Daten wurden zur besseren Auswertung einer Fehlerbereinigung und einem Sampling unterzogen. Ausgehend von den in dieser Art vorverarbeiteten Daten, sind für die Analyse mit einem wachsenden RBF-Netz die in den folgenden beiden Abschnitten diskutierten Punkte "Vollständigkeit von Mustern" und "Normalisierung" zu beachten.

2.2.1. Eliminierung von Mustern mit zu vielen fehlenden Werten

Bei vielen neuronalen Netzen werden vollständige Muster gefordert. Obwohl das neuronale Netz insoweit modifiziert wurde, dass ein Lernen und Abfragen auch mit unvollständigen Mustern funktioniert (Kap. 3), erhöht sich die Fehleranfälligkeit mit steigender Anzahl von fehlenden Werten. Voruntersuchungen deuten darauf hin, dass dieser Fehler bei einer Anzahl von fehlenden Werten, die größer als die Hälfte aller Variablen sind, nicht mehr akzeptiert werden kann. Diese Muster, bzw. Zustände der Patienten wurden nicht für die Analyse verwendet.

2.2.2. Normalisierung

Die Untersuchungen erfolgen mit einem neuronalen Netz, welches topologische Repräsentationen im euklidschen Raum erzeugt. Dazu müssen die Daten die gleiche Streuung und ggf. eine mittelwertfreie Verteilung aufweisen. Von jedem Muster $\tilde{x}$ wird daher der Mittelwert $\bar{x}$ subtrahiert und das Ergebnis durch die Standardabweichung σ_x geteilt:

$$x := \frac{\tilde{x} - \bar{x}}{\sigma_x}$$

2.3. Verwendete Variablen

Da zur Zeit kein weiteres Vorwissen über die Relevanz der gemessenen Variablen existiert, werden verschiedene Kombinationen untersucht. Zunächst werden weitgehend gleiche Parameter wie in einer Voruntersuchung [HANISCH ET AL., 1998] verwendet (Studie A). Eine darauffolgende Untersuchung (Studie B) verwendet die Parameter des APACHE II Scores [KNAUS ET AL., 1985]. Die dritte Untersuchung (Studie C) wurde mit Parametern durchgeführt, bei denen unterschiedliche Korrelationen zwischen Variablenpaaren bei überlebenden Patienten im Vergleich zu verstorbenen Patienten auftraten.

2.3.1. Studie A1

Die Studie A besteht aus zwei Varianten A1 und A2 und verwendet folgende 16 Variablen:

Variablen	Beschreibung
pO_2Ateriell	Sauerstoffpartialdruck im Blut
pCO_2Ateriell	Kohlendioxidpartialdruck im Blut
pH	pH-Wert des Blutes
Leukozyten	Anzahl weißer Blutkörperchen
TPZ	Prothrombinzeit (Blutgerinnungsparameter)
Thrombozyten	Anzahl Blutplättchen (für Blutgerinnung)
Laktat	Laktat (Milchsäure)
Kreatinin	Kreatininkonzentration im Blut
Herzfrequenz	Herzfrequenz
Urinmenge	Menge des Urins
BlutdruckSystolisch	systolischer Blutdruck
Beatmungsfrequenz	spontan oder von Maschine vorgegeben
O_2KonzentrationInspiratorisch	Sauerstoffgehalt der Einatemluft (FiO_2) in %
AT3	Antithrombin III (bei Blutgerinnungsstörungen)
Dopamin	Katecholamin, u.a. zur Steigerung des Blutdrucks
Dubutrex	Katecholamin, u.a. zur Steigerung des Blutdrucks

Tabelle 2.1. Verwendete Variablen in der Studie A1 und A2

Aufgrund der unterschiedlichen Messhäufigkeit, die aus der Abwägung des Arztes nach der augenblicklichen Relevanz einer Variable unter Berücksichtigung des Aufwandes und der Kosten resultiert, weisen einige Variablen (Tabelle 2.1) bei einer Sampling-Rate von einem Tag dennoch fehlende Werte auf (Abbildung 2.1). Die Medikamentengaben von Dobutrex und Dopamin weisen keine fehlenden Werte auf, da fehlende Angaben als keine Gabe des Medikaments gedeutet wurden. Ebenso wurden fehlende Angaben des FiO_2 mit der Konzentration des Sauerstoffs in der Raumluft (21%) ergänzt.

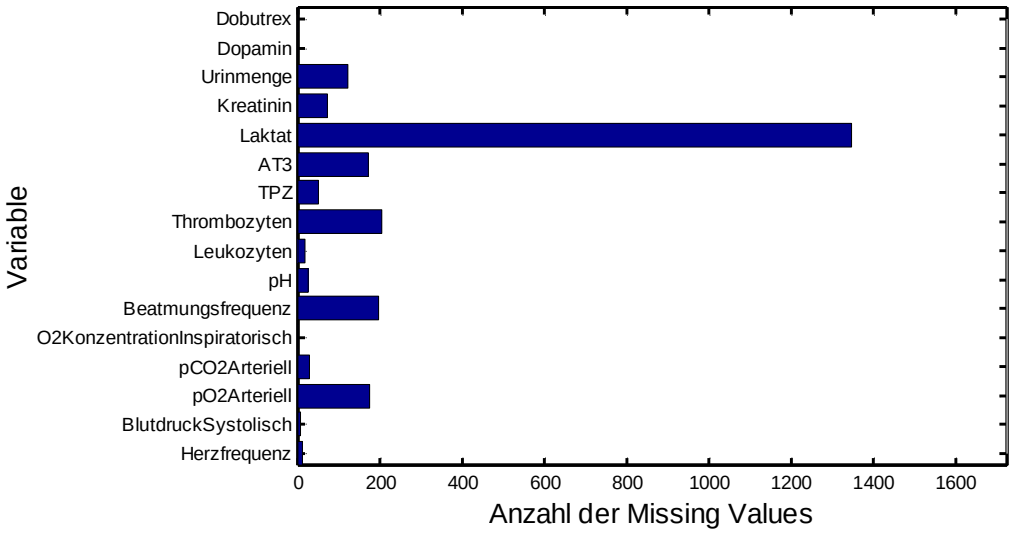


Abbildung 2.1. Fehlende Werte der in Studie A1 verwendeten Variablen bei 1720 möglichen Werten.

Als Folge des selten gemessenen Laktats finden sich bei der Betrachtung der Zustände nur wenige, die vollständig besetzt sind. Die meisten Zustände weisen einen oder mehr fehlende Werte auf (Abbildung 2.2). 345 Zustände wiesen mehr als acht fehlende Werte auf. Sie wurden vorher entfernt, so dass Abbildung 2.1 und Abbildung 2.2 Angaben über die tatsächlich analysierten Daten zeigen.

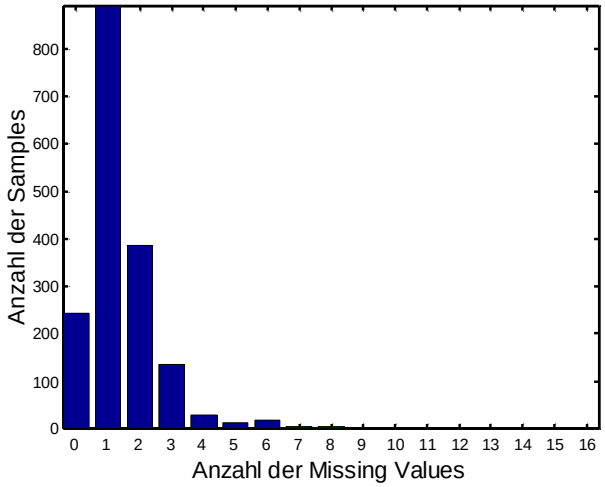


Abbildung 2.2. Histogramm über die Anzahl der fehlenden Werte in Studie A1.

2.3.2. Studie A2

Die Wahl der Zustände als Merkmale besitzt den Nachteil, dass die individuellen Normalzustände der Patienten schwer durch eine Trainingsmenge erfasst werden können. Beispielsweise besitzen Menschen von Natur aus eine andere Grundtemperatur. Dies bedeutet, wie eine Clusteranalyse in Abschnitt 6 aufzeigt, dass Häufungspunkte von Zuständen, sog. Cluster, vorwiegend jeweils aus den Messwerten nur eines Patienten bestehen. Eine Kompensation dieses Nachteils könnte durch eine Definition des Zustands als „Abweichung von den individuellen Mittelwerten“ erzielt werden. Die Verwendung des Mittelwerts über die Gesamtheit aller Patienten hilft hier nicht weiter, da die Abweichung der Patienten wieder unterschiedlich sein kann. Ein individueller Mittelwert, am besten der des gesunden Patienten, liegt allerdings nicht vor. So muss man sich leider damit behelfen, die Kurzzeitdynamik mit der Langzeitdynamik des Patienten zu vergleichen, d.h. die Abweichung vom gleitenden Mittelwert des Patienten als Merkmal heranzuziehen. Der gleitende Mittelwert $\bar{x}_t$ um t-ten Zeitpunkt eines Messwertes *x* wurde nach folgender Formel berechnet:

$$\bar{x}_t = k_T \cdot \bar{x}_{t-1} + (1 - k_T) \cdot x_t \quad \text{mit } k_T = 0{,}95$$

Frühere Messwerte klingen bei dem so berechneten gleitenden Mittelwert exponentiell in ihrem Einfluss zum Mittelwert ab. Da die Messwerte vermutlich einem nicht stationären Prozess entstammen, beeinflussen frühere Messwerte, wie gewünscht, den aktuellen gleitenden Mittelwert geringer als kürzlich aufgezeichnete Werte. Als Merkmal dient nun einfach die Differenz zwischen aktuellem Messwert und aktuellem gleitenden Mittelwert. Die ersten drei Merkmale wurden aus den Daten nach der Mittelwertbildung und der Differenzbildung gelöscht, da hier der Wert noch zu ungenau ist. Zusätzlich wiesen 323 Zustände mehr als acht fehlende Werte auf und wurden entfernt. Eine Übersicht über die Anzahl der vorhandenen Werte der Einzelvariablen gibt die Abbildung 2.3.

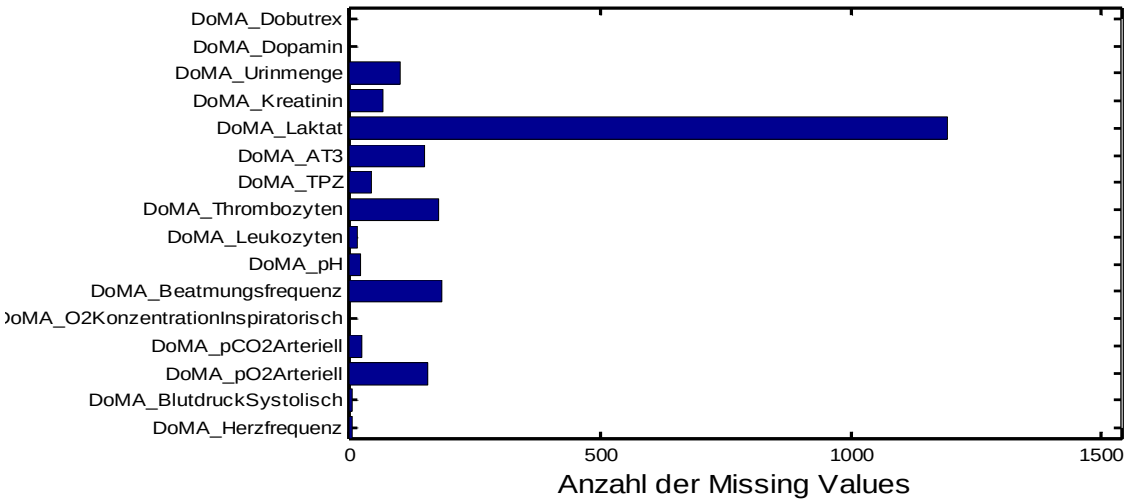


Abbildung 2.3. Fehlende Werte der in Studie A2 verwendeten Einzelvariablen bei 1720 möglichen Werten.

In Abbildung 2.4 ist die Lage für die Gesamtheit aller Messwerte dargestellt.

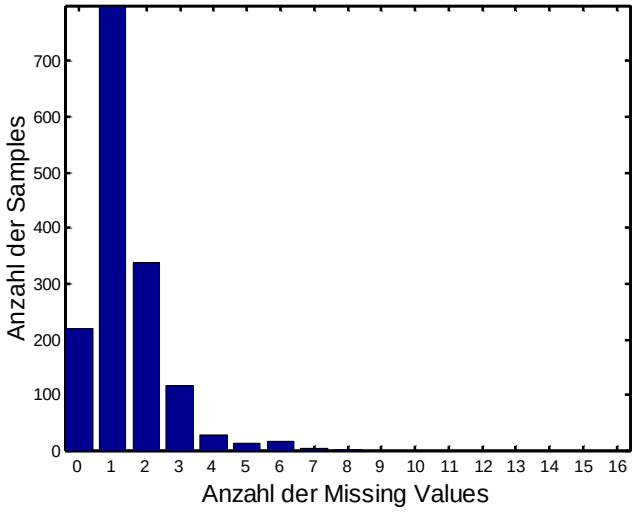


Abbildung 2.4. Histogramm über die Gesamtzahl der fehlenden Werte in Studie A2

2.3.3. Studie B

Die Variablen der Studie B wurden in Anlehnung an den in der Intensivmedizin wichtigen Apache II Score gewählt, da dieser oftmals in der Literatur als Indikator einer schweren Sepsis herangezogen wird.

Parameter	Beschreibung
PaO_2/FiO_2	Sauerstoffpartialdruck im Blut
Thrombozyten	Anzahl Blutplättchen (für Blutgerinnung)
Bilirubin	Bilirubinkonzentration im Blut
BlutdruckSystolisch	systolischer Blutdruck
Suprarenin	Katecholamin
Adrenalin	Suprarenin, Adrenalin
Noradrenalin	Arterenol, Noradrenalin
Kreatinin	Kreatininkonzentration im Blut

Tabelle 2.2: Verwendete Variablen in der Studie B.

Da alle Variablen recht häufig gemessen wurden, weisen die meisten Zustände keine fehlenden Werte auf (Abbildung 2.5).

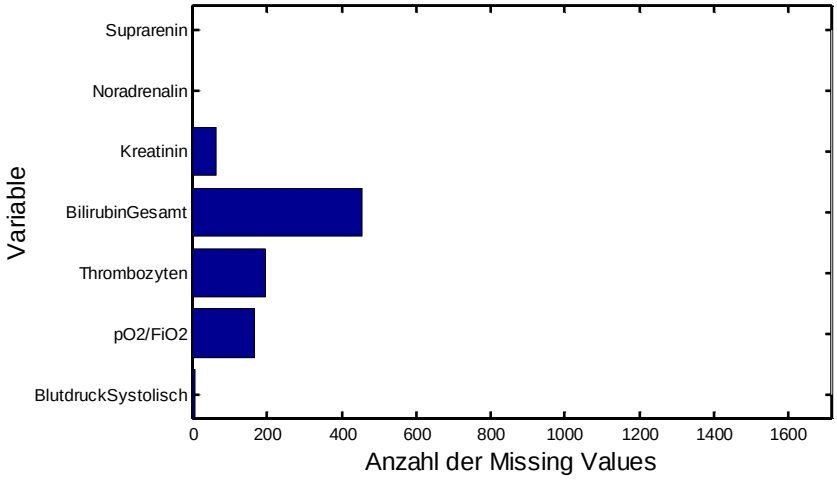


Abbildung 2.5. Fehlende Werte der in Studie B verwendeten Variablen bei 1717 möglichen Werten.

Die Medikamentengaben von Suprarenin und Noradrenalin weisen ebenfalls keine fehlenden Werte auf, da fehlende Angaben als keine Gabe des Medikaments gedeutet wurden. An dieser Stelle wird darauf hingewiesen, dass diese Annahme nur eingeschränkt sinnvoll ist, da insbes. beim Medikament Suprarenin häufiger von einer fehlenden Dokumentation der verabreichten Dosen ausgegangen werden muss. Ebenso wurden fehlende Angaben des FiO_2 mit der Konzentration des Sauerstoffs in der Raumluft (21%) ergänzt. 352 Zustände wiesen mehr als drei fehlende Werte auf. Sie wurden vorher entfernt, so dass Abbildung 2.5 und Abbildung 2.6 Angaben über die tatsächlich analysierten Daten zeigen.

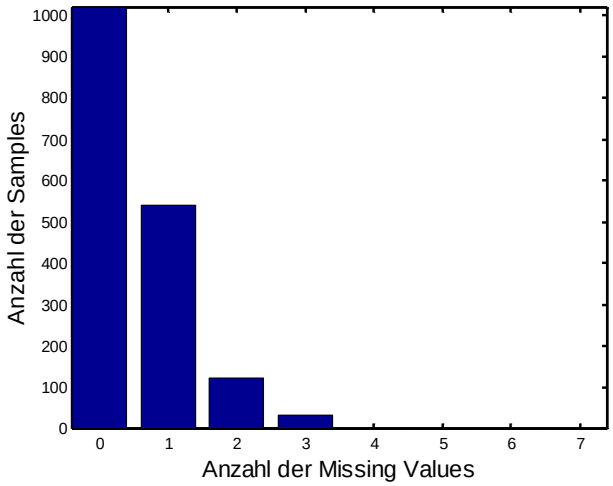


Abbildung 2.6. Histogramm über die Gesamtzahl der fehlenden Werte in Studie B.

2.3.4. Studie C

Während nennenswerte Korrelationen zwischen einer Variable und dem Ereignis *Exitus* nicht zu erwarten sind, verändern sich möglicherweise die Beziehungen der Variablen untereinander. Korrelationen zwischen zwei Variablen ergaben bei verstorbenen Patienten teilweise andere Werte als bei überlebenden Patienten (Tabelle 2.3), siehe [PAETZ ET AL., 2000] inkl. einer zusätzlichen Signifikanzbewertung.

Variable A	Variable B	Cor Überl.	Cor Verst.	Diff	Häufigkeit Var A (%)	Häufigkeit Var A (%)
O2KonzentrationInspiratorisch	pH	-0.03	-0.39	-0.36	57.3	89.0
Leukozyten	GGT	0.00	0.32	0.32	90.4	43.6
Eisen	GGT	0.31	0.01	-0.30	20.0	43.6
BilirubinGesamt	Harnstoff	0.26	-0.07	-0.33	58.1	72.9
Harnstoff	Kreatinin	0.14	0.57	0.43	72.9	83.9
pO2Arteriell	Kalium	-0.13	0.18	0.31	70.4	75.3
TZ	Chlorid	0.24	-0.07	-0.31	36.8	37.9
Fibrinogen	KreatininImUrin	0.05	-0.31	-0.35	80.8	25.8

Tabelle 2.3. Korrelationen bei denen die Differenz zwischen den Werten von überlebenden Patienten und verstorbenen Patienten mindestens 0,3 beträgt und die Häufigkeit der Messungen 20% überschreitet.

Die Korrelation zwischen zwei Variablen je Patient separat zu berechnen und als Maß heranzuziehen, scheitert zunächst einmal an der geringen Anzahl an Messwerten. Darüber hinaus ist es fraglich, ob es sich um ein stationäres Maß handelt. Viel eher ist zu vermuten, dass sich das Verhältnis von Variablen im Verlauf des Aufenthalts ändert und gerade diese Änderung eine wichtige Rolle bei der Detektion eines kritischen Zustands spielen könnte. Dieses könnte bei ausreichend häufigen Messungen durch die Verwendung eines gleitenden Fensters realisiert werden.
Statt nun die Korrelation selbst zu verwenden, wurde die Summe über alle Messwerte weggelassen und direkt die Abweichung des Messwertes der Variable A und B von ihren Mittelwerten als Maß verwendet, sozusagen relative Abhängigkeiten zwischen zwei Variablen berechnet. An dieser Stelle stellt sich erneut die Frage nach dem Mittelwert. Ist die Abweichung vom Mittelwert aller Patienten ein guter Indikator oder zu anfällig gegenüber individuellen Arbeitspunkten?

