Die Prädiktion des klinischen Funktionsniveaus mit Hilfe von neuronalen Netzen

Julia Maria Eder

München 2022

Aus der Klinik und Poliklinik für Psychiatrie und Psychotherapie
Klinikum der Ludwigs-Maximilians-Universität München

Vorstand: Prof. Dr. Peter Falkai

Die Prädiktion des klinischen Funktionsniveaus mit Hilfe von neuronalen Netzen

Julia Maria Eder

Dissertation
zum Erwerb des Doktorgrades
an der Medizinischen Fakultät der
Ludwig–Maximilians–Universität
München

vorgelegt von
Julia Maria Eder
aus Rotthalmünster

2022

Mit Genehmigung der Medizinischen Fakultät
der Universität München

Berichterstatter :	Prof. Dr. Nikolaos Koutsouleris
Mitberichterstatter :	Prof. Dr. Andreas Straube PD Dr. Daniela Hartmann PD Dr. Markus Pfirrmann
Mitbetreung durch :	Dr. David Popovic
Dekan :	Prof. Dr. med. Thomas Gudermann

Tag der mündlichen Prüfung : 17.11.2022

Inhaltsverzeichnis

Abkürzungsverzeichnis

APS	Abgeschwächte psychotische Symptome eng: attenuated psychotic symptoms
BAC	Die balancierte Genauigeit. Ein Maß zur Evaluation von Prognosemodellen eng: Balanced Accracy
BS	Basissymptome
BDI	Becks Depression Inventory II
BLIPS	kurze, limitierte, intermittierende psychotische Symptome eng: brief limited intermittend psychotic symptoms
CGI	Skala für den klinischen Gesamteindruck eng: clinical global impression
CHR	Probanden die Risikokriterien erfüllen und somit ein erhöhtes Psychoserisko haben; eng: clinical high risc
COGDIS	Spezifische Basissymptomkriterien, die vor allem kognitive Beeinträchtigung bewerten eng: cognitive disturbances
FG	Freiheitsgrad
GAF	Skala mit der die Funktionsfähigkeit beurteilt wird eng: general assesment of functioning
GAF D/I	Skala mit der die Funktionsfähigkeit vor allem mit dem Augenmerk auf die Einschränkungen und Einschränkungen beurteilt; eng: general assesment of functioning disability and impairment
GAF S	Skala mit der die Fuktionssfähigkeit anhand der Symptomschwere bewertet wird; eng: general assesment of functioning symptoms
GF	General functioning
GF R	General functioning: role
GF S	General functioning: social
GRDS	Eng: genetic risk and deterioration syndrome
HC	Gesunde Kontrollprobanden; eng: healty control
ICC	Interraterkorrelationskoeffizient eng: interrater correlation coefficient

KRT Kognitive Remediationstherapie

LSTM Eine spezielle Architektur eines rekurrenten Neuronalen Netzes eng: long short term memory

MRT Magnetresonanztomographie

NLR Eng: negative likelihood ratio

NN Neuronale Netze

NNT Die Anzahl an nötigen Menschen, die mit einer Intervention behandelt werden müssen, um einen positiven Ausgang detektieren zu können eng: Number Needed to treat

NPW Negativ prädiktiver Wert

PCA Hauptkomponentenanalyse eng: principal component analysis

PANSS Eng: positive and negative sydrome scale

PLR Eng: positive likelihood ratio

PPW Positiv prädiktiver Wert

RMSE Eng: rooted mean square error

ROD Eng: recent onset depression

ROP Eng: recent onset psychosis

SGD Stochastischer Gradientenabstieg eng: stochastic gradient descent

XOR Exklusives Oder

WR Wilcoxonrangsummentest

Abbildungsverzeichnis

Tabellenverzeichnis

Zusammenfassung

Durch eine Vorhersage des Funktionsniveaus können Patientinnen und Patienten vorzeitig erkannt werden, deren psychopathologischer Zustand sich im weiteren klinischen Verlauf verschlechtern wird. Modelle, die dazu publiziert wurden sind äußerst kompliziert, beinhalten MRT Daten und genetische Analysen, sowie aufwändige klinische Testverfahren, die speziell ausgebildetes Personal benötigen.
Im Rahmen dieser Arbeit konnte gezeigt werden, dass das Funktionsniveau von Patienten mit Hilfe von LSTM Netzen vorhergesagt werden kann. Es waren nur Daten aus Fragebögen nötig. Dabei ist herauszustellen, dass vor allem das soziale Funktionsniveau, welches durch den GF S gemessen wurde, die beste Prädiktionsgüte lieferte. Die aktuellen Ergebnisse übertreffen die Vorhersagegenauigkeit ähnlicher, bereits publizierter Modelle. Die aktuelle Arbeit kann neue Perspektiven für eine integrative und evidenzbasierte Medizin bieten.
Die Güte der einzelnen Prädiktion für die jeweiligen Subgruppen sollte in weiteren Studien mit Daten, welche nicht aus dem PRONIA Datensatz stammen, repliziert werden.

Kapitel 1

Einleitung

1.1 Schizophrene Psychosen

Die Wahrscheinlichkeit im Laufe des Lebens an einer schizophrenen Psychose zu erkranken liegt bei dem durchschnittlichen Menschen, wenn soziobiographische oder genetische Faktoren außer Acht gelassen werden, bei etwa 1% [1]. Das klassische Prädilektionsalter beläuft sich bei Männern auf 21 Jahre, während dies bei Frauen etwa fünf Jahre später anzusiedeln ist. Das Risiko, an einer Schizophrenie zu erkranken, setzt sich aus einer biologischen Veranlagung durch multigenetische Einflussfaktoren, die etwa zu einer Veränderungen im Neurotransmittersystem führen können, perinatale Risikofaktoren und Komplikationen während der Geburt, sowie insbesondere Cannabismissbrauch in der Adoleszenz und psychosozialen Risikofaktoren zusammen [2]. Die Diagnose einer Schizophrenie ist sowohl für die Betroffenen als auch für ihre Angehörigen niederschmetternd. Die Patienten sind meist arbeitslos und können ihr prä-psychotisches Potenzial nicht mehr ausschöpfen [3]. Nach chronischen Langzeitverläufen, konnte post mortem in obduzierten Gehirnen Erkrankter eine Erweiterung der Ventrikel, sowie der Abbau grauer und weißer Hirnsubstanz festgestellt werden [4].

1.1.1 Symptomatik und klinische Subtypen

Die diagnostische Gruppe der schizophrenen Psychosen ist sehr heterogen und kann in zahlreiche klinische Subtypen kategorisiert werden. Auf der symptomatischen Ebene lassen sich Positiv- und Negativsymptome unterscheiden. Studien weisen darauf hin, dass diese auch unterschiedlichen physiologischen Prozessen unterliegen [5].

Positivsymptome

Als Positivsymptome werden im Rahmen einer Psychose die Elemente der Erkrankung beschrieben, die im Vergleich zur Norm ein übersteigertes Erleben aufweisen [3, 6]

1. **Wahnvorstellungen**
   Wahn entsteht auf dem Boden eines veränderten Erlebens und führt beim Betroffenen zu einer Fehlbeurteilung der Realität. Nach Jaspers müssen drei Wahnkriterien erfüllt sein. Es muss mit subjektiver Gewissheit, trotz Vorhandensein von Widersprüchen, an der Wahnvorstellung festgehalten werden [7]. Der Patient ist auch bei Gegenargumenten nicht von seiner Haltung abzubringen. Der Wahninhalt ist unplausibel und kulturell nicht angemessen [8].

2. **Halluzinationen und Sinnestäuschungen**
   Ein Patient kann mittels aller seiner Sinnesorgane halluzinieren. Bei einer Illusion liegt anders als bei einer Halluzination eine Reizquelle vor, die fälschlicherweise verkannt wird [7].

3. **Ich -Störungen**
   Wenn es zu einer Beeinträchtigung der Ich-Umwelt-Grenze kommt, spricht man von einer Ich-Störung. Beispielsweise wird von einer Ich-Störung gesprochen, wenn aus Sicht des Patienten körperliche Vorgänge, wie Denken, Fühlen und Handeln von etwas Äußerlichem gesteuert erscheinen [7].

Negativsymptome

Negativsymptome beschreiben einen Zustand, der ein Defizit im Vergleich zur erlebten Norm aufweist [6, 9]. Nicht nur auf der symptomatischen Ebene zeichnen sich Negativsymptome von den Positivsymptomen ab, auch in der Bildgebung unterscheiden sich Negativ- und Positivsymptome. Negativsymptome werden mit einer sogenannten „*Hypofrontalität*" assoziiert [10].

1. **Affektverflachung**
   Diese zeigt sich durch einen Mangel an emotionaler Schwingungsfähigkeit, sowie herabgesetzte Gefühlsempfindungen. Betroffene wirken im Kontakt gleichgültig und indifferent [11].

2. **Apathie**
   Dieser Begriff beschreibt einen Zustand der Teilnahmslosigkeit. Apathie ist stärker mit dem funktionellen Ergebnis assoziiert als andere Symptome, und ist unabhängig von anderen Negativsymptomen mit Funktionsniveau der Patienten assoziiert [12].

3. **Anhedonie**
   Ist ein Ausdruck, der beschreibt, dass die Fähigkeit Freude zu empfinden reduziert ist. Harvey et al. fanden heraus, dass der Schweregrad der Anhedonie negativ mit dem Volumen des Nucleus caudatus und des ventralen Striatums korreliert [13].

4. **Alogie**
   Die Alogie, auch Wortarmut, ist ein Negativsymptom, welches unabhängig von der Intelligenz des Betroffenen auftreten kann [14].

5. **Sozialer Rückzug**
   Der soziale Rückzug oder auch die sogenannte Asozialität, konnte in einem Tiermodell auf die erschwerte soziale Interaktion bei einer Psychose zurückgeführt werden. Hierbei zeigen sich auch Überlappungen zum Autismus [15].

6. **Aufmerksamkeitsstörungen**
   Als besonders belastend werden kognitive Einschränkungen und Aufmerksamkeitsstörungen beschrieben, die häufig mit Psychosen einhergehen [16].

Paranoid-halluzinatorischer Typ

Im ICD 10 werden verschiedene Subtypen der Schizophrenie beschrieben. Die häufigste Untergruppe stellt die paranoide Schizophrenie dar. Hierbei sind sind Wahnphänomene und Halluzinationen vorherrschend. Veränderungen des Affekts oder Zerfahrenheit dominieren das klinische Bild nicht [17].

Katatoner Typ

Die katatone Schizophrenie zeichnet sich durch Stupor[1] oder Mutismus[2], Erregung, Haltungsstereotypien, sowie Negativismus aus [17]. Etwa 10% aller schizophrenen Patienten werden dem katatonen Typ zugeordnet [19].

Hebephrener Typ

Bei vordergründiger Veränderung der Affekte, also entweder dessen anhaltende Verflachung oder dessen Unangebrachtheit, sowie bei dem Vorhandensein einer Denkstörung, welche sich durch eine unzusammenhängende und zerfahrene Sprache äußert, spricht man vom einer hebephrenen Schizophrenie [17, 20].

[1] Der Stupor (lat: „Erstarrung“) ist charakterisiert durch einen Zustand verringerter Reaktionsfähigkeit auf externe Stimmuli [4].

[2] Der Begriff Mutismus (auch psychogenes Schweigen) leitet sich aus dem Lateinischen ab (lat: mutus = stumm). Es handelt sich um die Unfähigkeit zu sprechen, obwohl keine organische Ursache vorliegt [18].

Residualtyp

Als schizophrenes Residuum wird ein chronisches Stadium der Schizophrenie bezeichnet, welches deutlich durch Negativsymptome, verminderte Aktivität, Passivität, Mangel an Initiative und verminderte Körperpflege gekennzeichnet wird[17].

Schizophrenia simplex

Laut ICD-10 handelt es sich bei der Schizophrenia simplex um eine *„Störung mit schleichender Progredienz von merkwürdigem Verhalten, mit einer Einschränkung gesellschaftliche Anforderungen zu erfüllen und mit Verschlechterung der allgemeinen Leistungsfähigkeit“*. Die Stellung dieser Diagnose wird hingegen jedoch nicht empfohlen[4].

1.1.2 Die Psychose und das Prodromalstadium

Bereits in der ersten Hälfte des 20. Jahrhunderts gab es Überlegungungen zum Prodrom und dessen Existenz [21, 22]. Eine geläufige Definition des Prodroms wurde 1991 formuliert, die dieses als eine „*heterogene Gruppe von Verhaltensweisen, die zeitlich mit dem Beginn der Psychose zusammenhängen*“, bezeichnete [23]. Klosterkötter et al. waren die ersten, die anhand einer prospektiven Studie belegen konnten, dass das Vorhandensein eines Prodroms in 70% der Fälle nach durchschnittlich 4.3 Jahren zu einer Transition in eine Psychose führt [24]. Zuvor hatte es nur retrospektive Studien geben, welche durch das potentielle Vorhandensein eines Selektionsbias[3], eine geringere wissenschaftliche Validität aufwiesen. Fusar-Poli et al. zeigten in einer systematischen Aufarbeitung zahlreicher wissenschaftlicher Arbeiten den typischen Verlauf prodromaler Symptome bis zur manifesten Psychose auf und lieferten ein lineares Modell der Krankheitsentwicklung [26] (siehe Abbildung 1.1). Da ein Prodrom jedoch erst retrospektiv als solches bezeichnet werden kann, veränderte sich die Begrifflichkeit zugunsten des sogenannten „*klinischen Hoch-Risikostatus*“(clinical high risk - CHR), oder „*Ultra-Hoch-Risikostatus*“(ultra-high risk UHR)[4]. Der CHR Status kann als heterogener Zustand betrachtet werden, der in verschiedene Subgruppen unterteilt werden kann:

- **Basissymptome:**
  Gert Huber (*3. Dezember 1921 in Echterdingen;†8. April 2012) prägte den Begriff der „*Basissymptome*“(BS), die als am frühesten erlebten Anzeichen einer Psychose und als der unmittelbare symptomatische Ausdruck des neurobiologischen Korrelates der Erkrankung, gelten [27]. BS sind subtile, subjektiv erlebte, subklinische Störungen von Antrieb, Affekt, Denken, Sprache, (Körper-) Wahrnehmung, motorischem Handeln,

[3] Ein Selektionsbias kann immer dann vorhanden sein, wenn die Studienpopulation nicht aus zufällig ausgewählten Probanden besteht [25].

[4] In der folgenden Arbeit werden wir durchgehend die Abkürzung CHR nutzen, um diesen Zustand zu bezeichnen.

zentralen vegetativen Funktionen und der Stresstoleranz [28]. In einer bekannten Studie von Klosterkötter et al. in der eine Population aus komplexen BS Patienten über beinahe 10 Jahre regelmäßig untersucht wurde, konnte man feststellen, dass 77 der insgesamt 79 Patienten, welche durchgehend Basissymptome aufwiesen, im Mittel nach etwa 4,3 Jahren bei Frauen und nach 6,7 Jahren bei Männern in einen psychotischen Zustand transitionierten [24, 29].

- **Abgeschwächte Psychotische Symptome** (APS):
  APS, die von geschulten Klinikern erkannt werden, sind mit einem Risiko von bis zu 29% verbunden, in den folgenden 2 Jahren eine Psychose zu entwickeln[30]. 2013 wurde im DSM-5 die Diagnose des APS Syndroms eingeführt. Demnach wird dieses an Hand der Positivsymptome des SIPS Fragebogens (Structured Interview for Prodromal Syndromes)[31] diagnostiziert[32]. Kriterien sind beispielsweise unübliche Gedankeninhalte (P1), Paranoide Gedanken(P2), Größenideen(P3), abnorme Wahrnehmungsveränderungen(P4) und Sprachveränderungen(P5). Diese Symptome werden je nach Schweregrad auf einer Skala eingetragen. Solange diese Kriterien noch kein psychotisches Ausmaß erreicht haben, aber auch nicht mehr der Norm entsprechen, liegt ein APS Syndrom vor [31].

- **Kurze intermittierende Psychotische Symptome** (brief limited intermittent psychotic symptoms - BLIPS):
  Das BLIPS erfasst Menschen mit flüchtigen psychotischen Erfahrungen, die innerhalb einer Woche, ohne den Einsatz von Antipsychotika spontan remittierten [33]. Bei dem Vorhandensein von BLIPS, transitionieren nach 2 Jahren 30,3% der Patienten und nach 60 Monaten, also nach 5 Jahren, 54,3% der Patienten [34].

- **Genetisches Risiko und Leistungsknick**(genetic risk and deterioration syndrome - GRDS) Unter genetischem Risiko versteht man einen Angehörigen ersten Grades zu besitzen, der an einer Erkrankung aus dem schizophrenen Formenkreis leidet. Wer an einer schizotypen Persönlichkeitsstörung leidet, ist gemäß der Literatur auch dem GRDS zugehörig [35]. Unter dem Leistungsknickkriterium ('deterioration') versteht man, dass die Funktionsfähigkeit im Alltag signifikant einbricht. Dies kann sich etwa durch einen Jobverlust oder durch eine deutliche Verschlechterung der Schul- oder Arbeitsleistung äußern, die sich innerhalb der letzten zwölf Monate ereignete[36]. Gemessen wird diese Verschlechterung durch die GAF Skala (siehe 1.2.2 und 3.3.1).

Die Existenz eines CHR Stadiums als solches wird kontrovers diskutiert. Ein „typischer klinischer Verlauf" wird von bestimmten Forschungsgruppen angenommen, während andere diese Idee ablehnen. Van Os et al. behaupten, dass der Prozess des Übergangs selbst fragwürdig ist, und er nennt viele Gründe, warum Studien an sogenannten klinischen Hochrisikopatienten (CHR), kritisch ausgewertet werden müssen [37]. Aus den Ergebnissen der

ABC Studie schlussfolgerten Häfner et al, dass sich beim Auftreten eines Prodroms ohne Positivsymptome nicht vorhersagen lassen könne, ob der Patient später an einer Erkrankung des schizophrenen Formenkreises oder an einer Depression leiden wird. Daraus schlussfolgerte die Forschungsgruppe, dass psychotische Erkrankungen und schwere affektive Störungen keine eigenständigen Krankheiten sind, sondern wahrscheinlich verschiedene Stadien einer ähnlichen Psychopathologie, die durch verschiedene Grade von Hirnfunktionsstörungen hervorgerufen werden [38, 39]. Dennoch herrscht ein breiter Konsens in der Wissenschaft, dass die Schizophrenie als schwerwiegende psychiatrische Erkrankung, nicht nur die Lebensqualität, sondern auch das allgemeine Funktionsniveau maßgeblich einschränkt [40]. Diese Veränderung des Funktionsniveaus über die Zeit wollen wir vorhersagen.
Bei depressiven Patienten werden sehr unterschiedliche Krankheitsverläufe geschildert. Der Ausbruch einer Depression erfolgt in der Regel allmählich, kann aber auch abrupt sein. Bei den meisten Patienten ist der Krankheitsverlauf episodisch, die Krankheitsverläufe sind sehr unterschiedlich. Demnach ist die Dauer der Episoden, sowie die Anzahl der Episoden im Laufe des Lebens und deren Muster, in dem sie auftreten, äußerst variabel [41]. Patienten, die die Remissionskriterien einer Depression erfüllten, werden auch auf dem GAF höher eingestuft, was auf eine bessere Funktionsfähigkeit hinweist[42]. Bei vielen Therapiestudien, wird unter anderem der GAF herangezogen, um den Therapieerfolg zu beurteilen[43, 44]. Der intiale GAF scheint im Rahmen von Depressionen den Verlauf deutlich zu beeinflussen [45]. Dennoch gibt es auch wissenschaftliche Arbeiten, die die Prädiktionsfähigkeit des GAFs als widerlegt sehen [46].

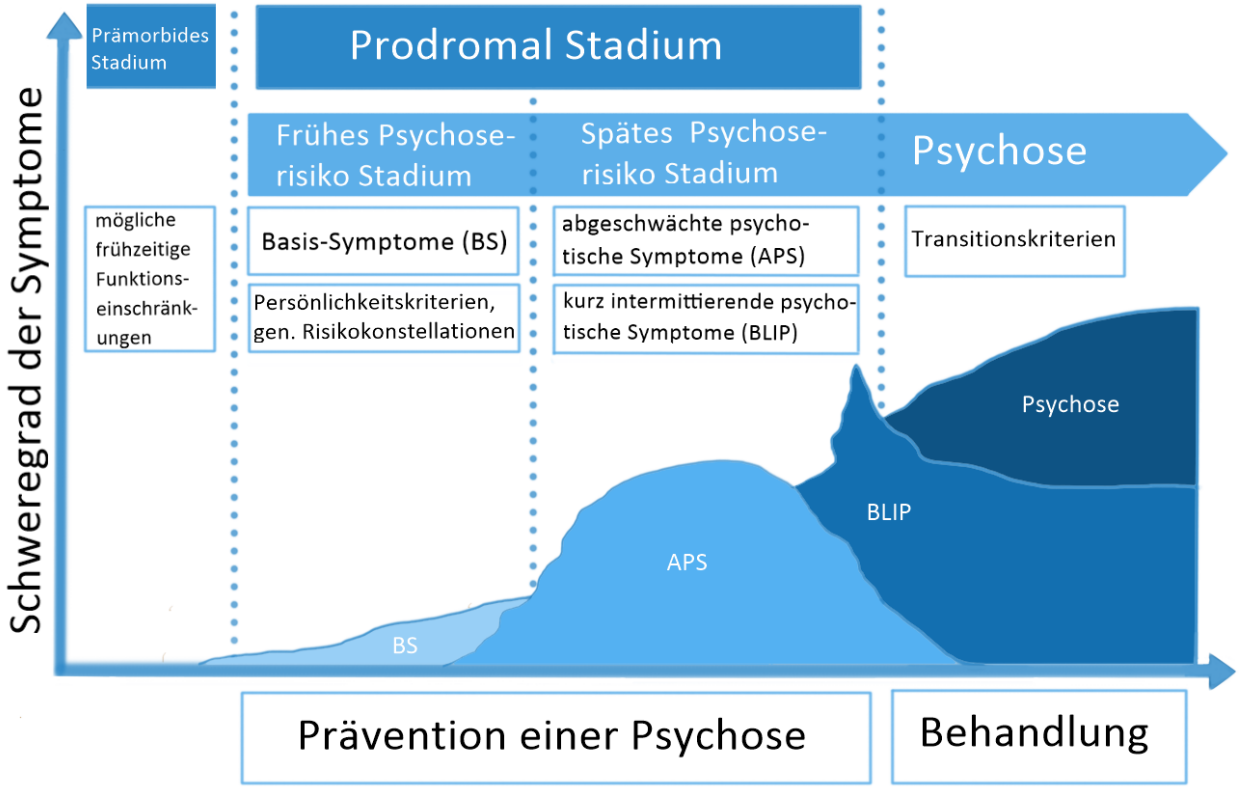

Abbildung 1.1: Darstellung des typischen Verlaufes eines Prodroms nach Paolo Fusar-Poli et al.[26] mit freundlicher Genehmigung des Verlages. Mit Hilfe der vorhandenen, wissenschaftlichen Daten konnte gezeigt werden, dass die Dauer bis zur Transition je nach CHR - Gruppe unterschiedlich ist (BLIP>APS>GRDS/BS) [47].

1.2 Die Vorhersage einer klinischen Prognose

Die Prognose des klinischen Zustandes eines Patienten ist zusammen mit der Diagnostik und Therapie ein essentieller Teil der ärztlichen Arbeit. Das Wissen um die Prognose eines Patienten vereinfacht nicht nur die Therapie, sondern wird auch von den Angehörigen und dem Patienten selbst erwartet und häufig aktiv gefordert.
Vor allem bei sehr einschneidenden Diagnosen, wie etwa Erkrankungen aus dem schizophrenen Formenkreis, ist eine akkurate Prognose wichtig.
Laut Literatur vermeiden Ärzte die Abgabe einer Prognose, unter anderem, weil dies sehr komplex und emotional belastend sein kann, aber auch weil eine Fehleinschätzung gravierende Auswirkungen auf den Patienten, sowie dessen Angehörigen, haben kann. Diese Vermeidung spiegelt sich sogar in der Anzahl von Publikationen wider, die sich der Pro-

gnose widmen. Lediglich 3% aller medizinischen Veröffentlichungen handeln davon [48]. Folglich gibt es häufig auch nur wenige Daten, die Kliniker zur Prognosefindung heranziehen können. Dies erschwert medizinische Prädiktionen oder macht sie unmöglich. Um diese Abschätzung für Ärzte zu vereinfachen, gibt es Publikationen, die Richtlinien und Regeln zur vereinfachten Prognoseabschätzung definieren[49, 50, 51, 52].

1.2.1 Prognoseabschätzung bei psychiatrischen Patienten

Psychosen machen 6,3 % der globalen Krankheitslast aus und verursachen allein in Europa Kosten von etwa 207 Milliarden Euro pro Jahr[53]. Damit sind sie die teuersten neurologischen Erkrankungen [54] und übertreffen die Kosten von Herz-Kreislauf-Erkrankungen (169 Milliarden Euro)[55]. Diese immense sozioökonomische Belastung begründet sich daraus, dass affektive und nicht-affektive psychotische Erkrankungen, häufig im Jugend- und frühen Erwachsenenalter beginnen und zu langfristigen Einschränkungen der Funktionalität und Lebensqualität führen. Beide Faktoren führen zu dauerhafter sozialer und beruflicher Ausgrenzung und tragen zu einer 8-20-fach höheren Suizidrate der Betroffenen bei [56, 57]. Nach der ersten Episode einer psychiatrischen Erkrankung bergen die nachfolgenden Krankheitsverläufe die größte Variabilität. Chronisch schizophrene Langzeitpatienten ohne adäquate Therapie haben meist ein kontinuierlich, schlechtes Niveau der Funktionsfähigkeit im Alltag. Die Arbeitslosigkeit bei diesem Kollektiv ist mit 80–90% sehr hoch[58, 59]. Folglich brächte die Auswertung der klinischen Daten von Ersterkrankten Patienten und eine Prädiktion deren Verläufe zahlreiche neue Erkenntnisse [52]. Therapiemaßnahmen, wie etwa die kognitive Remediationstherapie (KRT) zeigen einen bessere Wirkung, falls diese kurz nach der ersten psychotischen Episode durchgeführt wurden. Also während der Phase größter klinischer Variabilität [60]. Bei einer Anzahl von notwendigen Behandlungen (NNT) von neun können durch KRT, Transitionen zur ersten psychotischen Episode sogar ganz verhindert werden, wenn Patienten die Clinical High Risk (CHR) Kriterien (siehe Material und Methoden 3.2.1) erfüllen [61]. Daher ist es wichtig bereits früh eine Einschätzung abgeben zu können, welche Patienten ein erhöhtes Risiko haben, ihren Alltag auch langfristig nicht mehr adäquat meistern zu können. Diese Patienten benötigen mehr Zuwendung[62, 63]. Vor einer Transition in eine Psychose wurde bei CHR Patienten eine signifikant verringerte Verarbeitungsgeschwindigkeit, sowie Probleme beim Erlernen von Worten festgestellt, während CHR Patienten, die diese Einschränkungen nicht aufwiesen unwahrscheinlicher in einen psychotischen Zustand transitionierten [64]. Konkrete Anwendungen zur Vorhersage der Prognose gib es in der Psychiatrie nur vereinzelt und sind rein statischer Natur[65, 66]. Vor allem widmen sich diese allein dem CHR Status und machen keine Vorhersage zu erstmals depressiven und bereits psychotischen Patienten.
Das ursprüngliche Schicksal einer so genannten „*Dementia praecox*“[67], also einer vorzeitigen Demenz, wie Kraepelin die Verläufe der damals bekannten „*Hebephrenie*“ und „*Katatonie*“ zusammenfasste und popularisierte, wurde später von Bleuler, der den Begriff „*Schizo-*

phenie“ einführte, revidiert. Er beschrieb von der Dementia praecox abweichende Verläufe und beobachtete, dass sich die Zustände mancher Patientin im Laufe der Zeit verbesserten. Zu einer vollständigen „*restitutio ad integrum*“[5] kam es in seinen Ausführungen jedoch nicht [68, 69].

1.2.2 Abbildung der Funktionsfähigkeit mittels GAF / GF

Jones et al. führten im Jahr 1995 die Global Assessment of Funktioning Skala (GAF - Skala) ein, eine Einteilung, die das Funktionsniveau psychiatrischer Patienten messen soll (siehe Material und Methoden 3.3.1)[70]. Zahlreiche Studien haben seitdem den GAF verwendet, um das Outcome von psychiatrischen Patienten vorherzusagen[71]. Studien hatten gezeigt, dass der GAF als Wert an sich einen mangelnden prädiktiven Wert aufweist [46]. Zudem ist nachweislich die Interrater-Variabilität des GAFs in der Wissenschaft gut, allerdings ist dies nicht im klinischen Kontext der Fall [72]. Wir nutzen daher im Rahmen der gezeigten Experimente den GAF, aber auch dessen kategoriale Weiterentwicklung den GF (siehe Material und Methoden 3.3.2)[73].