In Anlehnung an die Studie A2 wurden wieder die gleitenden Mittelwerte je Patient berechnet, allerdings wird diesmal der Mittelwert aller Patienten $\bar{x}^P$ als Startwert herangezogen und dann adaptiert.

$$\bar{x}_t = k_T \cdot \bar{x}_{t-1} + (1 - k_T) \cdot x_t \quad \text{mit } k_T = 0{,}95 \quad \text{und } \bar{x}_0 = \bar{x}^P$$

Anschließend werden die Differenzen der Messwerte zu ihrem Mittelwert berechnet und dann die erhaltenden Werte je Variable normalisiert. Als Merkmal wird schließlich die Multiplikation der normalisierten Differenzen zweier ausgewählter Variablen verwendet. Um Zustände mit zu vielen fehlenden Werten zu vermeiden, die sich ja durch die Multiplikation nochmals erhöhen, wurden exemplarisch drei Korrelationen (O2KonzentrationInspiratorisch-pH; Leukozyten-GGT; pO2Arteriell-Kalium) herangezogen. Die Erweiterung des Datenraums mit BilirubinGesamt-Harnstoff brachte in einem zusätzlichen Test ein etwas schlechteres Ergebnis. Es wird daher nicht näher dokumentiert. Weitere Tests mit anderen Korrelationen wurden nicht durchgeführt. Eine Übersicht über die Anzahl der vorhandenen Werte geben Abbildung 2.7 und Abbildung 2.8. Dabei wiesen 704 Zustände mehr als zwei fehlende Werte auf und wurden daher entfernt.

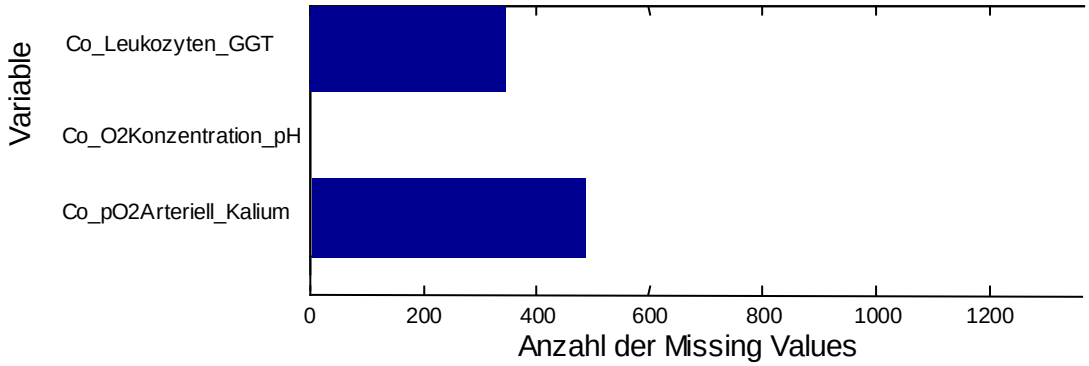


Abbildung 2.7. Fehlende Werte der in Studie C verwendeten Variablen bei 1365 möglichen Werten.

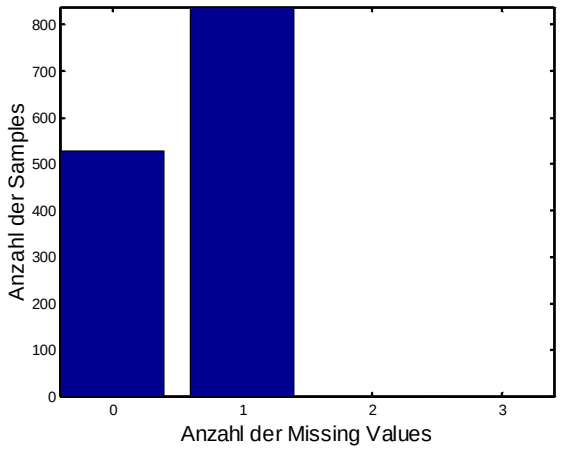


Abbildung 2.8. Histogramm über die Anzahl der fehlenden Werte in Studie C.

3. Das verwendete Neuronale Netz

Als neuronales Netz wird das Growing Neural Gas [FRITZKE, 1995] verwendet, das den Dynamic Cell Structures [BRUSKE und SOMMER, 1995] nahezu identisch ist.[7] Beide Netze basieren auf den Growing Cell Structures [FRITZKE, 1992; 1994] und verbinden Ideen des Neural Gas [MARTINETZ und SCHULTEN, 1991] und der Self Organizing Feature Maps [KOHONEN, 1982]. Im Folgenden wird daher oftmals die übergeordnete Gruppe der Netze als Zellstrukturen bezeichnet. Das Growing Neural Gas hat die Eigenschaft, Nachbarschaftsbeziehungen der Eingabedaten in der Netzstruktur, deren Dimension nicht vorher angegeben werden muss, sondern von den Daten gelernt wird, widerzuspiegeln. Der Algorithmus startet mit zwei Knoten und fügt an Positionen mit hohem Fehler neue Knoten ein. Jeder Knoten besitzt dafür einen Fehlerzähler τ, der den Ausgabefehler des gesamten Netzes akkumuliert. Eine Akkumulation erfolgt allerdings nur bei jeweils dem Knoten, der durch das angelegte Muster am stärksten aktiviert wurde.
Überwachtes Lernen erfolgt mit RBF-Funktionen, wie von Fritzke [1994] oder Bruske [1998] vorgeschlagen. Die Fähigkeit, die Anzahl der Knoten im Netz während des Trainings zu erhöhen sowie die regularisierende Wirkung der Nachbarschaften, machen den Algorithmus zu einem leistungsfähigen neuronalen Werkzeug wie einige Benchmarks zeigen [BRUSKE, 1998; HAMKER und HEINKE, 1997; HEINKE und HAMKER, 1998]. Gerade für die Erkennung kritischer Zustände ist die bei diesem Netztyp ausgeprägte Eigenschaft der Generalisierung sehr bedeutsam. Kritische Zustände werden erst erzeugt, wenn viele Zustände von Verstorbenen in einem Bereich vorkommen, während Zustände von Überlebenden dort wenig vertreten sind. Ausreißer führen nicht dazu, dass in diesem Bereich ein kritischer Zustand angelegt wird.

3.1. Kurzdarstellung des verwendeten Algorithmus

Im Folgenden wird kurz die Struktur und die Notation des verwendeten neuronalen Netzes vorgestellt. Gegenüber dem originalen Algorithmus wurden einige Verbesserungen vorgenommen.

3.1.1. Aufbau

Die Zellstrukturen bilden einen parametrisierten Graphen P=P(G,S), in dem jeden Knoten $v_i \in V$ eindeutig ein Gewichtsvektor $w_i \in S$ mit $S \subset R^n$ zugeordnet ist und die Nachbarschaftsverhältnisse über einen ungerichteten Graphen G definiert sind (vgl. [MARTINETZ und SCHULTEN, 1994; BRUSKE, 1998]).
Ein Graph G=G(V,E) besteht aus einer Menge V={v_1,...,v_n}, genannt „Menge der Knotenpunkte“ oder „Knoten des Graphen“, einer Menge E={e_1,...,e_n}, genannt „Menge der Kanten“, sowie einer sog. Inzidenzfunktion *f*, die jeder Kante e_r ein ungeordnetes Paar [v_i , v_j] von Knoten v_i, v_j, der Endpunkte oder Endknoten, zuordnet (vgl. [WALTHER und NÄGLER, 1987]). Als Nachbarn eines Knotens werden nur die Knoten bezeichnet, die mit dem Knoten eine gemeinsame Kante besitzen. Wenn G=G(V,E) der Graph des Netzes ist, dann ist die Menge der Nachbarn N_i des Knotens *i* über folgende Gleichung definiert:

$$N_i = \{ v_j \mid \exists e_k \quad f(e_k) = [v_i, v_j] \}$$

[7] Prinzipiell sind die folgenden Untersuchungen auch mit anderen, aus der Literatur bekannten Verfahren, denkbar.

Jede Kante besitzt zur ständigen Aktualisierung der Nachbarschaftsverhältnisse ein Alter. Die Gesamtanzahl der Knoten eines Graphen bestimmt sich zu $|S| = |V| = n_N$.

3.1.2. Adaption der Repräsentationsschicht

Die Zentren der radialen Basisfunktionen (Gewichtsvektor w_i) des *best-matching* Knotens *b* und seiner Nachbarn werden bei Präsentation des Eingabemusters *x* auf dieses zubewegt:

$$\Delta w_b = \eta_b \cdot (x - w_b)$$
$$\Delta w_c = \eta_n \cdot (x - w_c) \quad (\forall c \in N_b)$$

Die Bestimmung des *match* erfolgt durch Berechnung der geometrischen Abstände zwischen Eingabemuster *x* und Gewichtsvektoren w_i. Je kleiner der Abstand, desto größer ist die Ähnlichkeit.

3.1.3. Einfügen von Knoten in die Repräsentationsschicht

Nach einer bestimmten Anzahl von Adaptionsschritten λ wird der Knoten *q* mit dem maximalen Fehlerzähler τ gesucht. Zwischen diesem und einem weiteren Knoten wird ein neuer Knoten *r* eingefügt. Der weitere Knoten *s* ist der unter den direkten Nachbarn von *q* mit dem maximalen Fehler zu suchen. Darauffolgend müssen die neuen Eingangs- und Ausgangsgewichte nach dem arithmetischen Mittelwert sowie alle Zähler der an dem Einfügevorgang beteiligten Knoten durch Aufteilung der Fehlerwerte nach:

$$\tau_r := \beta \tau_q$$
$$\tau_q := (1 - \beta) \tau_q$$
$$\tau_s := (1 - \beta) \tau_s$$

berechnet werden.

3.1.4. Adaption der Ausgabeschicht

Zur Berechnung der Ausgabe muss zunächst die Aktivierung der Knoten in der Repräsentationsschicht mit einer Gaußfunktion bestimmt werden.

$$y_i = e^{-\frac{\|x - w_i\|^2}{\bar{\sigma}_i^2}} \quad \forall i \in G$$

Die Standardabweichung $\bar{\sigma}_i$ ist als die mittlere Kantenlänge aller vom Knoten *i* fortführenden Kanten $\bar{l}$ definiert. Im Gegensatz zu einer direkten Berechnung wie im ursprünglichen Algorithmus erfolgt die Berechnung hier als gleitender Mittelwert aus dem vorherigen Wert zum Zeitpunkt *k*-1 und der aktuellen mittleren Kantenlänge $\bar{l}$.

$$\bar{\sigma}_i^k = k_T \cdot \bar{\sigma}_i^{k-1} + (1 - k_T) \cdot \bar{l}_i \qquad \forall i \in G$$

Somit werden abrupte Änderungen in der Breite der Gaußfunktion vermieden und das nachfolgende Lernen weniger stark belastet.
Die Aktivierung der Ausgabeschicht o berechnet sich durch das Produkt der Ausgangsgewichte und der Aktivierung in der Repräsentationsschicht.

$$o_j = \sum_{i \in G} w_{ji}^{out} \cdot y_i \qquad \forall i \in G$$

Die Klasse erhält man durch die Bestimmung des Maximums unter Berücksichtigung des zusätzlichen Offsets L_S zur Steuerung der Sensitivität.

$$C_k = \max(o + L_S)$$

Der Fehler, der sich bei der Eingabe des Vektors *x* ergibt, ist der euklidsche Abstand der Ausgabe zum Zielvektor in der Ausgabeschicht ζ.

$$E_T(x) = \| \zeta - o(x) \|$$

Die Adaption der Gewichte erfolgt nach der Delta-Lernregel.

$$\Delta w_{j\,\mathrm{i}}^{\mathrm{out}} = \eta_o(\zeta_j - o_j)\, y_i \;; \quad \forall\ j \in \{1 \ldots m\},\ \forall\ i \in \mathrm{G}$$

3.1.5. Adaption der Zähler und Kanten der Knoten in der Repräsentationsschicht

Die Fehlerzähler aller Knoten des Graphen G werden um den Faktor α verringert.

$$\Delta\tau_i = -\alpha \cdot \tau_i \quad (\forall\, i \in \mathrm{G})$$

Lediglich der Fehlerzähler des Gewinners *b* erhöht sich um den aktuellen Fehler $E_T(x)$. Um nicht jeden noch so kleinen Fehler beim Einfügen zu berücksichtigen, kann eine Schwelle verwendet werden.

$$\Delta\tau_b = \begin{cases} E_T(x) & \text{falls } E_T(x) > \vartheta_C \\ 0 & \text{sonst} \end{cases}$$

Die Modifikation der Kanten erfolgt nach folgenden Regeln:

- Erhöhe das Alter aller vom Gewinner fortführenden Kanten um Eins.
- Setze das Alter der Kante zwischen *b* und *s* auf Null.
- Lösche alle Kanten, die das maximale Alter ϑ_{age} erreicht haben.
- Lösche alle Knoten, die keine Kante mehr besitzen.

Das Auftreten von fehlerhaften oder nicht gemessenen Werten (engl.: *missing value*, Abk.: MV) stellt für viele neuronale Netze ein Problem dar, da in der Regel eine Dateneingabe vollständig sein muss. Aus medizinischer Sicht ist dieses kaum zu gewährleisten, da (abgesehen von Standardparametern) die Messungen am Patienten nur nach Bedarf durchgeführt werden und einige Messungen auch hohe Kosten verursachen. RBF-Netze wie das GNG bieten einen Ausweg aus diesem Dilemma. Indem das Lernen nur innerhalb der im Datensatz vorhandenen Dimensionen erfolgt, können nicht-vollständige Messungen sowohl beim Training als auch beim Test genutzt werden. Allerdings fällt der Fehler bei zu vielen fehlenden Werten zu groß aus, so dass nur Messungen mit einer maximalen Anzahl an fehlenden Werten berücksichtigt werden können. Demnach berechnet sich der geometrische Abstand d_i mit

$$d_i = \frac{1}{\#I}\sqrt{\sum_{l \in I}\left(x_l - w_{il}\right)^2}\ , \quad I = \{l \mid x_l \neq \mathrm{MV}\}$$

Die Berechung der Aktivierung y_i erfolgt analog.

4. Vorgehensweise

4.1. Untersuchungen

Aus den oben genannten Fragestellungen lassen sich 5 Untersuchungen mit verschiedenen Datensätzen ableiten.

1. Zentral ist die Frage nach der Güte der Klassifikation, da sie die Grundlage für individuelle Warnungen bildet. Damit Warnungen möglichst verlässlich, aber auch ausreichend häufig sind, wird eine Sensitivität von ca. 80% angestrebt.
2. Ein weiterer interessierender Aspekt ist der Zusammenhang zwischen Sensitivität und Spezifität, der in der Regel durch eine *Receiver Operating Characteristic* (ROC-Kurve) illustriert wird.
3. Wichtig für die Akzeptanz der Warnungen wird sein, ob verstorbene Patienten im Mittel häufiger Warnungen erhielten als überlebende Patienten.
4. Weiterhin sind die Zeitpunkte der Warnungen von Interesse, besonders hinsichtlich einer späteren medizinischen Evaluation der Warnungen.
5. Besonders zu untersuchende Zeiträume sind jeweils die Tage
   - um der erstmaligen Diagnose des septischen Schocks
   - nach der Einlieferung
   - kurz vor der Entlassung bzw. dem Tod

4.2. Kriterien

Medizinische Untersuchungsergebnisse sind meist sehr differenziert. Für die Beurteilung hat sich ein Kriterienkatalog herausgebildet, der im Folgenden kurz diskutiert werden soll.

4.2.1. Gesamtprognose, Sensitivität und Spezifität

Das Ziel jeder Studie in diesem Bericht ist die Erkennung kritischer Zuständen, d.h. die Ausgabe einer Warnung falls das Ergebnis der Zustandsdiagnose in die Kategorie kritisch fällt. Dazu wird eine Klassifikation in *überlebt* (EXITUS=FALSE) oder *verstorben* (EXITUS=TRUE) durchgeführt. Dabei ist nicht nur allgemein die Güte der Vorhersage von Relevanz, sondern gleichzeitig die Angabe der Sensitivität (Wahrscheinlichkeit mit der ein Verstorbener tatsächlich als verstorben klassifiziert wird) und der Spezifizität (Wahrscheinlichkeit mit der ein Überlebender tatsächlich als überlebend klassifiziert wird).[8] Je besser die Gesamtprognose ausfällt, desto größer sind die Bereiche der kritischen Zustände. Um einen hohen Fehlalarm zu vermeiden, wird eine hohe Spezifizität gefordert.[9] Nur dann ist es dem Arzt zuzumuten, dass er die Warnung ernst nimmt. Ist die Gesamtprognose gering, liegen hohe Überlappungen der Klassen vor. In diesem Fall sind die Zustände von verstorbenen Patienten denen von überlebenen Patienten zu ähnlich und können daher ohne Erhöhung des Fehlalarms nicht als kritisch eingestuft werden. Prinzipell muss daher zwischen der Gefahr eines Fehla-

[8] Bei der Datenanalyse wird statt dieser Begriffe oft die Confusion-Matrix angegeben, die die Überlappung von Klassen aufzeigt (vgl. [GUÉRIN-DUGUÉ ET AL., 1995; HAMKER und HEINKE, 1997; HEINKE und HAMKER, 1998]).

[9] Eine Spezifität von 100% macht allerdings keinen Sinn, da zu erwarten ist, daß sich ein überlebender Patient während seines Aufenthalts auf der Intensivstation durchaus auch in einem kritischen Zustand befinden kann.

larms (abnehmende Spezifität), verbunden mit einer möglichst häufigen Warnung (zunehmende Sensitivität) sowie der möglichst sicheren Warnung (zunehmende Spezifität), verbunden mit der Gefahr von zu seltenen oder zu späten Warnungen (abnehmende Sensitivität), abgewogen werden. Dieser Zusammenhang wird oft durch eine Receiver Operating Characteristic (ROC)-Kurve, die die Spezifität über der Sensitivität aufträgt, illustriert.

4.2.2. Test und Validierung der Ergebnisse

Zur Gewinnung allgemeingültiger Aussagen reicht es nicht aus, alle Daten zum Training eines neuronalen Netzes zu verwenden, um im Anschluss den Trainingserfolg zu prüfen. Es müssen auch Daten zur Simulation einer realen Situation aus dem Trainingsdatensatz ausgelassen werden, um sicherzustellen, dass das gelernte Modell mit unbekannten, bisher nicht verwendeten Daten geprüft wird. Hier bieten sich zwei Möglichkeiten an, die als Extrema vieler Varianten gelten: die Bewertung mit einem gesonderten Testdatensatz sowie die Bewertung durch eine Kreuzvalidierung.

Besteht ein Datensatz S aus einer hinreichend großen Anzahl an Samples, so spiegelt das Ergebnis auf einem Testdatensatz die tatsächliche Generalisierung des gelernten neuronalen Netzes auf dem Datensatz S wider. An dieser Stelle wird aber noch einmal ausdrücklich darauf hingewiesen, dass der Datensatz der septischen Schock-Patienten eine nicht hinreichende Anzahl an Samples besitzt, wodurch die Aussagekraft beider Verfahren eingeschränkt ist.

Die einfachste Form der Kreuzvalidierung ist die *leave-one-out* Methode [MOSTELLER und TUKEY, 1968]. Anstatt mit allen Samples zu trainieren, wird eines herausgelassen und dem Klassifikator angeboten. Dieses wird für alle Samples wiederholt und der mittlere Fehler berechnet. Für neuronale Netze ist diese Form in der Regel zu aufwendig, so dass hier oft die *v-fold* Kreuzvalidierung verwendet wird [GEISSER, 1975; WAHBA und WOLD, 1975]. Nach dieser Methode wird der Datensatz in *D* Teile zerlegt und jeweils mit allen außer einem Teil trainiert. Als Ergebnis bekommt man den mittleren Fehler aller Anteile, die als Testmenge aus dem Training gelassen wurden.

Im Folgenden werden die jeweiligen $D-1$ Teile als Trainingsdatensatz und der D-te Teil als Testdatensatz bezeichnet. Der Testdatensatz darf keine Zustände enthalten, die bereits beim Training präsentiert werden. Darüber hinaus könnte es sein, dass Patienten an benachbarten Tagen ähnliche Zustände besitzen. Durch die zufällige Auswahl der Zustände des Testdatensatzes wird diese mögliche Überlappung nicht berücksichtigt.

Um der späteren realen Situation noch besser zu entsprechen, bietet es sich an, nicht nur unbekannte Zustände im Testdatensatz zu fordern, sondern unbekannte Patienten. Hier muss die Schnittmenge der Patienten von Trainings- und Testdatensatz eine leere Menge sein (Abbildung 4.1).

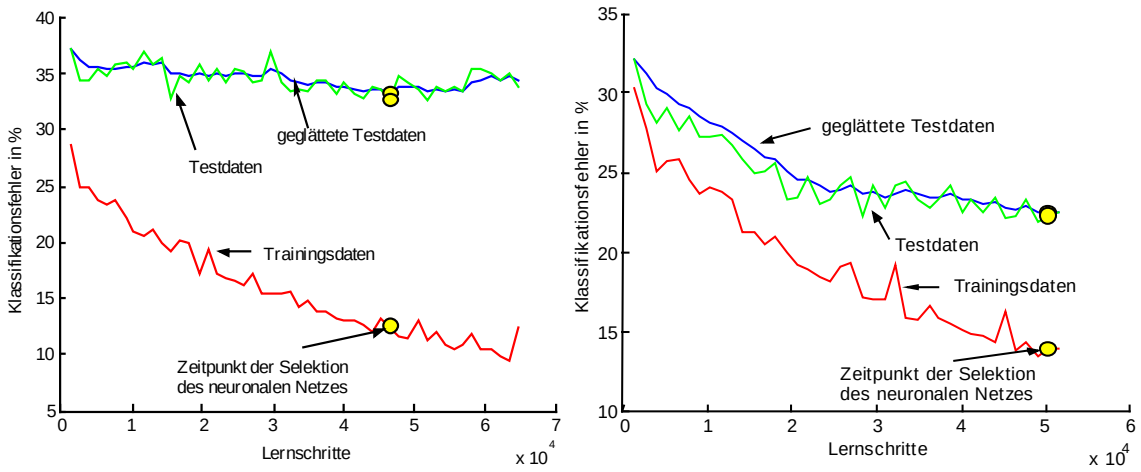


Abbildung 4.1. Vergleich zwischen einem Testdatensatz mit unbekannten Zuständen (*links*) und unbekannten Patienten (*rechts*). Kennt man bereits einige Zustände von Patienten, versehen mit dem Label *Exitus/not Exitus* wirkt sich dieses vorteilhaft auf das Ergebnis aus. Allerdings ist eine derartige Annahme unzulässig, da wir es in einem realen Einsatz mit neuen, unbekannten Patienten zu tun haben. Hier wird deutlich, dass eine beliebige Menge von Patienten nicht zwingend das Klassifikationsergebnis unbekannter Patienten verbessert. Die Patienten scheinen sich folglich stark in ihren Messwerten und somit in ihren individuellen Normalwerten zu unterscheiden. Eine Klassifikation auf Basis von Messwerten erfordert daher ein sehr großes Patientenkollektiv.

4.2.3. Selektion eines neuronalen Netzes

Da während des Trainings das neuronale Netz ständig Änderungen unterworfen ist, muss bestimmt werden, welches Modell das geeignete ist. Zur Auswahl eines geeigneten Modells mittels der Methode der Kreuzvalidierung bieten sich drei Möglichkeiten an.

- Bei der ersten wird ein weiterer Validierungsdatensatz verwendet. Hierzu wird pro Trainingsepoche ein Modell gewonnen und nach Beendigung des Trainings, das Modell mit dem besten Ergebnis auf dem Validierungsdatensatz gewählt. Typischerweise werden etwa 50% der Daten zum Training und jeweils 25% zur Validierung und zum Test verwendet. Diese Vorgehensweise garantiert, dass der Testdatensatz möglichst unabhängig ist, d.h. seine Muster dürfen weder zum Training noch zur Auswahl des Netzes herangezogen werden. Andernfalls sinkt das Vertrauen der Aussage bzw. die erzielten Ergebnisse besitzen eine geringere, allgemeingültige Aussagekraft. Nachteilig sind allerdings die Verwendung von 25% der Daten allein zur Auswahl eines Netzes. Insbesondere wenn wenig Daten zur Verfügung stehen, entspricht die gewählte Stichprobe von Validierungs- und Testdatensatz selten der tatsächlichen (aber unbekannten) Verteilung. Dieses macht sich besonders bemerkbar, wenn sich die einzelnen Patienten wie in dem hier vorliegenden Datensatz stark unterscheiden. Die Folge ist nicht nur ein unsicheres Abbruchkriterium, da sich Validierungs- und Testdatensatz zu sehr unterscheiden, sondern zusätzlich eine unsichere Aussage bezüglich der zu erwartenden Leistung, da der Testdatensatz nicht mehr repräsentativ ist.
- Als zweite Möglichkeit bietet sich die Verwendung des Testdatensatzes zur Validierung an. Strenggenommen bedeutet das ein Verzicht auf einen Testdatensatz, da die Unabhängigkeit bez. des Trainings und der Auswahl eines Netzes verletzt wird, was folglich eine Einschränkung der Verallgemeinerbarkeit der Aussagen zur Folge hat. Dieses Verfahren entspricht eher der Vorgehensweise, ein Netz für die praktische Anwendung in einem Prozess zu generieren, indem man so viele Daten wie möglich und notwendig zum Training verwendet, wobei der eigentliche Testdatensatz noch unbekannt ist. Der gravierendste Einwand gegen diese Verfahren ist dabei die Tatsache, dass man verallgemeinerbare Aussagen über die zu erwartende Leistung auf der Basis eines Testdatensatzes treffen will, den man allerdings vorher bereits verwendet hat, um ein optimales Netz auszuwählen. Man stelle sich z. B. vor, wie sich diese Vorgehensweise bei der *leave-one-out* Methode auswirken würde. Trotz dieser Verletzung ist diese Vorgehensweise bei einer ausreichenden Größe der Testmenge durchaus für die Abschätzung der Leistung geeignet, insbeson-

dere bei wenig Daten, da hier das Verfahren mit getrenntem Test- und Validierungsdatensatz unsicher wird. Der (unerlaubte) Vorteil, die Leistung auf dem Testdatensatz direkt als Stoppkriterium zu verwenden, kann verringert werden, indem nicht der Testdatensatz direkt als Stoppkriterium herangezogen wird, sondern sein gleitender Mittelwert. Zufällige, positive Ausreißer gehen damit nicht in die Bewertung ein und das zu erwartende Ergebnis wird realistischer. - Im Idealfall eines repräsentativen Datensatzes, der evtl. sogar zusätzliche Redundanz enthält, verursacht der Wegfall des Validierungsdatensatzes keine unerlaubte Verbesserung des Ergebnisses, da dann der Testdatensatz die gleichen probabilistischen Eigenschaften hat wie der Validierungsdatensatz. Ist ein Datensatz nicht repräsentativ, so sind die Ergebnisse aller Testmethoden nicht sehr vertrauenswürdig, insbes. für eine medizinische Anwendung, was aber auch in den Angaben über Standardabweichung, Maximum und Minimum bei Wiederholungen sichtbar wird. Unter diesen Gesichtspunkten ist die Angabe des *prototypischen* Klassifikationsergebnisses in [PAETZ ET AL., 2000] zu sehen, das ohne Validierungsdatensatz durchgeführt wurde.

- Ein Training bis zu einer maximalen Anzahl an Epochen bietet sich nur an, wenn sichergestellt ist, dass sich das resultierende Netz nicht zu sehr an die Trainingsdaten angepasst hat (*overfitting*) und somit noch ausreichend generalisiert. Durch die Beobachtung der Fehler beim Trainingsverlauf kann hier eine hinreichende Anzahl an Epochen abgeschätzt werden. Eine vorherige Abschätzung zufälliger Schwankungen, mit der ausgeschlossen wird, dass das Netz gerade am Ende des Trainings eine ungünstige Konstellation besitzt, ist nicht möglich.

Bei dem vorliegenden Datenmaterial von nur 70 Patienten mit der Diagnose "septischer Schock" und großen individuellen Unterschieden zwischen den Patienten ist es nicht sinnvoll, im Testdatensatz unbekannte Patienten zu verwenden und mit Hilfe eines Validierungsdatensatzes, wie in der ersten Variante erläutert, ein Netz auszuwählen. Um die große Vielfalt der Patienten möglichst weitgehend im Training abzudecken, erscheint es sinnvoll nach der *v-fold* Kreuzvalidierung vorzugehen, aber jeweils nicht eine beliebige Testmenge, sondern jeweils einen Patienten aus dem Trainingsatz herauszunehmen. Nach dieser *leave - one patient - out* Methode wird ein Netz wird mit allen außer einem Patienten trainiert - dieser Vorgang wird für alle Patienten wiederholt. Lässt man nur einen Patienten als Testmenge heraus, darf allerdings nicht mehr anhand des Ergebnisses auf dem Testdatensatz das Training gestoppt werden, so dass man gezwungen ist, das Training nach einer fest vorgegebenen maximalen Anzahl von Epochen abzubrechen und dieses Netz für den Test vorzusehen.

Eine derartige Vorgehensweise entspricht weitgehend dem Einsatzszenario in einem sequentiellen Test zur Evaluierung des Frühwarnsystems. Auch hier wird ein Netz auf Basis eines Patientenkollektivs trainiert. Der Test erfolgt dann auf unbekannten Patienten. Unter Vorbehalt aufgrund des geringen Patientenkollektivs, liefert die skizzierte Vorgehensweise Erwartungswerte für einen späteren realen Einsatz in einem medizinischen sequentiellen Test.

4.2.4. Gewinnung eines Erwartungsintervalls

Da die Trainings- und Testmenge in der Regel sehr begrenzt ist und den zugrunde liegenden Vorgang ggf. nur bedingt repräsentiert, muss zur Gewinnung allgemeingültiger Aussagen die Zufälligkeit bei der Auswahl von Trainings- Validierungs- und Testdatensatz sowie der Reihenfolge der Präsentation der Muster und der Anfangsinitialisierung relativiert werden. Davon abgesehen, verringert sich natürlich immer die Vertrauenswürdigkeit in das Ergebnis bei weniger oder nicht repräsentativen Daten. In der Regressionsanalyse werden daher in der Regel Konfidenzintervalle statistisch abgeleitet, die sich jedoch erschweren, wenn keine Aussagen über die Verteilung der Daten gemacht werden können (vgl. [HARTUNG, 1993]). Unglücklicherweise liegen für neuronale Netze kaum derartige Lösungen vor.[10] Alternativ lassen sich

[10] Überlegungen hinsichtlich der Zuweisung von Konfidenzintervallen zu neuronalen Netzen finden sich bei DYBOWSKI [1997], KINDERMANN ET AL. [1999] und TAGSCHERER ET AL. [1999].

Konfidenzen mit der Bootstrapping-Methode gewinnen, indem durch die Erzeugung von vielen verschiedenen Modellen eine repräsentative Verteilung von Modellen gewonnen wird. Zu diesem Zweck müssen die Parameter variiert werden, die das Ergebnis wesentlich mitbestimmen. Die *leave - one patient - out* Methode führt bereits zu der Erzeugung von 70 patientenbezogenen Ergebnissen. In dieser Untersuchung wird jeweils pro Patient nur ein Durchlauf mit zufälliger Initialisierung der Gewichte und zufälliger Reihenfolge der Präsentation von Mustern durchgeführt. Die patientenbezogenen Werte wie die jeweiligen Zeitpunkte der Warnungen haben daher strenggenommen nur einen illustrativen Charakter. Hier wären mehrere Durchläufe pro Patient notwendig. Vergleiche zwischen mehreren Durchläufen vermittelten allerdings den Eindruck, dass die patientenbezogenen Ergebnisse dennoch robust sind: Unabhängig von zufälligen Einflüssen stimmen die Zeitpunkte der Warnungen oftmals überein. Zahlen zum Gesamtkollektiv sowie die Vergleiche zwischen den überlebenden und verstorbenen Patienten erlauben unter der Einschränkung der insgesamt recht wenigen Patienten und der erläuterten Vorverarbeitung durchaus sinnvolle Aussagen.

In der Angabe der Ergebnisse wird auf die Berechnung eines speziellen Konfidenzmaßes verzichtet und statt dessen der Mittelwert, Median, Standardabweichung, oberes und unteres Quartil, sowie der minimale und maximale Wert der Klassifikationsperformanz angegeben.

Davon abgesehen kann die Güte der Klassifikation auch über die Eingabewerte variieren, so dass in manchen Bereichen eine hohe Sicherheit erzielt werden kann, während in Bereichen überlappender Klassen oder geringer Daten die Güte abfällt. Bei der Durchführung von individuellen Prognosen kann hier evtl. die Aktivierung der Ausgabeknoten des neuronalen Netzes eingesetzt werden, so dass für jede Prognose ein individuelles Gütemaß angegeben werden kann. Durch einen Vergleich der Übereinstimmung von der Richtigkeit der Klassifikationen mit der prognostizierten Güte wird überprüft, ob diese Form des individuellen Gütemaßes der Klassifikation angemessen ist.

4.3. Durchführung

In den folgenden Abschnitten werden die Bedingungen näher beschrieben, unter den die Ergebnisse von Abschnitt 5 entstanden.

4.3.1. Simulationsparameter und Training

Die Trainingsphase des neuronalen Netzes auf einem Datensatz wird hier nach einleitenden Versuchen auf den sinvollen Wert von 15 Epochen festgelegt. Das so gewonnene Modell wird auf einem Patienten, der sich nicht in der Trainingsmenge befindet, getestet. Zur Ermittlung allgemeingültiger Aussagen wird dieser Vorgang für jeden Patienten durchgeführt. Jeder Durchlauf erfolgt mit zufälliger Initialisierung der Gewichte, zufälliger Reihenfolge der Präsentation von Zuständen. Sinnvolle Simulationsparameter wurden anhand des Benchmarks von HAMKER und HEINKE [1997] gewählt (Tabelle 4.1).

Parameter	Wert	Beschreibung
η_b	0,1	Lernrate des Gewinners in der Repräsentationsschicht
η_n	0,01	Lernrate der Nachbarn in der Repräsentationsschicht
η_o	0,01	Lernrate im Ausgaberaum
λ	100	Schritte bis zur erneuten Anwendung des Einfügekriteriums
α	0,995	Multiplikative Verringerung der lokalen Fehlerzähler τ
β	0,5	Aufteilung der Fehlerzähler beim Einfügen
k_T	0,8	Faktor zur Berechnung des MA der Breite der Gaußglocke
υ_C	0,2	Schwellwert für den Fehlerzähler
υ_{age}	60	Maximales Alter der Kanten

Tabelle 4.1. Simulationsparameter des GNG. MA abk. für “Moving Average“.

4.3.2. Ein Gütemaß der Diagnose

Die Klassifikation erfolgt bei Darbietung eines Musters P anhand der Aktivierungen der Ausgabeknoten o. Das angelegte Muster gehört der Klasse C_i an: $P \in C_i$, wenn $\max_{k \in \{1...m\}}(o_k) = o_{\max} = o_i$. Die Berechnung des individuellen Gütemaßes g erfolgt auf Basis dieser Regel, indem es die Bekanntheit und die Trennschärfe eines Musters und seiner Klasse berücksichtigt (Abbildung 4.2).

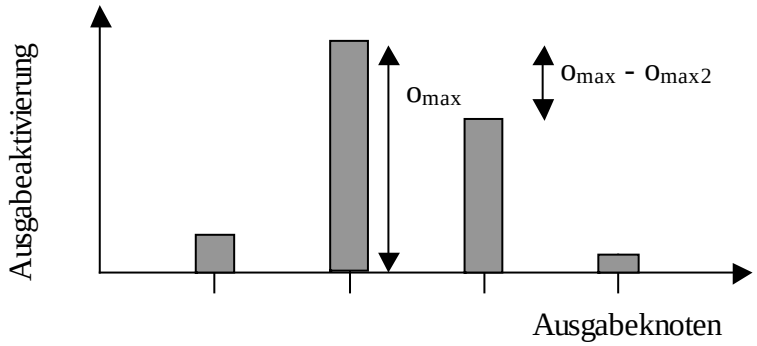


Abbildung 4.2. Illustration der Berechnung eines Gütemaßes der Klassifikation aus den beiden höchsten Aktivierungen o_{max} und $o_{max\,2}$ der Ausgabeknoten.

Bei nicht-normalisierten RBF-Netzen zeichnet sich eine gute Klassifikation im Wesentlichen durch zwei Kriterien aus. Da die Ausgabe o_i auf den Wert *Eins* trainiert wird, wenn das angelegte Muster der Klasse C_i angehört, und alle anderen Werte des Zielvektors *Null* betragen, geben höhere Ausgaben o_i eine höhere Sicherheit der Klassifikation an. In diesem Fall besitzt das angelegte Muster eine Ähnlichkeit zu gelernten Prototypen. Die Zielgröße für ein bekanntes Muster ist somit: $o_{\max} \to 1$. Es kann gelegentlich vorkommen, dass die Ausgabe eines Knotens etwas über der Zielgröße *Eins* liegt. Viel höhere Werte treten bei nicht zu großer Lernrate η_o in der Regel nicht auf. Dieser Fehler ist jedoch eher als ein Fehler beim Lernen der Ausgabe o_i = 1 einzuschätzen, der nicht prinzipiell die Sicherheit des Resultats eines Entscheidungskriteriums verringert. Der grundsätzliche Zusammenhang "hohe Aktivierung entspricht einer hohen Ähnlichkeit von Muster und Prototyp und geringe Aktivierung entspricht einer geringen Ähnlichkeit von Muster und Prototyp" bleibt erhalten.