1.3 Neuronale Netze

Vor allem in den letzten Jahren sind Neuronale Netze (NN) in das Licht der Öffentlichkeit gerückt, da diese eine sehr vielseitig explorative, statistische Methode darstellen.[74] Der erste, der ein Neuron auf mathematische Art und Weise darstellte, waren McCulloch et al. und markiert so die Geburtsstunde der neuronalen Netze[75]. Wenige Jahre später, 1949, proklamierte der Psychologe, Donald O. Hebb, erstmals ein Modell zur synaptischen Plastizität, sowie eine Theorie zum Lernverhalten biologischer Neurone. Demnach ist die klassische Konditionierung eine Ursache des jeweiligen Verhaltens der Neurone eines Individuums [76]. Die Hebb'sche Theorie ist theoretisches Fundament für die Annahme, dass Engrame mit Hilfe von Neuronalen Netzen dargestellt werden können[77]. Die erste praktische Anwendung von Neuronalen Netzen wurde fast zehn Jahre später von Rosenblatt et al. eingeführt. Mit Hilfe eines Perzeptron[6] Netzwerkes, konnte gezeigt werden, dass das Neuronale Netz die Fähigkeit besaß, Muster zu erkennen [78]. Marvin Minsky und Seymour Papert bewiesen 1969, dass ein einlagiges Neuronales Netz aus Perzeptronen den XOR-Operator[7] nicht darstellen kann[80]. Diese Entdeckung läutete unter anderem den ersten KI-Winter

[5] Der Begriff Restutio ad integrum kommt aus dem Lateinischen und bedeutet „Wiederherstellung“.

[6] Ein Perzeptron ist ein vereinfachtes Neuronenmodell, welches in der Grundversion aus einem einzelnen künstlichen Neuron mit Gewichtungen und einer binären Schwellwertsfunktion besteht (siehe 3.6.1).

[7] Der XOR - Operator (exklusives Oder; Kontravalenz) ist ein Junktor aus der klassischen Logik. Das Ergebnis eines XOR-Operators ist genau dann wahr, wenn beide durch das XOR verbundene Aussagen unterschiedliche Wahrheitswerte aufweisen [79].

ein. Neue Erkenntnisse und eine Zunahme der Rechenleistung ermöglichten die Wiedergeburt der Neuronalen Netze [81]. Mit der Erfindung des Fehlerrückführungsalgorithmus (Backpropagation- Algorithm) war es erstmals möglich Neuronale Netze (NN) zu trainieren, die mehrere verborgene Schichten (hidden layers) besaßen. In der Folge konnte mit mehreren Perzeptronen den XOR-Operator dargestellt werden, da hierfür mindestens zwei Schichten nötig sind [82]. Mit einem Neuronalen Netz, das aus zwei Schichten mit einer sigmoiden Aktivierungsfunktion und einer linearen Aktivierungsfunktion (siehe 3.6.1) in der Aussganssignalschicht besteht, können alle Funktionen angenähert werden - vorausgesetzt die Anzahl der versteckten Neurone (hidden units) ist groß genug. Demnach könnte ein Neuronales Netz bei ausreichender und qualitativ guter Datenzufuhr jeden Zusammenhang erlernen [81, 83, 84].

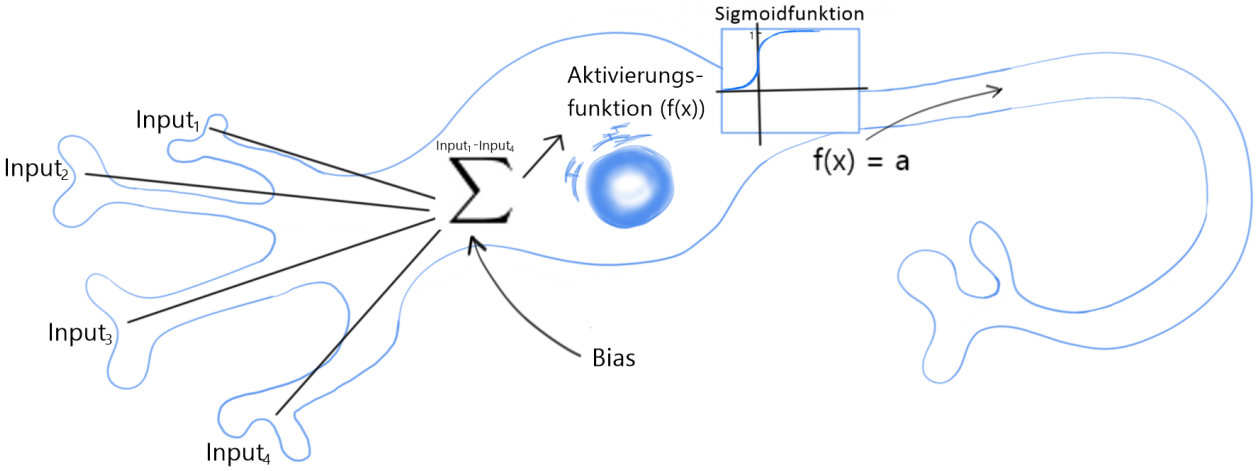

Abbildung 1.2: Darstellung eines Neurons in einem Neuronalen Netz mit mehreren Inputs. Die Inputvektoren werden elementweise addiert und anschließend in die Aktivierungsfunktion f(x) eingespeißt. Ähnlich eines echten Neurons wird durch die sigmoidale Aktivierungsfunktion (welche in diesem Beispiel f(x) repräsentiert) eine annähernde 1 oder 0 Reaktion ausgelöst.

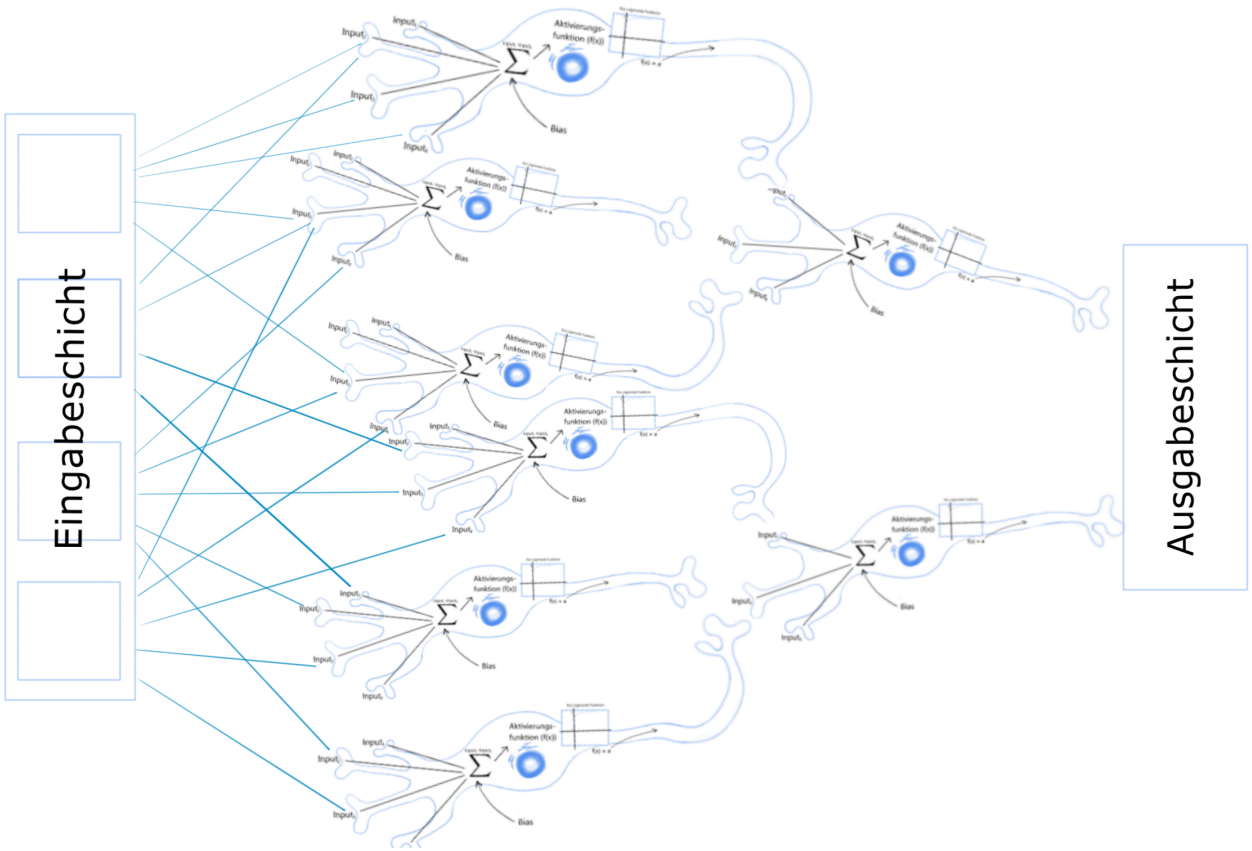

Abbildung 1.3: Abstrakte Darstellung eines Neuronalen Netz mit mehreren Inputs. Die einzelnen Neurone der zwei „hidden layer" entsprechen der Darstellung des Neurons aus Abbildung 1.2.

Obwohl es viele verschiedene NN Architekturen gibt, ist der Aufbau eines einzelnen Neurons stets gleich. In Abbildung 1.2 ist ein einzelnen Neuron mit mehreren Inputs zu sehen. In Abbildung 1.3 kann beobachtet werden, wie ein solches artifizielle Neuron in einem neuronalen Netz verknüpft sein kann.
Ein Neuronales Netz ist ein sogenannter supervidierter Lernalgorithmus. Das bedeutet, dass wir das optimale Ergebnis, welches das Netz prädizieren sollte, kennen. Durch eine zentrale Fehlerfunktion (siehe 3.6.2 Formel 3.6) wird berechnet inwieweit das neuronale Netz ‚*falsch*' oder ‚*richtig*' lag. Anhand dieser Information optimiert sich das Neuronale Netz durch Fehlerrückführungsalgorithmen wie etwa den SGD selbst (siehe 3.6.3).

1.4 Rekurrente Neuronale Netze

Rekurrente Neuronale Netze sind eine Weiterentwicklung der Neuronalen Netze, die einen Rückfluss aufweisen. Diese Netzarchitekturen konnten bereits erfolgreich für Spracherkennung und zeitliche Vorhersagen verwendet werden. Typischerweise werden rekurrente Netze für sequentielle Daten verwendet [85]. Durch die Erweiterung des Fehlerrückführungsalgorithmus (siehe 3.6.3) über die zeitliche Dimension (Backpropagation through time), können auch diese Netze zuverlässig optimiert werden werden[86, 87]. Ein Problem bei Rekurrenten Neuronalen Netze (RNN) ist der so genannte verschwindende Gradienteneffekt (vanishing gradient). Das bedeutet, dass während des Lernprozesses die Gradienten im Neuronalen Netz sehr kleine Werte annehmen können (vanishing gradient). Wenn etwa ein Ereignis mit einer sehr späten Latenz auftritt, kann dieses kaum von einem RNN prädiziert werden [88]. Umgekehrt wird auch ein explodierender Gradienteneffekt (exploding gradient) beschrieben, bei den die Gradienten sehr große Werte annehmen, sodass die nachgeschalteten Neurone immer aktiviert werden. Beide Effekte führen zu ungenaueren Prädiktionen [89].

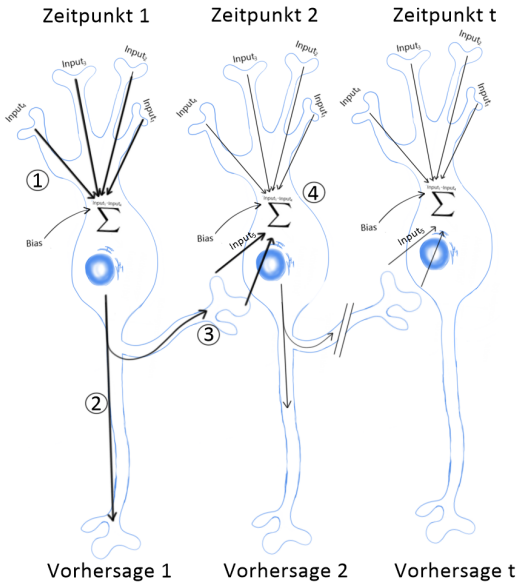

Abbildung 1.4: Die Abbildung zeigt ein einlagiges RNN. Zunächst werden alle Eingaben zum Zeitpunkt 1 mit dem Bias verrechnet (1). Nach Durchlauf durch die Aktivierungsfunktion wird eine Vorhersage zum Zeitpunkt 1 ausgegeben (2). Durch die zusätzliche, eingezeichnete Verbindung (3) wird veranschaulicht, dass die Prädiktionen der vorhergehenden Zeitpunkte in die darauffolgenden Vorhersagen beeinflussen. Zum nächsten Zeitpunkt 2 werden die Inputs mit der Vorhersage zum Zeitpunkt 1 verrechnet (4).

1.4.1 Long Shortterm Memory – LSTM

Long Shortterm Memory Netze sind eine äußerst beliebte Variante der RNN Netze, die erstmal 1997 von Hochreiter und Schmidhuber et al. eingeführt wurden [90, 91]. LSTMs stellen eine Weiterentwicklung der klassischen RNN dar und es konnte gezeigt werden, dass diese eine überwiegende Mehrzahl von Problemen besser und zuverlässiger lösen können[92]. Eine entscheidende Veränderung war es, dass die zusätzlich eingebauten Schranken (siehe Abbildung 1.5), ein LSTM Netz stärker auf den Kontext der sequentiellen Daten konditionieren. Indem das Gewicht dieser selbstständigen Schleifen über weitere Neurone gezielt beschränkt werden kann, ist die dynamische Integration über den Zeitrahmen möglich und unterschiedlich lange Zeitketten können verarbeitet werden [93, 94]. Dies führt auch dazu, dass bei LSTM Netzen das Phänomen der explodierenden oder der verschwindenden Gradienten kaum auftritt [88].

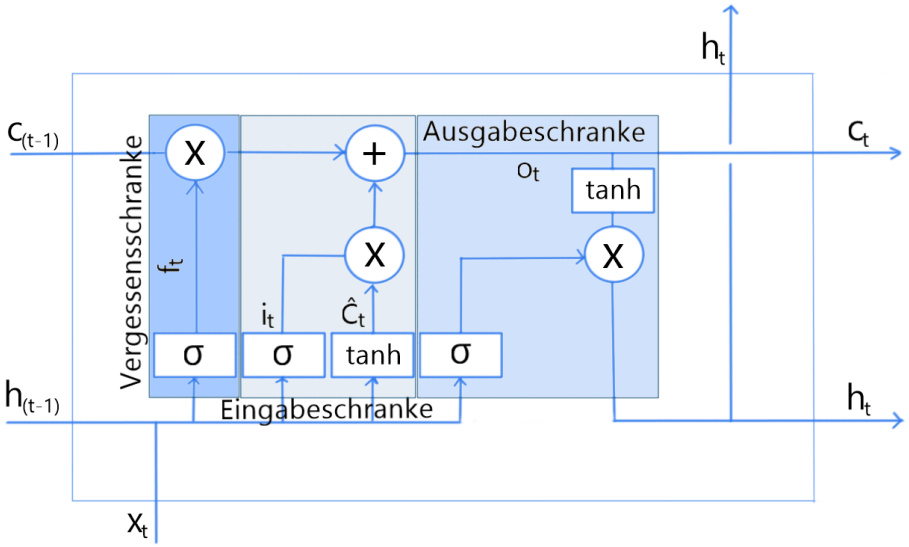

Abbildung 1.5: Darstellung eines einzelnen LSTM Neurons. Innerhalb des Kastens sind die beschriebenen Schranken zu erkennen. Von links nach rechts sind die Vergessenschranke (f_t), die Eingabeschranke ($i_t, \hat{C}_t$) und die Ausgabeschranke (o_t) abgebildet. Bei dem Zeichen ‚+', sowie ‚x' handelt es sich jeweils um die elementweise Addition oder Multiplikation. Das Symbol σ steht für die sigmoidale Aktivierungsfunktion. Die Grafik wurde von [95] adaptiert. Zum Besseren Verständnis der intern ablaufenden Berechnungen siehe Material und Methoden 3.6.4.

1.5 Schwierigkeiten beim Trainieren Neuronaler Netze

1.5.1 Lernrate (Learning Rate)

Die Lernrate bestimmt in wie großen Schritten der quadratische Fehler (SE) der Fehlerfunktion (siehe 3.6.2 Formel 3.6) minimiert wird und beeinflusst so die Geschwindigkeit in der die Konvergenz des Algorithmus erreicht wird. Die Lernrate ist somit einer der wichtigsten Parameter, den es gilt im Rahmen der Konstruktion eines Neuronalen Netzes zu optimieren[94]. Ist die Lernrate zu groß, so kann es sein, dass der Gradientenabstieg zum Optimum scheitert, da die einzelnen Sprünge pro Iteration zu groß sind. Man schießt bildlich gesprochen über das Ziel hinaus und überspringt das Optimum (siehe Abbildung 1.6). Ist die Lernrate zu gering, wird das globale Minimum zwar zuverlässig erreicht, allerdings nach sehr langer Zeit, was bei tiefen Neuronalen Netzen ein Problem darstellen kann.

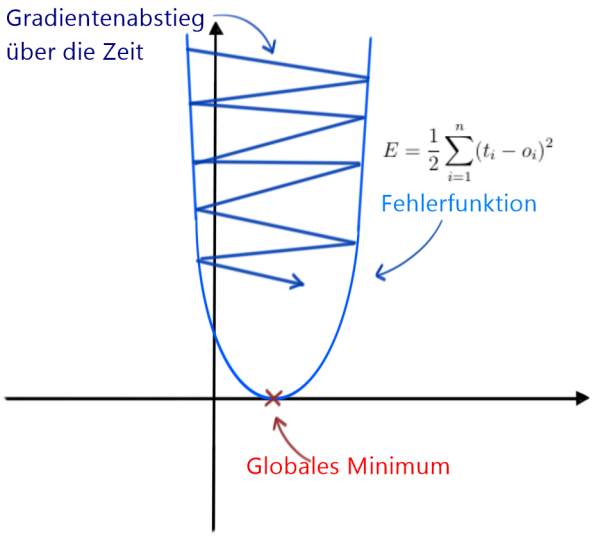

Abbildung 1.6: Die Abbildung zeigt, dass bei einer zu groß gewählten Lernrate der SE immer wieder über das Globale Minimum hinausschießt, welches am Ende nicht erreicht wird, was zu einer schlechteren Performance des NN führen wird.

Daher ist die Lernrate einer der kritischen Hyperparameter, die bei einem neuronalen Netz sorgfältig bestimmt werden müssen. Die nötige Optimierung der Lernrate kann durch die Nutzung von Algorithmen die wie etwa dem RMSProp Algorithmus, der ein Momentum verwendet, etwas vereinfacht werden. Dennoch muss auch hier eine Hyperparameteroptimiertung stattfinden. So können etwa zu kleine Lernschritte auch hier die Performance beeinträchtigen, da sonst die Lernrate zu schnell wachsen könnte [96].

1.5.2 Langzeitabhängigkeiten

Eine weitere Schwierigkeit kann im Speziellen bei besonders tiefen Neuronalen Netzen, aber auch bei langen Rekurrenten Neuronalen Netzen auftreten, da diese ebenso einen tiefen mathematischen Graphen darstellen.
Dies führt dazu, dass wiederholt die gleichen mathematischen Operationen durchgeführt werden. Nehmen wir an ein Graph beinhaltet einen Pfad, der wiederholt eine Matrix W multipliziert, nach t Schritten, ist dies äquivalent zu W^t. Jeder Eigenwert λ, der nicht nahe an einem Wert von 1 liegt, wird entweder explodieren, wenn $\lambda > 1$ (Exploding Gradient) oder verschwinden, wenn $\lambda < 1$ (Vanishing Gradient). Verschwindende Gradienten (vanishing gradients) machen es sehr schwierig, den Gradienten korrekt an das an das globale Minimum anzunähern. Bildhaft gesprochen wird die Fehlerfunktion (auch Kostenfunktion) so flach, dass eine Veränderung des Gradienten von dem Algorithmus weder bemerkt wird wenn sich dieser von dem Minimum entfernt oder sich diesem annähert. Bei einem explodierenden Gradienten, wird der Lernprozess sehr instabil. Das heißt, dass bereits eine sehr kleine Veränderung des Gradienten zu einer sehr großen Veränderung des Ausgabewertes führen wird[94].Da rekurrente Netze für jeden Zeitschritt auf der gleichen Matrix rechnen, ist dort das Problem verschwindender oder explodierender Gradienten gravierender als bei klassischen (feedforward) NNs. [97] Dementsprechend sind bei Langzeitabhängigkeiten die jeweiligen Gewichte exponentiell kleiner, da mit jedem Schritt erneut Matrizen multipliziert werden müssen. Trotz intensiver Forschung, bleibt das Erlernen von Langzeit Abhängigkeiten ein Problem tiefer und rekurrenter neuronaler Netze [98, 99, 100].

1.6 LSTM Netzwerke in der Psychiatrie

Bisher finden LSTMs in der Psychiatrie nur vereinzelt Anwendung etwa zur Detektion autistischer Kinder mittels der Analyse von Augenbewegungen über die Zeit[101]. Zumeist werden LSTMs verwendet, um automatisierte Textanalysen durchzuführen[102, 74]. Prinzipiell können solche Netzwerke auch für dynamische Vorhersagen verwendet werden. So wurde zum Beispiel, das Verhalten von Verbrauchern bereits mit LSTM Netzwerken prädiziert[103]. Im Rahmen dieser Doktorarbeit werden wir daher die GAF Werte der Patienten dynamisch nach jeweils drei Monaten vorhersagen. Bei erneuten Follow-up Terminen, kann der Algorithmus eine erneute Vorhersage zum Folgezeitpunkt treffen, dabei lernt das LSTM dynamisch aus Abweichungen und trifft mit den bereits präsentierten Angaben eine erneute, verbesserte Prädiktion. Zur Vorhersage des psychiatrischen Funktionszustandes wurden LSTM Netzwerke nach unserem Wissenstand bisher nicht verwendet. Somit wurde im Rahmen dieser Arbeit ein bereits bewährtes mathematisches Modell verwendet, um ein vollständig neuartiges Screeningverfahren zu entwickeln.
Ähnliche Verfahren existieren nach unserem Kenntnisstand bisher nicht.

Kapitel 2

Zielsetzung

Diese Arbeit widmet sich der Vorhersage und Verbesserung der funktionalen Prognose verschiedener Krankheitsbilder der Psychiatrie.
Viele Modelle, die dazu publiziert wurden, sind äußerst kompliziert, beinhalten MRT Daten und genetische Analysen[104, 105], sowie aufwändige klinische Testverfahren, die speziell ausgebildetes Personal benötigen [106]. Das Ziel dieser Arbeit ist es das Funktionsniveau von unterschiedlichen Patienten möglichst einfach und dennoch treffsicher vorherzusagen, um so eine einfache und breit verfügbare Screening Methode zu schaffen. Diese Methodik sollte äußerst dynamisch funktionieren und im drei-Monats-Schritt ihre Prädiktionen anpassen können. Bis zum heutigen Tage gibt es unseres Wissens nur statische Anwendungen und keine Prädiktionsmodelle, die sowohl nach einem Vorstellungstermin, als auch nach mehreren Follow-up Terminen dynamische Prädiktionen des Funktionsniveaus zum nächsten Vorstellungstermin liefern können. Der Vorhersagealgorithmus soll im klinischen und ambulanten Setting Allgemeinmedizinern, Psychotherapeuten und Psychiatern helfen, besser den Krankheitsverlauf ihrer Klienten abschätzen zu können, ohne dabei teure und aufwändige Tests durchführen zu müssen.
Indem man sich auf das Outcome der Patienten und nicht auf die Krankheit konzentriert, kann die Modellierung der klinischen Risikovorhersage zum Eckpfeiler einer wissenschaftlichen und personalisierten Psychiatrie werden [66]. So können Patienten, die zu einen ungünstigen Verlauf tendieren würden, präventiv besser therapiert werden. Die Daten, die im Rahmen dieser Doktorarbeit herangezogen werden, stammen aus der PRONIA Studie [107, 57].

Kapitel 3

Material und Methoden

3.1 Digitale Infrastruktur und Architektur

Alle Daten und der hier vorgestellte Code befinden sich auf dem Hochleistungsrechner des medizinischen Lehrstuhls für Psychiatrie. Der Server befindet sich physikalisch in Großhadern und besteht aus einem HPE StoreEasy 48TB SAS Speichermedium, das mit dem LRZ Sicherungsmanager verbunden ist. Dieser basiert auf auf der IBM Tivoli Storage Manager Software.[108] Dieses generiert täglich um zehn Uhr abends ein Backup. Die Daten werden auf vier HP DL560 Gen9 E5-4667v3 Knoten und einen Dell PowerEdge-R820 gespeichert. Der Hochleistungsrechner des Lehrstuhls soll dazu dienen dem Benutzer ein leicht bedienbares Frontend zu bieten, sowie die Möglichkeit rechenintensive Prozeduren schnell durchführen zu können.
Als Betriebssystem wird Linux (CentOS) verwendet, dabei bildet KDE aktuell das Frontend.
Linux 3.10.0-1062.18.1.el7.x86_64 (Kernelversion)
Auf den Server kann nur über das neurobiologische Subnetz zugegriffen werden, das von Sven Wichert geschaffen wurde. Dies ist zum Beispiel über einen VPN möglich.[109]
Zur Analyse und Bearbeitung der Daten wird Matlab verwendet.
MATLAB Version: 9.5.0.944444 (R2018b)
MATLAB Lizenz Nummer: 708505 [110]

3.2 PRONIA Datensatz

Im Rahmen der PRONIA Studie, wurden Patienten aus 15 Zentren rekrutiert. Die Studie startete am 1. Oktober 2013 und beendete die Rekrutierung Ende Oktober 2018.
Die Studie bestand aus elf Arbeitspaketen. Diese hatten spezielle Zuständigkeiten, wie zum Beispiel „*Bildgebung*" oder „*genetische Daten*".
PRONIA ist eine longitudinale Beobachtungsstudie, in der Patienten, bis 36 Monate nach

dem Einschluss, von ausgebildetem Personal regelmäßig befragt wurden. Die Untersuchungsintervalle finden bis 18 Monate nach Einschluss alle drei Monate statt. Danach wurden zwei weitere Erhebungen durchgeführt, die erst nach jeweils neun Monaten erfolgten.
Insgesamt fanden neun verschiedene Untersuchungen statt, die je nach Größe der jeweiligen Untersuchung benannt wurde. Ein einfaches Intervall in dem nur klinisch, behavoriale Daten erhoben werden, wurde kurz als 'IV' bezeichnet. Die Anzahl der verstrichenen Monate nach Einschluss, werden durch die darauf folgende Zahl angedeutet. Eine große Untersuchung wird mit 'T' abgekürzt und jeweils von null bis zwei chronologisch nummeriert. Üblicherweise laufen die Untersuchungen in folgender Reihenfolge ab: T0, IV3, IV6, T1, IV12, IV15, T2, IV27 und IV36.

3.2.1 Einschuss- und Ausschlusskriterien

Im Rahmen der PRONIA Studie wurden vier Studiengruppen definiert:
Eine gesunde Kontrollgruppe (HC), eine Population mit einer erstmaligen depressiven Episode (ROD), eine Gruppe, die Prodromal-Symptomatik aufweist (CHR), sowie eine Patientengruppe, die das erste Mal an einer Erkrankung des psychotischen Formenkreises litt (ROP). Keiner der Patienten durfte länger als 2 Jahre an der jeweiligen Symptomkonstellation leiden, die zum Einschluss führte.
Die gesunden Kontrollen durften an keinen psychiatrischen Vorerkrankungen leiden.
Diese Arbeit widmete sich allen erkrankten Patienten, also den ROD, CHR and ROP Gruppen. In den folgenden Zeilen finden sich die jeweiligen Subgruppen, sowie eine Auswahl zu Aus- und Einschlusskriterien. Wir verwendeten nur Daten der erkrankten Probanden. Es wurden demnach keine gesunden Kontrollen (HC) für das Training der Neuronalen Netze verwendet. Insgesamt wurden 904 Datensätze aus dem PRONIA Projekt für die Analysen in dieser Dissertation verwendet.

Generelle Einschlusskriterien

Unabhängig von der letztlichen Gruppenzuordnung mussten die Patienten bei Einschluss zwischen 15 und 40 Jahren alt sein. Außerdem mussten sie Einwilligungsfähig sein und über ausreichende Sprachkenntnisse verfügen.

Generelle Ausschlusskriterien

1. IQ unter 70
2. Schwerhörigkeit, die die Teilnahme an der neuro - kognitiven Testung erschwert
3. Aktuelles oder vergangenes Schädelhirntrauma mit Bewusstseinsverlust

4. Aktuelle oder vergangene somatische Störungen des Gehirns, welche die Struktur, sowie das Funktionieren des Gehirn beeinträchtigen.