Neben der Tatsache, dass ein Muster dem Netz bekannt sein sollte, um eine hohe Sicherheit der Klassifikation anzugeben, ist die Trennschärfe von Bedeutung. Das Muster soll möglichst wenig Ähnlichkeit zu den Prototypen anderer Klassen besitzen. Da es sich bei der Klassifikation um eine *1 aus n* Entscheidung handelt, reicht es aus, den Knoten mit der zweithöchsten Aktivierung $o_{\max 2}$ zu betrachten. Es gilt $o_{\max 2} \le o_{\max}$. Die Zielgröße für eine gute Trennschärfe ist somit $o_{\max} - o_{\max 2} \to 1$. Verwendete man nur dieses Kriterium, erhielte man keinen Unterschied zwischen den beiden Fällen *i*) $o_{\max} = 1,\ o_{\max 2} = 0{,}5$ und *ii*) $o_{\max} = 0{,}5,\ o_{\max 2} = 0$. Mit dem Ziel einer möglichst hohen Bekanntheit des Musters und einer möglichst guten Trennung der Klassen lässt sich ein Gütemaß der Klassifikation angeben:

$$g = 2 \cdot o_{\max} - o_{\max 2} - 1 \ .$$

Regionen, in denen keine Trainingsmuster liegen und daher nicht von Knoten repräsentiert werden, führen wie gewünscht nur zu geringen Aktivierungen und damit einem geringen Gütemaß. Bei Werten von g um *Null* und kleiner ist die Klassifikation als unsichere Aussage einzustufen. Werte um *Eins* zeigen eine sehr sichere Aussage an.

Unterteilt man die Menge aller durchgeführten Klassifikationen auf einem Datensatz in *False*, falls die Klassifikation nicht richtig war und *True*, falls die Klassifikation der Vorgabe entspricht, kann man durch den Vergleich der Gütemaße beider Mengen Rückschlüsse über den Lernerfolg auf den Daten ziehen. Bei Testdaten, die der Trainingsmenge gut entsprechen, sollte die Sicherheit bei falscher Klassifikation wesentlich geringer ausfallen als bei richtiger Klassifikation. Um das Gütemaß in Ergänzung zu der eigentlichen Klassifikation auf den Daten sinnvoll anwenden zu können, sollten die Sicherheiten von falschen Klassifikationen möglichst unter einem Wert g_{limit} und die von Richtigen möglichst über dem Wert liegen. In diesem Fall lassen sich viele falsche Klassifikationen, die eine geringe Sicherheit aufweisen, als neutral oder nicht genauer spezifizierbar korrigieren.

4.3.3. Sensitivität und Spezifität

Sensitivität und Spezifität hängen sowohl von der Häufigkeitsverteilung der Klassen in der Trainingsmenge als auch von Schwellwertoperationen der Ausgabeschicht ab. Grundsätzliche Vor- und Nachteile, die gewünschte Sensitivität und Spezifität durch die Zusammensetzung des Traingsdatensatzes oder durch eine Modifikation der Schwelle in der Ausgabe zu erzielen, sind nicht bekannt. Da die Daten der Frankfurter Vorstudie mehr Zustände von überlebenden als sterbenden Patienten aufweist, tritt im Falle von Überlappungen bereits, wie gewünscht eine höhere Spezifität auf.[11] Zum Feintuning bietet sich die Anpassung der Schwelle in der Ausgabe an. Hierzu wird die Klassifikationsschwelle in der Ausgabeschicht so verschoben, dass auf dem Trainingsdatensatz eine Spezifität zwischen 85-90% herrscht. Damit werden eher weniger, aber verlässliche Warnungen erzeugt.

[11] Strenggenommen gilt dieses jedoch nur, wenn in den Bereichen der Überlappungen tatsächlich die überlebenden Patienten überwiegen. Die Betrachtung der Gesamtzahl läßt diesen Sachverhalt lediglich vermuten.

5. Ergebnisse des neuronalen Netzes

In den folgenden Abschnitten werden die Ergebnisse des Trainings des neuronalen Netzes unter den Vorgaben der Studien A1, A2 und B beschrieben.

5.1. Studie A1

Die Ergebnisse von Studie A1 werden nun unter den Kriterien aus Abschnitt 4 dargelegt und diskutiert werden. Die Ergebnisse aller Patienten der Studie sind unter der Webadresse *http://www.medan.de/Klassifikation/Vorstudie/Klassifikation_sepsA/Klassifikation.html* einsehbar.

5.1.1. Gesamtprognose

Betrachtet man alle 70 Patienten zusammen, so ergeben sich aus den in Studie A1 ausgewählten Zuständen die in Tabelle 5.1 dargestellten Werte. Das neuronale Netz kann auf Basis jedes einzelnen Zustandes eines neuen Patienten im Mittel zu 70% eine richtige Prognose (*exitus*/*not exitus*) abliefern. Wie bereits vorher erwähnt, wollen wir eine derartige Prognose so nicht durchführen, da es unrealistisch ist, eine verlässliche Prognose für jeden einzelnen Zustand zu geben. Das Zahlenmaterial ist nur insofern von Interesse, als dass mehr richtige Vorhersagen und verlässliche Warnungen an den Arzt ermöglicht werden.

Richtige Vorhersagen: 70,46%	Stdabw: 26,55	Min: 5,00%	Max: 100.00%
Sensitivität: 54,64%	Stdabw: 30,52	Min: 5,00%	Max: 100.00%
Spezifität: 80,39%	Stdabw: 17,87	Min: 14,29%	Max: 100.00%

Tabelle 5.1. Übersicht der Klassifikation aller Patienten. Die (nicht nach der Länge der Patientenzeitreihen gewichteten) Mittelwerte und die Standardabweichung beziehen sich auf die Patienten und auf alle Zustände. Im Mittel kann man für jeden Patienten mit 70% richtigen Vorhersagen rechnen. Die hohen Schwankungen treten dadurch auf, dass das System z.B. für den Patienten mit der ID 152, der 20 Tage anwesend war, nur am ersten Tag eine Warnung ausgab. Infolgedessen ergibt sich sofort ein Minimum von 5% und eine entsprechend hohe Standardabweichung.

Weiterhin erhält man im Mittel eine Spezifität von 80% und eine Sensitivität von 55%. Die alleinige Betrachtung der Mittelwerte verdeckt allerdings die individuellen Unterschiede der Klassifikation von Patienten. Während bei einigen Patienten alle Zustände in die richtige Klasse eingeteilt werden können, misslingt dieses bei anderen Patienten fast vollständig wie durch Boxplots zu erkennen ist (Abbildung 5.1). Die Gründe dafür werden später erläutert.

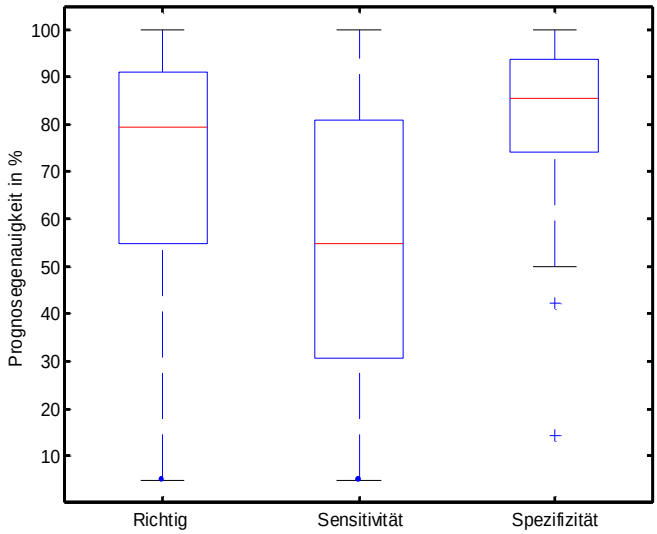


Abbildung 5.1. Boxplot der richtigen Vorhersagen, Sensitivität und Spezifität in Studie A1

5.1.2. Gegenüberstellung von Sensitivität und Spezifität

Receiver Operating Characteristic (ROC)-Kurven bieten die Möglichkeit, den Zusammenhang zwischen Sensitivität und Spezifität zu visualisieren (Abbildung 5.2). Um den Arzt mit möglichst verlässlichen Warnungen zu versorgen, ist ein Arbeitsbereich mit einer hohen Spezifität sinnvoll. Werte nahe bei 100% machen allerdings keinen Sinn, da auch die Zustände von überlebenden Patienten in einem kritischen Bereich liegen können.

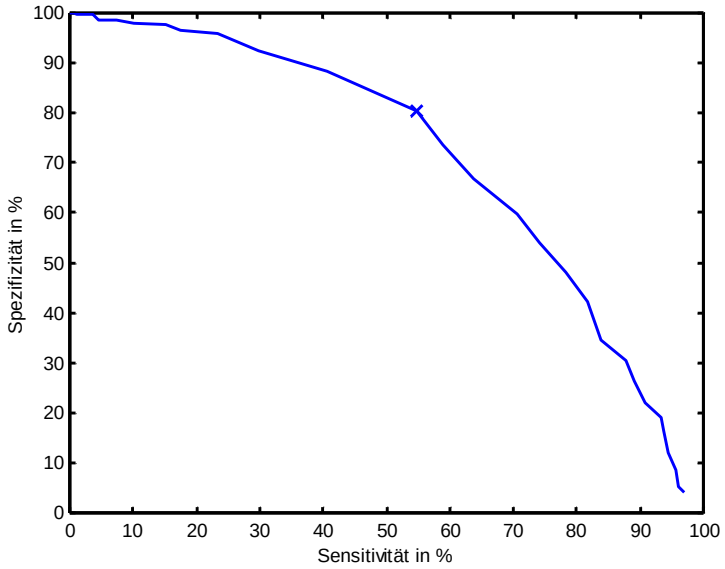


Abbildung 5.2. Receiver Operating Characteristic (ROC)-Kurve, berechnet aus den Mittelwerten aller Patienten in Studie A1. Das x kennzeichnet den mittleren Arbeitspunkt aller Netze auf den getesteten Patienten.

5.1.3. Individuelle Warnungen bei verstorbenen Patienten

Betrachtet man lediglich die verstorbenen Patienten, so sollte der Arzt während ihrer Liegedauer möglichst viele und frühe Warnungen erhalten. Tatsächlich hätte der Arzt bei vielen

verstorbenen Patienten der Frankfurter Vorstudie mehrere (auch frühzeitige) Warnungen erhalten, die eventuell zu einer erhöhten Aufmerksamkeit des Arztes geführt hätten (Abbildung 5.3). So wurde beispielsweise beim Patienten mit der ID 998 an allen vier Tagen eine Warnung ausgesprochen. Es kam sogar bei allen Patienten zu einer Warnung; bei manchen, wie dem Patienten mit der ID 661, wurde diese aber sehr spät ausgesprochen.

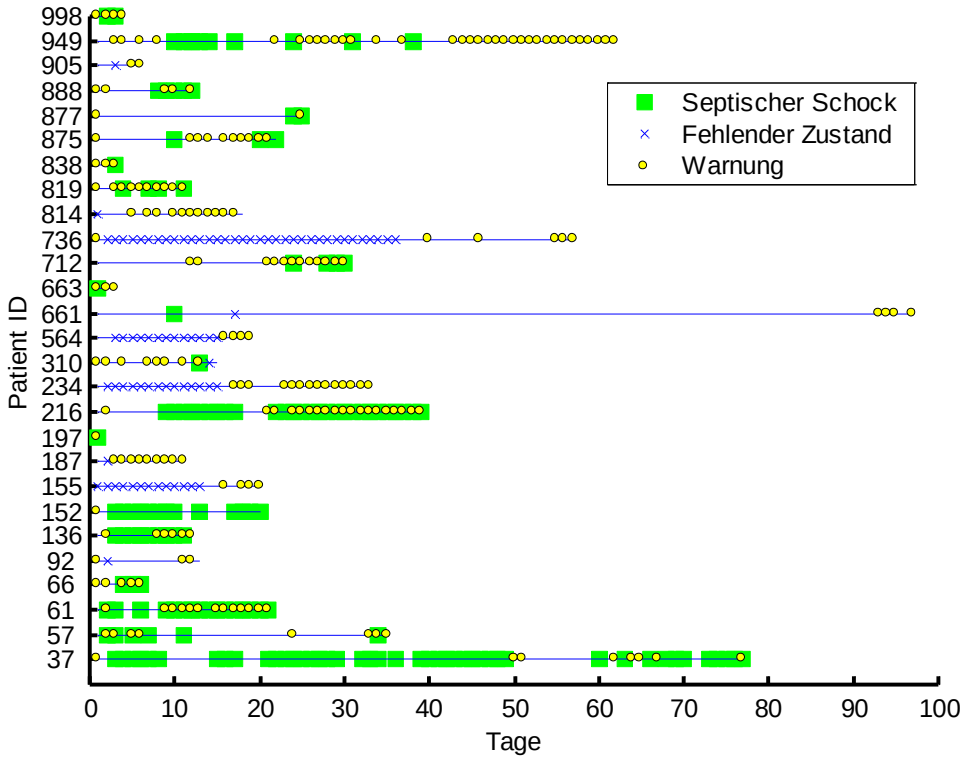


Abbildung 5.3. Septischer Schock und die Ausgabe von Warnungen bei verstorbenen Patienten vom ersten bis zum letzten Tag auf der Intensivstation in Studie A1. Das x kennzeichnet Zustände, die aufgrund von zu vielen fehlenden Werten nicht in die Analyse eingehen konnten.

5.1.4. Individuelle Warnungen bei überlebenden Patienten

Auch Zustände von überlebenden Patienten können zu Warnungen führen, schließlich liegen diese auf der Intensivstation und sind von vornherein nicht von den verstorbenen zu unterscheiden. Aus gutem Grund gibt daher das neuronale Netz ebenfalls zahlreiche Warnungen bei kritischen Zuständen aus und könnte so die Aufmerksamkeit des Arztes für den Patienten erhöhen (Abbildung 5.4). Durch den Vergleich von Abbildung 5.3 mit Abbildung 5.4 kann man jedoch schon erkennen, dass die Anzahl der Warnungen bei den Überlebenden geringer ist. Offenbar gelingt es dem neuronalen Netz anhand der jeweils 69 Patienten, Zustände zu finden, die öfter von verstorbenen Patienten aufgesucht werden als von überlebenden. Dennoch muss man auch hier ergänzend fragen, warum überhaupt die Patienten mit der ID 28 und 999 überlebt haben, da während ihrer Liegedauer sehr viele Warnungen ausgesprochen wurden.

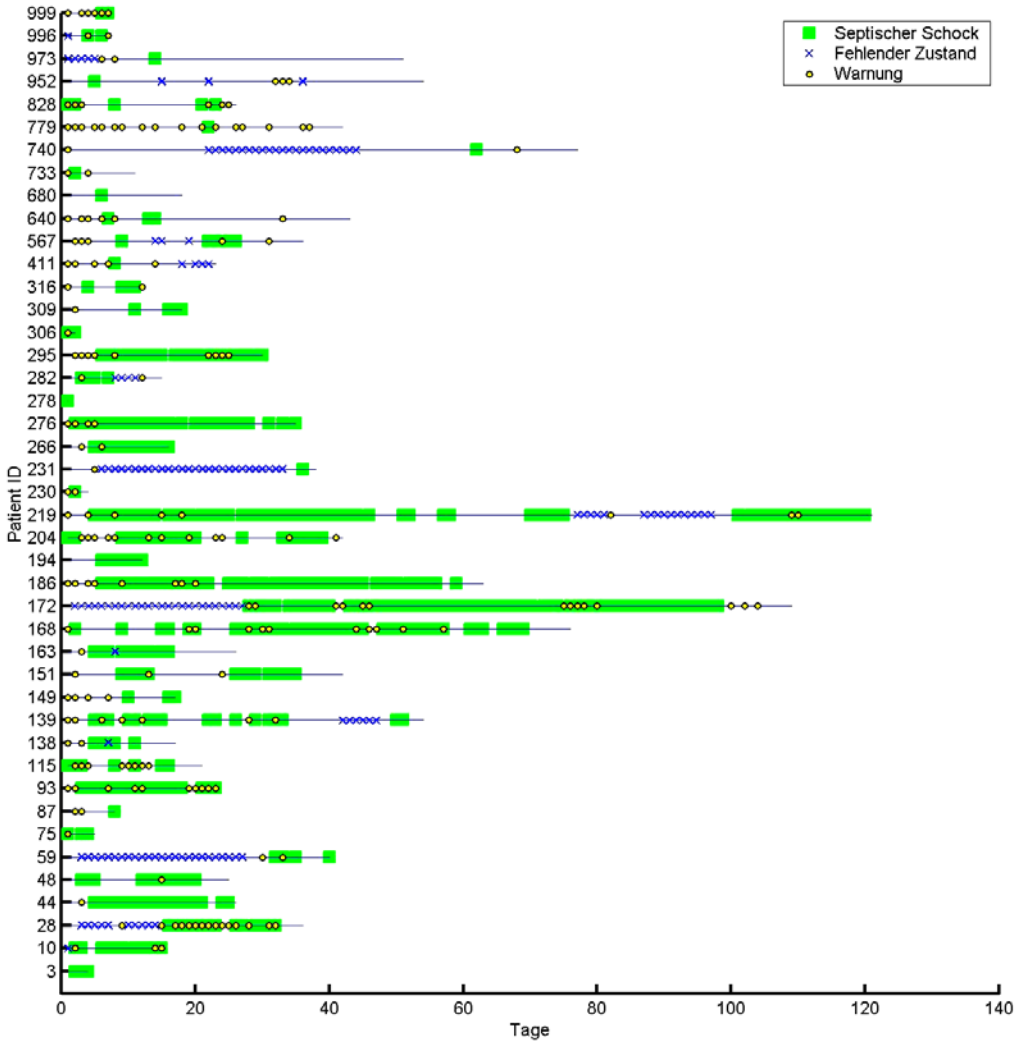


Abbildung 5.4. Septischer Schock und die Ausgabe von Warnungen bei überlebenden Patienten vom ersten bis zum letzten Tag auf der Intensivstation in Studie A1. Das x kennzeichnet Zustände, die aufgrund von zu vielen fehlenden Werten nicht in die Analyse eingehen konnten.

5.1.5. Mittlere Anzahl an Warnungen

Trotz aller individuellen Unterschiede sollte man doch annehmen, dass Patienten, die versterben, häufiger Zustände besitzen, bei denen eine Warnung ausgesprochen wird, als Patienten, die überleben. Wie von einem sinnvollen Analysverfahren anzunehmen, ergibt sich dieses auch für die trainierten neuronalen Netze. Normiert man die Liegedauer aller Patienten und bildet das Mittel der Warnungen innerhalb der entsprechenden Bereiche, so zeigt sich deutlich bei verstorbenen Patienten eine gegenüber überlebenden Patienten erhöhte mittlere Anzahl von Warnungen (Abbildung 5.5).

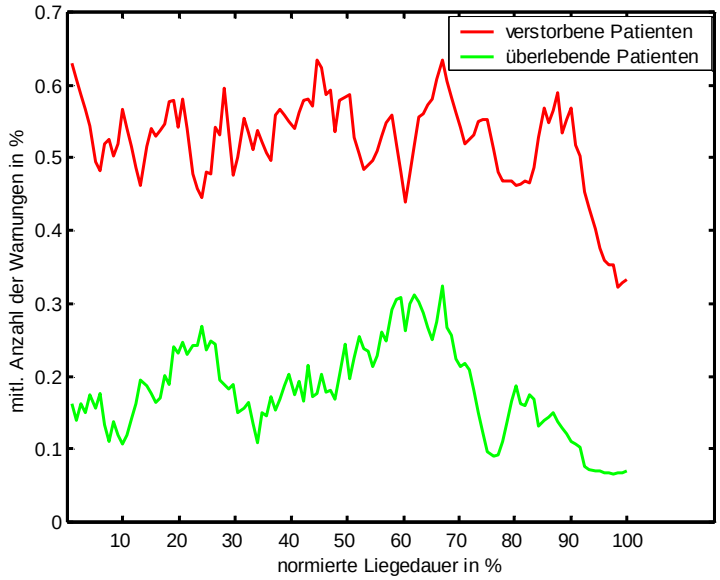


Abbildung 5.5. Mittlere Anzahl der Warnungen bei überlebenden und verstorbenen Patienten in Studie A1.