5. Aktuelle oder vergangene Alkoholabhängigkeit

6. Aktuelle Polytoxikomanie oder während der letzten 6 Monate

7. Bestehende Kontraindikationen für eine MRT Untersuchung

ROD - Recent Onset Depression
Patienten mit einer ersten depressiven Episode

Eine der Studiengruppen in der PRONIA Studie umfasst die Population der Recent-Onset-Depression (ROD) Patienten. Eine Population an Probanden, die erstmals an einer Depression erkrankt sind. Die Dauer der ersten depressiven Episode darf 24 Monate nicht übersteigen. Die Kriterien einer Depression müssen während den letzten drei Monaten erfüllt sein. Rezidivierende Depressive Störungen stellten ein Ausschlusskriterium dar.In den letzten drei Monaten durften diese Patienten keine Anti-Psychotika einnehmen. Auf die gesamte Lebenszeit betrachtet, durften die Patienten keine antipsychotische Medikation für mehr als 30 Tage einnehmen, oder eine Dosis, die über den Grenzwerten der DGPPN S3 Leitlinien liegt[111].

CHR - clinical high risc
Patienten die gefährdet sind eine Psychose zu erleiden

Die CHR Patienten wurden in vier Unterkategorien eingeteilt. Falls die Symptome die Beschreibung einer der Kategorien entsprachen, konnten die Patienten eingeschlossen werden.

1. Basis Symptome - COGDIS [28]

2. Brief Intermittent Psychotic (BLIPS) Symptom Psychose Risiko Syndrom

3. Attenuated Positive Symptoms (APS) Psychose Risiko Syndrom

4. Genetisches Risiko, Leistungsknick und, oder Schizotypische Persönlichkeitsstörung

[112]

ROP - Recent Onset Psychosis
Patienten mit der ersten psychotischen Episode

Um als ROP Patient klassifiziert zu werden, mussten folgende Punkte erfüllt werden:

1. DSM-IV-TR affektive oder nicht affektive psychotische Episode während der letzten drei Monate

2. Beginn der Psychose innerhalb der letzten 24 Monate

3.3 Verwendete Fragebögen

3.3.1 GAF - Global Assessment of Functioning

Der GAF ist eine Metrik, die auch im klinischen Alltag sehr präsent ist. Der Wert zwischen 0 und 100 gibt Auskunft über das Funktionsniveau der jeweiligen Person. Der GAF besitzt eine Achse, die die Einschränkung im Alltag beschreibt (GAF D/I - Disability) und eine Achse mit der die Symptomschwere der Probanden bewertet wird (GAF S - Symptoms)[46]. Die traditionelle GAF-Messung, die auf dem Punktwert von Symptom (GAF S) und Funktionsebene (GAF D/I) basiert, dient als guter globaler Indikator für subjektive Beschwerden und soziale Dysfunktion der Patienten.[113]. Die Interrater-Reliabilität des GAFs wird mit einer Intraklassenkorrelation (ICC) [1] von 0.81 bis 0.85 bei Einschluss von 0.94 nach 6 Monaten und von 0.95 nach 12 Monaten angegeben. Die Interrater-Reliabilität erwies sich bei klinischen Routineerhebungen jedoch als niedrig (ICC Koeffizienten zwischen r =0.39 und 0.59), während sich diese im Rahmen von Studien als sehr gut herausstellte[72]. Der GAF zeigte sich zudem als äußerst reliabel. Bei mehreren Messungen der Symptomschwere, dem sozialen Verhalten, sowie dem jeweiligen GAF, zeigten sich gute Korrelationen mit anderen klinischen Metriken, deren Güte nach weiteren Follow-up Terminen zunahm [116]. Der GAF kann, aufgeteilt in seine Subdomänen GAF D/I und GAF S, im Anhang unter Abschnitt A.1 eingesehen werden.

3.3.2 GF - Global Functioning

Ähnlich wie der GAF, trifft auch der GF eine Aussage über das Funktionsniveau. Der GF ist jedoch eine neuere Metrik und daher noch nicht im klinischen Alltag etabliert. Der GF, der von Cornblatt et al. entwickelt wurde, besteht aus zwei Achsen. Nämlich dem GF Social und dem GF Functioning. Diese Metrik der Funktionsäbschätzung wurde speziell im Hinblick auf die Entdeckung des Prodroms geschaffen [73]. Der ICC zwischen Ratern in acht verschiedenen Zentren belief sich auf Werte zwischen 0.94 und 1 [117]. Der GF findet sich im Anhang unter dem Abschnitt A.2.

[1] Die Intraklassenkorrelation ist ein Verfahren zur Quantifizierung der Übereinstimmung der Testergebnisse (Interrater-Reliabilität) zwischen mehreren Begutachtern (Ratern) Das dazugehörige Maß, der sogenannte Intraklassen-Korrelationskoeffizient[114, 115].

3.4 Datenschutz und -sicherheit in PRONIA

Das Projekt hielt sich strikt an die Good Clinical Practice (GCP)-Richtlinien der Europäischen Kommission (2005/28/EG), die EU-Richtlinie 95/46/EG, den internationalen GCP-Standard CMPM/ICH/139/95, sowie an alle relevanten Regelwerke, um die Vertraulichkeit und den Schutz der Privatsphäre aller gesammelten Informationen vollständig zu gewährleisten. Die Pseudonymisierung aller Daten, die im Rahmen der Projektaktivitäten erfasst, gespeichert und ausgetauscht werden, wurde automatisch durch das zentrale Datenbanksystem ausnahmslos gewährleistet. Zum Zeitpunkt des Screenings wurde ein eindeutiger Identifikator (PSN), der keine Informationen bezüglich Namen und Geburtsdatum des Probanden enthält, automatisch vom zentralen Datenbanksystem erstellt und an den Screening-Standort übermittelt. Die Daten mit denen der jeweilige Proband identifiziert werden könnte, wie etwa Name oder Adresse, wurden nicht digitalisiert. Um weitere Sicherheit im Bezug auf den Datenschutz für die Probanden zu garantieren, mussten die Mitarbeiter der PRONIA Studie eine Verschwiegenheitserklärung unterschreiben. Da es sich bei gesundheitsbezogenen Daten um sensible Informationen handelte, haben alle Teilnehmer ein Recht auf Löschung ihrer Daten [57].

3.5 Einverständniserklärung

Das PRONIA Projekt führte eine nicht-interventionelle diagnostische Studie mit sicheren und etablierten Untersuchungsverfahren durch, so dass keine größeren somatischen Risiken und Nebenwirkungen durch die Untersuchungsmethoden von PRONIA zu erwarten waren. Alle potenziell in die Studie aufzunehmenden Hilfesuchenden und gesunden Kontrollpersonen waren in der Lage, eine selbstbestimmte Entscheidung bezüglich der Studienteilnahme zu treffen, nachdem sie über alle Verfahren, Vorteile und potenziellen Risiken der PRONIA-Studie informiert wurden. Durch diesen Prozess wurde sichergestellt, dass die Studienteilnehmer wirklich verstehen, wozu sie sich bereit erklären und dass sie nach ihrem freien Willen handeln. Alle Teilnehmer der PRONIA-Studie wurden darüber informiert die Studie jederzeit beenden oder bestimmte Untersuchungen ablehnen zu können. Die Einverständniserklärungen werden so verfasst, dass sie auch für Laien verständlich sind [57].

3.5.1 Forschung an Minderjährigen

Auch Kinder zwischen dem 15. und dem 18. Lebensjahr konnten in die PRONIA-Studie aufgenommen werden. In diesem Zeitraum kann bereits eine CHR-Symptomatik auftreten, die im Verlauf in einer Erkrankung des psychotischen Formenkreises münden kann. In solchen Fällen muss die schriftliche Zustimmung von den Erziehungsberechtigten eingeholt werden. Das PRONIA-Konsortium stellte sicher, dass die Einwilligungserklärungen für Minderjährige

sowie das Studiendesign von PRONIA die Richtlinien der „Ethical Considerations For Clinical Trials On Medicinal Products Conducted With The Paediatric Population Recommendations“ der Ad-hoc-Gruppe zur Entwicklung von Durchführungsrichtlinien für die Richtlinie 2001/20/EG umsetzten[118, 57].

3.6 Neuronale Netze

Idealerweise eigenen sich für Neuronale Netze große Datensätze ab etwa 1000 Stichproben. Besonders im medizinischen Kontext und sind die Kohorten kleiner und von unterschiedlicher Größe. Besonders bei kleinen Stichproben unterschiedlicher Größe, verschlechtert sich die Prädiktion durch Neuronale Netze [119].
Dennoch bietet die Verwendung von neuronalen Netzen in diesem Kontext einige Vorteile, da diese vor allem sehr komplexe, nicht-lineare Zusammenhänge detektieren können. Angenommen wir haben **ein** Neuron mit den Eingangssignalen $p_1, p_2, ..., p_R$ und den skalaren Gewichten $w_{1,1}, w_{1,2}, ...w_{1,R}$, die die korrespondierende Matrix W bilden, dann beschreiben die Indices der Gewichte $w_{1,R}$ mit der ersten Ziffer die Zugehörigkeit zu dem einen Neuron und mit der zweiten Ziffer die Zugehörigkeit zu dem korrespondierenden Signal p_R. Die Eingangssignale p werden mit den skalaren Gewichten w multipliziert und bilden wp. All diese einzelnen Inputs werden addiert ($\sum$) und schließlich mit dem Bias b verrechnet.

$$x = (w_{1,1}p_1) + (w_{1,2}p_2) + (w_{1,3}p_3) + ... + (w_{1,R}p_R) + b \tag{3.1}$$

Der skalare Wert x, der sich aus dieser Operation ergibt, wird dann einer Aktivierungsfunktion f übergeben, die dann den Output a erzeugt ($f(x) = a$). Formel 3.1 kann in der Matrixschreibweise notiert werden:

$$a = f(W_p + b) \tag{3.2}$$

W_p ist die Matrix aus $(w_{1,1}p_1) + (w_{1,2}p_2), ..., (w_{1,R}p_R)$. Im Bezug auf das biologische Neuron, entspräche a dem elektrischen Potential, das am Axon gemessen werden kann [81]. Veranschaulicht wird dies in Abbildung 1.2.

3.6.1 Aktivierungsfunktion

Die Aktivierungsfunktion ein essentieller Bestandteil eines künstlichen Neurons und somit aller NN-Architekturen.

Binäre Schwellenwertfunktion (hardlim)

Einem Rechteckimpuls ähnlich, ist die binäre Schwellenwertfunktion eine sehr naheliegende Aktivierungsfunktion, die an die „alles oder nichts“ Reaktion aus der Biologie erinnert. Das Ausgangssignal a nimmt bei diesem Modell den Wert 1 an, sobald ein Schwellenwert θ überschritten wird. Die hardlim Funktion bei manchen NN-Architekturen nach wie vor verwendet [120].

$$f(n) = \begin{cases} 0, & \text{für } n < \theta \\ 1, & \text{für } n \geq \theta \end{cases} \tag{3.3}$$

Sigmoide Aktivierungsfunktion

Die Sigmoide Aktivierungsfunktion ist eine sehr häufig verwendete Funktion, die sehr häufig verwendet wird. Vor allem bei tiefen, also „mehrschichtigen“ Neuronalen Netzen erfreut sie sich großer Beliebtheit, da ‘das Sigmoid’ vollständig differenzierbar ist [81]. Der Fehlerrückführungsalgorithmus kann nur bei vollständig differenzierbaren Funktionen angewandt werden [87, 121].
Sei s das Steigungsmaß :

$$f_s(n) = \frac{1}{1 + exp(-sn)} \tag{3.4}$$

Lineare Aktivierungsfunktion

Lineare Aktivierungsfunktionen finden sich meistens in der Ausgabeschicht (Output Layer) eines neuronalen Netzwerkes. Wären die Aktivierungfunktionen der Hidden Layers linear, könnte das neuronale Netz weniger Funktionen annähern.

Der Tangens Hyperbolicus (TanH) Die TanH Aktivierungsfunktion ähnelt dem Sigmoid. Anders als die sigmoide Aktivierungsfunktion, kann diese jedoch auch negative Werte ausgeben. Bei sehr tiefen Neuronalen Netzen mit vielen Schichten, wird vermutet, dass durch die TanH Funktion das Training des Netzes vereinfacht wird[122], Die TanH Funktion wird in der Standardarchitektur der LSTM Netze regelhaft verwendet.

$$f_{tanh}(n) = \frac{e^x - e^{-n}}{e^n + e^{-n}} \tag{3.5}$$

3.6.2 Der Fehlerrückführungsalgorithmus (Backpropagation)

Fehlerfunktion

Ein neuronales Netz soll eine möglichst zuverlässige Abbildung von gegebenen Eingabevektoren auf gegebene Ausgabevektoren anstreben. Dazu wird die Qualität der Abbildung durch eine Fehlerfunktion beschrieben, die hier durch den quadratischen Fehler definiert wird:

$$E = \frac{1}{2}\sum_{i=1}^{n}(t_i - o_i)^2 \tag{3.6}$$

Dabei ist E der berechntete quadratische Fehler (SE), n ist die Anzahl der Inputs, die dem Netz zum Trainieren gegeben werden, t_i ist, das „target“, also die uns bekannte Zielvariable und o_i ist die, durch das NN, errechnete Ist-Ausgabe (output)[123]. Die Multiplikation mit $\frac{1}{2}$ wird nur zur Vereinfachung der Ableitung im Rahmen der Backpropagation herangezogen. Bei einem klassischen neuronalen Netz, kennen wir die wahren Outputlabels, weswegen es sich bei diesem Algorithmus, um einen überwachten Lernalgorithmus (supervised algorithm) handelt [74].

Die Kettenregel

Um das Grundprinzip der Backpropagation verstehen zu können, muss man sich die Kettenregel aus der Analysis zu Gemüte führen, da das Prinzip auf Ableitungen basiert. Daher ist es wichtig, dass es sich bei den Aktivierungsfunktionen um vollständig differenzierbare Funktionen handelt.

$$\begin{aligned} f(x) &= u(v(x)) \\ f'(x) &= u'(v(x))v'(x) \end{aligned} \tag{3.7}$$

Im Rahmen der Backpropagation wird die Fehlerfunktion abgeleitet und so die Steigung der Funktion angenähert. Wir wissen, dass die Fehlerfunktion ein globales Minimum besitzt. Das heißt eine Stelle, an der die Ableitung null ist.
Prinzipiell werden bei der Backpropagation die Gewichte w solange aktualisiert, bis die Ableitung der Fehlerfunktion minimal ist. Allerdings kann der Algorithmus nicht unterscheiden, ob er das tatsächliche globale Minimum oder ein lokales Minimum gefunden hat. Das Prinzip gleicht dem Gradientenabstieg [124].

3.6.3 Reduktion der Fehlerfunktion durch unterschiedliche Lösungsstrategien

Stochastischer Gradientenabstieg - Stochastic Gradientdescent (SGD)

Der Stochastische Gradientenabstieg (SGD) und seine Varianten sind für das maschinelle Lernen essentiell. Algorithmus 1 zeigt schematisch die Funktionsweise des SGD. Dieser ermöglicht, eine unvoreingenommene Schätzung des Gradienten unter Verwendung des durchschnittlichen Gradienten auf einem Minibatch der Größe m von k Beispielen, aus der datengenerierenden Verteilung, zu erhalten. Ein entscheidender Parameter für den SGD-Algorithmus ist die Lernrate ϵ.
Die wichtigste Eigenschaft der SGD und der damit verbundenen Minibatch- oder der Gradienten - basierten Optimierung ist, dass die Berechnungszeit pro Update nicht mit der Anzahl der Trainingsbeispiele wächst. Dies ermöglicht eine Konvergenz, auch wenn die Anzahl der Trainingsbeispiele sehr groß wird. Bei einem ausreichend großen Datensatz kann es vorkommen, dass SGD, bei Verwendung eines statischen Fehlertoleranzwertes, konvergiert, bevor es den gesamten Trainingssatz verarbeitet hat [125, 94, 126].

Algorithmus 1 SGD Lösungsalgorithmus nach [94]

benötigt: Lernrate $\epsilon_1(, \epsilon_2)$ (bzw. Veränderung der Lernrate pro Schritt)
den Parameter Θ, der die Anfangsgewichte w initialisiert
$k \leftarrow 1$:
while Während das Stopkriterium nicht erfüllt ist **do**

Berechne den geschätzten Gradienten aus einem
Minibatch des Trainingssets mit der Größe m

set: x^1, x^2, ... x^m mit den jeweiligen Zielgrößen y^1, y^2, ... y^m
Berechne den geschätzten Gradienten: $\hat{g} \leftarrow \frac{1}{m} \nabla_\Theta \sum_i L(f(x^{(i)}; \Theta), y^{(i)})$
Update:$\Theta \leftarrow \Theta - \epsilon_k \hat{g}$
$k \leftarrow k + 1$
close;

RMSProp – Root mean square propagation

Der eben vorgestellte SGD Lösungsalgorithmus findet seinen Weg ins Optimum ohne dabei ein Momentum zu verwenden. Das heißt, dass die Lernrate im Laufe des Gradientenabstiegs nicht verändert wird.
Bei RMSProp handelt sich um einen sehr berühmten Algorithmus, da er niemals publiziert

wurde, sich aber dennoch sehr großer Beliebtheit erfreut. Der RMSProp Algorithmus kann die Lernrate adaptiv anpassen. Mit zunächst großen Schritten nähert er sich dem Optimum an und verkleinert dann sukzessive die Lernrate, um sich möglichst genau dem Optimum annähern zu können. Auf Diese Art und Weise ist seine Performance zumeist besser als die des SGD Algorithmus[96].

3.6.4 LSTM – Long Short Term Memory

Im Rahmen dieser Disseration wurde jeweils neuronale Netze mit LSTM Neuronen für die Klassifikation und für die Regression entwickelt. Dafür wurde jeweils die DeepLearning Toolbox in Matlab 2018b genutzt [127].
Als Lernalgorithmus wurde stets der RMSProp Algorithmus verwendet. Für die jeweiligen Modelle wurden jeweils der SquareGradientDecay', der die Lernrate beeinflusst, die Anzahl der verschiedenen Neurone, sowie die L2 Regularisierung und die initiale Lernrate mit Hilfe des Bayesianischen Optimierungsalgorithmus bestimmt (siehe Abschnitt 3.7.4).

Die klassische LSTM Architektur:

Sigma (σ) entspricht der sigmoiden Aktivierungsfunktion (siehe Kapitel 3.6.1)

Definition der Schranken (gates) nach [128, 129] :

Vergessensschranke (forget gate):
$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f)$
Eingabeschranke (input gate layer):
$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)$
$\hat{C}_t = tanh = (W_C[h_{t-1}, x_t] + b_C)$
der Output dieser beiden Schranken innerhalb des LSTM Neurons entspricht C_t. Dieser ist folglich definiert durch:
$C_t = f_t \cdot C_{t-1} + i_t \cdot \hat{C}_t$
Diese Information wird in die Ausgabeschranke o_t(Outputgate) eingespeist. Hier wird letztendliche Output h_t generiert.
$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$
$h_t = o_t \cdot tanh(C_t)$

3.7 Versuchsaufbau

3.7.1 Datenmenge

Insgesamt wurden im Rahmen der Proniastudie 904 Datensätze generiert. Unter anderem handelte es sich dabei auch um Testdatensätze, die angelegt wurden, um die Funktionalität des Systems zu testen. Zehn Datensätze wurden im Zuge dessen aus dem Datenpool entfernt. 22 weitere Patienten legten keine Einverständniserklärung vor. Ein Patient aus dem Patientendatenpool war fälschlicherweise als gesunde Kontrolle (healthy control - HC) deklariert. Somit blieben 881 Datensätze, die verwendet werden konnten.
Davon waren 286 CHR Patienten, 321 ROP Patienten und 267 ROD Patienten. 7 Datensätzte blieben ohne Gruppenzugehörigkeit.
Im Sinne einer Leave - Site Out Validation [65, 130] wurden alle Datensätze, die im Uniklinikum Köln erhoben wurden nicht in den Test-/ Trainingsdatensatz aufgenommen. Der Datensatz, welcher in Köln erhoben wurde enthielt 141 Probanden. Die verfügbaren Datensätze wurden demnach in einen Validierungsdatensatz zu 141 Stichproben und einen Test-/ Trainingsdatensatz zu 470 Datensätze aufgeteilt. Im Validierungsdatensatz befanden sich keine Datensätze ohne Gruppenzugehörigkeit. Erst nach der endgültigen Abspaltung des Validierungsdatensatzes erfolgten weitere Präprozessierungsschritte, um somit ein Datenleck (data leakage), welches die Prädiktionen der Validierungsdaten fälschlicherweise verbessern könnte, zu verhindern. Die Zeitreihen wurden, wie in 3.7.3 beschrieben gekappt, sodass der letzte prädizierbare Wert auch die letzte tatsächlich stattgefundene Verlaufsbeobachtung des Patienten darstellte. Nach der Entfernung von Datensätzen, bei denen 40% oder mehr Daten aus den Verlaufbeobachtungen fehlte, blieben zur Validierung 96 Stichproben und 398 Datensätze zur Etablierung des LSTM Modells übrig.
Die restlichen Daten des Test und des Trainingdatensatzes wurden mit Hilfe des K-nearest neighbor (KNN) Algorithmus vervollständigt. Im Rahmen einer 5-fachen Kreuzvalidierung (siehe Abschnitt 3.7.2) unterteilten wir die Daten in einen Trainingsdatensatz (4/5) und einen Testdatensatz (1/5) und bestimmten aus den fünf Durchläufen das beste Modell anhand der Ergebnisse des Testdatensatzes (siehe Abbildung 3.1). Mittels einer Modulo-Operation wechseln sich die Gruppen jeweils ab, sodass jede Stichprobe genau einmal Teil des Testdatensatzes war. Dabei erfolgte die Vervollständigung durch den KNN-Algorithmus jeweils erst, nachdem ein Fünftel des Datensatzes abgespalten war, um in der internen Schleife der Kreuzvalidierung kein Datenleck zu ermöglichen. Das beste Modell aus der Kreuzvalidierung wird an den Validierungsdaten, den Daten aus Köln, erprobt, welche bis dahin für das Neuronale Netz noch gänzlich unbekannt sind [131]. Diese Ergebnisse können im Ergebnisteil (siehe Kapitel 4) nachgelesen werden.
Die Anwendung, welche im Rahmen dieser Arbeit programmiert wurde, interagiert mit dem Nutzer. Über die Kommandozeile kann ausgewählt werden, ob man im Folgenden eine Regression oder eine Klassifikation rechnen lassen möchte. Je nach Einstellung unterscheidet

sich die Ausgabeschicht des Neuronalen Netzes (siehe Abbildung 3.2 und Abbildung 3.3). Dort wird, falls man den Modus Klassifikation ausgewählt hat, eine 2, für eine gute Prognose beziehungsweise eine 1, für eine schlechte Prognose, zum nächsten Zeitabschnitt ausgegeben. Im Modus Regression wird der nächste zukünftige Wert direkt berechnet. In unserem konkreten Fall wurden für die vereinfachte Vergleichbarkeit zwischen GAF und GF alle Ergebnisse zwischen 0-10 skaliert. Um also den „wahren“ prädizierten GAF Wert zu ermitteln müsste das Ergebnis einer GAF Regression mit 10 multipliziert werden.

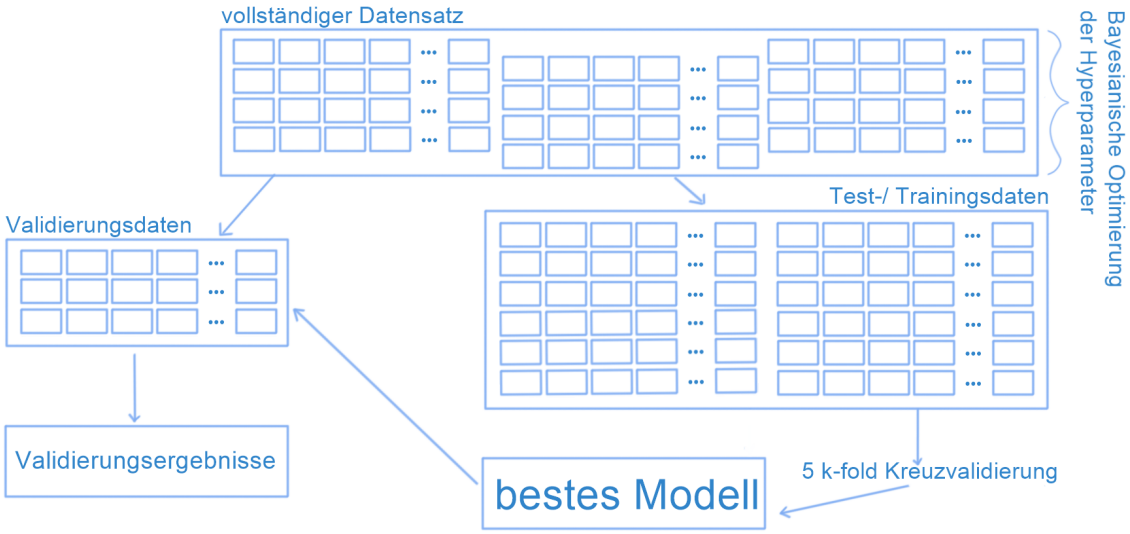

Abbildung 3.1: Die Abbildung zeigt schematisch welche Schritte durchlaufen werden, bis das beste LSTM Modell gefunden wurde, welches schlussendlich an den Validierungsdaten erprobt wird. Diese Daten bestehen aus den Stichproben des Studienzentrums Köln. Die Daten der anderen Studienzentren sind Teil des Test-/Trainingsdatensatzes.

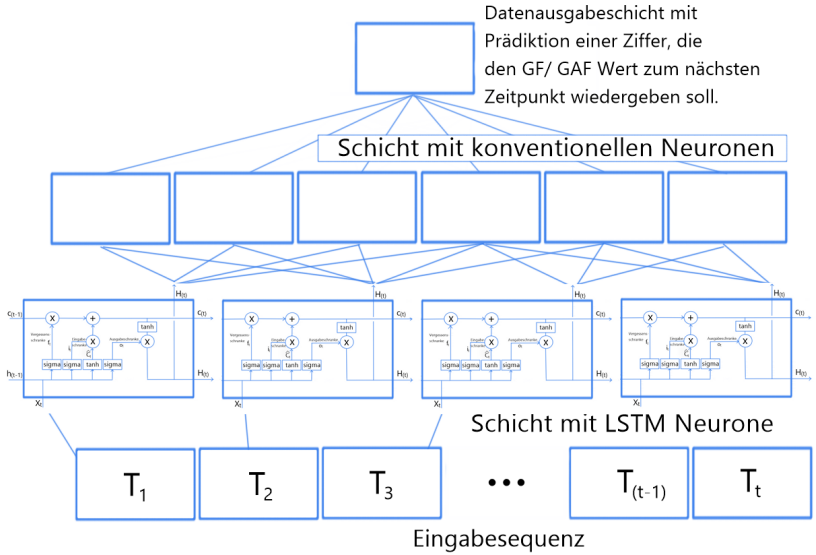

Abbildung 3.2: Grober Aufbau des LSTM Netzwerkes welches zur Regression verwendet wird.

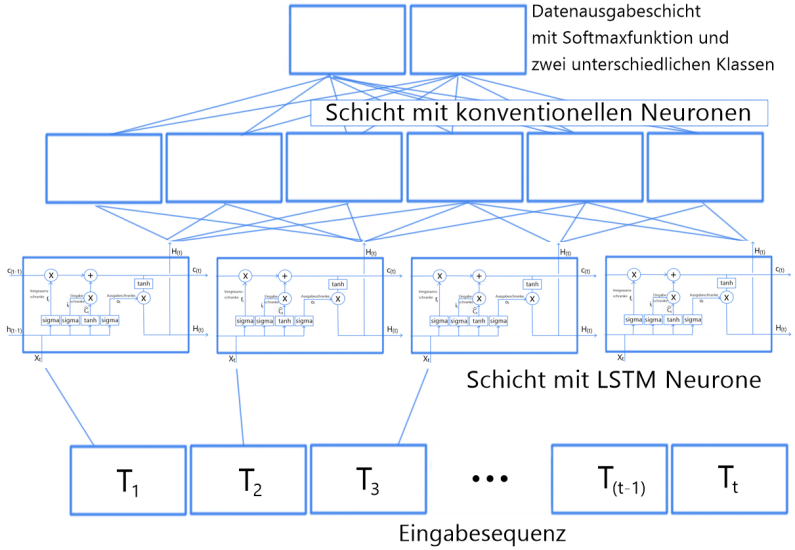

Abbildung 3.3: Grober Aufbau des LSTM Netzwerkes, welches zur Klassifikation herangezogen wird.

3.7.2 Kreuzvalidierung und Leave-Site-Out Validierung

Der Zweck der Kreuzvalidierung ist, die Auswahl eines Modells zu erleichtern, das eine generalisierbare Vorhersage aus der verfügbaren Stichprobe liefert. Ein Overfitting[2] an den Trainingsdatensatz kann so vermieden werden. Wenn die Kalibrierungsstichprobe klein ist, werden die genauesten Schätzungen tendenziell von Modellen mit wenigen Parametern geliefert [132]. In dieser Arbeit wird eine fünffache Kreuzvalidierung durchgeführt. Hierbei wird bei jedem Durchlauf der Test-/Trainingsdatensatz in 5 gleich große Anteile aufgeteilt. Diese anfänglich bestimmten Gruppenzugehörigkeiten ändern sich während der fünf Durchläufe nicht. Vier dieser fünf Anteile werden zum Trainingsdatensatz, während der fünfte Datensatz zum Testen zurück gehalten wird. Die Rolle des Testdatensatzes rotiert mittels einer Modulo-Operation während der fünf Durchläufe durch die fünf Gruppen hindurch. Dabei entstehen fünf verschiedene LSTM Modelle, die jeweils an komplett unterschiedlichen Testdatensätzen getestet wurden. Abschließend wird das LSTM Modell ausgewählt, welches den geringsten RMSE oder die höchste Genauigkeit auf dem Testdatensatz zeigte [133]. Dieses Modell wird abschließend an den Daten, welche in Köln erhoben wurden, validiert. Da die Daten dieses Rekrutierungs-Ortes (Site) nicht in die Erstellung des Modells involviert wurde, handelt es sich um eine Leave-Site-Out Validierung. Dies führt zu einer verbesserten Abschätzung der Gerneralisierbarkeit des jeweiligen Modells [130].

3.7.3 Aufbau der Zeitreihe

Die Zeitreihen wurden zunächst mit Hilfe der K-nearest Neighbor Methode vervollständigt [134]. Dabei wurde diese Technik immer auf unterster Ebene angewendet. Das bedeutet immer nachdem die Aufteilung durch die Kreuzvalidierung bereits abgeschlossen war, um so ein mögliches Datenleck zu vermeiden. Die Training und Testdaten wurden also immer separat vervollständigt. Auch die Validierungsdaten durchliefen diesen Prozess ohne mit den Trainings- und Testdaten in Kontakt zu geraten.
Nicht alle Probanden haben den kompletten Beobachtungszeitraum von 36 Monaten durchlaufen. Daher sind die Zeitreihen auf den letzten tatsächlich erhobenen Wert gekürzt worden. Insgesamt ergaben sich Zeitketten in der Länge von 4 bis 10 Einheiten. Somit können maximal 8 Punkte in die Zukunft prädiziert werden (IV3-IV36). Die ersten beiden Werte einer Zeitkette bestehen immer aus einem retrospektiv erhobenen GAF/GF Wert und dem aktuellen GAF/GF Wert.

[2]Mit dem Begriff „Overfitting“ wird eine Überanpassung beschrieben. Dies bedeutet, dass das Modell, welches anhand der Trainingsdaten ein bestimmtes prädiktives Verhalten erlernt hat, welches an nicht generalisierbar ist und bei einem Validierungsdatensatz deutlich schlechtere Ergebnisse liefert [81].

GAF/GF

Der GAF oder der GF stellen das Herzstück der jeweiligen Prädiktion dar. Einer dieser beiden Metriken wird mit Hilfe der vorherigen Werte für den nächsten Zeitabschnitt prädiziert. Zusätzliche dynamische, behavoriale Parameter (BDI, PANSS) können herangezogen werden. Um einen Leistungsknick abbilden zu können, der im Rahmen der Basissymptome häufig beschrieben wird, verwendeten wir zusätzlich zum aktuellen GAF Wert, den niedrigsten GAF Wert zum letzten Monat, der retrospektiv zur Erstuntersuchung (T0) erhoben wurde. So generierten wir aus einer Kette, die aus neun Erhebungen bestand, zehn Zeitpunkte. Als Grenzwert wurde gemäß der aktuellen wissenschaftlichen Literatur bei den GAF Klassifikationen stets ein Grenzwert von 65 herangezogen. Patienten mit einem guten Funktionsniveau hatten demnach einen GAF, der größer gleich 65 war [65]. Bei den GF Klassifikationen wurde ein Grenzwert von 7 herangezogen. Patienten mit einem GAF kleiner 7 hatten ein schlechtes Funktionsniveau. Auch hier stützten wir uns auf die verfügbare Literatur [107].

K- nearest neighbor (K–NN)

Bei der Mustererkennung ist der K-nearest neighbor (k-NN) Algorithmus eine von Thomas Cover vorgeschlagene nichtparametrische Methode, die zur Klassifizierung und Regression verwendet werden kann. In beiden Fällen besteht die Eingabe aus den k-nächsten Trainingsbeispielen im Merkmalsraum (Trainingsdaten). Eine Besonderheit des k-NN-Algorithmus ist, dass er empfindlich auf die lokale Struktur der Daten reagiert[134]. Der K-nearest neighbor Algorithmus kann auch zur Vervollständigung von Daten verwendet werden. Bei Letzterem werden alle Datensätze im Raum miteinander verglichen und die fehlenden Daten mit den Werten des Datensatzes ergänzt, der dem zu vervollständigenden Datensatz im diskreten Raum am meisten ähnelt [135].

3.7.4 Hyperparameteroptimierung

Nach der Aufbereitung der Datensätze, wurde eine Hyperparameteroptimierung durchgeführt. In unserem Fall eine Bayesianische Optimierung. Diese Methode wurde 1975 erstmals publiziert [136]. In Matlab läuft diese Methode mit folgenden Schlüsselelementen ab:

1. Erstellung eines Gaussischen Prozess Models der unbekannten Funktion $f(x)$.
2. Ein Bayesianischer Aktualisierungsprozess, der das Gaussische Prozess Model der Funktion $f(x)$ bei jeder Evaluierung modifiziert.
3. Eine Erfassungsfunktion $a(x)$ (basierend auf dem Gaußschen Prozessmodell von $f(x)$), dessen globales Optimum bestimmt wird, um den nächsten Punkt x für die nächste Evaluierung zu bestimmen.

Bei $f(x)$ handelt es sich um eine sogenannte Black Box Funktion. Der Algorithmus weiß zunächst nie bei welcher Eingabe, welche Ergebnisse resultieren werden. Bei der Eingabe von bestimmten Werten kann er aber den Verlauf der Funktion $f(x)$ von Iteration zu Iteration sukzessive annähern[137]. Mit Hilfe dieser Methode optimieren wir die Anzahl der LSTM Neurone, der „*fully-connected*" Neurone, die wir auch als konventionelle Neurone bezeichnen, sowie die L2 Regularisierung und den Gradient Decay, der im Rahmen der RMSProp Lernalgorithmus unsere Lernrate eklatant beeinflusst (siehe Einleitung 1.5.1). Zudem optimieren wir mit dieser Methode die initiale Lernrate, sowie die Größe der Mini-Batches, die für den RMSProp Algorithmus benötigt werden.
Es konnte gezeigt werden, dass der Bayesianische Optimierungsalgorithmus bessere Ergebnisse liefert, als etwa Gridsearch[3], eine Zufallssuche oder eine Auswahl der Parameter durch Experten [139] [140].

3.7.5 L2 Regularisierung

Wie die Kreuzvalidierung dient die L2 Regularisierung der Reduktion eines möglichen Overfittings, welche sich durch eine schlechte Generalisierbarkeit unseres Modells äußern würde. Eine L2 Regularisierung sorgt dafür, dass die Gewichte w unwichtiger Parameter ein Wichtung von beinnahe Null annehmen. Durch eine L2 Regularisierung kann das LSTM somit bessere Vorhersagen bei bisher unbekannten Daten treffen [141].

3.7.6 Evaluierung

Balancierte Genauigkeit (Balanced Accuracy-BAC)

In den Experimenten, in denen wir zwischen guten und schlechten Outcomes klassifizieren, verwenden wir die balancierte Genauigkeit, um die Korrektheit der Prädiktionen möglichst realistisch abzubilden. Falls der Datensatz nicht balanciert ist, kann die regulär berechnete Genauigkeit, fälschlicherweise zu gute Werte angeben. Dafür verwenden wir im Zuge dieser Arbeit die balancierte Genauigkeit - kurz BAC [142]. Intern für das LSTM berechnet MATLAB die Genauigkeit [127].

[3] Unter Gridsearch versteht man einen Hyperparameteroptimierungsalgorithmus, der verschiedene Hyperparameterkombinatioenen durch ein gitterartiges Muster (jeweils an den Kreuzungen des Gitters) im mathematischen Raum, den die Hyperparameter bilden, testet[138].

Genauigkeit

Die Genauigkeit beschreibt die Wahrscheinlichkeit, dass dem Probanden das korrekte Label zugewiesen wurde.

$$Genauigkeit = SE \cdot PR + SP \cdot (1 - PR) \tag{3.8}$$

Die Genauigkeit wird von der Prävalenz (PR) beeinflusst. Bei Datensätze in denen eine Gruppe deutlich seltener vertreten ist, kann die BAC deutlich abweichen, weswegen beide Werte wichtige Parameter zur Abschätzung der Qualität eines Modells sind [143, 144]. In Formel 3.8 bezeichnet PR die Prävalenz, SE die Sensitivität und SP die Spezifität unter Punkt 3.7.7 wird diese Notation fortgeführt.

Der mittlere quadratische Fehler (Rooted Mean square error - RMSE)

Der mittlere quadratische Fehler (RMSE) ist ein häufig verwendetes Maß für die Genauigkeit eines Schätzers wiedergeben soll. Der RMSE kondensiert, die Abweichungen des Schätzers in den Vorhersagen für verschiedene Zeitpunkte zu einem einzigen Wert.
Der RMSE ist maßstabsabhängig, daher kann er nicht zwischen unterschiedlich skalierten Datensätzen verwendet werden [145].

Operationscharakteristik (ROC) und AUC

Da im Rahmen der Regressionsanalyse post - hoc erneute Klassifizierungen durch verschiedene Grenzwerte durchgeführt werden können, können jeweils die Richtig-Positiv-Rate (auch bekannt als Sensitivität) und die Falsch-Positiv-Rate der jeweiligen Grenzwerte auf der x- und y- Koordinate eingetragen werden, was die sogenannte ROC Kurve ergibt [146]. Die AUC (eng: area under the curve) ist das Integral, welches die Fläche unterhalb der ROC Kurve bildet. Im Rahmen dieser Arbeit wurde die AUC mit der Matlab internen „*trapz*" Funktion ermittelt [147, 148]. Die Berechnung der AUC ist ein beliebtes Verfahren um Machine - Learning - Algorithmen zu bewerten [149].

3.7.7 Klassifikation

Zunächst haben wir versucht mit Hilfe des LSTMs vorherzusagen, ob die Prognose der Patienten innerhalb des nächsten Zeitintervalls gut oder schlecht sein wird. Als Cutoff haben wir beim GAF die 65 gewählt. Also GAF Werte, welche kleiner 65 waren, entsprachen einer schlechten Prognose, während Werte größer gleich 65 mit einer guten Prognose gleichgesetzt wurden [65, 150, 151].

Beim GF entsprachen alle Werte kleiner sieben einer schlechten Prognose, während Werte, die einer sieben entsprachen oder höhere Werte mit einer guten Prognose gleichgesetzt worden sind. Beim GF kann ein Wert von 6,5 nicht vom Rater eingetragen werden. Gemäß der Literaur[107] wurde ein Grenzwert von 7 herangezogen.

Sensitivität

Die Sensitivität (SE) beschreibt die Wahrscheinlichkeit, dass ein Testverfahren ein positives Resultat ausgibt, wenn der zu detektierende Zustand tatsächlich vorliegt.

$$SE = \frac{RPG}{RPG + FNG} \tag{3.9}$$

Zur Berechnung werden die richtig positiv getesteten Probanden (RPG), sowie die falsch negativ Getesteten (FNG) benötigt [152].

Spezifität

Die Spezifität (SP) beschreibt die Wahrscheinlichkeit, dass ein Testverfahren ein negatives Ergebnis ausgibt, wenn der zu detektierende Zustand nicht vorhanden ist.

$$SP = \frac{RNG}{RNG + FPG} \tag{3.10}$$

Zur Berechnung die richtig negativ Getesteten (RNG), sowie die falsch positiv Getesteten (FPG) herangezogen.

Positiv Prädiktiver Wert (PPW)

Der positiv prädiktive Wert gibt wieder, wie wahrscheinlich es ist, dass der zu detektierende Zustand bei einem positiven Ergebnis auch tatsächlich vorhanden ist.

$$PPW = 100 \cdot \frac{PR \cdot SE}{PR \cdot SE + ((1 - PR) \cdot (1 - SP))} \tag{3.11}$$

Wenn mehr sich der Wert den 100% annähert, desto besser ist das untersuchte Verfahren. Um den PPW zu berechnen, werden die Prävalenz (PR), die Spezifität, sowie die Sensitivität genutzt [143, 153].

Negativ Prädiktiver Wert (NPW)

Der negativ Prädiktive Wert gibt die Wahrscheinlichkeit an, dass ein Merkmal tatsächlich nicht vorhanden ist, wenn der betreffende Test ein negatives Ergebnis zeigt.

$$NPW = 100 \cdot \frac{SP \cdot (1 - PR)}{PR \cdot (1 - SE) + ((1 - PR) \cdot (SP))} \tag{3.12}$$

Der positive prädiktive Wert, sowie der negative prädiktive Wert sind von der Prävalenz der gesuchten Merkmale abhängig. [143].

Positive - Likelihood - Ratio (PLR)

Die Positive Likelihood Ratio (PLR) ist ein Wert, der umso höher ist, je wahrscheinlicher ein positives Testergebnis auch für das Vorhandensein des zu detektierenden Zustandes spricht.

$$PLR = \frac{SE}{1 - SP} \tag{3.13}$$

Eine Likelihood-Ratio größer als 1 bedeutet, dass das Testergebnis mit dem Vorliegen eines Zustandes assoziiert ist [154, 152].

Negative - Likelihood - Ratio (NLR)

Die negative Likelihood Ratio beschreibt das Verhältnis zwischen der von 1 subtrahierten Sensitivität und der Spezifität eines Merkmals.

$$NLR = \frac{1 - SE}{SP} \tag{3.14}$$

Eine Likelihood-Ratio kleiner als 1 suggeriert, dass das Testergebnis mit dem Nichtvorliegen der Krankheit in Verbindung steht. Je weiter das Likelihood-Verhältnis von 1 entfernt ist, desto stärker ist die Evidenz für das Vorliegen oder Nichtvorliegen des zu untersuchenden Merkmals. Bei der NLR strebt man daher Werte an, die sich der 0 annähern. Likelihood-Ratios über 10 und unter 0,1 gelten in den meisten Fällen als starker Beweis für den Ausschluss bzw. die Ablehnung einer Diagnose [155, 152].

Konfidenzintervalle

Die NLR, PLR, der PPW, NPW, sowie die Spezifität und Sensitivität werden im Ergebnisteil gemeinsam mit den 95% Konfidenzintervallen (CI) angegeben. Ein Konfidenzintervall schätzt ab in welchem Bereich sich ein Parameter bei einer erneuten Testung anhand eines anderen Datensatzes befinden könnte [152]. Der geschätzte Bereich soll bei erneuten Testungen, die zu erwartenden Ergebnisse in 95% der Fällen abdecken. Zur Bestimmung der Konfidentintervalle von Sensitivität und Spezifität wurden die exakten Clopper-Pearson Konfidenzintervalle berechnet [156]. Zur Berechnung der Konfidenzintervalle für die Likelihood - Ratios wurde die „Log Methode" verwendet [155]. Für die prädiktiven Werte wurden die Konfidenzintervalle mit Hilfe des Logit kalkuliert [157]. Um die berechneten Konfidenzintervalle abschließend zu überprüfen, wurde die Medcalc Software verwendet [158].

3.7.8 Rekursives Prädiktionsmodell

Anders als das oben vorgestellte LSTM Modell kann das rekursive Prädiktionsmodell, welches auch im Rahmen dieser Dissertation erarbeitet wurde, mit Hilfe eines Regressions LSTMs rekursiv nur mit den T0 Daten, die Verläufe eines Patienten bis IV 36 vorhersagen. Dieses Modell basiert auf dem beschriebenen Modell zur Regression (siehe Abbildung 3.4). Da eine Regression seinen Eingabewert zu IV3 vorherzusagen versucht. Diese Vorhersage kann wiederum erneut in das trainierte LSTM Netz gefüttert werden.

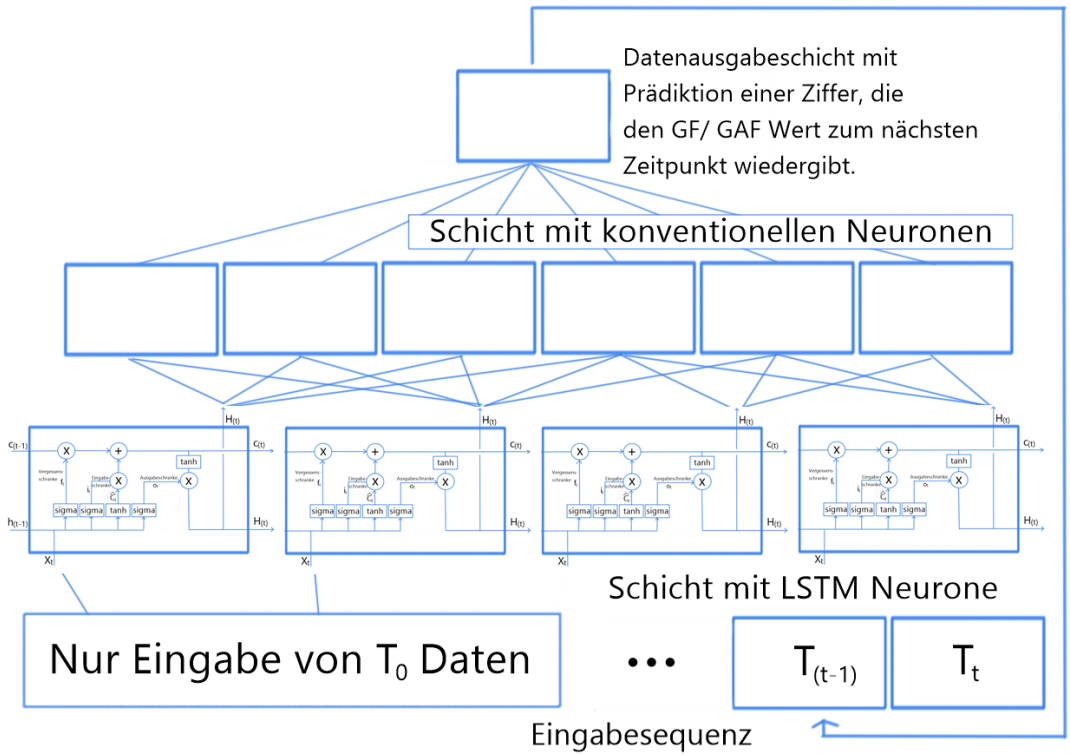

Abbildung 3.4: Auf der Grafik ist zu erkennen, dass dem rekursiven Prädiktionsmodell lediglich die Daten aus T0 (zwei Datenpunkte) gegeben werden. Da die darauffolgenden Prädiktionen dem LSTM immer wieder neu in die Eingabeschicht hinzugegeben werden, können so Prädiktionen bis IV36 erfolgen. Und man kann somit auch die GAF Verläufe bis IV36 mit den bloßen T0 Daten vorhersagen.

3.8 Statistik

Im Rahmen dieser Arbeit wurde nicht nur deskriptive Statistik, also etwa die Berechnung von Durchschnitten und Standardabweichungen durchgeführt, sondern es wurden auch zahlreiche statistische Tests zum Gruppenvergleich herangezogen, die im Verlauf kurz erläutert werden.

3.8.1 Chi-Quadrat-Test

Der Chi-Quadrat-Test ist ein nicht-parametrischer statistischer Test, der zur Analyse von Gruppenunterschieden herangezogen werden kann. Um ihn anwenden zu können, wird keine Varianzgleichheit oder Homoskedastizität der Daten benötigt [159]. In der Familie der Chi-Quadrat-Tests können drei Modi unterschieden werden, nämlich der Chi-Quadrat Verteilungstest, der Chi-Quadrat Unabhängigkeitstest, sowie der Chi-Quadrat Homogenitätstest. Im Zuge dieser Arbeit wird der Chi-Quadrat-Unabhängigkeitstest verwendet, um zu überprüfen, ob zwei kategoriale Variablen in einer Stichprobe voneinander unabhängig sind.

$$X^2 = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(N_{ij} - n_{ij})^2}{n_{ij}} \tag{3.15}$$

X^2 symbolisiert die Testprüfgröße des Chi-Quadrat-Tests. Es gibt insgesamt n paarweise Beobachtungen, die sich auf $m \cdot k$ Kategorien verteilen. In der Kontingenztabelle verdeutlicht m die Summe der Zeilen, k die Summe der Spalten. N_{ij} entspricht den beobachteten Häufigkeiten der Kategorie ij. Die Variable n_{ij} ergibt sich aus $n_{ij} = p_{ij} \cdot n$. Die Wahrscheinlichkeit, dass die Kategorie ij auftritt entspricht p_{ij}. Erst bei einem ausreichend großen N_{ij} kann von einer Chi-Quadrat Verteilung mit $(m-1)(k-1)$ Freiheitsgraden ausgegangen werden [152]. Ein signifikanter Chi-Quadrat-Test teilt uns mit, dass die Verteilung der beobachteten Werte von den erwarteten Werten abweicht, sagt uns aber nicht, wo die Diskrepanz in der Kontingenztabelle liegt [160].

3.8.2 Exakter Fisher Test

Der exakte Test von Fisher ist ein statistischer Signifikanztest, der bei der Analyse von Kontingenztabellen verwendet wird. Er ist selbst für sehr kleine Stichproben oder Kategorien mit niedriger Frequenz anwendbar [161]. Daher wird dessen Anwendung empfohlen, wenn die Häufigkeit von 5 einer Kategorie unterschritten wird. Er ist für alle Strichprobengrößen gültig [162].

$$P(H_a = a) = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!} = \frac{\binom{a+c}{a}\binom{b+d}{b}}{\binom{n}{a+b}} \tag{3.16}$$

Die Buchstaben a,b,c,d entsprechen den jeweiligen Werten in der 2x2 Kontingenztabelle. n entspricht der Summe aller beobachteten Kategorien innerhalb der Tabelle. Ursprünglich war der exakte Fisher Test nur für 2x2 Kontingenztabellen ausgerichtet. Diese Formel ist unter 3.15 abgebildet [163].

3.8.3 Wilcoxon Rangsummen Test

Der Wilcoxon-Rangsummentest ist ein statistischer Test, der für nicht parametrische, ordinalskalierte Daten herangezogen werden kann. Dieser wurde von Frank Wilcoxon im Jahre 1945 entwickelt. Der Mann-Whitney-U-Test ist ein äquivalenter Test [164]. Der Wilcoxon Rangsummen Test wird in dieser Arbeit durch die Matlab Methode $ranksum(x, y)$ berechnet. W_R entspricht der Wilcoxon-Rangsumme.

$$W_R = \sum_{i=1} R(X_i) \tag{3.17}$$

$R(X_i)$ entspricht dem Rang der i-ten X in der gepoolten, geordneten Stichprobe[152]. Durch eine z-Statistik wird der p-Wert durch $ranksum$ angenähert. Die Variable μ_{W_R} entspricht dem Durchschnitt und σ_{W_R} der Varianz der berechneten Wilcoxon-Rangsummen.

$$z = \frac{W_R - \mu_{W_R}}{\sqrt{\sigma^2_{W_R}}} \tag{3.18}$$

Matlab nutzt zur Berechnung der z-Statistik folgende Formel:

$$z = \frac{W_R - [\frac{n_1 n_2 + n_1(n_1+1)}{2}] - 0.5 * sign(W_R - \mu_{W_R})}{\sqrt{\frac{n_1 n_2 (n_1+n_2+1-tiescor)}{12}}} \tag{3.19}$$

Wir nehmen an, dass zwei unabhängige Stichproben der Größen n_1 und n_2 vorliegen und $n_1 < n_2$. Die Matlab interne Formel $tiescor$ dient der Adjustierung der Rangsummen. Die Methode $sign$ gibt eine Matrix der gleichen Dimensionen des Inputs aus, welche je nach Wert der jeweiligen Elemente nur die Werte 1, 0, -1 annimmt, falls es sich bei keiner Zahl um eine komplexe Zahl handelt [165].

3.8.4 Zweistichproben T-Test

Zum Vergleich der Altersverteilungen zwischen den Gruppen wurde auch der T-Test im Rahmen dieser Arbeit verwendet. Der Zweistichproben-t-Test prüft mit Hilfe der Mittelwerte x hier dargestellt als: $\overline{x}_1$ und $\overline{x}_2$ zweier Stichproben, ob die Mittelwerte μ_1 und μ_2 der zugehörigen Grundgesamtheiten verschieden sind. Es wird angenommen, dass die Varianz beider Stichproben gleich ist. Zudem sind die untersuchten Daten metrisch verteilt. Der Zweistichproben T- Test kann mit folgender Formel dargestellt werden.

$$t = \frac{\overline{x}_1 - \overline{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{3.20}$$

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_1 - 1)s_2^2}{n_1 + n_2 - 2}} \tag{3.21}$$

Die Stichprobengröße der jeweiligen Gruppe entsprechen den Variablen n_1 und n_2. Die jeweiligen Stichprobenvarianzen werden mit dem Symbol s^2 dargestellt [152, 166].

3.8.5 Spearman Rangkorrelation

Die Rangkorrelationsanalyse nach Spearman kann den linearen Zusammenhang zweier mindestens ordinalskalierter Variablen berechnen. Diese kann auch bei Ausreißern reliabel angewendet werden [167]. Der Spearmankorrelationskoeffizient kann positive, sowie negative Korrelationen anzeigen. Bei Ersterem korrelieren hohe Werte der Variable A mit hohen Werten der Variable B.
Für eine Probe der Größe n, werden die n Rohwerte aus den zwei zu vergleichenden Variablen X_i und Y_i zu den Rängen rg_{X_i} und rg_{Y_i} konvertiert. Der Spearmankorrelationskoeffizient r_s oder ρ ist definiert als der Korrelationskoeffizient nach Pearson zwischen den Rangvariablen und wird demnach aus folgender Formel berechnet:

$$r_s = \rho_{rg_X, \mathrm{rg}_Y} = \frac{\mathrm{cov}(\mathrm{rg}_X, \mathrm{rg}_Y)}{\sigma_{\mathrm{rg}_X} \sigma_{\mathrm{rg}_Y}} \tag{3.22}$$

Mit σ_{rg_X} und σ_{rg_Y} werden die Standardabweichungen der Rangvariablen dargestellt [168]. Wir korrelieren die durch das Regressions-LSTM berechneten GAF/ GF Vorhersagen mit den tatsächlichen Werten mit Hilfe der Spearman Rangkorrelation. Bei den Ausgabewerten des LSTM Netzwerkes handelt es nicht um ganze Zahlen. Die tatsächlichen GAF / GF Werte sind ganzzahlig.

Kapitel 4

Ergebnisse

Für unsere Analyse haben wir die Daten aus 36 Monaten, die im Rahmen des PRONIA Projekts erhoben wurden, verwendet. Insgesamt 881 Datensätze sind im August 2020 zur Verfügung gestellt worden, nachdem leere und Datensätze ohne Einverständniserklärung entfernt worden waren. Davon konnten 16% dem Herkunftsort Köln zugeordnet werden. 310 der Datensätze gehörten zu ROP Patienten, 283 Datensätze zu ROD Patienten und 286 waren CHR Patienten. Dies ergab 879 Patientendaten mit eindeutiger Gruppenzugehörigkeit. Der Datensatz bestand aus 420 Frauen und 459 Männern. 34 Datensätze beinhalteten keine Angabe im Bezug auf das Geschlecht. Das Durchschnittsalter der Patienten belief sich auf 24,96 Jahre (SD = 5,80).