5.1.6. Tagesanalyse

Es wurden weiterhin die Klassifikationsraten der ersten 12 Tage ab der Einlieferung, den fünften Tag vor bis zum sechsten Tag nach dem erstmaligen Auftreten des septischen Schocks und die letzten 13 Tage bis zum Exitus bzw. der Entlassung dokumentiert, allerdings fanden sich keine Auffälligkeiten oder besondere Trends.

5.1.7. Gütemaß

In Studie A1 ist die Sicherheit der auf dem nicht trainierten Patienten (Testdatensatz) durchgeführten Klassifikationen eher gering (Abbildung 5.6). Darüber hinaus kann keine Grenzlinie gefunden werden, die die Sicherheiten der richtigen und der falschen Klassifikation trennt. Die Verteilungen überlappen sich zu stark. Ein ergänzender Einsatz des Gütemaßes zur Korrektur der Klassifikation macht daher auf diesen Daten wenig Sinn. Durch die Zuweisung von $g_{\text{limit}} = 0$ ließen sich zwar, aufgrund zu geringer Sicherheit, zahlreiche falsche Entscheidungen als neutrale Entscheidungen relativieren, aber zu viele richtige Entscheidungen wären dann ebenfalls als neutral eingestuft.

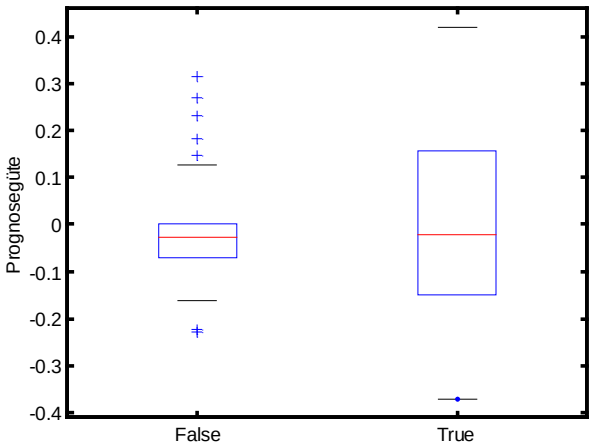


Abbildung 5.6. Prognosegüte der richtigen und falschen Klassifikationen in Studie A1.

5.2. Studie A2

Die Ergebnisse von Studie A2 werden im folgenden Abschnitt unter den Kriterien aus Abschnitt 4 dargelegt und diskutiert. Die Ergebnisse aller Patienten der Studie sind einsehbar unter *http://www.medan.de/Klassifikation/Vorstudie/Klassifikation_sepsA_ma/Klassifikation.html*.

5.2.1. Gesamtprognose

Auf Basis der Differenzen vom gleitenden Mittelwert der sonst mit Studie A1 identischen Parameterauswahl erreicht das neuronale Netz nicht ganz das Ergebnis der Studie A1 (Tabelle 5.2). Trotzdem ist eine Generalisierung der Trainingsmenge hinsichtlich des überprüften Patienten erkennbar.

Richtige Vorhersagen: 66,62%	Stdabw: 28,58	Min: 0,00%	Max: 100.00%
Sensitivität: 44,05%	Stdabw: 25,61	Min: 0,00%	Max: 100.00%
Spezifität: 79,83%	Stdabw: 21,09	Min: 25,00%	Max: 100.00%

Tabelle 5.2. Übersicht der Klassifikation aller Patienten in Studie A2. Die Mittelwerte und die Standardabweichung beziehen sich auf die Patienten und auf alle Zustände. Im Mittel kann man für jeden Patienten bei Verwendung der Differenzen vom gleitenden Mittelwert mit knapp 67% richtigen Vorhersagen rechnen. Die Schwankungen konnten nicht signifikant reduziert werden.

Allerdings wird das erhoffte Ziel, eine größere Unabhängigkeit von individuellen Patienten zu erzielen, nicht erreicht – die Standardabeichung der richtigen Vorhersagen wird nicht kleiner. Lediglich die Quartile des Boxplots der Sensitivität sind geringer (Abbildung 5.7).

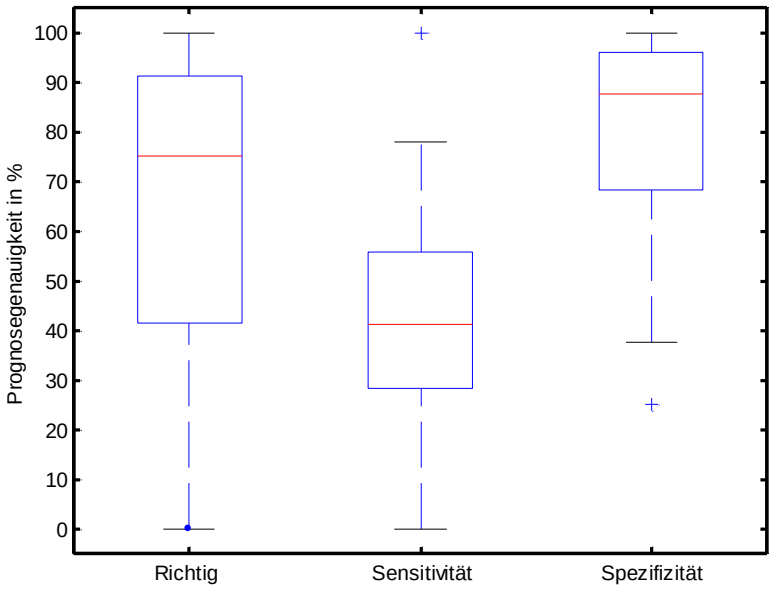


Abbildung 5.7. Boxplot der richtigen Vorhersagen, Sensitivität und Spezifität in Studie A2

5.2.2. Gegenüberstellung von Sensitivität und Spezifität

Die geringere Klassifikationsgenauigkeit zeigt sich auch an der schwächer ausgeprägten ROC-Kurve (Abbildung 5.8).

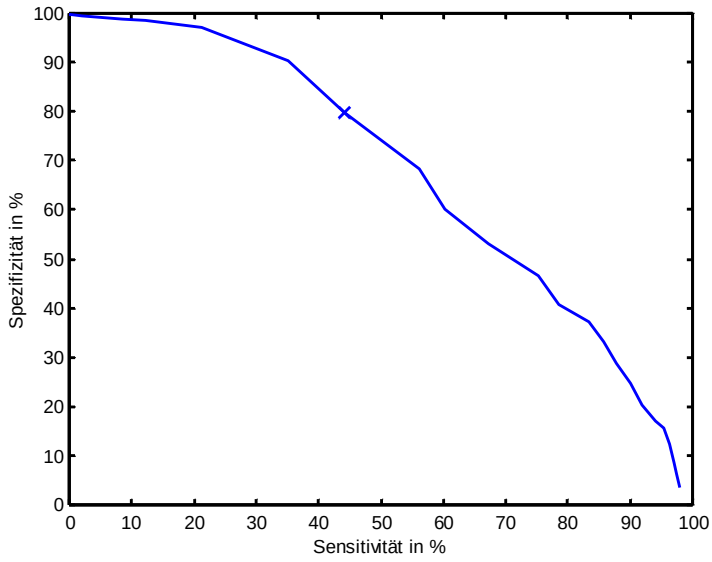


Abbildung 5.8. Receiver Operating Characteristic (ROC)-Kurve in Studie A2, berechnet aus den Mittelwerten aller Patienten. Das x kennzeichnet den mittleren Arbeitspunkt aller Netze auf den getesteten Patienten.

5.2.3. Individuelle Warnungen bei verstorbenen Patienten

Die Betrachtung der individuellen Warnungen bei Studien A2 bringt an sich keine neue Erkenntnis gegenüber A1. Doch vergleicht man den Zeitpunkt der Warnung in Studie A1 mit dem in A2, so stellt man fest, dass bei einigen Patienten oft an gleichen Zeitpunkten die Warnungen gegeben wurden. Allerdings treten in Studie A2 deutlich weniger Warnungen auf.

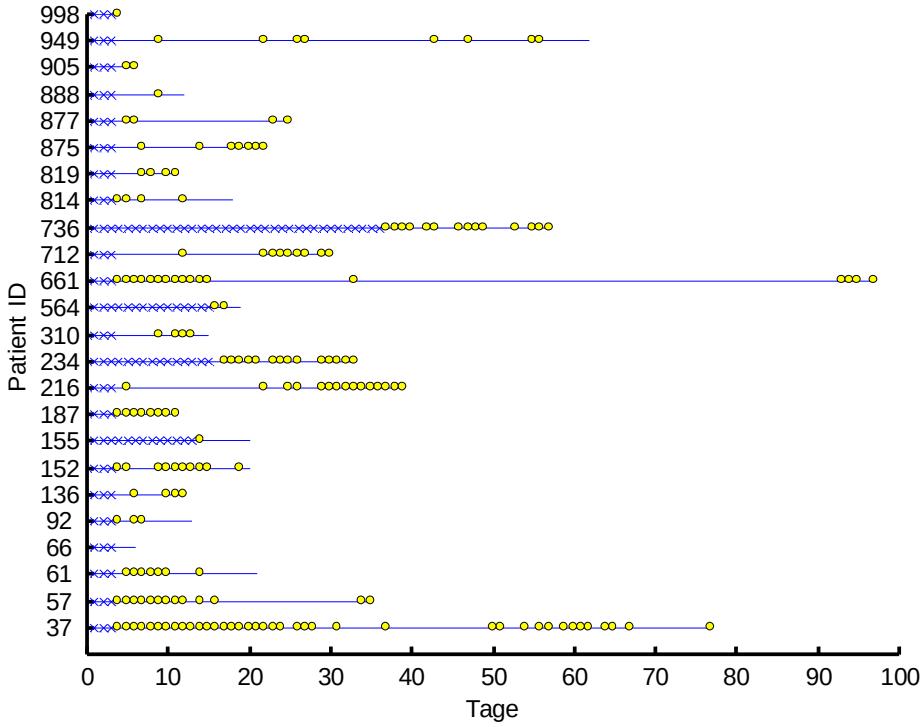


Abbildung 5.9. Septischer Schock und die Ausgabe von Warnungen bei verstorbenen Patienten vom ersten bis zum letzten Tag auf der Intensivstation in Studie A2. Das x kennzeichnet Zustände, die nicht in die Analyse eingehen konnten. Der septische Schock wurde bei diese Studie nicht dokumentiert.

5.2.4. Individuelle Warnungen bei überlebenden Patienten

Auch bei den überlebenden Patienten (Abbildung 5.10) zeigen sich einige Übereinstimmungen zu Studie A1, doch scheinen diese insgesamt geringer auszufallen.

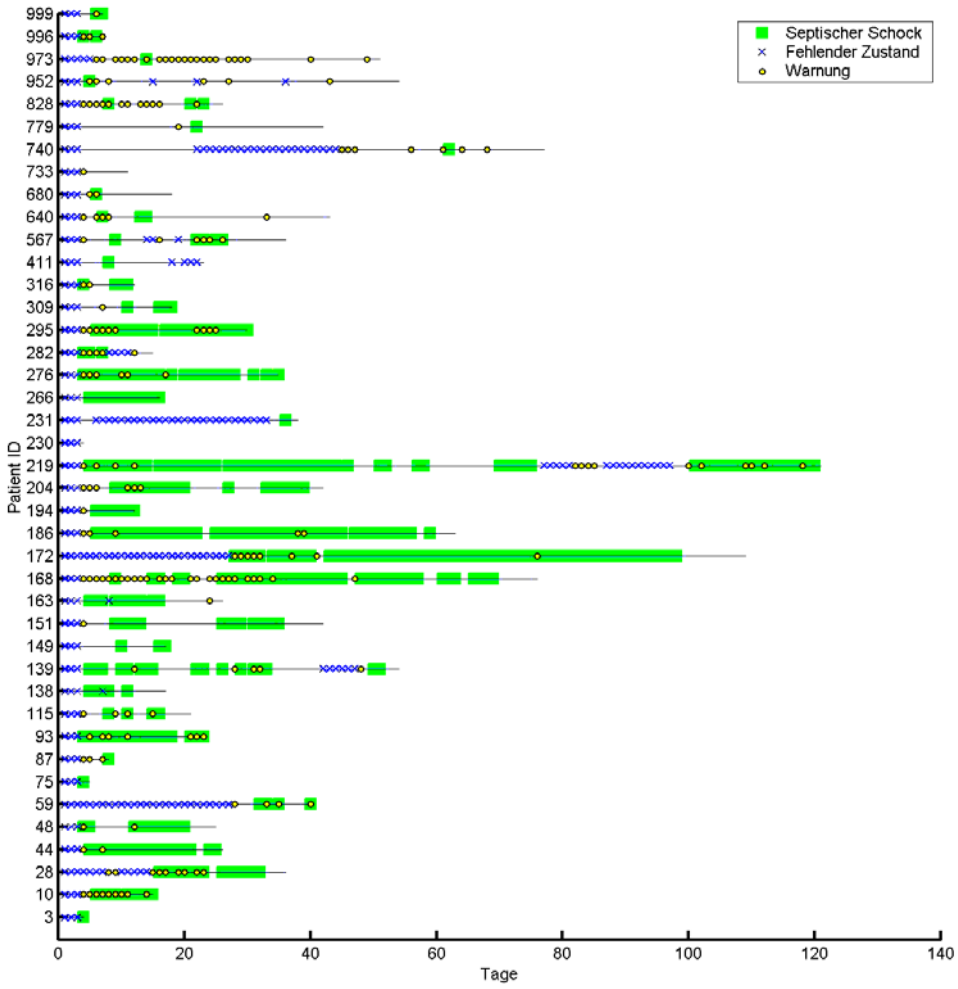


Abbildung 5.10. Septischer Schock und die Ausgabe von Warnungen bei überlebenden Patienten vom ersten bis zum letzten Tag auf der Intensivstation. Das x kennzeichnet Zustände, nicht in die Analyse eingehen konnten.

5.2.5. Mittlere Anzahl an Warnungen

Wie auch in der Studie A1 liegt die mittlere Anzahl an Warnungen bei verstorbenen Patienten höher als bei überlebenden (Abbildung 5.11). Allerdings ist hier der Abstand deutlich geringer.

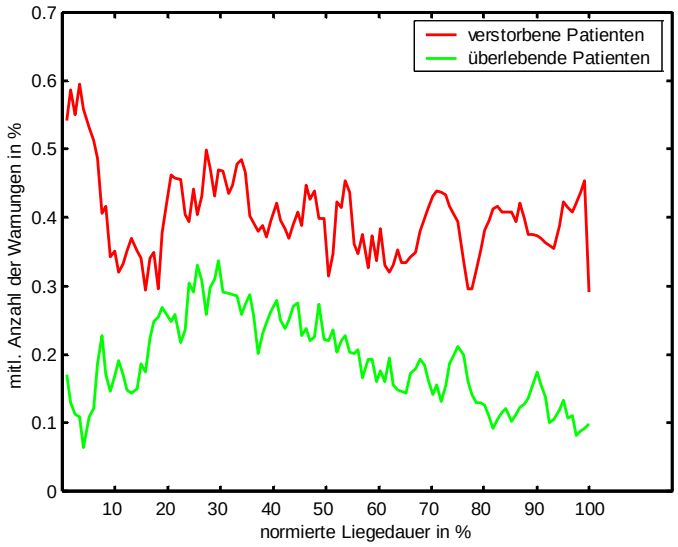


Abbildung 5.11. Mittlere Anzahl der Warnungen bei überlebenden und verstorbenen Patienten in Studie A2.

5.2.6. Tagesanalyse

Es fanden sich keine Auffälligkeiten oder besondere Trends bei den Klassifikationsraten der ersten 12 Tage ab der Einlieferung, den fünften Tag vor bis zum sechsten Tag nach dem erstmaligen Auftreten des septischen Schocks und den letzten 13 Tagen bis zum Exitus bzw. der Entlassung.

5.2.7. Gütemaß

In Studie A2 ist die Sicherheit der auf dem nicht trainierten Patienten durchgeführten Klassifikationen ebenfalls eher gering (Abbildung 5.12), aber eine Grenzlinie bei $g_{\text{limit}} \approx 0$ kann die Klassifikation etwas korrigieren. So würden, aufgrund zu geringer Sicherheit, ca. 75% falsche Entscheidungen als neutral eingestuft, aber nur ca. 50% der richtigen Entscheidungen.

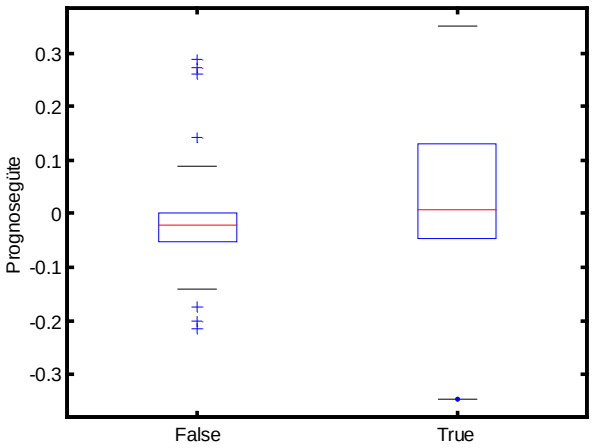


Abbildung 5.12. Prognosegüte der richtigen und falschen Klassifikationen in Studie A2.

5.3. Studie B

In Studie B wurden speziell die Variablen ausgewertet, die zur Bildung des Apache II Score benötigt werden in der Hoffnung, besonders gute Diagnoseergebnisse zu erhalten. Die Ergebnisse aller Patienten der Studie sind einsehbar unter *http://www.medan.de/Klassifikation/Vorstudie/Klassifikation_sepsB/Klassifikation.html*.

5.3.1. Gesamtprognose

In Anlehnung an den Apache II Score erreicht das neuronale Netz keine besseren Ergebnisse als bisherige Untersuchungen (Tabelle 5.3).

Richtige Vorhersagen: 64,56%	Stdabw: 33,95	Min: 0,00%	Max: 100.00%
Sensitivität: 31,68%	Stdabw: 25,35	Min: 0,00%	Max: 100.00%
Spezifität: 85,20%	Stdabw: 19,10	Min: 0,00%	Max: 100.00%

Tabelle 5.3. Übersicht der Klassifikation aller Patienten in Studie B. Die Mittelwerte und die Standardabweichung beziehen sich auf die Patienten und auf alle Zustände. Im Mittel kann man für jeden Patienten bei Verwendung der Differenzen vom gleitenden Mittelwert mit knapp 65% richtigen Vorhersagen rechnen. Die Schwankungen konnten nicht signifikant reduziert werden.

Die weitgehende Berücksichtigung von Variablen des Apache II Scores führt somit nicht zu besseren Ergebnissen als andere medizinisch plausible Parameter. Deutlich findet sich dieses in der breiten Verteilung der richtigen Vorhersagen und in der geringen Sensitivität bestätigt (Abbildung 5.13).