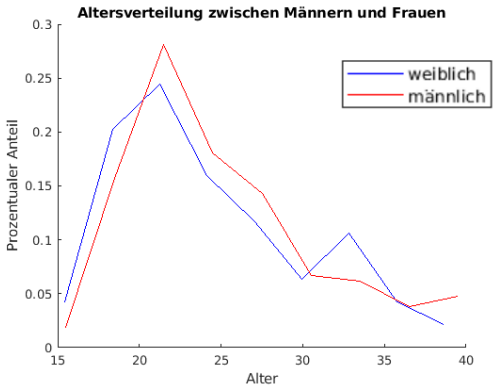

Abbildung 4.1: Die Abbildung zeigt die Altersverteilung bei Männern und Frauen im Test-/Trainingsdatensatz (n=398)

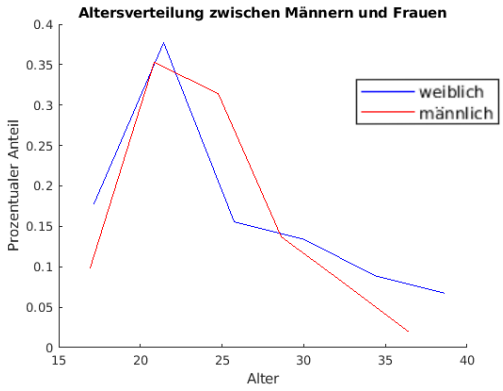

Abbildung 4.2: Die Abbildung zeigt die Altersverteilung bei Männern und Frauen im Validierungsdatensatz (n=96)

Etwas mehr Frauen (n_{CHR} = 147 N_{ROD} = 145) als Männer (n_{CHR} = 139, N_{ROD} = 138) gehörten der CHR Gruppe, sowie der depressiven Gruppe an, während mehr Männer mit

der Gruppenzugehörigkeit ROP (n_{Frauen} = 128, $N_{Männer}$ = 182) eingeschlossen worden waren. Insgesamt stammten 141 Datensätze aus dem Studienzentrum in Köln. Die restlichen 470 Datensätze stammten aus den anderen Studienzentren. Nachdem zu lückenhafte oder fehlerhafte Datensätze gemäß Abschnitt 3.7.1 entfernt worden waren, blieben insgesamt 96 Datensätze übrig, die in Köln rekrutiert wurden. Zur Entwicklung der LSTM Modelle konnten abschließend 398 Datensätze herangezogen werden, die aus neun unterschiedlichen Studienzentren stammten. Im Validierungsdatensatz befanden sich 40 ROD, 26 CHR und 36 ROP Patienten. Im Test-/Trainingsdatensatz waren 118 ROD, 140 CHR, und 140 ROP Patienten. Ein Chi-Quadrat Test zeigte, dass keine signifikanten Verteilungsunterschiede zwischen den beiden Datensätzen vorhanden waren (χ^2: FG = 2; Chi^2 Statistik = 4,65; P = 0,10). Die 96 Stichproben des Validierungsdatensatzes beinhalteten 51 Männer und 45 Frauen. Die Verteilung bezüglich der jeweiligen Häufigkeiten war verglichen mit dem Test-/Trainingsdatensatz nicht signifikant unterschiedlich. Mit 210 Männern und 188 Frauen belief sich der Qui-Quadrat Wert auf 0,01. und der p-Wert auf 0,95 (χ^2: FG = 1; N = 398; Chi^2 Statistik = 0,01; P = 0,95). Das Alter der Probanden unterschied sich im Validierungs-

		Test-/Trainingsdatensatz				
Bei Einschluss		Mittelwert	Standardabweichung	Median	Minimum	Maximum
GAF D/I	letztes Jahr	66,65	14,63	69	7	95
	letzten Monat	51,53	14,71	50	6	90
GAF S	letztes Jahr	65,38	14,76	65	6	99
	letzten Monat	49,81	14,06	51	5	91
GF R	bester Wert letztes Jahr	8,05	0,99	8	4	10
	aktueller Wert	5,70	1,82	6	1	10
GF S	bester Wert letztes Jahr	7,07	1,33	7	2	10
	aktueller Wert	6,20	1,42	6	2	10

Tabelle 4.1: Die Tabelle zeigt die GAF und GF Werte des Test-/Trainingsdatensatzes. N=398

und im Test-/ Trainingsdatensatz nicht signifikant voneinander (T-Test: t= 0,16; FG = 505; SD = 5,90; P = 0,88). Das durchschnittliche Alter in Test-/Trainingsdatensatz belief sich bei einer Standardabweichung von 6,02 auf 24,54. Der älteste Proband war 40,92 Jahre alt - der Jüngste 15,00. Im Validierungsdatensatz betrug das Durchschnittsalter 24,21 Jahre (n = 96; SD = 5,44; min = 17; max = 40,7). Die Durchschnitte des GAF D/I bei Einschluss beliefen sich auf $66{,}64_{letztesJahr}$ (n = 398; SD = 14,63; min = 7; max = 95) und $51{,}53_{letztenMonat}$ (n = 398; SD = 14,71; min = 6; max = 90), sowie auf $66{,}31_{letztesJahr}$ (n = 96; SD = 12,19; min = 31; max = 91) und $51{,}04_{letztenMonat}$ (n = 96; SD = 13,95; min = 21; max = 85) im Validierungsdatensatz. Die zugehörigen Wilcoxon Rangsummentests(WR) waren jeweils nicht statistisch signifikant (WR GAF D/I $_{letztesJahr}$: Z = 0,54; P = 0,59) - (WR GAF

D/I $_{letztenMonat}$: Z = -0,88; P = 0,38). Beim Vergleich der GAF S$_{letztesJahr}$ Validierungs- mit den Test-/Trainingsdaten waren die Durschnittswerte mit Werten von 63,51 (n = 96; SD = 12,31; min = 25; max = 90) und 65,38 (n = 398; SD = 14,76; min = 6; max = 99) ähnlich. Dementsprechend fand sich kein statistisch signifikanter Unterschied (WR GAF S$_{letztesJahr}$: Z = 1,40; P = 0,16). Die GAF S Werte, welche den Zeitraum innerhalb des letzten Monats bei Einschluss bewerten sollen, waren mit durchschnittlich 49,80 (n = 398; SD = 14,06; min = 5; max = 91) und 49,93 (n = 96; SD = 12,75; min = 15; max = 70) jeweils niedriger. Auch hier fand sich kein statistisch signifikanter Unterschied (WR GAF S$_{letztenMonat}$: Z = -0,54; P = 0,59). Auch bei dem GF R zeigten sich keine Unterschiede zwischen den Gruppen (WR GF R$_{besterWertletztesJahr}$:Z = 0,79; P = 0,43) - (WR GF R$_{aktuellerWert}$: Z = -0,37; P = 0,71). Die Standardabweichungen, der Median, sowie Minima und Maxima des GF R, aber auch des GF S und der GAF Werte können in Tabelle 4.1 und 4.2 eingesehen werden. Beim GF S verhielt es sich wie mit den vorhergehenden Metriken. Sowohl bei Betrachtung des besten Wertes innerhalb des letzten Jahres, (WR GF S $_{besterWertletztesJahr}$: Z = 1,33; P = 0,19) sowie bei Betrachtung des aktuell besten Wertes (WR GF $_{s_{aktuellerWert}}$: Z = 0,60; P = 0,55), zeigten sich keine statistisch signifikanten Unterschiede.

		Validierungsdatensatz					p-Wert
Bei Einschluss		Mittel-wert	Standard-abweichung	Median	Minimum	Maximum	Wilcoxon Rangtest
GAF D/I	letztes Jahr	66,31	12,19	65	31	91	0,56
	letzten Monat	52,49	13,96	55	21	81	0,32
GAF S	letztes Jahr	63,51	12,31	63	25	90	0,13
	letzten Monat	49,94	12,75	51	15	70	0,57
GF R	bester Wert letztes Jahr	7,12	1,15	7	2	9	0,46
	aktuell bester Wert	5,88	1,34	6	2	8	0,57
GF S	bester Wert letztes Jahr	6,85	1,32	7	3	9	0,18
	aktuell bester Wert	6,03	1,40	6	3	8	0,58

Tabelle 4.2: Die Tabelle zeigt die GAF und GF Werte des Validierungsdatensatzes bei Einschluss. Mittels eines Wilcoxon Rang Tests wurden die Daten mit dem Test- und Trainingsdatensatz verglichen (n=96).

4.1 Prädiktionen durch neuronale Netze

Bei allen Prädiktionen, welche im Rahmen dieser Arbeit präsentiert werden, optimierten wir die Hyperparameter mit Hilfe des Bayesianischen-Optimierungs-Algorithmus (siehe Material und Methoden 3.7.4). Die ermittelten Werte, werden bei den folgenden Modellen jeweils angegeben, um die Reproduzierbarkeit zu vereinfachen.
Theoretisch wäre es im Rahmen des implementierten Skriptes möglich LSTM Modelle für den GAF D/I, den GAF S, den GF R, sowie den GF S optimieren zu lassen. Dabei können als weitere Variablen der GF R negativ, GF S negativ, der BDI[1] Summenwert oder der PANSS[2] hinzugewählt werden. In dieser Arbeit zeigen wir die Ergebnisse der Modelle, welche lediglich die Funktionsniveauskalen verwendeten. Im Folgenden zeigen wir eine sinnvolle Auswahl möglicher Modelle.

4.2 Prädiktion mit LSTM Netzen

4.2.1 GF S

Klassifikation

Orientierend an der wissenschaftlichen Arbeit aus unserem Haus [107] wurde für die GF S Klassifikation ein Grenzwert von 7 herangezogen (siehe Material und Methoden). Folgende Einstellungen wurden im Rahmen der Bayesianischen Hyperparameter Optimierung gefunden:

- initiale Lernrate : 0,0013
- Größe der Minibatches : 13
- L2 Regularisierung: 2,0532e^{-5}
- Gradient Decay: 0,8537
- Anzahl LSTM Neurone: 14
- Anzahl konventioneller Neurone: 22

[1]Das Beck-Depressions-Inventar [169] ist ein psychologisches Testverfahren, das depressive Symptome erfasst. Dieser kann zwischen depressiven und nicht-depressiven Probanden unterscheiden. 1996 wurde die revidierte Version des BDI, der BDI-II veröffentlicht. Diese Kriterien wurden passend zu den DSM-IV-Kriterien umformuliert und verbessert [170].

[2]Die Positiv-Negativ-Syndrom-Skala (PANSS) ist ein Fragebogen, der zur Messung der Symptomschwere bei Patienten mit Schizophrenie verwendet wird und aus 30 Elementen besteht. Dieser wurde 1987 erstmals publiziert[171].

Die BAC über den gesamten Datensatz belief sich auf 85,04% (siehe Tabelle 4.3) und wich leicht von der Genauigkeit, die 86,04% betrug, ab. Der Algorithmus detektierte zukünftige schlechte GF S Werte mit einer guten Sensitivität von 81,25% (95% CI: 75,51% - 86,14%) und einer Spezifität von 88,83%(95% CI: 85,25% - 91,80%) (χ^2: FG=1, N = 609; Chi^2 Statistik = 298,55; P $<$ 0,00001). Die positive Likelihood Ratio (PLR) belief sich auf 7,27 (95% CI: 5,45 - 9,71) und der positiv prädiktive Wert (PPW) könnte mit 80,89% (95% CI: 76,03% - 84,96%) beziffert werden. Der negativ prädiktive Wert (NPW) war mit einem Wert von 89,06% (95% CI: 86,08% - 91,47%) etwas geringer. Die negative Likelihood Ratio (NLR) war 0,21 (95% CI: 0,16 - 0,28). Die Prävalenz einer schlechten Prognose lag bei einem Grenzwert von 7 bei 36,78% (95% CI: 32,94% - 40,75%).
Die Performance unterschied sich zwischen den Testgruppen. Die ROD Patienten wiesen eine Sensitivität von 84,26% (95% CI: 76,00% - 90,55%) und einer Spezifität von 91,76% (95% CI: 86,57% - 95,42%) auf (χ^2: FG=1, N = 278; Chi^2 Statistik = 162,40; P $<$ 0,00001). Eine schlechter GF S Wert war bei den ROD Patienten mit einer Prävalenz von 38,85% (95% CI: 32,05% - 43,75%) vorzufinden. Hieraus kann der PPW berechnet werden, der 86,67% (95% CI: 79,63% - 91,53%) betrug. Der NPW war mit 90,17% (95% CI: 85,54% - 93,43%) niedriger. Die Genauigkeit des Modells erzielte an den Validierungsdaten einen Wert von 88,85%. Die BAC betrug 88,01%. Bei den Probanden aus der ROD Gruppe konnte für die NLR ein Wert von 0,17 (95% CI: 0,11 - 0,27) und für die PLR ein Wert von 10,23 (95% CI: 6,15 - 17,01) ermittelt werden.

	alle Gruppen	ROD	CHR	ROP	N
über alle Zeitpunkte	85,04%	88,01%	82,48%	82,40%	
IV3	73,08%	75,96%	76,79%	67,50%	96
IV6	88,25%	94,23%	80,77%	88,28%	96
T1	91,83%	91,58%	87,50%	100,00%	92
IV12	83,01%	91,48%	75,83%	77,27%	84
IV15	91,05%	91,18%	93,33%	94,12%	82
T2	88,43%	87,73%	94,12%	87,06%	81
IV27	78,01%	82,06%	80,00%	65,00%	47
IV36	85,71%	83,33%	100,00%	83,33%	31

Tabelle 4.3: Die Tabelle zeigt die balancierte Korrektheit (balanced accuracy) Ergebnisse der GF S Prädiktionen. Die Zeilen zeigen die Zeitabschnitte, während die Spalten die Gruppen der Teilnehmenden definieren. Die BAC über alle Zeitpunkte und Gruppen war: 85,04%.

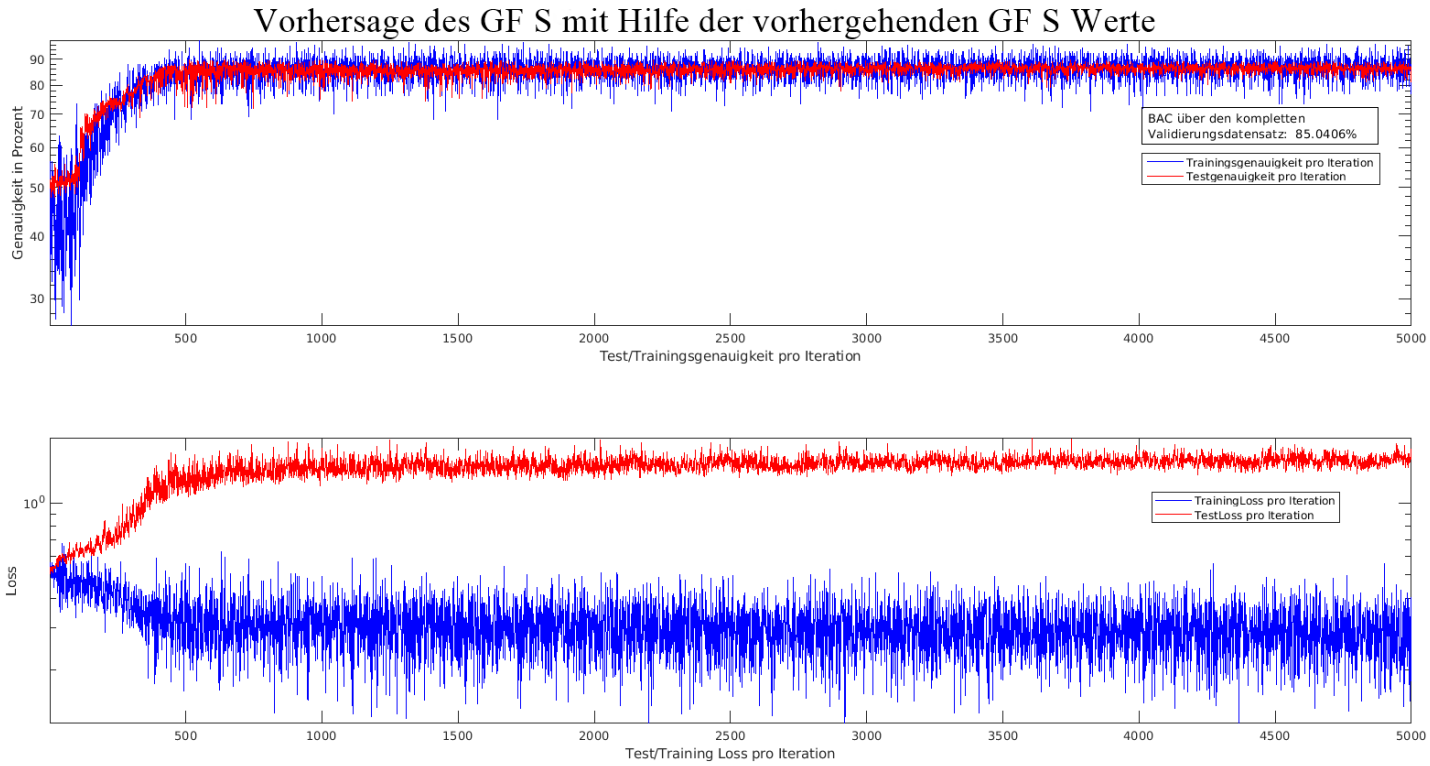

Abbildung 4.3: Die Abbildung zeigt den Trainingsverlauf des Neuronalen Netzes. In Blau ist die Genauigkeit (accuracy) der Trainingdaten im Verlauf zu erkennen. In Rot die Genauigkeit der Testdaten. Nach Beendigung des Trainings wurden die Werte der Validierunngsdaten prädiziert, die im Mittel in 85,04% (balanced accuracy) der Fälle korrekt eingeschätzt wurden.

Bei den CHR Patienten belief sich die Sensitivität auf 79,66% (95% CI: 67,17% - 89,02%), sowie die Spezifität auf 85,29%(95% CI: 76,91% - 91,53%) (χ^2: FG=1; N = 161; Chi^2 Statistik = 63,88; P $<$ 0,00001). Die Prävalenz lag bei 36,65% (95% CI: 29,20% - 44,59%). Für den PPW und den NPW konnten die Werte 75,81% (95% CI 65,86% - 83,57%) und 87,88% (95% CI: 81,30% - 92,36%) ermittelt werden. Die PLR ist 5,42 (95% CI: 3,34 - 8,80) und die NLR betrug 0,24 (95% CI: 0,14 - 0,40). Unter den ROD Patienten zeigte sich für die Genauigkeit ein Wert von 83,23%. Die BAC konnte mit einem Wert von 82,48% beziffert werden.
Bei den ROP Patienten konnte der Algorithmus mit einer Sensitivität von 77,19% (95% CI: 64,16% - 87,26%) ein schlechtes Outcome vorhersagen. Die Spezifität bei psychotischen Patienten ergab einen Wert von 87,61% (95% CI: 80,09% - 93,06%). Ein Chi-Quadrat Test zeigte auch hier, dass das GF S LSTM Modell keine zufälligen Entscheidungen traf (χ^2: FG=1; N = 170; Chi^2 Statistik = 70,79; P $<$ 0,00001). Die Prävalenz war mit 33,53% (95% CI: 26,48% - 41,16%) etwas niedriger. Auch hier war der NPW mit 88,39% (95% CI: 82,46% - 92,50%) höher als der PPW 75,86% (95% CI: 65,36% - 83,96%). Die BAC Werte des Validierungsdatensatzes können in Tabelle 4.3 eingesehen werden. Für die Probanden aus der ROP Gruppe betrug die BAC innerhalb dieses Modells 82,40%. Die Genauigkeit belief sich auf 84,12%. Die PLR betrug 6,23 (95% CI: 3,74 - 10,38) und die NLR konnte mit einem Wert von 0,26 (95% CI: 0,16 - 0,42) beziffert werden. Abbildung 4.3 zeigt den Trainingsverlauf des LSTM Modells, welches für die beschriebenen Ergebnisse herangezogen wurde.

Regression

Bei der Regression wurde jeweils versucht den möglichst genauen GF S Wert zum nächsten Zeitpunkt zu prädizieren. Zur Berechnung der Güte des ermittelten LSTM Modells wurde hierfür der RMSE (siehe Material und Methoden 3.7.6) herangezogen. Da hierfür ein erneutes LSTM Modell trainiert werden muss, wurden die Hyperparameter mit Hilfe der Bayesianischen Optimierung folgendermaßen ausgewählt:

- initiale Lernrate : 0,0013
- Größe der Minibatches : 13
- L2 Regularisierung: 2,0532e^{-5}
- Gradient Decay: 0,8537
- Anzahl LSTM Neurone: 14
- Anzahl konventioneller Neurone: 22

Der RMSE betrug bei Betrachtung aller Gruppen und Zeitpunkte 0,72 (siehe Tabelle 4.4). Zur Berechnung wie stark die prädizierten und die tatsächlichen Werte miteinander korrelieren, wurde der Rangkorrelationskoeffizient nach Spearman berechnet. Dieser betrug S_ρ= 0,81 (N=609; P $<$ 0,00001) (siehe Abbildung 4.4).

	alle Gruppen	ROD	CHR	ROP
über alle Zeitpunkte	0,72	0,69	0,73	0,72
IV3	0,98	0,96	1,07	0,91
IV6	0,74	0,57	1,01	0,67
T1	0,72	0,80	0,66	0,67
IV12	0,60	0,61	0,61	0,58
IV15	0,61	0,71	0,59	0,43
T2	0,57	0,62	0,60	0,44
IV27	0,72	0,40	0,81	1,20
IV36	0,55	0,48	0,79	0,44

Tabelle 4.4: Die Tabelle zeigt die RMSE Ergebnisse des GF S Regressions Modells. Die Zeilen zeigen die Zeitabschnitte, während die Spalten die Gruppen der Teilnehmenden definieren. Der RMSE über alle Zeitpunkte und Gruppen war: 0,72.

Die Area under the curve (AUC) des GF S Regressions Modells betrug 0,90 (siehe Abbildung 4.5). Post-hoc wurden die Ergebnisse der Regression in gute ($\geq$ 7) und in schlechte GF S ($<$7) Prognosen unterteilt. Das sich hieraus ergebende Modell konnte auf dem Validierungsdatensatz, welches nur Daten aus dem Uniklinikum Köln beinhaltete, eine Sensitivität von 88,84% (95% CI: 83,97%- 92,65%) und eine Spezifität von 80,52% (95% CI: 76,20%- 84,36%) erreichen. Die BAC belief sich somit auf 84,68% und die Genauigkeit auf 83,58%. Die jeweiligen BAC Werte, über die Zeitabschnitte und die Probandengruppen hinweg, können in Tabelle 4.5 eingesehen werden. Der Positiv-Prädiktive-Wert (PPW) der Post-Hoc Klassifikation konnte mit 72,63% (95% CI: 68,30% - 76,57%) beziffert werden, während die PLR 4,56 (95% CI: 3,70 - 5,62) betrug. Der Negativ-Prädiktive-Wert (NPW) war mit 92,54% (95% CI: 89,52% - 94,74%) höher als der PPW. Die NLR betrug 0,14 (95% CI: 0,10 - 0,20). Die Prävalenz einer schlechten Prognose lag bei einem Grenzwert von 7 bei 36,78% (95% CI: 32,94% - 40,75%). Ein Chi-Quadrat Test zeigte, dass die Prädiktionen des LSTMs nicht zufällig waren (χ^2: FG=1; N = 609; Chi^2 Statistik = 275,25; P $<$ 0,00001).

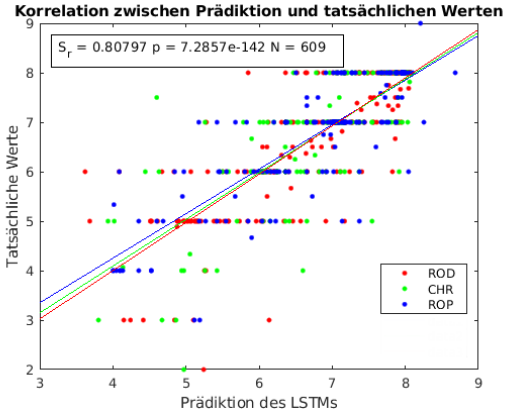

Abbildung 4.4: Der Graph zeigt die die Ergebnisse der Korrelation nach Spearman, wenn tatsächliche Werte und Prädiktionen des GF S LSTMs auf der x und y Achse gegenübergestellt werden. S_ρ wurde mit 0,81 berechnet. Der p-Wert belief sich auf $7{,}29e^{-142}$.

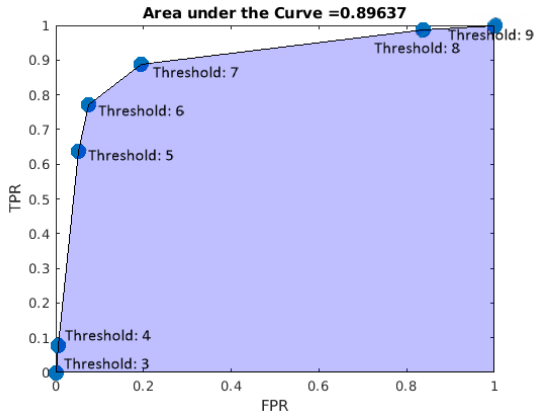

Abbildung 4.5: Die Abbildung zeigt die ROC Kurve des GF S LSTM Regressionsmodells. Anhand verschiedener Grenzwerte wurde jeweils die TPR und die FPR berechnet und auf den Achsen aufgetragen. Die AUC ist 0,90.

Bei den ROD Patienten war die BAC mit 88,19%, aber auch die Genauigkeit mit 87,41% höher. Die Sensitivität rangierte mit 91,67% (95% CI: 84,77% - 96,12%) oberhalb der 90% Marke. Die Prävalenz eines niedrigen GF S betrug unter den ROD Patienten im Validierungsdatensatz 38,85% (95% CI: 33,09% - 44,85%). Hieraus können der PPW und der NPW berechnet werden, die 79,20% (95% CI: 72,69% - 84,49%) und 94,12% (95% CI: 89,51% - 96,77%) betrugen. Die PLR und die NLR beliefen sich auf 5,99 (95% CI: 4,19 - 8,58) und 0,10 (95% CI: 0,05 - 0,18). Ein hoch signifikanter Chi-Quadrat Test belegte, dass es sich um keine zufälligen Vorhersage von Seiten des Modells handelt (χ^2: FG=1; N = 278; Chi^2 Statistik = 155,66; P $<$ 0,00001).
Sowohl eine leicht niedrigere Genauigkeit (78,88%) als auch BAC (80,48%) lag bei den CHR Patienten vor. Die Sensitivität war mit 86,44% (95% CI: 75,02% - 93,96%) höher als die Spezifität bei 74,51% (95% CI: 64,92% - 82,62%). Der NPW war mit 90,48% (95% CI: 83,16% - 94,81%) deutlich höher als der PPW, der 66,23% (95% CI: 58,10% - 73,51%) betrug. Die PLR war 3,39 (95% CI: 2,40 - 4,80) und die NLR 0,18 (95% CI: 0,09 - 0,35). Die Chi-Quadrat Test war erneut hoch signifikant (χ^2: FG=1; N = 161; Chi^2 Statistik = 55,65; P $<$ 0,00001). Bei den ROP Patienten war die Sensitivität mit 85,96% (95% CI: 74,21% - 93,74%) ebenfalls höher als die Spezifität mit 79,65% (95% CI: 71,04% - 86,64%). Die BAC und die Genauigkeit betrugen 82,80% und 81,76%. Aus einer Prävalenz von 33,53% (95% CI: 26,48% - 41,16%) ergab sich ein PPW von 68,06% (95% CI: 59,31% - 75,69%) und ein NPW von 91,84% (95% CI: 85,46% - 95,56%). Die PLR konnte man auf 4,22 (95% CI: 2,89 - 6,17) beziffern. Die NLR war mit 0,18 (95% CI: 0,09 - 0,34) niedrig. Ein Chi-Quadrat Test

zeigte auch bei ROP Patienten einen hoch-signifikanten p-Wert (χ^2: FG=1; N = 170; Chi^2 Statistik = 64,14; P $<$ 0,00001).

	alle Gruppen	ROD	CHR	ROP
über alle Zeitpunkte	84,68%	88,19%	80,48%	82,80%
IV3	75,17%	81,84%	77,38%	65,00%
IV6	87,89%	90,38%	80,77%	92,10%
T1	87,85%	90,35%	80,00%	94,74%
IV12	85,51%	92,32%	73,33%	86,36%
IV15	86,17%	89,36%	86,67%	85,24%
T2	88,86%	86,52%	91,18%	94,12%
IV27	81,48%	89,12%	80,00%	65,00%
IV36	81,55%	77,08%	100,00%	83,33%

Tabelle 4.5: Die Tabelle zeigt die balancierte Korrektheit (balanced accuracy) Ergebnisse der GF S Vorhersagen. Die Regressionsergebnisse wurden post hoc in gute und in schlechte Prognosen (<7) unterteilt. Die Zeilen zeigen die Zeitabschnitte, während die Spalten die Gruppen der Teilnehmenden definieren. Die BAC über alle Zeitpunkte und Gruppen war: 84,68%.

Rekursives Modell

Das rekursive Modell verwendete lediglich die erhobenen GF S Werte bei T0 um eine Vorhersage bis IV36 zu treffen.

	alle Gruppen	ROD	CHR	ROP
über alle Zeitpunkte	0,93	0,95	0,86	0,99
IV3	0,98	0,96	1,07	0,91
IV6	1,05	0,82	1,39	0,95
T1	1,10	1,04	1,14	1,12
IV12	1,08	1,07	1,29	0,89
IV15	1,02	1,16	1,10	0,78
T2	1,10	1,16	1,24	0,90
IV27	1,23	1,27	1,40	1,03
IV36	0,80	1,03	0,69	0,66

Tabelle 4.6: Die Tabelle zeigt die RMSE Ergebnisse des rekursiven GF S Modells. Die Zeilen zeigen die Zeitabschnitte, während die Spalten die Gruppen der Teilnehmenden definieren. Der RMSE über alle Zeitpunkte und Gruppen war: 0,93.

Mittels des Spearman Korrelationskoeffizienten wurde ein $S\rho$ von 0,57 ermittelt (N=609; P <0,00001). Der RMSE über alle Zeitpunkte war mit 0,93 höher (siehe Tabelle 4.6) und die AUC mit 0,78 niedriger (siehe Abbildung 4.6) als bei dem vorhergehenden Modell. Nachträglich wurden die Regressionsergebnisse zwischen guten ($\geq$ 7) und schlechten (<7) GF S Werten separiert. Danach konnte eine Sensitivität von 75,89% (95% CI: 69,75% - 81,35%) eine Spezifität von 65,71% (95% CI: 60,74% - 70,45%), ein PPW von 56,29% (95% CI: 52,40% - 60,10%) und eine PLR von 2,21 (95% CI: 1,89 - 2,59) ermittelt werden. Dem gegenüber belief sich der NLR und der NPW auf 0,37 (95% CI: 0,29-0,47) und 82,41% (95% CI: 78,60% - 85,66%). Die Prävalenz eines eines schlechten GF S Wertes betrug 36,78% (95% CI: 32,94% - 40,75%). Die BAC dieses Modells bezifferte, wie auch in Tabelle 4.7 eingesehen werden kann über alle Gruppen und Zeitpunkte eine BAC von 70,80%. Die Genauigkeit war mit einem Wert von 69,46% etwas niedriger. Der Chi-Quadrat Test zeigt einen signifikante Ergebnisse bei den vorhersagen des Modells (χ^2: FG=1; N = 609; Chi^2 Statistik = 98,07; P < 0,00001). Es handelte sich somit um keine zufälligen Prädiktionen.

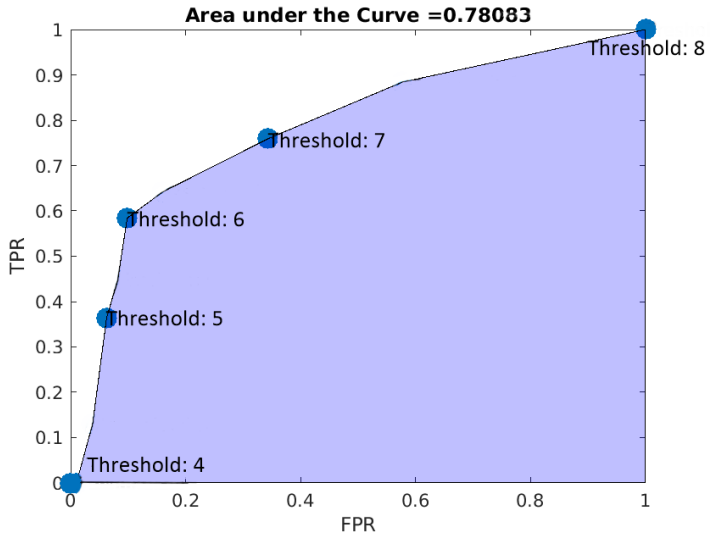

Abbildung 4.6: Die Abbildung zeigt die ROC Kurve des rekursiven GF S LSTM Modells. Es wurden mit verschiedenen Grenzwerten die TPR, sowie FPR berechnet und so eine dem Regressionsmodell zugehörige ROC Kurve gebildet. Die AUC war 0,78.

	alle Gruppen	ROD	CHR	ROP	N
über alle Zeitpunkte	70,80%	75,36%	64,16%	70,00%	
IV3	75,17%	81,84%	77,38%	65,00%	
IV6	70,92%	75,55%	65,38%	72,49%	
T1	69,78%	75,41%	63,33%	66,45%	
IV12	71,88%	73,86%	55,00%	83,92%	
IV15	67,34%	71,01%	65,71%	66,47%	
T2	66,36%	67,42%	69,41%	66,47%	
IV27	68,23%	74,12%	46,67%	75,00%	
IV36	82,44%	93,75%	70,00%	83,33%	

Tabelle 4.7: Die Tabelle zeigt die balancierte Korrektheit (balanced accuracy) Ergebnisse der GF S Prädiktionen. Die Zeilen zeigen die Zeitabschnitte, während die Spalten die Gruppen der Teilnehmenden definieren. Die BAC über alle Zeitpunkte und Gruppen war: 70,80%.

Die Sensitivität und die Spezifität rangierte bei den ROD Patienten bei 77,78% (95% CI: 68,76% - 85,21%) und 72,94% (95% CI: 65,61% - 79,46%). Die Genauigkeit der ROD Vorhersagen des rekursiven Modells betrug 74,82% und die BAC 75,36%. Die Prävalenz schlechter GF S Werte war 38,85% (95% CI: 31,01%- 42,65%) und entsprach auch der vorbeschriebenen Prävalenz bei ROD Patienten im Rahmen der GF S Analyse. Hieraus konnte der PPW mit einem Wert von 64,62% (95% CI: 58,31% - 70,45%) und der NPW mit 83,78% (95% CI: 78,21% - 88,15%) berechnet werden. Die PLR und die NLR lagen bei 2,87 (95% CI: 2,20 - 3,75) und 0,30 (95% CI: 0,21 - 0,44) (Chi^2: FG=1; N = 278; Chi^2 Statistik = 68,24; P $<$ 0,00001). Bei den CHR Patienten konnte die Genauigkeit der GF S Analyse auf 62,73% und die BAC auf 64,16% (siehe Tabelle 4.7) beziffert werden. Die Sensitivität war mit 69,49% (95% CI: 56,13% - 80,81%) höher als die Spezifität bei 58,82% (95% CI: 48,64% - 68,48%). Der NPW betrug 76,92% (95% CI: 68,70% - 83,51%) und der PPW belief sich auf 49,40% (95% CI: 42,28% - 56,54%). Die NLR war mit 0,52 (95% CI: 0,34 - 0,79) hoch, aber niedriger als die PLR mit 1,69 (95% CI: 1,27 - 2,25). Auch hier zeigte ein Chi-Quadrat Test statistisch signifikante Ergebnisse (χ^2: FG=1; N = 161; Chi^2 Statistik = 12,00; P $<$ 0,00001). Die ROP Patienten zeigten mit einer BAC von 70,00% und einer Genauigkeit von 67,06% bessere Ergebnisse als die Probanden der CHR Gruppe. Die Sensitivität betrug innerhalb dieser Subpopulation 78,95% (95% CI: 66,11% - 88,62%) und die Spezifität 61,06% (95% CI: 51,44% - 70,09%). Der PPW lag bei 50,56% (95% CI: 43,92% - 57,19%), während der NPW mit einem Wert von 85,19% (95% CI: 77,30% - 90,66%) höher war. Die NLR bezifferte 0,34 (95% CI: 0,20 - 0,58) und die PLR 2,03 (95% CI: 1,55 - 2,65). Der Chi-Quadrat Test war statistisch signifikant (Chi^2: FG=1; N = 170; Chi^2 Statistik = 24,31; P $<$ 0,00001).

4.2.2 GF R

Klassifikation

Wie auch bei der vorhergehenden GF S Klassifikation wird hier ein Grenzwert von 7 herangezogen. Folgende Einstellungen sind im Rahmen der Bayesianischen Hyperparameter Optimierung gefunden worden:

- initiale Lernrate : 0,0056322
- Größe der Minibatches : 8
- L2 Regularisierung: 5,3013e^{-8}
- Gradient Decay: 0,81505
- Anzahl LSTM Neurone: 50
- Anzahl konventioneller Neurone: 51

Das GF R LSTM Klassifikations Modell konnte im Köln Datensatz eine Sensitivität von 85,77% (95% CI: 80,69% - 89,94%) erzielen. Die Spezifität war mit 79,73% (95% CI: 75,27% - 83,71%) knapp unter 80%. Die PLR und die NLR beliefen sich jeweils auf 4,23 (95% CI: 3,43 - 5,21) und 0,18 (95% CI: 0,13 - 0,24). Der PPV betrug 73,21% (95% CI: 68,82% - 77,19%), während der NPV mit einem Wert von 92,07% (95% CI: 89,46% to 94,08%) über der 90% Schwelle eingeordnet werden konnte. Der PPW und der NPW bezifferte sich auf 76,99% (95% CI: 69,93% - 82,80%) und 90,91% (95% CI: 86,19% - 94,13%). In der Validierungskohorte war ein schlechter GF R in 39,24% (95% CI: 35,34% - 43,25%) der Fälle vorhanden (95% CI: 30,23% - 37,32%). Die BAC des Modells ergab einen Wert von 82,75%. Die Genauigkeit war bei 83,58% etwas höher. Somit zeigte sich ein hoch signifikantes Ergebnis im Chi-Quadrat Test (χ^2: FG=1; N = 609; Chi^2 Statistik = 250,84; P $<$ 0,00001). Ähnlich sensitiv zeigte sich das GF R LSTM Modell bei den ROD Patienten mit einer Sensitivität von 85,29% (95% CI: 76,91% - 91,53%). Die Spezifität war mit 85,23% (95% CI: 79,11%- 90,12%) geringfügig niedriger. Unter den ROD Patienten zeigte sich zudem, dass ein günstiger GF R Wert etwas prävalenter war als in der gesamten Kohorte. So kamen GF R Werte, welche kleiner als 7 waren in 36,69% (95% CI: 31,01%- 42,65%) der Fälle vor. Der PPW und der NPW betrugen 76,67% (95% CI: 69,24% - 82,75%), sowie 87,32% (95% CI: 78,63% - 92,80%). Für die Genauigkeit des Modells konnte ein Wert von 81,37% ermittelt werden. Die BAC belief sich auf 85,26%. Der Chi - Quadrat Test zeigt, dass auch hier das Modell keine zufälligen Entscheidungen traf (χ^2: FG = 1; N = 278; Chi^2 Statistik = 133,12; P $<$ 0,00001). Die BAC Werte der jeweiligen Gruppen und Vorhersagezeitpunkte können in Tabelle 4.8 eingesehen werden.

	alle Gruppen	ROD	CHR	ROP	N
über alle Zeitpunkte	82,75%	85,26%	81,58%	78,91%	
IV3	72,03%	76,73%	72,02%	65,84%	96
IV6	79,93%	82,76%	83,75%	79,55%	96
T1	88,34%	89,81%	85,71%	85,00%	92
IV12	84,90%	87,50%	78,33%	85,56%	84
IV15	84,71%	84,66%	95,45%	75,89%	82
T2	85,49%	89,91%	77,50%	85,42%	81
IV27	83,27%	86,36%	81,67%	75,00%	47
IV36	86,61%	90,00%	NaN	87,50%	31

Tabelle 4.8: Die Tabelle zeigt die balancierte Korrektheit (balanced accuracy) Ergebnisse der GF R Prädiktionen. Die Zeilen zeigen die Zeitabschnitte, während die Spalten die Gruppen der Teilnehmenden definieren. Die BAC über alle Zeitpunkte und Gruppen betrug : 82,75%.

Auch die Sensitivität der CHR Prädiktionen war mit 88,46% (95% CI: 79,22% - 94,59%) über der 80% Marke anzusiedeln. Für die Spezifität konnte ein Wert von 74,70% (95% CI: 63,96% - 83,61%) ermittelt werden. Dies führte bei den CHR Patienten zu einem PPW von 76,67% (95% CI: 69,24% - 82,75%) und einem NPW von 87,32% (95% CI: 78,63% - 92,80%). Innerhalb dieser Kohorte sind mit 48,45% (95% CI: 40,51% - 56,44%) niedrige GF R Werte am wahrscheinlichsten. Die Genauigkeit des LSTM Modells belief sich bei den CHR Patienten auf 81,37%, die BAC auf 81,58%. (χ^2: FG = 1; N = 161; Chi^2 Statistik = 62,53; P $<$ 0,00001).
Die BAC der ROP Patienten betrug 78,91% und war somit höher als die Genauigkeit, für die man einen Wert von 77,65% ermittelte. Die Sensitivität und die Spezifität betrugen 83,05% (95% CI: 71,03% - 91,56%) und 74,77% (95% CI: 65,65% - 82,54%). Niedrigere GF R Werte waren bei einer Prävalenz von 34,71% (95% CI:27,58% - 42,37%) in dieser Kohorte verhältnismäßig selten vertreten. Sowohl der PPW bei 63,64% (95% CI: 55,46% - 71,10%), als auch der NPW mit 89,25% (95% CI: 82,36% - 93,65%) war bei den ROP Patienten niedriger als bei den anderen Probanden der GF R Analyse. Die NLR lag bei 0,23 (95% CI: 0,13 - 0,04) und die PLR bei 3,29 (95% CI: 2,34 - 4,63). Auch hier ergaben sich bei einem Chi-Quadrat Test signifikante Ergebnisse (χ^2: FG=1; N = 170; Chi^2 Statistik = 49,68; P $<$

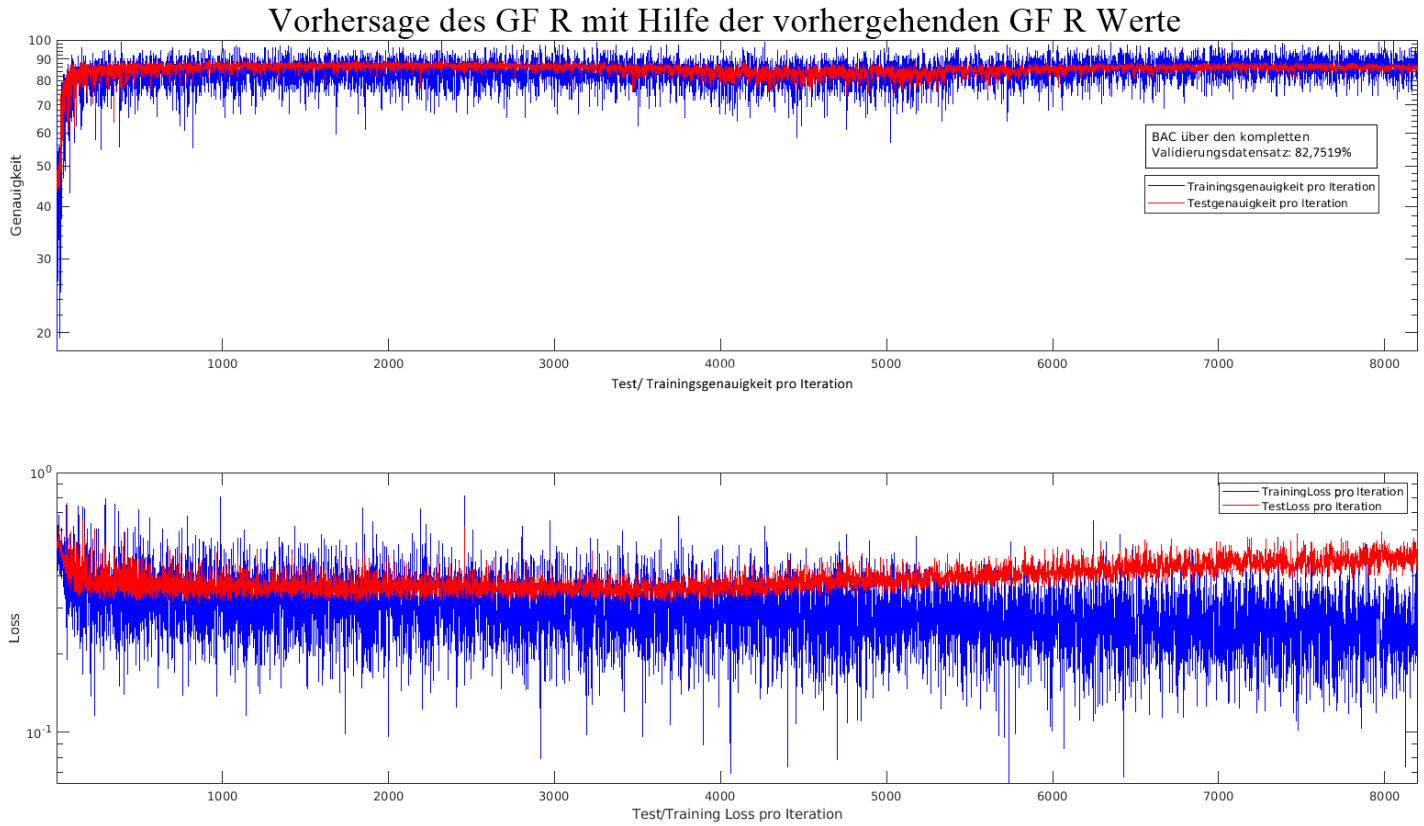

Abbildung 4.7: Die Abbildung zeigt den Trainingsverlauf des Neuronalen Netzes. In Blau ist die Genauigkeit (accuracy) der Trainingdaten im Verlauf zu erkennen. In Rot die Genauigkeit der Testdaten. Nach Beendigung des Trainings wurden die Werte der Validierungsdaten prädiziert, die im Mittel in 82,75% (balanced accuracy) der Fälle korrekt eingeschätzt wurden.

Regression

Bei der GF R Regression konnte in der Post-Hoc Korrelationsanalyse ein Spearmans ρ von 0,81 erzielt werden, was darauf hindeutet, dass eine starke Korrelation zwischen tatsächlichen Werten und Vorhersagen des LSTMs vorhanden ist. Dementsprechend zeigte sich ein hoch signifikanter p-Wert (siehe Abbildung 4.8) S_ρ=(0,81; N = 609; P $<$ 0,00001). Für das Regressions-Modell wurden folgende Hyperparameter gefunden:

- initiale Lernrate : 4,8266e^{-4}
- Größe der Minibatches : 40
- L2 Regularisierung: 6,7849e^{-10}
- Gradient Decay: 0,9402
- Anzahl LSTM Neurone: 13
- Anzahl konventioneller Neurone: 29

Nach einer 5-fachen Kreuzvalidierung konnte am Validierungsdatensatz ein Gesamt RMSE von 0,76 ermittelt werden. Die jeweiligen RMSE Werte je nach Probandengruppen und Zeitpunkte kann in Tabelle 4.9 eingesehen werden. Nachträglich wurden anhand verschiedener Grenzwerte die TPR und die FPR berechnet und so die ROC Kurve aufgetragen, welche in Abbildung 4.9 einzusehen ist. Daraus ergab sich eine AUC von 0,91 und weist somit auf ein exzellentes Modell hin. Wir haben einen Grenzwert von 7 verwendet, um nachträglich die Vorhersagen eines tatsächlichen guten (GF R $\geq$7)/schlechten ($<$ 7) Verlaufes zu vergleichen. Hierbei ergab sich eine BAC von 79,28% und eine Genauigkeit von 76,35%. Die Sensitivität und die Spezifität betrugen 92,89% (95% CI: 88,86 - 95,80%) und 65,68% (95% CI: 60,58% - 70,51%). Die Prävalenz eines niedrigen GF R ($<$7) konnte mit 39,24% (95% CI: 35,34% - 43,25%) beziffert werden. Die PLR belief sich auf 2,71 (95% CI: 2,34 - 3,13), während die NLR 0,11 (95% CI: 0,07 - 0,17) betrug. Der NPW war mit 93,49% (95% CI: 89,99% - 95,79%) höher als der PPW 63,61% (95% CI: 60,19% - 66,90%). Über dem kompletten Validierungsdatensatz konnte ein Chi-Quadrat Wert von 203,55 ermittelt werden, sowie ein hoch signifikanter p-Wert (χ^2: FG = 1; N = 609; Chi^2 Statistik = 203,55; P $<$ 0,00001).
Die Sensitivität und die Spezifität rangierten bei den ROD Patienten bei 91,18% (95% CI: 83,91% - 95,89%) und 73,86% (95% CI: 66,72% - 80,19%). Die Genauigkeit der ROD Vorhersagen betrug 80,22% und die BAC 82,52%. Die Prävalenz schlechter GF R Werte war 36,69% (95% CI: 31,01%- 42,65%). Hieraus konnte der PPW mit einem Wert von 66,91% (95% CI: 61,02% - 72,30%) und der NPW mit 93,53% (95% CI: 88,50% - 96,44%) berechnet werden. Die PLR und die NLR lagen bei 3,49 (95% CI: 2,70 - 4,50) und 0,12 (95% CI:

0,06 - 0,22). Ein Chi-Quadrat Test wies statistisch signifikante Ergebnisse hin (χ^2: FG=1; N = 278; Chi^2 Statistik = 109,27; P $<$ 0,00001).

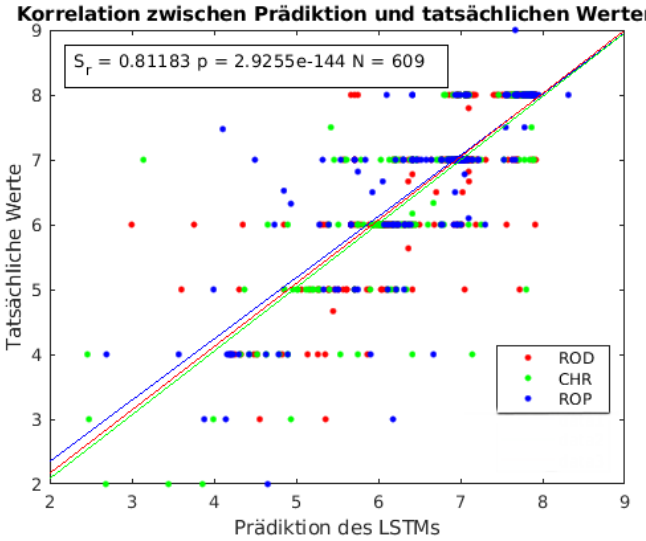

Abbildung 4.8: In der Grafik können die Ergebnisse der Korrelationsanalyse nach Spearman, sowie die zugehörigen Regressionslinien eingesehen werden. Das stufenartige Erscheinungsbild der Grafik ergibt sich aus der ordinalen Verteilung der GF R Skala. S_ρ wurde mit 0,81 berechnet. Der p-Wert belief sich auf 2,93e^{-144}.

	alle Gruppen	ROD	CHR	ROP
über alle Zeitpunkte	0,76	0,68	0,74	0,76
IV3	1,09	1,04	1,14	1,12
IV6	0,82	0,54	0,92	1,02
T1	0,65	0,54	0,67	0,76
IV12	0,73	0,77	0,88	0,48
IV15	0,45	0,44	0,38	0,51
T2	0,58	0,59	0,60	0,56
IV27	0,60	0,63	0,37	0,75
IV36	0,96	0,71	1,64	0,72

Tabelle 4.9: Die Tabelle zeigt die RMSE Werte der LSTM GF R Prädiktionen. Die Zeilen zeigen die Zeitabschnitte, während die Spalten die Gruppen der Teilnehmenden definieren. Der RMSE über alle Zeitpunkte und Gruppen war 0,76.

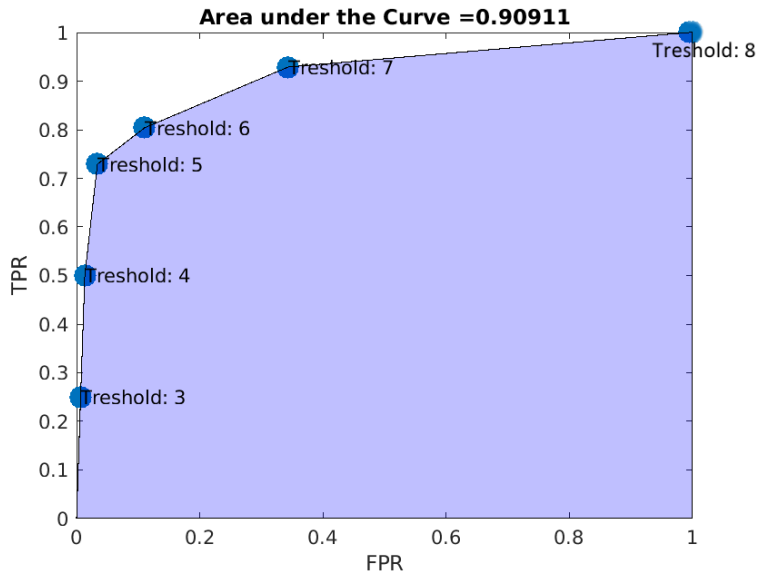

Abbildung 4.9: Die Abbildung zeigt die ROC Kurve des LSTM Modells. Es wurden mit verschiedenen Grenzwerten die TPR, sowie FPR berechnet und so eine dem Regressionsmodell zugehörige ROC Kurve gebildet. Die AUC belief sich auf 0,91.

Die Vorhersagen der CHR Patienten waren mit einem Wert von 94,87% (95% CI: 87,39% - 98,59%) sehr sensitiv und detektierten fast alle schlechten GF R Verläufe. Währenddessen belief sich die Spezifität auf 61,45% (95% CI: 50,12% - 71,93%). Somit ergab sich eine Gesamt BAC von 78,16%. Die BAC Ergebnisse nach Zeitpunkten und Gruppen können in Tabelle 4.9 eingesehen werden. Auch hier zeigte ein Chi-Quadrat Test statistisch signifikante Ergebnisse (χ^2: FG=1; N = 161; Chi^2 Statistik = 54,23; P $<$ 0,00001).
Bei den ROP Patienten war die Sensitivität mit 93,22% (95% CI: 83,54% - 98,12%) ebenfalls hoch. Die Spezifität betrug 55,86% (95% CI: 46,12% - 65,27%). Aus der Prävalenz von 34,71% (95% CI: 27,58% - 42,37%), konnte ein PPW von 52,88% (95% CI: 47,38% - 58,32%) und ein NPW von 93,94% (95% CI: 85,57% - 97,59%), sowie die Genauigkeit berechnet werden. Diese belief sich auch 68,82%. Die BAC war mit 74,54% in der ROP Gruppe am niedrigsten. Auch hier zeigten sich statistisch signifikante Ergebnisse (χ^2: FG=1; N = 170; Chi^2 Statistik = 37,03; P $<$ 0,00001).

	alle Gruppen	ROD	CHR	ROP
über alle Zeitpunkte	79,28%	82,52%	78,16%	74,54 %
IV3	72,03%	76,73%	72,02%	65,84%
IV6	76,44%	81,03%	76,88%	72,73%
T1	82,15%	81,02%	82,14%	82,35%
IV12	77,96%	79,26%	79,17%	74,44%
IV15	82,61%	85,51%	81,82%	78,57%
T2	80,69%	89,13%	70,00%	78,13%
IV27	85,81%	89,20%	91,67%	66,67%
IV36	82,44%	82,86%	NaN	87,50%

Tabelle 4.10: Die Tabelle zeigt die balancierte Genauigkeit (balanced accuracy) nachdem die GF R Regressionsvorhersagen post hoc in gute ($\geq$ 7) und schlechte Prognosen ($<$7) eingeteilt wurden . Die Zeilen zeigen die Zeitabschnitte, während die Spalten die Gruppen der Teilnehmenden definieren. Die BAC über alle Zeitpunkte und Gruppen war: 79,28%.

Rekursives Modell

Wie in 3.7.8 in Material und Methoden beschrieben, können mit den alleinigen Werten zum Einschluss unter Verwendung des rekursiven Modells Prädiktionen bis IV36 gemacht werden. Es sind die Einstellungen des obigen GF R Regressions Modell herangezogen worden. Hierbei ergibt sich ein S_ρ von 0,51 S_ρ=(0,51; N=609; P $<$ 0,00001). Dementsprechend ist der Gesamt RMSE mit 1,12 auch höher. Die RMSE Ergebnisse über alle Zeitpunkte und Gruppen können in Tabelle 4.11 eingesehen werden.

	alle Gruppen	ROD	CHR	ROP
über alle Zeitpunkte	1,12	1,06	1,08	1,21
IV3	1,09	1,04	1,14	1,12
IV6	1,23	0,97	1,33	1,40
T1	1,13	0,91	1,36	1,11
IV12	1,07	1,08	1,16	0,98
IV15	1,15	1,11	1,20	1,14
T2	1,21	0,99	1,38	1,27
IV27	1,34	1,25	1,46	1,30
IV36	1,30	1,22	1,01	1,66

Tabelle 4.11: Die Tabelle zeigt die RMSE Ergebnisse des rekursiven GF R Modells. Die Zeilen zeigen die Zeitabschnitte, während die Spalten die Gruppen der Teilnehmenden definieren. Der RMSE über alle Zeitpunkte und Gruppen war: 1,12.