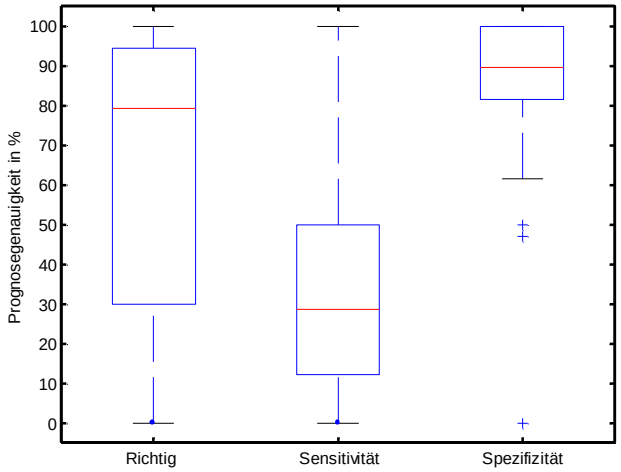


Abbildung 5.13. Boxplot der richtigen Vorhersagen, Sensitivität und Spezifität in Studie B

5.3.2. Gegenüberstellung von Sensitivität und Spezifität

Die geringere Klassifikationsgenauigkeit zeigt sich auch an der schwächer ausgeprägten ROC-Kurve (Abbildung 5.14) im Vergleich zur Studie A1 (Abbildung 5.2).

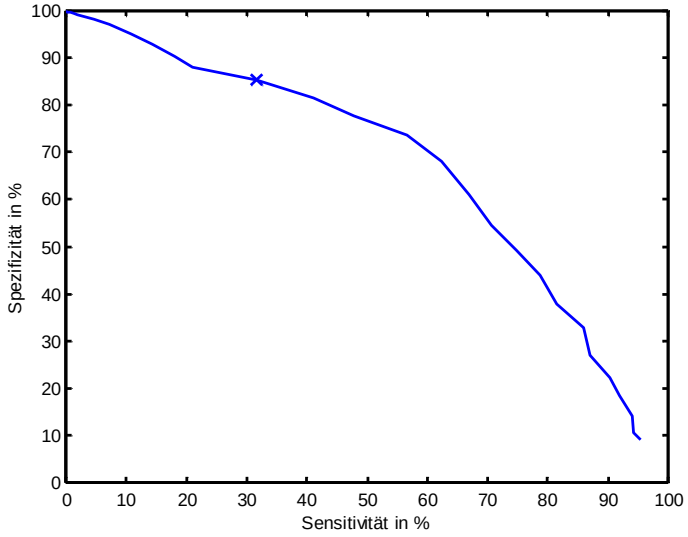


Abbildung 5.14. Receiver Operating Characteristic (ROC)-Kurve in Studie B, berechnet aus den Mittelwerten aller Patienten. Das x kennzeichnet den mittleren Arbeitspunkt aller Netze auf den getesteten Patienten.

5.3.3. Individuelle Warnungen bei verstorbenen Patienten

Bei vielen Patienten erfolgt eine frühe Warnung, jedoch findet sich nicht bei jedem verstorbenen Patienten eine Warnung. Andere Parameter führen womöglich auch zu anderen Zeitpunkten zu einer Warnung. Trotz einiger Übereinstimmungen unterscheiden sich ebenso oft die Zeitpunkte einer Warnung von Studie A1 und Studie B (Abbildung 5.15).

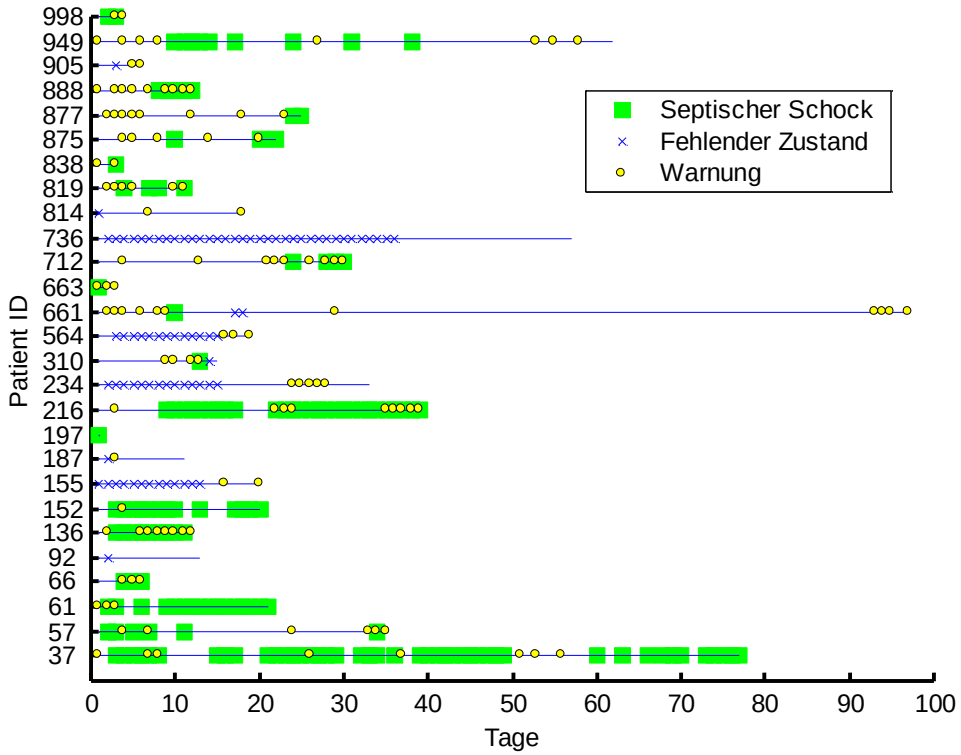


Abbildung 5.15. Septischer Schock und die Ausgabe von Warnungen bei verstorbenen Patienten vom ersten bis zum letzten Tag auf der Intensivstation in Studie B. Das x kennzeichnet Zustände, die nicht in die Analyse eingehen konnten.

5.3.4. Individuelle Warnungen bei überlebenden Patienten

Auch bei den überlebenden Patienten (Abbildung 5.16) stellt sich wie in Studie A1 die Frage, wieso der Patient mit der ID 999 überleben konnte. Der Patient mit der ID 28 ist in dieser Studie nicht mehr so auffällig.

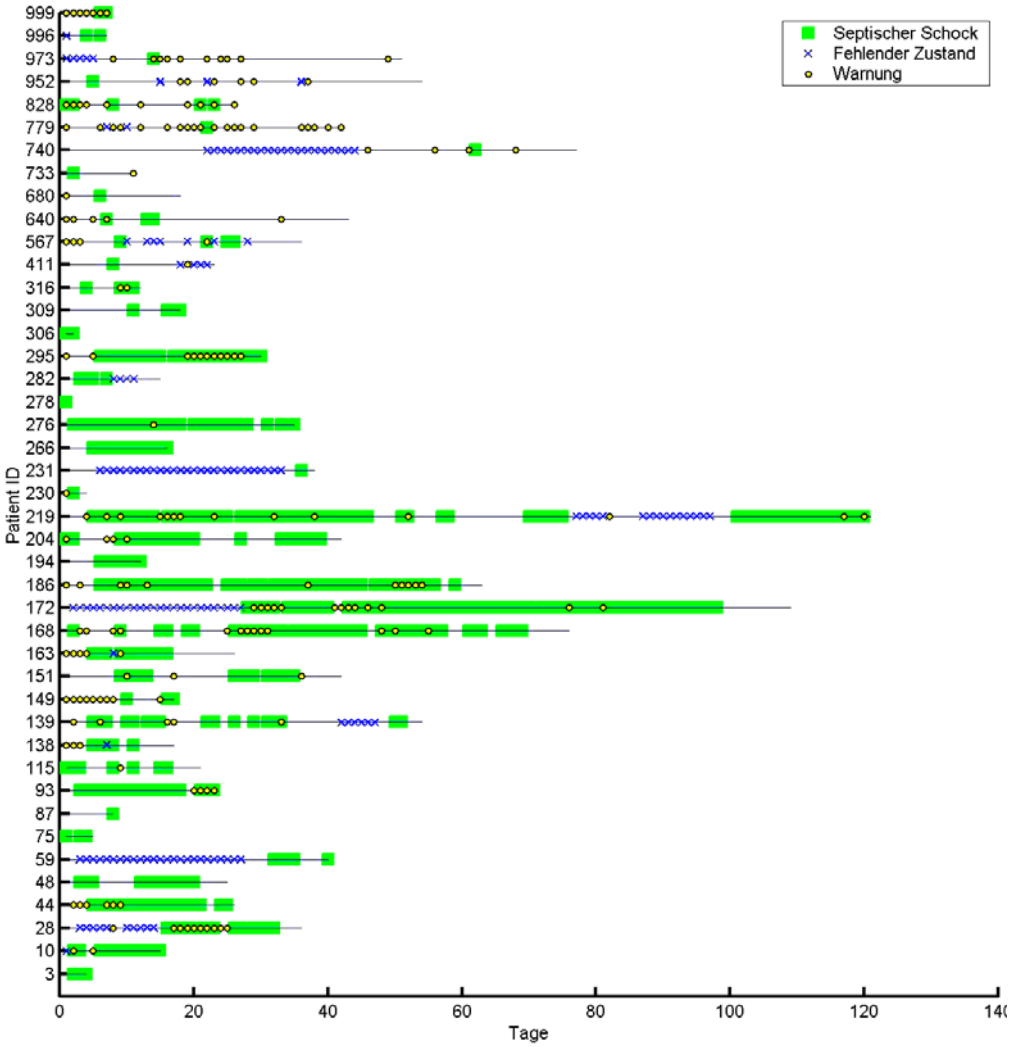


Abbildung 5.16. Septischer Schock und die Ausgabe von Warnungen bei überlebenden Patienten vom ersten bis zum letzten Tag auf der Intensivstation in Studie B. Das x kennzeichnet Zustände, nicht in die Analyse eingehen konnten.

5.3.5. Mittlere Anzahl an Warnungen

Die mittlere Anzahl an Warnungen liegt zwar bei verstorbenen Patienten höher als bei Überlebenden (Abbildung 5.17), allerdings nicht immer deutlich genug.

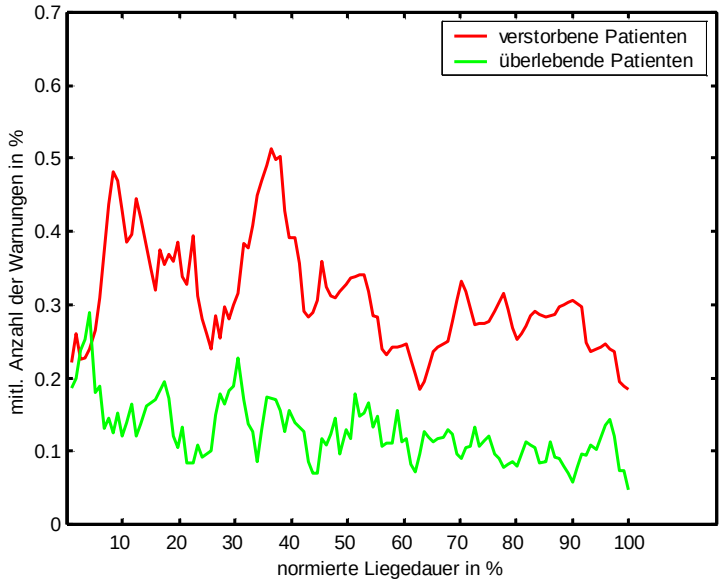


Abbildung 5.17. Mittlere Anzahl der Warnungen bei überlebenden und verstorbenen Patienten in Studie B.

5.3.6. Tagesanalyse

Es fanden sich keine Auffälligkeiten oder besondere Trends bei den Klassifikationsraten der ersten 12 Tage ab der Einlieferung, den fünften Tag vor bis zum sechsten Tag nach dem erstmaligen Auftreten des septischen Schocks und den letzten 13 Tagen bis zum Exitus bzw. der Entlassung.

5.3.7. Gütemaß

In Studie B ist die Sicherheit der richtigen Klassifikationen etwas höher als in Studie A1 und A2 (Abbildung 5.18). Ähnlich wie in Studie A2 kann eine Grenzlinie bei $g_{\text{limit}} \approx 0$ die Klassifikation etwas korrigieren. So würden, aufgrund zu geringer Sicherheit, ebenfalls ca. 75% falsche Entscheidungen und ca. 50% der richtigen Entscheidungen als neutral eingestuft.

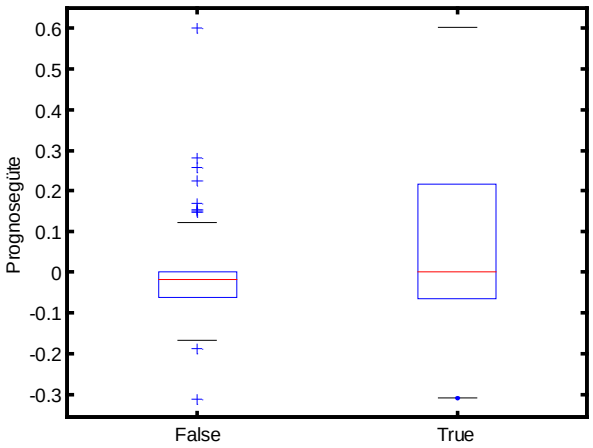


Abbildung 5.18. Prognosegüte der richtigen und falschen Klassifikationen in Studie B.

5.4. Studie C

In Studie C wurden speziell die Variablen ausgewählt, die höhere Korrelationsunterschiede bei *Exitus*/*nichtExitus* aufweisen. Anstatt die Korrelationen selbst heranzuziehen wurden tageweise, relative Abhängigkeiten zwischen den Variablen berechnet.
Die Ergebnisse aller Patienten der Studie sind einsehbar unter
http://www.medan.de/Klassifikation/Vorstudie/Klassifikation_sepsQC1_ma/Klassifikation.html.

5.4.1. Gesamtprognose

Betrachtet man die Beziehungen von Variablen untereinander, statt lediglich ihre Zustände ergibt sich ein ähnlich gutes Bild wie in Studie A1 (Tabelle 5.4). Zwar ist der Mittelwert der richtigen Vorhersagen leicht geringer, aber der Median liegt leicht über dem von Studie A1 (Abbildung 5.19). Die Schwankungen der Sensitivität und Spezifizität sind ebenfalls geringer und weisen auf eine klarere Trennung hin.

Richtige Vorhersagen: 67,50%	Stdabw: 31,57	Min: 0,00%	Max: 100.00%
Sensitivität: 34,51%	Stdabw: 27,32	Min: 0,00%	Max: 100.00%
Spezifität: 87,45%	Stdabw: 9,38	Min: 60,00%	Max: 100.00%

Tabelle 5.4. Übersicht der Klassifikation aller Patienten. Die Mittelwerte und die Standardabweichung beziehen sich auf die Patienten und auf alle Zustände. Im Mittel kann man für jeden Patienten bei Verwendung der Differenzen vom gleitenden Mittelwert mit gut 67% richtigen Vorhersagen rechnen. Die Schwankungen konnten nicht signifikant reduziert werden.

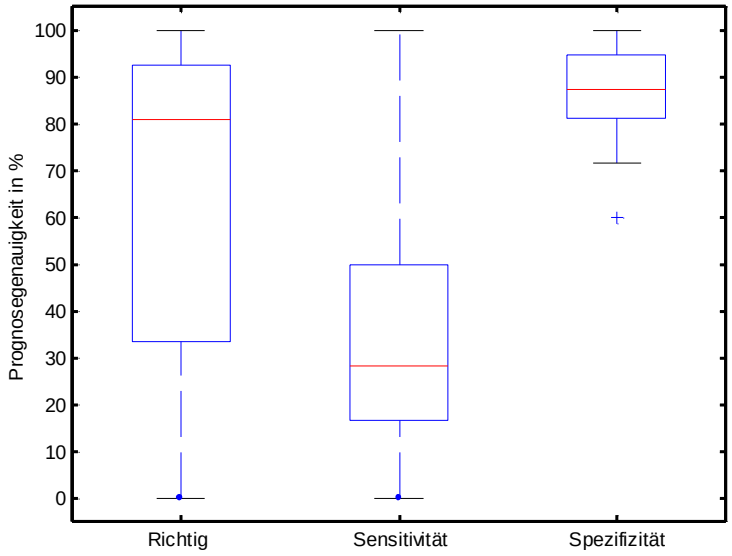


Abbildung 5.19. Boxplot der richtigen Vorhersagen, Sensitivität und Spezifität in Studie C.

5.4.2. Gegenüberstellung von Sensitivität und Spezifität

Verglichen mit der Studie A1 zeigt die ROC-Kurve eine nicht so gute Spezifität bei einer mittleren Sensitivität an (Abbildung 5.20). Bei einer Spezifität von 85%-100% lässt sich jedoch wie in Studie A1 eine recht hohe Sensitivität erzielen.

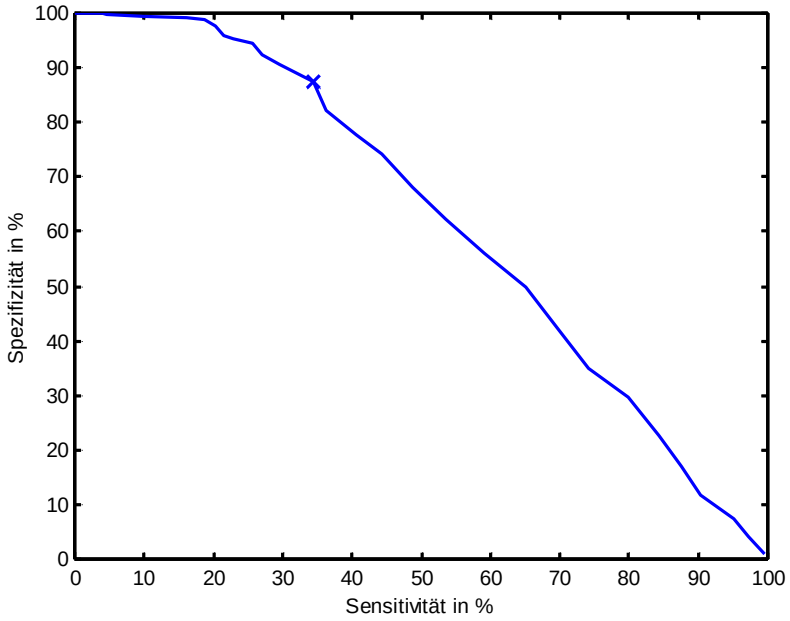


Abbildung 5.20. Receiver Operating Characteristic (ROC)-Kurve in Studie C, berechnet aus den Mittelwerten aller Patienten. Das x kennzeichnet den mittleren Arbeitspunkt aller Netze auf den getesteten Patienten.

5.4.3. Individuelle Warnungen bei verstorbenen Patienten

Da der Arbeitspunkt innerhalb der ROC-Kurve von Studie A1 und Studie C zu weit auseinanderliegt, treten in Studie C weniger Warnungen auf. Die Zeitpunkte der Warnungen liegen auch hier oft in ähnlichen Bereichen wie in Studie A1.

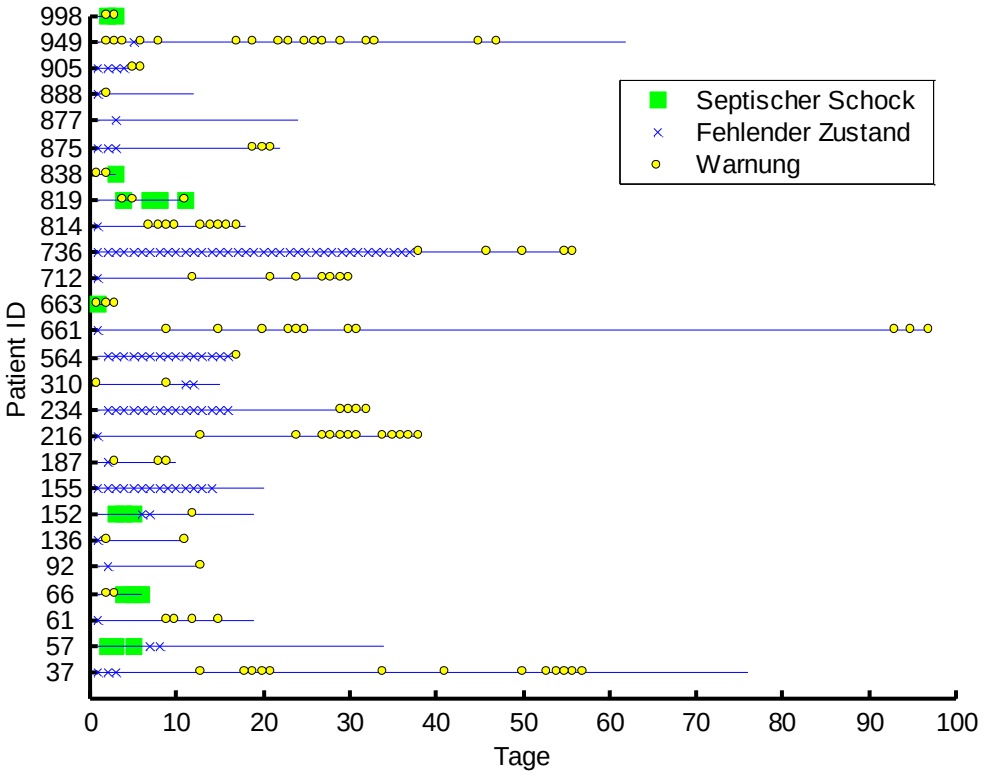


Abbildung 5.21. Septischer Schock und die Ausgabe von Warnungen bei verstorbenen Patienten vom ersten bis zum letzten Tag auf der Intensivstation in Studie C. Das x kennzeichnet Zustände, die nicht in die Analyse eingehen konnten. Der septische Schock wird hier nicht bei allen Patienten angezeigt.

5.4.4. Individuelle Warnungen bei überlebenden Patienten

Auch bei den überlebenden Patienten (Abbildung 5.10) zeigen sich einige Übereinstimmungen zu Studie A1, doch scheinen diese insgesamt geringer auszufallen. Die sonst festgestellten extrem häufigen Warnungen bei PID 999 treten in dieser Studie nicht auf. Die Gesamtzahl der Warnungen ist allerdings ebenfalls geringer.

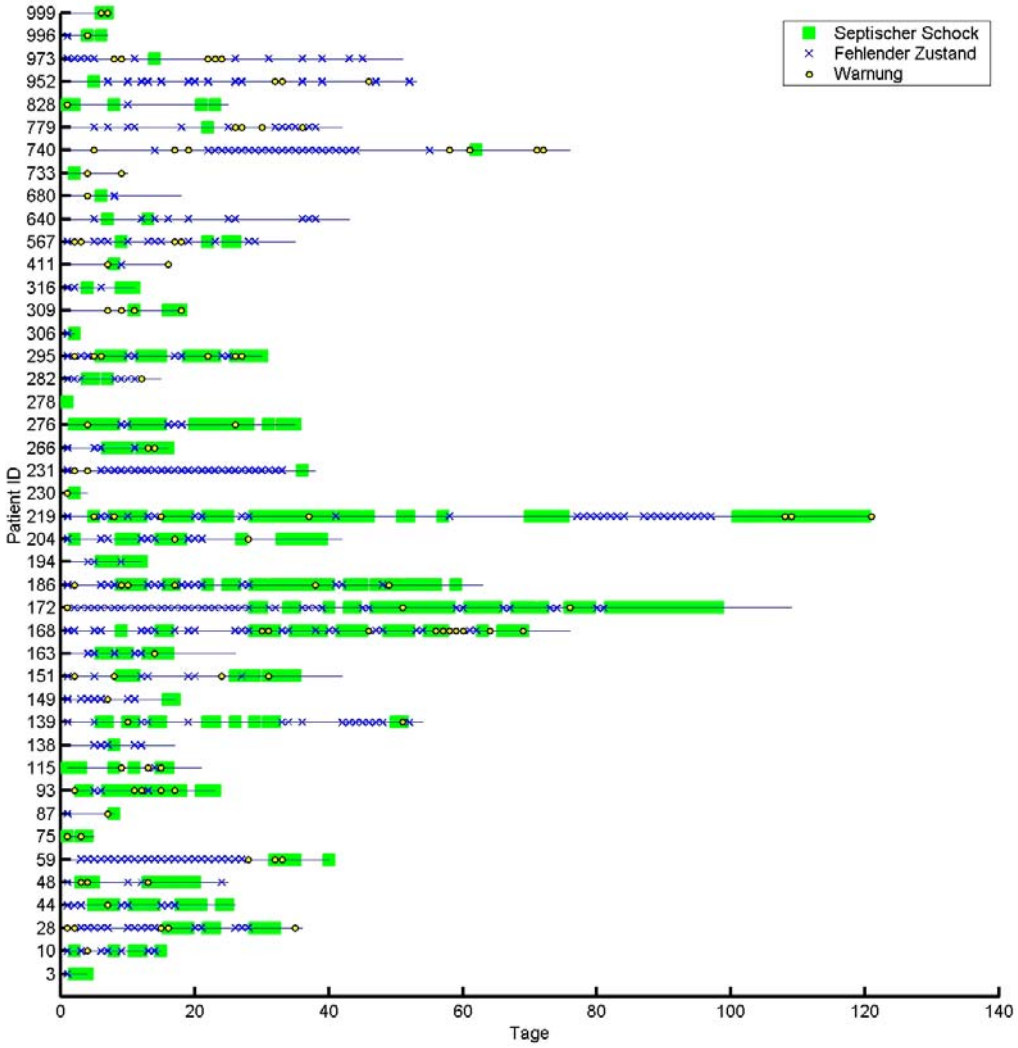


Abbildung 5.22. Septischer Schock und die Ausgabe von Warnungen bei überlebenden Patienten vom ersten bis zum letzten Tag auf der Intensivstation in Studie C. Das x kennzeichnet Zustände, nicht in die Analyse eingehen konnten.

5.4.5. Mittlere Anzahl an Warnungen

Wie auch in der Studie A1 liegt die mittlere Anzahl an Warnungen bei verstorbenen Patienten höher als bei überlebenden (Abbildung 5.23). Allerdings ist hier der Abstand deutlich geringer. In weiten Bereichen der Liegedauer gibt es häufigere Warnungen bei verstorbenen Patienten als bei überlebenden.

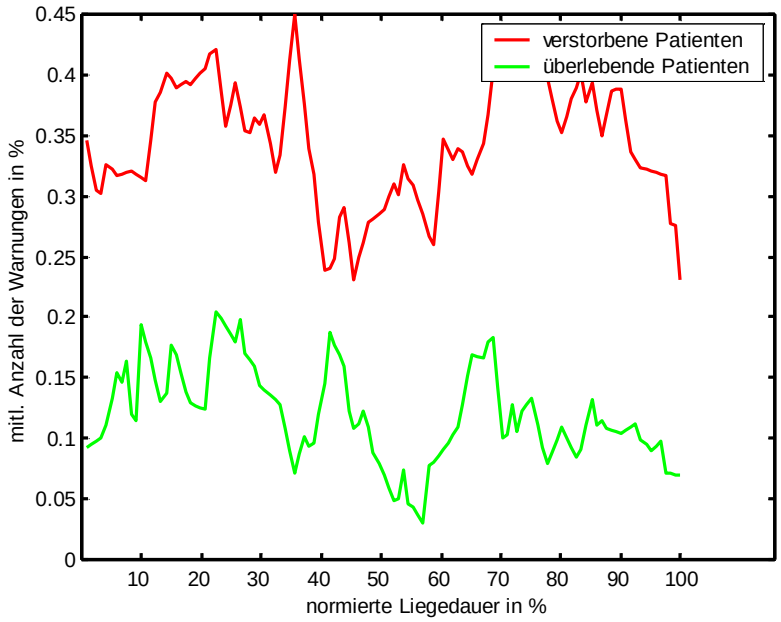


Abbildung 5.23. Mittlere Anzahl der Warnungen bei überlebenden und verstorbenen Patienten in Studie C.

5.4.6. Tagesanalyse

Es allerdings fanden sich keine Auffälligkeiten oder besondere Trends bei den Klassifikationsraten der ersten 12 Tage ab der Einlieferung, den fünften Tag vor bis zum sechsten Tag nach dem erstmaligen Auftreten des septischen Schocks und den letzten 13 Tagen bis zum Exitus bzw. der Entlassung.

5.4.7. Gütemaß

Studie C zeigt als einzige Studie eine deutlich unterschiedliche Sicherheit von richtigen und falschen Klassifikationen. Das obere Quartil der falschen Klassifikationen liegt bei ca. 0.02, während der Median der richtigen Klassifikationen einen Wert von ca. 0.05 und das obere Quartil einen Wert von ca. 0.2 besitzt (Abbildung 5.24). Obwohl damit die beste Korrektur der Klassifikation aller durchgeführten Studien erfolgen kann, ist dieses absolut bewertet noch nicht vollends zufriedenstellend, da die Sicherheit der richtigen Klassifikationen noch zu gering ist, und noch zu viele falsche Klassifikationen mit hoher Sicherheit vorhanden sind. So wird die höchste Sicherheit sogar von einer falschen Klassifikation erzielt.

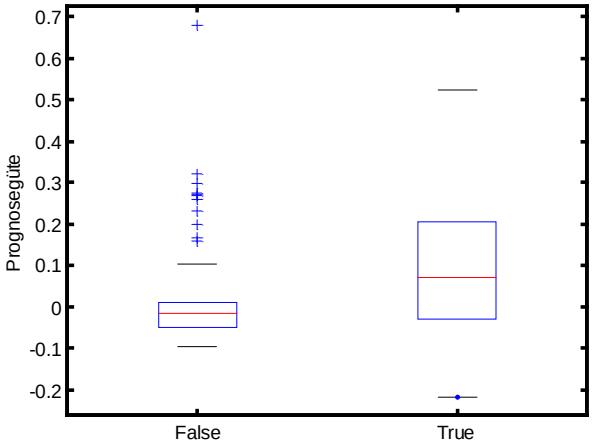


Abbildung 5.24. Prognosegüte der richtigen und falschen Klassifikationen in Studie C.

5.5. Fazit

Die Klassifikation (einhergehend mit der Ausgabe von Warnungen bei kritischen Zuständen) erweist sich prinzipiell als ein guter Ansatz zur Unterstützung einer Therapie. Die Häufigkeit der Warnungen kann den Erfordernissen angepasst werden, prinzipiell ist aber ein eher konservatives Maß zu empfehlen, so dass die Warnungen eher selten, aber dafür verlässlich sind. Besonders hervorzuheben sind die klar häufigeren Warnungen bei verstorbenen Patienten. Allerdings können die erzielten Ergebnisse nicht als ein Durchbruch gewertet werden. Die Anzahl der Warnungen bei verstorbenen Patienten ist noch zu niedrig, um den Arzt deutlich zu unterstützen. Weiterhin konnten keine Variablen bzw. weitergehende Merkmalsextraktionsverfahren ermittelt werden, die sich deutlich in ihrem Ergebnis von anderen abheben. Die dokumentierten Studien erwiesen sich, verglichen mit anderen, hier nicht dokumentierten Variablen, bereits als brauchbarer.

Zur Beurteilung der Qualität einer Warnung (z.B. neue Erkenntnis / Bestätigung einer Vermutung / dem Arzt ohnehin schon bekannt / Widerspruch zur aktuellen med. Einschätzung) hilft nur eine rückblickende medizinische Analyse der Patientenkurve. Dabei wäre zu klären, ob sich die Patienten wirklich so oft in einem kritischen Zustand befanden oder ob hier die zufällige Ähnlichkeit mit verstorbenen Patienten in unkritischen Zuständen den Ausschlag gegeben hat. Falls letzteres der Fall ist müssen diese Zustände aus der Trainingsmenge entfernt werden. Wie man an diesem Beispiel erkennt, hat eine vollautomatische Vorgehensweise Grenzen. Mit Hilfe einer Begutachtung der Entscheidung des neuronales Netzes durch einen erfahrenden Mediziner und einem erneuten Training können evtl. bessere Ergebnisse erwartet werden.

Die zu hohe Rate der Fehlklassifikation und die teilweise hohen Sicherheiten bei falschen Klassifikationen weisen auf eine starke Überlappung der Patientendaten hin. Um den Grund der geringen Trennung zwischen überlebenden und verstorbenen Patienten näher zu betrachten, wird im Folgenden exemplarisch für die Daten der Studie A1 und C eine Clusteranalyse durchgeführt.

6. Ergebnisse der Clusteranalyse

Bei der unüberwachten Clusteranalyse wird den Daten zunächst keine Bedeutung wie Verstorben/nicht Verstorben zugewiesen. Allein die Verteilung der Daten im Raum ist hierbei von Interesse. Die Wahl der Clusterzentren erfolgt daher unabhängig von speziellen Kategorisierungen. Ähnliche Zustände von Patienten, die überlebten und von denen die verstarben,

werden hier nicht getrennt, sondern in gemeinsamen Clustern zusammengefasst. Erst durch anschließende Interpretationen hinsichtlich der Verteilung der Daten und des zeitlichen Verlaufs von Patienten lassen sich medizinisch relevante Aussagen gewinnen.
Um herauszufinden warum zwar im Mittel vernünftige Ergebnisse auf der Frankfurter Vorstudie erzielt wurden, aber die Unsicherheit bzgl. der Patienten sehr hoch ist, wurden die aus den Daten gebildeten Cluster näher in ihrer Zusammensetzung untersucht. Diese Analyse wurde exemplarisch für die Daten der Studie A1 durchgeführt. Nach den Klassifikationsergebnissen beurteilt, ergeben sich vermutlich ähnliche Ergebnisse bei den anderen Studien. Die Analyse erfolgte mit einem hierarchischen Clusterverfahren (vgl. [SL77]) auf Basis der euklidschen Metrik. Die Abstände wurden nach dem *Complete Linkage* Maß berechnet. Anhand dieser Abstandsmaße lassen sich Zustände zu Clustern zusammenfassen. Jeder Cluster enthält damit, bezogen auf das Abstandsmaß, ähnliche Zustände, d.h. ähnliche Vektoren aus Messwerten der Patienten.

6.1. Zustände der Patienten

Bei einer groben Clusterung der Zustände in Studie A1 lassen sich die 1724 Zustände, von denen 1134 von überlebenden und 533 von verstorbenen Patienten stammen, in 35 Cluster aufteilen. Will man Warnungen bei kritischen Zuständen ausgeben, sind besonders die Cluster interessant, in denen sich mehrheitlich verstorbene Patienten befinden. Bei einer Aufteilung in 35 Cluster sind dieses 9 Cluster (Abbildung 6.1). Drei der 9 Cluster enthalten nur Zustände von jeweils einem Patienten. Diese Cluster sind für eine Prognose nicht brauchbar, da so von einem Patienten kaum auf einen anderen geschlossen werden kann. Der interessanteste Cluster enthält immerhin Zustände von jeweils vier verschiedenen Patienten und nur 20% der Zustände sind von überlebenden Patienten. Andere Cluster enthalten zwar Zustände von noch mehr verstorbenen Patienten, aber der Anteil der Zustände von überlebenden Patienten wächst ebenfalls. Die 9 Cluster decken 52 % der verstorbenen Patienten, aber nur 10% ihrer Zustände ab. D.h. 48% der verstorbenen Patienten haben nie einen Zustand besessen, der in diese 9 Cluster, sondern in Clustern mit mehrheitlich überlebenden Patienten fällt, und 90% der Zustände liegen in Clustern mit mehrheitlich überlebenden Patienten.

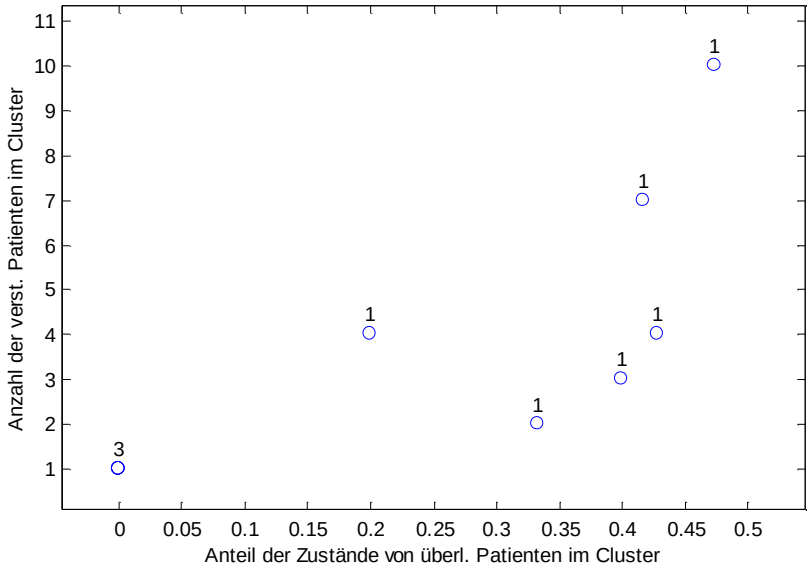


Abbildung 6.1. Anteil der verstorbenen Patienten in 9 Clustern mit überwiegendem Anteil von Zuständen verstorbener Patienten bei insgesamt 35 Clustern in Studie A1. Die Zahlen geben die Anzahl der Cluster mit diesen Charakteristika an.

Mit diesem Grad an Detail ist die Überlappung zwischen den beiden Kategorien noch sehr hoch, so dass Bereiche der Zustände von verstorbenen Patienten noch nicht ausreichend zu Tage treten. Um verstorbene Patienten und ihre Zustände stärker von überlebenden Patienten zu trennen, muss man die Clusterung verfeinern, um so aus den großen Clustern mit mehrheitlich überlebenden Patienten kleinere mit mehrheitlich verstorbenen Patienten abzuspalten.

Erhöht man die Anzahl der Cluster auf 125, so ergeben sich 44 Cluster mit mehrheitlich verstorbenen Patienten (Abbildung 6.2).

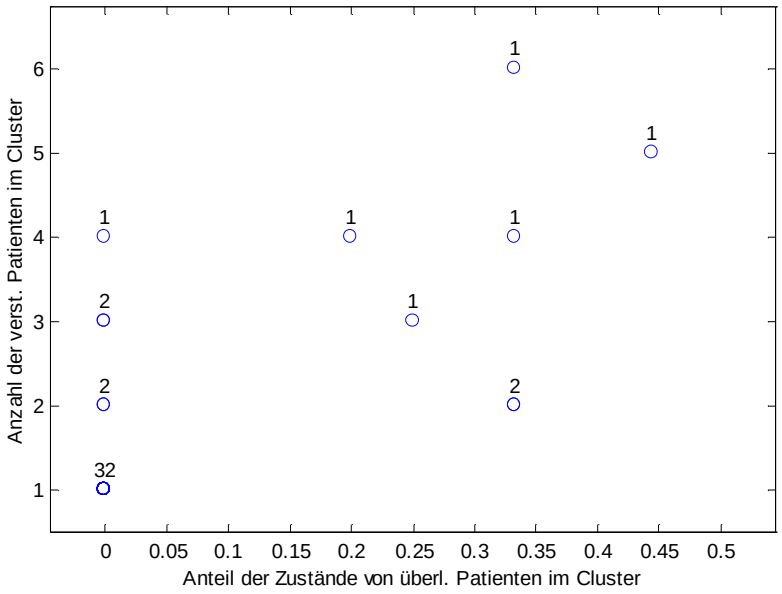


Abbildung 6.2. Anteil der verstorbenen Patienten in 44 Clustern mit überwiegendem Anteil von Zuständen verstorbener Patienten bei insgesamt 125 Clustern. Die Zahlen geben die Anzahl der Cluster mit diesen Charakteristika an.