Nach Berechnung der ROC Kurve (siehe Material und Methoden 3.7.6), lag die AUC (siehe Abbildung 4.10) bei 0,74. Bei einem Grenzwert von 7 betrug die BAC 72,52% und die Genauigkeit 69,13%. Die Sensitivität belief sich auf 88,28% (95% CI: 83,51% - 92,07%), während die Spezifität mit 56,76% (95% CI: 51,54% - 61,87%) niedriger war. Da sich im Validierungsdatensatz die Anzahl der Probanden mit einem tatsächlich niedrigen GF R (Werte $<$7) nicht ändert, betrug auch wie zuvor die Prävalenz 39,24% (95% CI: 35,34% - 43,25%). Der NPW war mit einem Wert von 88,24% (95% CI: 83,97% - 91,48%) gut, während der PPW mit 56,87% (95% CI: 53,77% - 59,92%) niedriger war. Die NLR und die PLR betrugen 0,21 (95% CI: 1,80 - 2,31) und 2,04 (95% CI: 1,80 - 2,31). Ein Chi-Quadrat Test zeigte auch hier statistische Signifikanz (χ^2: FG=1; N = 609; Chi^2 Statistik = 123,85; P $<$ 0,00001).
Bei ROD Patienten konnte eine hohe Sensitivität von 92,16% (95% CI: 85,13% - 96,55%) festgestellt werden. Die Spezifität betrug 67,05% (95% CI: 59,57% - 73,93%). Hieraus konnte eine BAC von 79,60% und eine Genauigkeit von 76,26% berechnet werden. Der PPW war 61,84% (95% CI: 56,58% - 66,84%) und der NPW 93,65% (95% CI: 88,27% - 96,66%) und wurden mit der Prävalenz von 36,69% (95% CI: 31,01%- 42,65%) berechnet. Die PLR und NLR konnte mit 2,80 (95% CI: 2,25 - 3,48) und 0,12 (95% CI: 0,06 - 0,23)

beziffert werden. (χ^2: FG=1; N = 278; Chi^2 Statistik = 91,33; P < 0,00001)

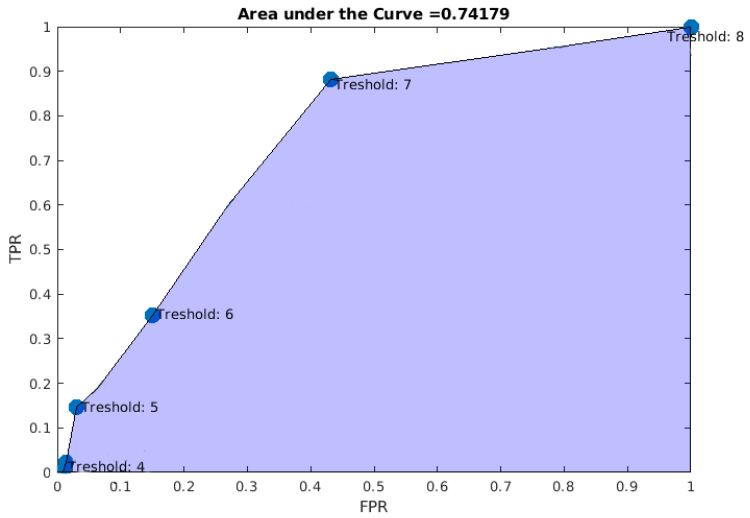

Abbildung 4.10: Die Abbildung zeigt die ROC Kurve des rekursiven GF R LSTM Modells. Es wurden mit verschiedenen Grenzwerten die TPR, sowie FPR berechnet und so eine dem Regressionsmodell zugehörige ROC Kurve gebildet. Die AUC war 0,74.

	alle Gruppen	ROD	CHR	ROP
über alle Zeitpunkte	72,52%	79,60%	69,49%	63,60 %
IV3	72,03%	76,73%	72,02%	65,84%
IV6	72,76 %	79,31%	70,63%	72,73%
T1	74,04%	81,48%	66,67%	71,47%
IV12	71,43%	77,84%	70,00%	62,22%
IV15	74,07%	83,24%	72,73%	58,93%
T2	73,43%	86,96%	70,00%	52,08%
IV27	70,68%	75,28%	75,00%	50,00%
IV36	69,94%	75,71%	NaN	62,50%

Tabelle 4.12: Die Tabelle zeigt die balancierte Genauigkeit (balanced accuracy) nachdem die GF R Regressionsvorhersagen post hoc in gute ($\geq$ 7) und schlechte Prognosen (<7) eingeteilt wurden. Die Zeilen zeigen die Zeitabschnitte, während die Spalten die Gruppen der Teilnehmenden definieren. Die BAC über alle Zeitpunkte und Gruppen war: 72,52%.

Die Genauigkeit der CHR Patienten belief sich auf 68,94% und die BAC auf 69,49%. Die Sensitivität betrug mit einem Wert von 87,18% (95% CI: 77,68% - 93,68%) mehr als 80%,

während die Spezifität 51,81% (95% CI: 40,56% - 62,92%) betrug. Gemeinsam mit der Prävalenz konnte der PPW mit einem Wert von 62,96% (95% CI: 57,25% - 68,34%) und der NPW mit 81,13% (95% CI: 69,93% - 88,83%) berechnet werden. Die PLR und die NLR lagen bei 1,81 (95% CI: 1,42 - 2,39) und 0,25 (95% CI: 0,13 - 0,46). (χ^2: FG=1; N = 161; Chi^2 Statistik = 27,68; P $<$ 0,00001).
Bei den ROP Patienten war die Sensitivität mit 93,22% (95% CI: 83,54% - 98,12%) hoch. Die Spezifität betrug 55,86% (95% CI: 46,12% - 65,27%). Aus der Prävalenz von 34,71% (95% CI: 27,58% - 42,37%), konnte ein PPW von 52,88% (95% CI: 47,38% - 58,32%) und ein NPW von 93,94% (95% CI: 85,57% - 97,59%), sowie die Genauigkeit berechnet werden. Diese belief sich auch 68,82%. Die BAC war mit 74,54% in der ROP Gruppe am niedrigsten. (χ^2: FG=1; N = 170; Chi^2 Statistik = 12,57; P $<$ 0,00001). Die jeweiligen BAC Werte können in Tabelle 4.12 eingesehen werden.

4.2.3 GAF D/I

Klassifikation

Wie auch in der Arbeit von [65] wurde für die GAF D/I Klassifikation ein Grenzwert von 65 herangezogen (siehe Material und Methoden 3.7.3). Folgende Einstellungen wurden im Rahmen der Bayesianischen Hyperparameter Optimierung gefunden:

- initiale Lernrate : 0.0046
- Größe der Minibatches : 32
- L2 Regularisierung: 1,067e^{-6}
- Gradient Decay: 0.9497
- Anzahl LSTM Neurone: 146
- Anzahl konventioneller Neurone: 34

Im Rahmen dieses Durchlaufes kann eine Vier-Felder Tafel erstellt werden, die in Tabelle 4.13 abgebildet ist. Ein mit diesen Daten berechneter Chi-Quadrat Test zeigt einen hoch signifikanten p-Wert (χ^2: FG = 1; N = 609; Chi^2 Statistik = 268,09; P $<$ 0,00001). Bei den GAF D/I Analysen war mit 42,69% (95% CI: 38,73% - 46,73%) eine hohe Prävalenz von schlechten GAF D/I Werten vorhanden. Die BAC über den gesamten Datensatz belief sich auf 83,53% (siehe Tabelle 4.14) und wich leicht von der Genauigkeit, die 82,92% betrug, ab. Der Algorithmus detektierte zukünftige schlechte GAF D/I Werte mit einer guten Sensitivität von 87,69% (95% CI: 83,07% - 91,43%) und einer Spezifität von 79,37% (95% CI: 74,74% - 83,49%). Zudem konnte die PLR mit einem Wert von 4,25 (95% CI: 3,44 - 5,25) und die NLR mit einem Wert von 0,16 (95% CI: 0,11 - 0,22) berechnet werden. Unter Hinzuziehung der Prävalenz, Sensitivität und Spezifität ergab sich ein PPW von 76,00% (95% CI: 71,95% - 79,63%) und ein NPW von 89,64% (95% CI: 86,17% - 92,32%)

		Wahre GAF D/I Werte		
		Schlechter GAF D/I	Guter GAF D/I	Total
LSTM Vorhersage	Schlechter GAF D/I	277	72	349
	Guter GAF D/I	32	228	260
	Total	309	300	609

Tabelle 4.13: Die Vierfeldertafel zeigt alle GAF D/I Werte, die im Validierungsdatensatz (über alle Zeitpunkte und Gruppen) prädiziert wurden und vergleicht diese mit den tatsächlichen Werten.

Bei ROD Patienten konnte eine Sensitivität vom 89,15% (95% CI: 82,46% - 93,94%) festgestellt werden. Die Spezifität betrug 79,87% (95% CI: 72,52% - 85,98%). Hieraus konnte eine BAC von 84,51% und eine Genauigkeit von 84,17% berechnet werden. Die PPW war 79,31% (95% CI: 73,46% - 84,15%) und die NPW betrug 89,47% (95% CI: 83,74% - 93,35%). Dieser Werte wurden mit einer Prävalenz von 46,40% (95% CI: 40,43%- 52,46%) ermittelt.

	alle Gruppen	ROD	CHR	ROP	N
über alle Zeitpunkte	83,53%	84,51%	84,68%	76,50%	
IV3	73,25%	69,17%	88,46%	66,74%	96
IV6	81,78%	81,07%	88,46%	76,98%	96
T1	86,41%	88,18%	87,50%	82,06%	92
IV12	86,08%	89,72%	82,50%	88,24%	84
IV15	93,96%	92,11%	100,00%	90,18%	82
T2	88,80%	89,18%	96,43%	80,36%	81
IV27	72,37%	85,16%	81,67%	22,50%	47
IV36	77,67%	86,67%	NaN	62,50%	31

Tabelle 4.14: Die Tabelle zeigt die balancierte Korrektheit (balanced accuracy) Ergebnisse der GAF D/I Prädiktionen. Die Zeilen zeigen die Zeitabschnitte, während die Spalten die Gruppen der Teilnehmenden definieren. Die BAC über alle Zeitpunkte und Gruppen war: 83.53%.

Das GAF D/I Modell konnte auf den CHR Patienten des Validierungsdatensatz, der nur Daten aus dem Studienzentrum Köln beinhaltete, eine Sensitivität von 88,84% (95% CI: 83,97%-92,65%) und eine Spezifität von 80,52% (95% CI: 76,20%- 84,36%) feststellen. Die BAC belief sich somit auf 84,68% und die Genauigkeit auf 88,20%. Der PPW der GAF D/I CHR Klassifikation konnte mit 82,28% (95% CI: 74,06% - 88,30%) beziffert werden, während die PLR 6,04 (95% CI: 3,71 - 9,81) betrug. Der Negativ-Prädiktive-Wert (NPW) war mit 93,90% (95% CI: 86,82% - 97,30%) höher als der PPW. Auch bei der CHR Kohorte ergaben sich statistisch signifikante Ergebnisse (χ^2: FG=1; N = 161; Chi^2 Statistik = 95,02; P $<$ 0,00001).
Bei genauer Betrachtung der ROP Kohorte innerhalb der GAF D/I Klassifikationsanalyse zeigte sich eine verhältnismäßig niedrige Prävalenz eines schlechten Funktionsniveaus von 35,88% (95% CI: 28,68% - 43,58%). Die BAC der ROP Patienten betrug 76,50% und war somit höher als die Genauigkeit, für die man einen Wert von 75,88% ermittelte. Die Sensitivität und die Spezifität betrugen 78,69% (95% CI: 66,32% - 88,14%) und 74,31% (95% CI: 65,06% - 82,20%). Der PPW betrug hierbei 63,16% (95% CI: 54,84% - 70,76%) und es konnte ein NPW von 86,17% (95% CI: 79,16% - 91,09%) berechnet werden. Die PLR und die NLR beliefen sich auf 3,06 (95% CI: 2,17 - 4,33) und 0,29 (95% CI: 0,17 - 0,47).

Regression

Mit Hilfe der Bayesianischen Hyperparameteroptimierung, wurden folgende Werte für das GAF D/I Modell gefunden:

- initiale Lernrate : 0,0035
- Größe der Minibatches : 60
- L2 Regularisierung: 4,7615e^{-10}
- Gradient Decay: 0.8756
- Anzahl LSTM Neurone: 8
- Anzahl konventioneller Neurone: 55

Der Gesamt-RMSE betrug 0,89. Für weitere RMSE Werte je nach Probandenzugehörigkeit und Zeitpunkt siehe Tabelle 4.15. Nach Anfertigung einer ROC Kurve konnte eine AUC von 0,90 ermittelt werden (Abbildung 4.11). Zudem wurde die tatsächlichen Werte mit den jeweiligen Vorhersagen verglichen. Dies kann in Abbildung 4.12 eingesehen werden. Es erfolgte eine Korrelationsanalyse nach Spearman, die eine starken Zusammenhang zwischen den Vorhersagen und den wahren GAF D/I Werten zeigte S_ρ=(0,81; N = 609; P <0,00001). Post-Hoc wurden die Ergebnisse der Regression in günstige und ungünstige ($<$65) Werte eingeteilt (siehe Material und Methoden 3.7.3). Danach konnte eine Sensitivität von 71,54% (95% CI: 65,64% - 76,94%) und eine Spezifität von 90,83% (95% CI: 87,30% - 93,64%), sowie eine Genauigkeit von 82,59% und eine BAC von 81,18% ermittelt werden. Bei einer Prävalenz von 42,69% (95% CI: 38,73% - 46,73%) betrug der PPW 85,32% (95% CI: 80,55% - 89,08%) und der NPW 81,07% (95% CI: 77,89% - 83,89%). Die PLR und die NLR waren 7,80 (95% CI: 5,56 - 10,95) und 0,31 (95% CI: 0,26 - 0,38). Ein Chi-Quadrat Test zeigte statistisch signifikante Ergebnisse (χ^2: FG=1; N = 609; Chi^2 Statistik = 252,19; P $<$ 0,00001).
Bei den ROD Patienten war die BAC mit 79,36%, aber auch die Genauigkeit bei 80,22% höher. Die Sensitivität rangierte mit 67,28% (95% CI: 58,64% - 75,43%) lediglich oberhalb der 60% Marke. Die Prävalenz eines niedrigen GAF D/I betrug unter den ROD Patienten im Validierungsdatensatz 46,40% (95% CI: 40,43% - 52,46%). Hieraus konnte der PPW und der NPW berechnet werden, die 87,00% (95% CI: 79,70% - 91,94%) und 76,40% (95% CI: 71,54% - 80,66%) betrugen. Die PLR und die NLR bezifferten 7,73 (95% CI: 4,54 - 13,17) und 0,36 (95% CI: 0,28 - 0,46). Die Ergebnisse waren signifikant (χ^2: FGe=1; N = 278; Chi^2 Statistik = 103,50; P $<$ 0,00001).

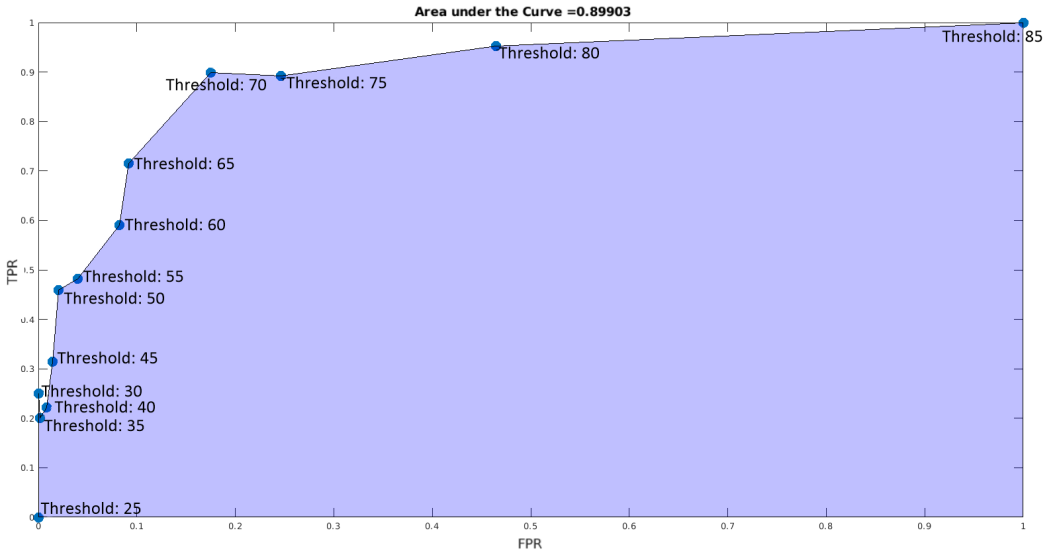

Abbildung 4.11: Die Abbildung zeigt die ROC Kurve des GAF D/I LSTM Modells. Es wurden mit verschiedenen Grenzwerten die TPR, sowie FPR berechnet und so eine dem Regressionsmodell zugehörige ROC Kurve gebildet. Die AUC belief sich auf 0,90.

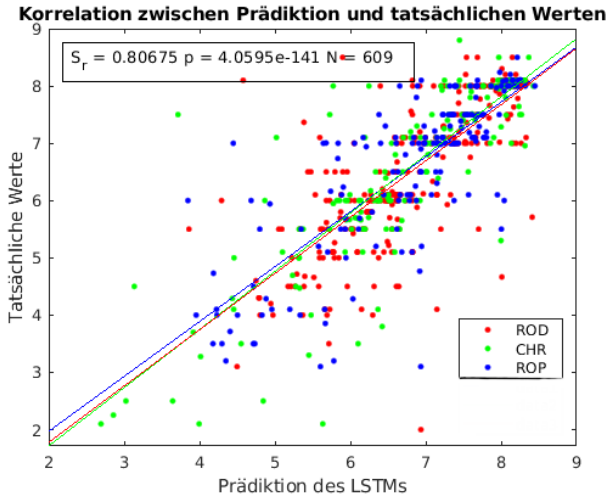

Abbildung 4.12: Der Graph zeigt den Rangkorrelationskoeffizienten nach Spearman. Tatsächliche Werte und Prädiktionen des GAF D/I werden auf der x und y Achse gegenübergestellt. S_ρ wurde mit 0,81 berechnet. Der p-Wert belief sich auf $4{,}0595e^{-141}$.

Die CHR Probanden erreichten eine Sensitivität von 82,86% (95% CI: 71,97% - 90,82%) Spezifität von 91,21% (95% CI: 83,41% - 96,13%) und weisen so die beste Performance aller Subpopulationen der GAF D/I Regressionsanalyse auf. Die Prävalenz konnte mit 43,48% (95% CI: 35,69% - 51,51%) bei den CHR Patienten beziffert werden.
Auch hier war der NPW mit 87,37% (95% CI: 80,45% - 92,08%) höher als der PPW 87,88% (95% CI: 78,76% - 93,41%). Die BAC betrug 87,03% (siehe Tabelle 4.16) und die Genauigkeit 87,58%. Ein Chi-Quadrat Test zeigt, dass der Algorithmus keine zufälligen Vorhersagen vornahm (χ^2: FG=1; N = 161; Chi^2 Statistik = 89,73; P $<$ 0,00001).
Bei den ROP Patienten war die Sensitivität mit 67,21% (95% CI: 54,00% - 78,69%) ebenfalls höher als die Spezifität mit 89,91% (95% CI: 82,66% - 94,85%). Die BAC und die Genauigkeit betrugen 78,58% und 81,76%. Aus einer Prävalenz von 35,88% (95% CI: 28,68% - 43,58%) ergab sich ein PPW von 78,85% (95% CI: 67,45% - 87,02%) und ein NPW von 83,05% (95% CI: 77,28% - 87,59%). Die PLR konnte man auf 6,66 (95% CI: 3,70 - 11,98) beziffern. Die NLR war mit einem Wert von 0,36 (95% CI: 0,25 - 0,53) niedrig. Ein Chi-Quadrat Test zeigte auch bei ROP Patienten einen signifikanten p-Wert (χ^2: FG=1; N = 170; Chi^2 Statistik = 60,11; P $<$ 0,00001). Der Trainingsprozess des LSTM Netzes, kann in Abbildung 4.13 betrachtet werden.

	alle Gruppen	ROD	CHR	ROP
über alle Zeitpunkte	0,89	0,89	0,89	0,89
IV3	1,07	1,11	1,23	0,86
IV6	0,83	0,59	0,98	0,97
T1	0,84	0,60	0,90	1,07
IV12	0,97	1,16	0,81	0,76
IV15	0,79	1,03	0,42	0,57
T2	0,77	0,98	0,48	0,60
IV27	0,80	0,60	0,51	1,39
IV36	0,89	0,61	1,59	0,65

Tabelle 4.15: Die Tabelle zeigt die RMSE Ergebnisse des Regressions GAF D/I Modells. Die Zeilen zeigen die Zeitabschnitte, während die Spalten die Gruppen der Teilnehmenden definieren. Der RMSE über alle Zeitpunkte und Gruppen war: 0,89. Die identischen Zahlen über die jeweiligen Probandengruppen ergeben sich durch Rundung (alle= 0,885; ROD= 0,891; CHR= 0,890; ROP= 0,885).

	alle Gruppen	ROD	CHR	ROP
über alle Zeitpunkte	81,18%	79,36%	87,03%	78,58%
IV3	72,73%	70,68%	84,62%	66,06%
IV6	83,60%	76,60%	88,46%	89,68%
T1	80,75%	77,85%	85,63%	80,00%
IV12	84,36%	82,50%	90,83%	82,77%
IV15	87,02%	84,21%	90,91%	87,50%
T2	82,10%	80,70%	90,18%	77,68%
IV27	76,27%	82,14%	80,00%	55,00%
IV36	87,67%	96,67%	NaN	75,00%

Tabelle 4.16: Die Tabelle zeigt die balancierte Korrektheit (balanced accuracy) Ergebnisse der GAF D/I Prädiktionen. Vorab wurde ein Regressionsmodell erstellt und nachträglich die BAC nach Verwendung des Grenzwertes 65 berechnet. Hieraus ergab sich eine Gesamt - BAC von 81,18%.

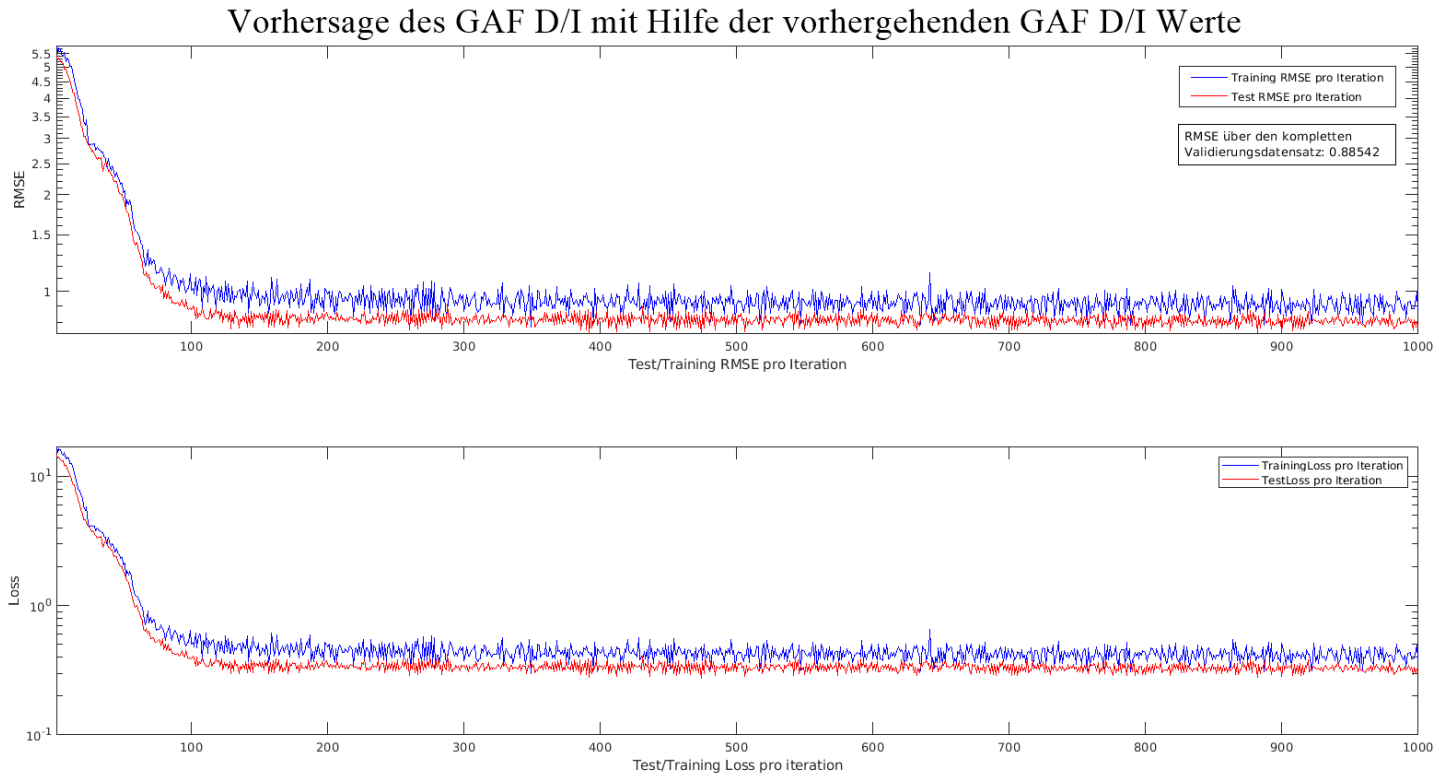

Abbildung 4.13: Die Abbildung zeigt den Trainingsverlauf des Neuronalen Netzes. In Blau ist der RMSE der Trainingdaten im Verlauf zu erkennen. In Rot der RMSE der Testdaten. Nach Beendigung des Trainings wurden die Werte der Validierunngsdaten prädiziert, was einen RMSE von 0,89 ergab.

Rekursives Modell

Für das rekursive Modell wurden die Parameter des Regressionsmodells verwendet. Hierbei ergab sich eine positive Korrelation zwischen tatsächlichen und Prädizierten GAF D/I Werten S_ρ=(0,51; N=609; P $<$ 0,00001).

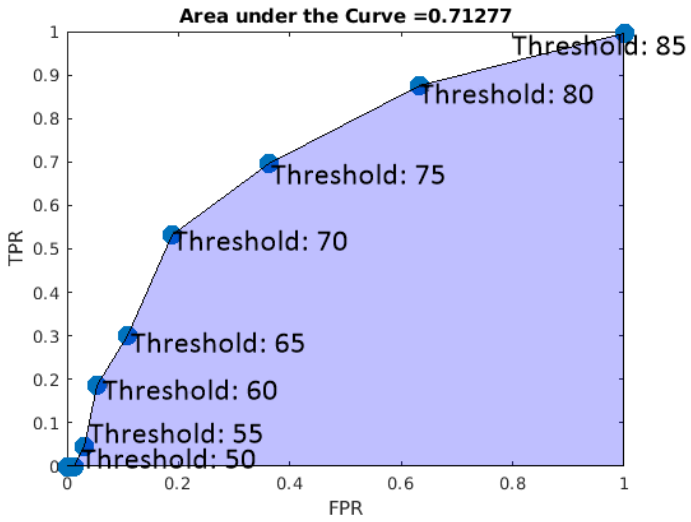

Abbildung 4.14: Die Abbildung zeigt die ROC Kurve des rekursiven GAF D/I LSTM Modells. Es wurden mit verschiedenen Grenzwerten die TPR, sowie FPR berechnet und so eine dem Regressionsmodell zugehörige ROC Kurve gebildet. Die AUC belief sich auf 0,71.

	alle Gruppen	ROD	CHR	ROP
über alle Zeitpunkte	1,26	1,13	1,35	1,30
IV3	1,07	1,11	1,23	0,86
IV6	1,16	0,98	1,36	1,13
T1	1,39	1,04	1,63	1,50
IV12	1,41	1,53	1,53	1,16
IV15	1,67	1,88	1,66	1,46
T2	1,66	1,72	1,82	1,45
IV27	1,92	1,90	2,05	1,81
IV36	1,32	1,41	0,62	1,92

Tabelle 4.17: Die Tabelle zeigt die RMSE Ergebnisse des rekursiven GAF D/I Modells. Die Zeilen zeigen die Zeitabschnitte, während die Spalten die Gruppen der Teilnehmenden definieren. Der RMSE über alle Zeitpunkte und Gruppen war: 1,26.