Bei dieser Aufteilung enthalten die Cluster mit Zuständen, die mehrheitlich von verstorbenen Patienten stammen 89% der Patienten und 23% ihrer Zustände. D.h. 11% der verstorbenen Patienten haben nie einen Zustand besessen, der in diese 44 Cluster, sondern in Clustern mit mehrheitlich überlebenden Patienten fällt und 77% der Zustände liegen in Clustern mit mehrheitlich überlebenden Patienten. Obwohl sich die Zustände von verstorbenen und überlebenden Patienten weiterhin stark überlappen, finden sich jedoch die meisten verstorbenen Patienten in Clustern wieder, die mehrheitlich verstorbene Patienten enthalten. Die 23% der Zustände könnten demnach durchaus kritisch für die Patienten sein.

Betrachtet man nun die Zusammensetzung der Cluster genauer, so fällt auf, dass 32 Cluster wiederum nur jeweils einen Patienten enthalten. Die bessere Trennung der Zustände von überlebenden und verstorbenen Patienten geht demnach mit einer Abspaltung einzelner Patienten einher. Von den Clustern mit keinen Zuständen überlebender Patienten weisen zwei Cluster Zustände von zwei verstorbenen Patienten, weitere zwei Cluster Zustände von drei verstorbenen Patienten und ein Cluster Zustände von vier verstorbenen Patienten auf. Derartige Cluster sind ganz besonders interessant für die Analyse, da sie anscheinend typische Konstellationen aufweisen, die mehrere verstorbene Patienten besaßen (und somit prognostizierbar sind) aber nicht von überlebenden Patienten aufgesucht werden. Selbstverständlich findet diese auch das neuronale Netz. Da diese Cluster aber im Vergleich zu selten sind, und viele Patienten ihre eigenen Cluster definieren ist eine individuelle Prognose auf Basis der Zustände so unsicher – je nachdem ob ein anderer Patient der gleichen Kategorie ebenfalls im Clu-

ster vorhanden ist.

6.2. Korrelierte Variablen der Patienten

In Studie C sieht das Bild etwas anders aus. Die Cluster beschreiben jetzt nicht mehr direkt den Zustand von Patienten hinsichtlich ihrer Messungen, sondern relative Abhängigkeiten der Variablen untereinander. Um wiederum neun Cluster mit mehrheitlich verstorbenen Patienten zu erhalten, sind 38 Cluster erforderlich (Abbildung 6.3).

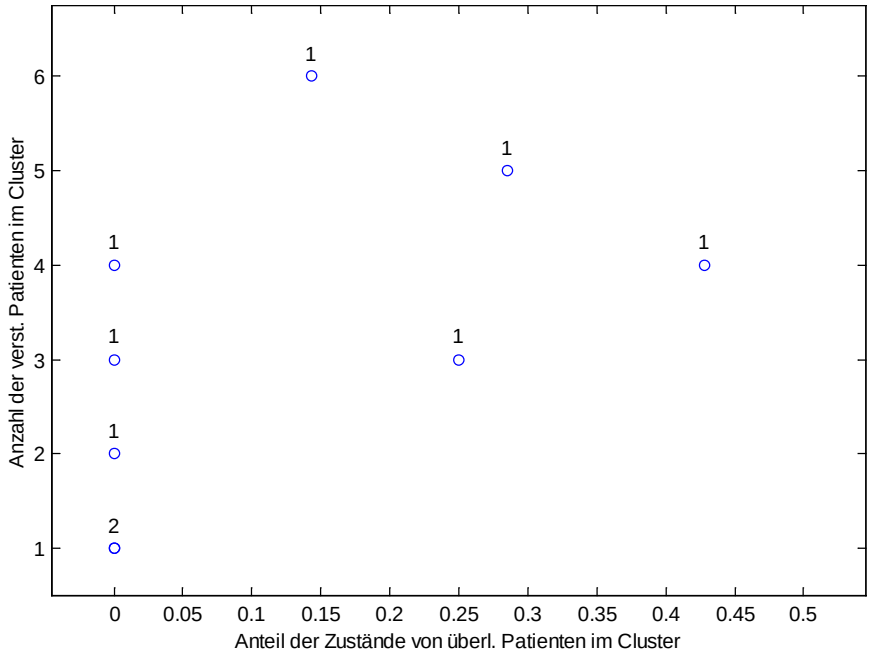


Abbildung 6.3. Anteil der verstorbenen Patienten in 9 Clustern mit überwiegendem Anteil von Zuständen verstorbener Patienten bei insgesamt 38 Clustern in Studie C. Die Zahlen geben die Anzahl der Cluster mit diesen Charakteristika an.

Es finden sich allerdings mehr Cluster in dem oberen linken Bereich, der für eine Vorhersage optimal ist. Während in Studie A1 alle Cluster mit Zuständen, die ausschließlich von verstorbenen Patienten stammen nur einen einzelnen Patienten enthielten, finden sich hier auch Cluster mit zwei, drei und vier verstorbenen Patienten. Auch Cluster mit mehr verstorbenen Patienten weisen im Mittel eine geringere Quote von Zuständen überlebender Patienten als in Studie A1 auf. Genauso wie in Studie A1 beinhalten diese neun Cluster 10% der Zustände von allen verstorbenen Patienten, allerdings immerhin 65% der verstorbenen Patienten, gegenüber nur 52% in Studie A1.

Erhöht man nun die Anzahl der Cluster derart, dass wiederum 44 Cluster mehrheitlich verstorbene Patienten aufweisen, so sind hierfür 153 Cluster erforderlich (Abbildung 6.4), gegenüber 125 in Studie A1.

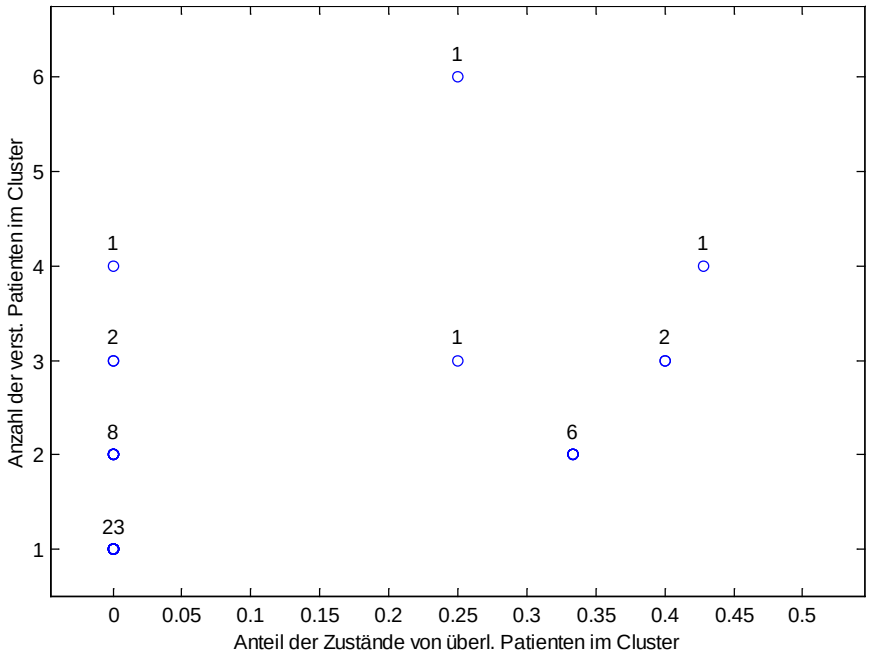


Abbildung 6.4. Anteil der verstorbenen Patienten in 44 Clustern mit überwiegendem Anteil von Zuständen verstorbener Patienten bei insgesamt 153 Clustern in Studie C. Die Zahlen geben die Anzahl der Cluster mit diesen Charakteristika an.

Bei dieser Aufteilung erhalten die Cluster mit mehrheitlich verstorbenen Patienten 88% der verstorbenen Patienten und 26% ihrer Zustände. Die Zahl der Zustände ist damit leicht höher als in Studie A1. Betrachtet man nur die reinen Cluster, so gibt es wie in Studie A1 einen Cluster, der vier verschiedene Patienten und zwei, die drei verschiedene Patienten enthalten. Die Anzahl der reinen Cluster mit zwei Patienten ist mit acht Clustern um sechs höher als in Studie A1. Weiterhin ist die Anzahl der reinen Cluster mit jeweils nur einem Patienten deutlich geringer. Allerdings erhöhen sich im wesentlichen nur die Anzahl der Cluster in den sich zwei verschiedene verstorbene Patienten befinden.

7. Zusammenfassung

Es wurde gezeigt, dass der Arzt mit dem wachsenden RBF-Netz durch die Ausgabe von verlässlichen Warnungen unterstützt werden kann. Wie in der Clusteranalyse erläutert, leiden die Ergebnisse jedoch unter den wenigen Patienten und unter der ungenauen zeitlichen Erfassung der Daten. Da jeder Patient sehr individuelle Zustände annimmt, ist ein größeres Patientenkollektiv notwendig, um eine umfassende Wissensbasis zu lernen. Eine medizinische Nachbearbeitung der Wissensbasis durch die Analyse der Fälle ließe eine weitere Verbesserung des Ergebnisses erwarten. Somit könnten unbekannte Zusammenhänge durch das Lernen aus Beispielen und medizinisches Fachwissen kombiniert werden.

Abstraktere Merkmale, die weniger abhängig von individuellen Zuständen sind, könnten eine Klassifikation noch weiter verbessern. Ein Ansatzpunkt ist z.B. die Abweichung der Messwerte vom gleitenden Mittelwert. Dieses Maß ist unempfindlicher gegenüber den individuellen Arbeitspunkten der Patienten und bildet auch die Basis von relativen Abhängigkeiten zwischen zwei Variablen, die in einem weiteren Schritt ebenfalls als Merkmal herangezogen wurden. Obwohl die Verwendung der relativen Abhängigkeiten zwischen zwei Variablen als Merkmal nicht deutlichere oder häufigere Warnungen hervorbringen konnte, weist doch die Clusteranalyse auf eine bessere Verteilung der Patienten hin. Einige Cluster sind besser für die Vorhersage geeignet, als dieses bei einer Clusterung auf Basis der Zustände erreicht werden kann. Unterstützt wird dieses Ergebnis auch durch den größeren Unterschied der Sicherheiten von falschen und richtigen Klassifikationen. Neben den bisher untersuchten Merkma-

len scheinen auch die Variablen interessant zu sein, bei denen festgestellt wurde, dass sie sich trotz Medikamentengabe und adäquater Behandlung schwer stabilisieren lassen. Durch den behandelnden Arzt werden diese Werte üblicherweise in einem gewissen Bereich gehalten. Falls sich das Paar Medikament/physiologischer Parameter nicht mehr in einem sinnvollen Verhältnis befindet, kann dieses ein wichtiger Indikator sein.

Nach dem Aufbau der grundlegenden Funktionalität der hier untersuchten Methoden ist die Suche nach geeigneten Merkmalen als Eingabe für ein neuronales Netz ein wesentlicher Bestandteil folgender Arbeiten. Abgesehen von dem generell anspruchsvollen Vorhaben aus Klinikdaten deutliche Hinweise für die Mortalität septischer-Schock-Patienten zu erhalten, liegen die wesentlichen Probleme in dem Umfang und der Messhäufigkeit der Frankfurter Vorstudie begründet, so dass eine Anwendung von Klassifikationsverfahren auf das umfassendere Patientenkollektiv der MEDAN Multicenter-Studie klarere Ergebnisse erwarten lässt.

Eine weitere, für medizinische Anwendungen interessante, Analysemöglichkeit ist die Regelgenerierung, die zur Zeit in einem anderen Teilprojekt in der MEDAN-Arbeitsgruppe bearbeitet wird. Hier können im Fall metrischer Daten zusätzliche Hinweise für die Leistung eines reinen Klassifikationsverfahrens gewonnen werden mit dem Vorteil einer expliziten Regelausgabe. Zum anderen werden in diesem Teilprojekt auch Verfahren zur Regelgenerierung eingesetzt, die ordinale und nominale Variablen wie Diagnosen, Operationen, Therapien und Medikamentenangaben (binär, ohne genaue Dosis) auswerten können. Diese werden in den Multicenter-Daten vorhanden sein. Durch Kopplung der Regelgeneratoren für metrische Daten auf der einen Seite und für diskrete Variablen auf der anderen Seite, besteht durchaus die Hoffnung bessere Ergebnisse zu erzielen.

Da der Regelgenerator für metrische Daten auf dem RBF-DDA (Abk. für: Dynamic Decay Adjustment)-Netz [BERTHOLD und DIAMOND, 1995] beruht, bietet es sich innerhalb des MEDAN-Projekts an, einen (bislang nicht durchgeführten) Vergleich mit dem hier verwendeten Netztyp durchzuführen. Der Vergleich ist allerdings nur von prinzipiellem Interesse und kann auf den hier betrachteten Daten kein grundsätzlich besseres Ergebnis liefern als die bislang durchgeführten Analysen; er kann aber zu einer umfangreichen Bewertung der Ergebnisse beitragen.

Literatur

[BERTHOLD und DIAMOND, 1995] Berthold, M. R.; Diamond, J.: *Boosting The Performance of RBF Networks with Dynamic Decay Adjustment*. Advances In Neural Information Processing Systems, vol. 7, S. 521-528, 1995.

[BONE ET AL., 1992] Bone, R.C.; Balk, R.A.; Cerra, F.B.; Dellinger, R. P.; American College of Chest Physician/Society of Critical Care Medicine Consensus Conference Committee: *Definition for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis*. Critical Care Medicine, vol. 20 (1992), S. 864-874.

[BRAUSE, 1999] Brause, R.: *Revolutionieren Neuronale Netze unsere Vorhersagefähigkeiten?* Zentralblatt für Chirurgie, vol. 124 (1999), S. 692-698.

[BRUSKE und SOMMER, 1995] Bruske, J.; Sommer, G.: *Dynamic cell structure learns perfectly topology preserving map*. Neural Computation, vol. 7 (1995), S. 845-865.

[BRUSKE, 1998] Bruske, J.: *Dynamische Zellstrukturen.Theorie und Anwendung eines KNN-Modells*. Dissertation, Technische Fakultät der Christian-Albrechts-Universität zu Kiel 1998.

[DYBOWSKI, 1997] Dybowski, R.: *Assigning confidence intervals to neural network predictions*. Technical Report, Division of Infection, UMDS (St Thomas‘ Hospital), London 1997.

[FRITZKE, 1992] Fritzke, B.: *Wachsende Zellstrukturen – ein selbstorganisierendes neuronales Netzwerkmodell*. Dissertation, Technische Fakultät der Universität Erlangen-Nürnberg 1992.

[FRITZKE, 1994] Fritzke, B.: *Growing cell structures – A self-organizing network for unsupervised and supervised learning*. Neural Networks, vol. 7 (1994), S. 1441-1460.

[FRITZKE, 1995] Fritzke, B.: *A growing neural gas network learns topologies*. Advances in Neural Information Processing Systems (NIPS 7), Hrsg. von G. Tesauro, D. S. Touretzky, T. K. Leen. Cambridge, MA: MIT Press 1995, S. 625-632.

[GEISSER, 1975] Geisser, S.: *The predictive sample reuse method with applications*. Journal of The American Statistical Association, vol. 70 (1975).

[GUÉRIN-DUGUÉ ET AL., 1995] Guérin-Dugué, A. et al.,: *Enhanced learning for evolutive neural architecture*. Technischer Bericht, Deliverable R3-B4-P - Task B4: Benchmarks: Elena- NervesII, Juni 1995. Anonymous FTP: on ftp.dice.ucl.ac.be/pub/neural-nets/ELENA /Benchmarks.ps.Z.

[HAMKER und HEINKE, 1997] Hamker, F.; Heinke, D.: *Implementation and Comparison of Growing Neural Gas, Growing Cell Structures and Fuzzy Artmap*. Technischer Bericht (Report 1/97), Schriftenreihe des FG Neuroinformatik der TU Ilmenau 1997, ISSN 0945-7518.

[HANISCH ET AL., 1998] Hanisch, E.; Büssow, M.; Brause, R.; Encke, A.: *Individuelle Prognose bei kritisch kranken Patienten mit septischem Schock durch ein neuronales Netz?*. Der Chirurg, vol. 69 (1998), S. 77-81.

[HARTUNG, 1993] Hartung, J.: *Statistik: Lehr- und Handbuch der angewandten Statistik*. Oldenbourg-Verlag 1993.

[HEINKE und HAMKER, 1998] Heinke, D.; Hamker, F. H.: *Comparing Neural Networks: A Benchmark on Growing Neural Gas, Growing Cell Structures, and Fuzzy ARTMAP*. IEEE Transactions on Neural Networks, vol. 9 (1998), S. 1279-1291.

[KINDERMANN ET AL., 1999] Kindermann, L., Lewandowski, A., Tagscherer, M., Protzel, P.: *Computing Confidence Measures and Marking Unreliable Predictions by Estimating Input Data Densities with MLPs*. Proceedings of the Sixth International Conference on Neural Information Processing (ICONIP'99), Perth, Australia 1999.

[KNAUS ET AL., 1985] Knaus, W.A., Draper, E., Wagner, D.P., Zimmerman, J.E.: *APACHE II: A Severity of Disease Classification System*, Critical Care Medicine, vol. 13(10), 1985, S. 818-829.

[KOHONEN, 1982] Kohonen, T.: *Self-organized formation of topologically correct feature maps*. Biological Cybernetics, 43 (1982), S. 59-69.

[MARTINETZ und SCHULTEN, 1991] Martinetz, T. M.; Schulten, K. J.: *A "neural gas" network learns topologies.* In: Proceedings of the International Conference on Artificial Neural Networks (ICANN '91 in Espoo, Finland), Amsterdam: Elsevier Science Publishers 1991, S. 397-402.

[MARTINETZ und SCHULTEN, 1994] Martinetz, T. M.; Schulten, K. J.: *Topology representing networks.* Neural Networks, vol. 7 (1994), S. 507-522.

[MOSTELLER und TUKEY, 1968] Mosteller, F.; Tukey, J.W.: *Data analysis, including statistics*. In: Handbook of Social Psychology, Vol.2. Hrsg. von G. Lindzey und E. Aronson. Addison-Wesley 1968.

[PAETZ ET AL., 2000] Paetz, J.; Hamker, F. H.; Thöne, S.: *About the Analysis of Septic Shock Patient Data.* In: Proceedings of the First International Symposium of Medical Data Analysis ISMDA 2000, Lecture Notes in Computer Science, vol. 1933, S. 130-137, Springer-Verlag Heidelberg 2000.

[TAGSCHERER ET AL., 1999] Tagscherer, M., Kindermann, L., Lewandowski, A., Protzel, P.: *Overcome neural limitations for real world applications by providing confidence values for network predictions*. Proceedings of the Sixth International Conference on Neural Information Processing (ICONIP'99), Perth, Australia 1999.

[WADE ET AL., 1998] Wade, S.; Büssow, M.; Hanisch, E.: *Epidemiologie von SIRS, Sepsis und septischem Schock bei chirurgischen Intensivpatienten*. Der Chirurg, vol. 69 (1998), S. 648-655.

[WAHBA und WOLD, 1975] Wahba, G.; Wold, S.: *A completely automatic french curve: Fitting spline functions by cross-validation*. Communications in Statistics, vol. 4 (1975), S. 1-17.

[WALTHER und NÄGLER, 1987] Walther, H.; Nägler, G.: *Graphen Algorithmen Programme*. Leipzig: Fachbuchverlag 1987.