Der RMSE über den kompletten Validierungsdatensatz betrug 1,26. Die restlichen RMSE Ergebnisse können in Tabelle 4.17 betrachtet werden. Post-Hoc wurde erneut eine Unterteilung in hohe (≥ 65) und in niedrige (< 65) GAF D/I Werte durchgeführt. Die BAC über den gesamten Validierungsdatensatz beliefen sich auf 59,56%. Die BAC Ergebnisse je nach Gruppen und Zeitpunkte sind in Tabelle 4.18 einzusehen. Das rekursive Modell (siehe Material und Methoden 3.7.8) detektiert zukünftige schlechte GAF D/I Werte mit einer niedrigen Sensitivität von 30,00% (95% CI: 24,49% - 35,97%) und einer Spezifität von 89,11% (95% CI: 85,36% - 92,18%). Die Genauigkeit betrug 63,88%.

	alle Gruppen	ROD	CHR	ROP
über alle Zeitpunkte	59,56%	53,26%	65,44%	66,17%
IV3	72,73%	70,68%	84,62%	66,06%
IV6	64,71%	53,84%	73,08%	73,81%
T1	60,02%	51,63%	63,75%	69,12%
IV12	60,61%	51,94%	65,83%	75,63%
IV15	51,85%	44,74%	50,00%	68,75%
T2	50,41%	47,37%	50,00%	56,25%
IV27	47,92%	50,00%	50,00%	40,00%
IV36	50,00%	50,00%	NaN	50,00%

Tabelle 4.18: Die Tabelle zeigt die balancierte Korrektheit (balanced accuracy) Ergebnisse der GAF D/I Prädiktionen. Die Zeilen zeigen die Zeitabschnitte, während die Spalten die Gruppen der Teilnehmenden definieren. Die BAC über alle Zeitpunkte und Gruppen war: 59,56%.

Ein Chi-Quadrat Test blieb trotz schlechterer Performance hoch signifikant (χ^2: FG=1; N = 609; Chi^2 Statistik = 35,29; P $<$ 0,00001). Der PPW und der NPW rangierten jeweils im Bereich zwischen 60% und 70% mit 67,24% (95% CI: 61,02% - 65,10%) und 63,08% (95% CI: 61,08% - 65,10%). Die PLR (95% CI: 59,05% - 74,50%) und die NLR (95% CI: 0,72 - 0,86) konnte mit 2,76 und 0,79 beziffert werden.
Die Performance unterschied sich zwischen den verschiedenen Kohorten. Die ROD Patienten wiesen eine Sensitivität von 18,60% und einer Spezifität von 87,92% auf. Die Genauigkeit und die BAC betrugen 55,76% und 53,26%. Bei dieser Probandengruppe zeigte der Chi-Quadrat Test erstmals keine statistisch signifikanten Ergebnisse (χ^2: FG=1; N = 278; Chi^2 Statistik = 2,29; P = 0,13). Dementsprechend bezifferten der PPW mit 57,14% (95% CI: 43,14% - 70,09%) und der NPW mit 55,51% (95% CI: 52,98% - 58,00%) Werte, die nur geringfügig oberhalb der 50% Marke anzusiedeln waren. Die NLP betrug 0,93 (95% CI: 0,84 - 1,02) und die PLR 1,54 (95% CI: 0,88 - 2,71).
Bei den CHR Patienten belief sich die Sensitivität auf 38,57% (95% CI: 27,17% - 50,97%), sowie die Spezifität auf 92,31% (95% CI: 84,79% - 96,85%). Die Genauigkeit betrug 68,94%,

die BAC 65,44%. Der PPW konnte auf 79,41% (95% CI: 64,09% - 89,29%) und der NPW auf 66,14% (95% CI: 61,65% - 70,36%) beziffert werden. (χ^2: FG=1; N = 161; Chi^2 Statistik = 22,65; P $<$ 0,006).
Die Vorhersagen der ROP Patienten wiesen eine Genauigkeit von 72,35% und eine BAC von 66,17% auf. Die Sensitivität war mit 44,26% (95% CI: 31,55% - 57,55%) nach wie vor niedrig, jedoch etwas höher bei zu vergleichenden Probandengruppen. Die Spezifität nahm mit 88,07% (95% CI: 80,47% - 93,49%) einen etwa doppelt so hohen Wert an. Wie auch bei den vorhergehenden GAF D/I Analysen betrug die Prävalenz eines niedrigen GAF D/I, innerhalb der ROP Kohorte, 35,88% (95% CI: 28,68% - 43,58%). Der PPW betrug 67,50% (95% CI: 53,70% - 78,81%) und der NPW war 73,85% (95% CI: 69,08% - 78,11%). Die PLR konnte auf 3,71 beziffert werden (95% CI: 2,07 - 6,65). Die NLR war 0,63 (95% CI: 0,50 - 0,80). Innerhalb der ROP Population ergab sich bei der rekursiven GAF D/I Analyse der niedrigste p-Wert (χ^2: FG=1; N = 170; Chi^2 Statistik = 22,73; P $<$ 0,00001).

4.2.4 GAF S

Klassifikation

Anders als der GAF D/I, fokussiert sich der GAF S auf die Symptomausprägung, während der GAF D/I sich eher auf die Beeinträchtigungen stützt, die durch die zu Grunde liegende Erkrankung entsteht. Durch die Bayesianische Optimierung ergaben sich folgende Hyperparameterkonstellationen bei der GAF S Klassifikationsanalyse:

- initiale Lernrate : 0,045351
- Größe der Minibatches : 56
- L2 Regularisierung: 4,7238e^{-6}
- Gradient Decay: 0,9799
- Anzahl LSTM Neurone: 139
- Anzahl konventioneller Neurone: 24

Die BAC über den kompletten Validierungsdatensatz der GAF S Vorhersagen betrug 78,28%; die Genauigkeit 78,33%. Die Sensitivität und die Spezifität beliefen sich auf 87,26% (95% CI: 82,99% - 90,78%) und 69,31% (95% CI: 63,78% - 74,45%). Aus der Prävalenz von 50,35% (95% CI: 46,20% - 54,29%) ergab sich ein PPW von 74,17% (95% CI: 70,68% - 77,37%) und ein NPW von 84,34% (95% CI: 79,91% - 87,93%). Die PLR war 2,84 (95% CI: 2,39 - 3,38) und die NLR betrug 0,18 (95% CI: 0,14 - 0,25). Ein Chi- Quadrat Test zeigte statistisch signifikanten Ergebnisse (χ^2: FG=1; N = 609; Chi^2 Statistik = 201,52; P $<$ 0,00001). Die zu Grunde liegende Vier - Felder Tafel, kann in Tabelle 4.19 betrachtet werden. Die durchschnittliche BAC der ROD Patienten im Validierungsdatensatz betrug 84,10% und war etwas geringer als die Genauigkeit, welche mit einem Wert von 84,17% beziffert werden konnte. Die Sensitivität betrug 89,36 % (95% CI: 83,06% - 93,92%) und die Spezifität 75,56% (95% CI: 71,03% - 85,34%). Die Berechnung der PLR ergab 4,22 (95% CI:

		Wahre GAF S Werte		
		Schlechter GAF S	Guter GAF S	Total
LSTM Vorhersage	Schlechter GAF S	210	93	303
	Guter GAF S	39	267	306
	Total	249	360	609

Tabelle 4.19: Die Vierfeldertafel zeigt alle GAF S Werte, die im Validierungsdatensatz (über alle Zeitpunkte und Gruppen) prädiziert wurden und vergleicht diese mit den tatsächlichen Werten. Somit lag eine Sensitivität von 87,25% und eine Spezifität von 69,31% vor.

3,04 - 5,86) - die NLR 0,13 (95% CI: 0,08 - 0,22). Ein Chi-Quadrat Test wies auf statistisch signifikante Ergebnisse hin (χ^2: FG=1; N = 278; Chi^2 Statistik = 130,99; P $<$ 0,00001). Bei einer Prävalenz von 50,72% (95% CI: 44,68% - 56,74%) wurde ein PPW von 81,29% (95% CI: 75,78% - 85,78%) und ein NPW von 87,80% (95% CI: 81,80% - 92,13%) berechnet.

	alle Gruppen	ROD	CHR	ROP	N
über alle Zeitpunkte	78,28%	84,10%	74,75%	72,15%	
IV3	64,97%	68,92%	61,88%	61,11%	96
IV6	78,78%	82,50%	75,15%	76,67%	96
T1	77,86%	85,32%	72,12%	73,90%	92
IV12	74,90%	79,27%	75,83%	67,14%	84
IV15	85,33%	97,62%	77,27%	72,73%	82
T2	81,95%	89,33%	81,67%	72,32%	81
IV27	89,18%	92,50%	81,67%	90,00%	47
IV36	75,25%	77,38%	NaN	75,00%	31

Tabelle 4.20: Die Tabelle zeigt die balancierte Korrektheit (balanced accuracy) Ergebnisse der GAF S Prädiktionen. Die Zeilen zeigen die Zeitabschnitte, während die Spalten die Gruppen der Teilnehmenden definieren. Die BAC über alle Zeitpunkte und Gruppen war: 76,12%.

Die CHR Patienten wiesen bei 87,95% (95% CI: 78,96% - 94,07%) eine höhere Sensitivität und eine niedrigere Spezifität bei 61,54% (95% CI: 49,83% - 72,34%) als die ROD Patienten auf. Dies spiegelte sich in einer BAC von 74,75% und einer Genauigkeit von 75,16% wieder. Bei den CHR Patienten waren GAF S Werte kleiner 65 mit 51,55% etwas prävalenter. Der PPW war mit 70,87% (95% CI: 64,51% - 76,51%) und der NPW mit 82,76% (95% CI: 72,34% - 89,81%) etwas niedriger als der Durchschnitt über alle Probandengruppen. Die PLR betrug 2,29 (95% CI: 1,17 - 3,06) und die NLR war 0,20 (95% CI: 0,11 - 0,36)(χ^2: FG=1; N = 161; Chi^2 Statistik = 42,73; P $<$ 0,00001). Der GAF S der ROP Probanden deutete mit einer BAC von 77,96% und einer Genauigkeit von 71,76% - die GAF S Klassifikationsanalyse betreffend - eine schlechtere Performance an. Die Sensitivität war mit 82,93% (95% CI: 73,02% - 90,93%) höher als die Spezifität, welche 61,36% (95% CI: 50,38% - 71,56%) betrug. Bei einer Prävalenz von 48,24% (95% CI: 40,52% - 56,01%) belief sich der PPW auf 66,67% (95% CI: 60,16% - 72,60%) und der NPW auf 79,41% (95% CI:69,95% - 86,47%) Die PLR (95% CI: 1,62 - 2,84) war mit einem Wert von 2,15 niedrig. Die NLR war mit 0,28 (95% CI: 0,17 - 0,46) im Vergleich hoch. Ein Chi-Quadrat Test war statistisch signifikant (χ^2: FG=1; N = 170; Chi^2 Statistik = 34,70; P $<$ 0,00001).

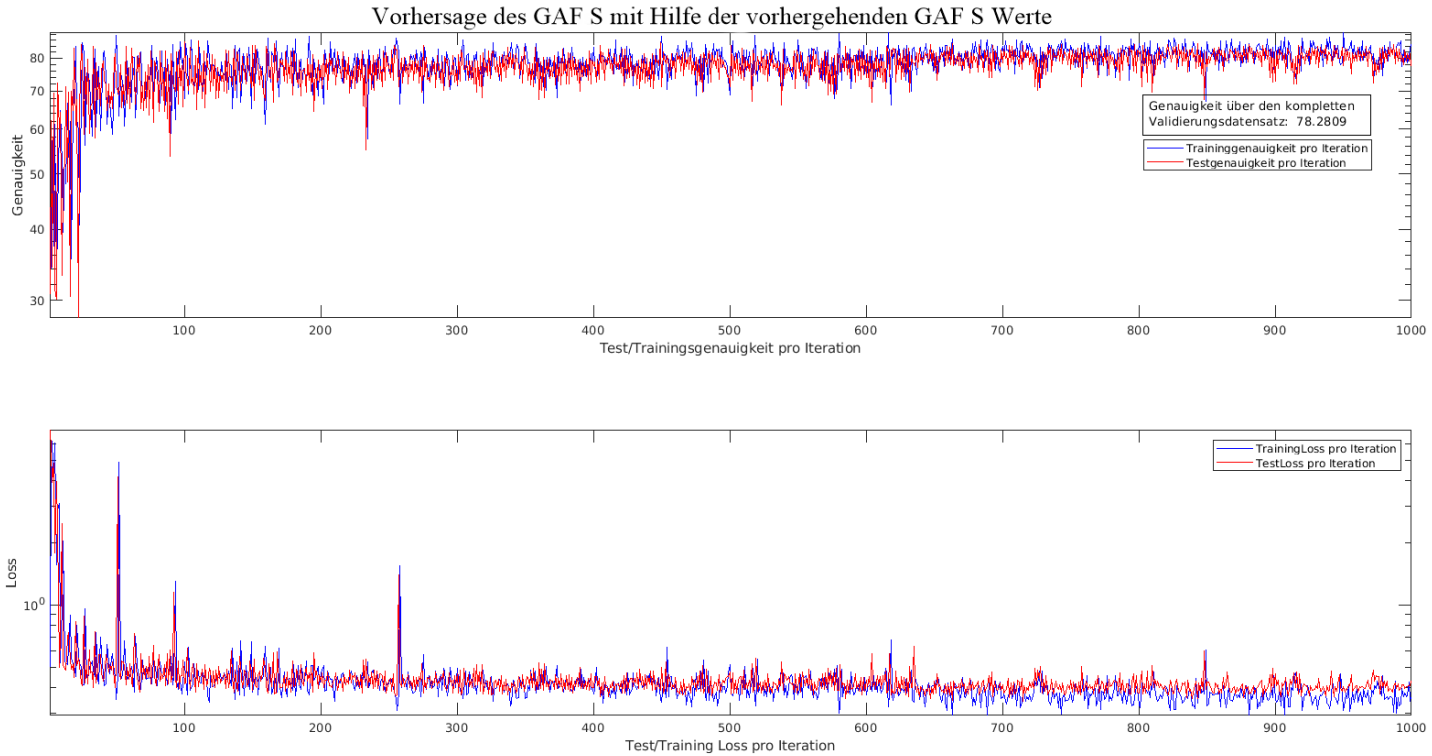

Abbildung 4.15: Die Abbildung zeigt den Trainingsverlauf des Neuronalen Netzes. In Blau ist die Genauigkeit (accuracy) der Trainingdaten im Verlauf zu erkennen. In Rot die Genauigkeit der Testdaten. Nach Beendigung des Trainings wurden die Werte der Validierunngsdaten prädiziert, die im Mittel in 78,28% (balanced accuracy) der Fälle korrekt eingeschätzt wurden.

Regression

Bei der GAF S Regression versucht man den möglichst genauen zukünftigen GAF S Wert vorherzusagen. Zur Bestimmung der Güte des ermittelten LSTM Modells wird hierfür der RMSE (siehe Material und Methoden 3.7.6) herangezogen. Die Hyperparameter wurden mit Hilfe der Bayesianischen Optimierung folgendermaßen ausgewählt:

- initiale Lernrate : 0,0013
- Größe der Minibatches : 58
- L2 Regularisierung: 1,2728e^{-4}
- Gradient Decay: 0,8461
- Anzahl LSTM Neurone: 13
- Anzahl konventioneller Neurone: 9

Der RMSE betrug bei Betrachtung aller Gruppen und Zeitpunkte 0,85. Zur Berechnung wie sehr die prädizierten und die tatsächlichen Werte miteinander korrelierten, wurde der Rangkorrelationskoeffizienten nach Spearman berechnet. Dieser betrug S_ρ = 0,76 (N = 609; P <0,00001) (siehe Abbildung 4.16).

	alle Gruppen	ROD	CHR	ROP
über alle Zeitpunkte	0,85	0,88	0,87	0,85
IV3	1,12	1,13	1,14	1,08
IV6	0,81	0,88	0,92	0,58
T1	0,80	0,67	0,82	0,93
IV12	0,77	0,85	0,73	0,68
IV15	0,84	0,96	0,80	0,65
T2	0,82	0,93	0,66	0,78
IV27	0,60	0,61	0,49	0,71
IV36	0,87	0,82	1,06	0,82

Tabelle 4.21: Die Tabelle zeigt die RMSE Ergebnisse des Regressions GAF S Modells. Die Zeilen zeigen die Zeitabschnitte, während die Spalten die Gruppen der Teilnehmenden definieren. Der RMSE über alle Zeitpunkte und Gruppen war 0,85.

Die Area under the curve (AUC) des GAF S Regressions Modells war 0,88 (siehe Abbildung 4.17). Post-hoc wurden die Ergebnisse der Regression in gute ($\geq$ 65) und in schlechte GAF S (<65) Werte unterteilt. Die hieraus berechneten BAC Werte können in Tabelle 4.22 eingesehen werden. Auf den Daten des Uniklinikums Köln, was dem Validierungsdatensatz

entspricht, konnte eine Sensitivität von 79,08% (95% CI: 74,09% - 83,50%) und eine Spezifität von 83,17% (95% CI: 78,47%- 87,20%) berechnet werden. Die BAC belief sich somit auf 81,13% und ähnelt der Genauigkeit, die 81,12% betrug. Der PPW der Post-Hoc Klassifikation konnte mit 82,59% (95% CI: 78,59% - 85,98%) beziffert werden, während die PLR mit einem Wert von 4,70 (95% CI: 3,63 - 6,07) berechnet wurde. Der NPW war mit 79,75% (95% CI: 75,89% - 83,12%) höher als der PPW. Die NLR betrug 0,25 (95% CI: 0,20 - 0,31). Die Prävalenz einer schlechten Prognose lag bei einem Grenzwert von 65 bei 50,25% (95% CI: 46,20% - 54,29%). Ein Chi-Quadrat-Test zeigte, dass die Gruppenzuteilungen von Seiten des Modells nicht zufällig erfolgten (χ^2: FG=1; N = 609; Chi^2 Statistik = 236,35; P < 0,00001).

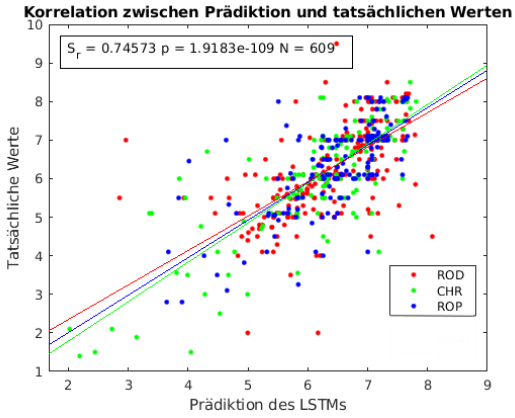

Abbildung 4.16: Der Graph zeigt den Rangkorrelationskoeffizienten nach Spearman. Tatsächliche Werte und Prädiktionen des GAF S werden auf der x und y Achse gegenübergestellt. S_ρ wurde mit 0,75 berechnet. Der p-Wert belief sich auf $1{,}82e^{-109}$.

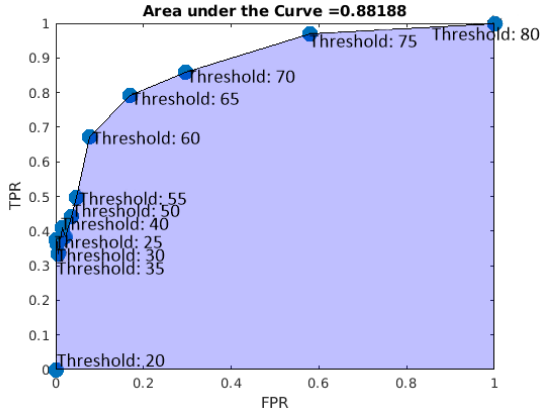

Abbildung 4.17: Die Abbildung zeigt die ROC Kurve des GAF S LSTM Regressionsmodells. Anhand verschiedener Grenzwerte wurde jeweils die TPR und die FPR berechnet und auf den Achsen aufgetragen. Die AUC ist 0,88.

Bei den ROD Patienten war die BAC mit 86,37%, aber auch die Genauigkeit bei 86,33% höher. Die Sensitivität rangierte mit 83,69% (95% CI: 76,54% - 89,37%) oberhalb der 80% Marke. Die Prävalenz eines niedrigen GAF S betrug unter den ROD Patienten im Validierungsdatensatz 50,72% (95% CI: 44,68% - 56,74%). Hieraus konnte der PPW und der NPW berechnet werden, die 88,72% (95% CI: 82,91% - 92,73%) und 84,14% (95% CI: 78,42% - 88,56%) betrugen. Die PLR und die NLR ergaben 7,64 (95% CI: 4,72 - 12,39) und 0,18 (95% CI: 0,13 - 0,27). Ein hoch signifikanter Chi-Quadrat Test belegte, dass es sich um keine zufälligen Vorhersage von Seiten des Modells handelte (χ^2: FG = 1; N = 278; Chi^2 Statistik = 155,66; P < 0,00001).
Sowohl eine leicht niedrigere Genauigkeit (79,50%) als auch BAC (79,50%) lag bei den

CHR Patienten vor. Die Sensitivität war mit 79,52% (95% CI: 69,24% - 87,59%) höher als die Spezifität bei 79,49% (95% CI: 68,84% - 87,80%). Der NPW war mit 78,48% (95% CI: 70,17% - 84,97%) deutlich höher als der PPW, der 80,49% (95% CI: 72,45% - 86,62%) betrug. Die PLR bezifferte 3,88 (95% CI: 2,47 - 6,08) und die NLR belief sich auf 0,28 (95% CI: 0,17 - 0,40). Die Chi-Quadrat Statistik belief sich auf 55,65 (χ^2: FG = 1; N = 161; Chi^2 Statistik = 55,65; P $<$ 0,00001).
Bei den ROP Patienten war die Sensitivität mit 70,73% (95% CI: 59,65% - 80,26%) ebenfalls höher als die Spezifität mit 77,27% (95% CI: 67,11% - 85,53%). Die BAC und die Genauigkeit betrugen 74,00% und 74,12%. Aus einer Prävalenz von 48,24% (95% CI: 40,52% - 56,01%) ergab sich ein PPW von 74,36% (95% CI: 65,82% - 81,37%) und ein NPW von 73,91% (95% CI: 66,52% - 80,16%). Die PLR konnte man mit 3,11 (95% CI: 2,07 - 4,69) beziffern. Die NLR war mit einem Wert von 0,38 (95% CI: 0,27 - 0,54) niedrig. Ein Chi-Quadrat Test zeigt auch bei ROP Patienten einen hoch-signifikanten p-Wert (χ^2-Test: FG = 1; N = 170; Chi^2 Statistik = 64,14; P $<$ 0,00001).

	alle Gruppen	ROD	CHR	ROP	N
über alle Zeitpunkte	81,13%	86,37%	79,50%	74,00%	
IV3	68,12%	68,05%	65,00%	70,83%	
IV6	80,39%	82,50%	79,70%	76,67%	
T1	87,19%	92,46%	86,67%	81,59%	
IV12	82,22%	85,15%	86,67%	75,71%	
IV15	84,17%	94,68%	81,82%	68,18%	
T2	86,60%	94,74%	86,67%	73,21%	
IV27	91,45%	100,00%	81,67%	77,50%	
IV36	69,70%	73,21%	NaN	62,50%	

Tabelle 4.22: Die Tabelle zeigt die balancierte Korrektheit (balanced accuracy) Ergebnisse der GAF S Prädiktionen. Es wurde Post-Hoc auf die Regessionsergebnisse der Grenzwert von 65 angewendet. Hieraus ergab sich eine Gesamt BAC von 81,13%

Rekursives Modell

Das rekursive Modell verwendete um eine Vorhersage bis IV36 zu treffen lediglich die erhobenen GAF S Werte bei T0. Mittels des Spearman Korrelationskoeffizienten wurde ein ρ von 0,47 ermittelt S_ρ=(0,47; N=609; P $<$ 0,00001). Der RMSE über alle Zeitpunkte betrug 1,16 (siehe Tabelle 4.23) und die AUC mit 0,66 niedriger (siehe Abbildung 4.19).

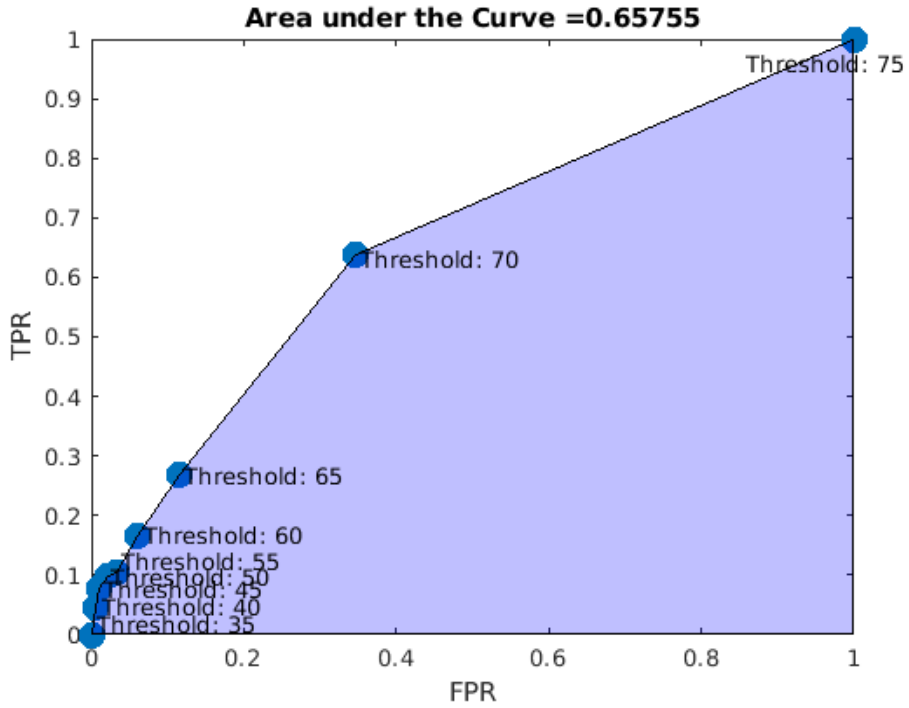

Abbildung 4.18: Die Abbildung zeigt die ROC Kurve des GAF D/I LSTM Modells. Es wurden mit verschiedenen Grenzwerten die TPR, sowie FPR berechnet und so eine dem Regressionsmodell zugehörige ROC Kurve gebildet. Die AUC belief sich auf 0,66.

Nachdem Post-Hoc auf die tatsächlichen GAF S Werte, sowie die Regressionsergebnisse ein Grenzwert von 65 angewandt wurde, ergab sich im rekursiven Modell eine BAC von 57,62%, während die Genauigkeit mit einem Wert von 55,04% etwas geringer war. Die Sensitivität betrug 26,80% (95% CI: 21,92% - 32,13%) und die Spezifität belief sich auf 88,45% (95% CI: 84,30% - 91,82%). Die PLR und die NLR bezifferten 2,32 (95% CI: 1,61 - 3,33) und 0,83 (95% CI: 0,76 - 0,90). Aus der Prävalenz konnte ein PPW von 70,09% (95% CI: 61,99% - 77,10%) und ein NPW von 54,47% (95% CI: 52,50% - 56,42%) berechnet werden. Ein Chi - Quadrat Test zeigte statistisch signifikante Ergebnisse (χ^2-Test: FG=1; N = 609; Chi^2

Statistik = 22,80; P $<$ 0,00001).
Bei den ROD Patienten zeigte sich eine niedrigere Sensitivität bei 20,57% (95% CI: 14,23% - 28,18%) und eine höhere Spezifität von 90,51% (95% CI: 84,32% - 94,85%). Die BAC war mit 55,54% geringfügig höher als die Genauigkeit bei 55,04%. Die Prävalenz eines niedrigen GAF S betrug 50,72% (95% CI: 44,68% - 56,74%). Der PPW konnte bei den ROD Patienten mit einem Wert von 69,05% (95% CI: 54,78% - 80,42%) beziffert werden. Der NPW war bei einem Wert von 52,54% (95% CI: 50,05% - 55,03%) niedriger. Die PLR betrug 2,17 (95% CI: 1,18 - 3,99) und die NLR war 0,88 (95% CI: 0,79 - 0,97). Ein Chi-Quadrat Test zeigte signifikante Ergebnisse (χ^2: FG=1; N = 278; Chi^2 Statistik = 6,65; P $<$ 0,01). Die Sensitivität der CHR Vorhersagen betrug im Validierungsdatensatz 36,14% (95% CI: 25,88% - 47,43%). Die Spezifität war bei 84,62% (95% CI: 74,67% - 91,79%) über der 80% Marke. Die PLR ergab 2,35 (95% CI: 1,30 - 4,25) und die NLR 0,75 (95% CI: 0,63 - 0,91). Aus den Vorhersagen der CHR Probanden konnte eine Genauigkeit von 56,63% und eine BAC von 60,38% berechnet werden. Aus einer Prävalenz von 51,55% (95% CI: 43,56% - 59,49%) ergab sich ein PPW von 71,43% (95% CI: 57,99% - 81,91%) und ein NPW von 55,46% (95% CI: 50,80% - 60,03%). (χ^2: FG=1; N = 161; Chi^2 Statistik = 8,99; P $<$ 0,003). Bei Betrachtung der ROP Kohorte zeigte sich eine Sensitivität von 28,05% (95% CI: 18,68% - 39,06%) und eine Spezifität von 88,64% (95% CI: 80,09% - 94,41%). Die Genauigkeit innerhalb dieser Gruppe betrug 59,41%. Die BAC kann in Tabelle 4.24 eingesehen werden. Der NPW und der PPW betrugen 56,93% (95% CI: 53,11% - 60,67%) und 69,70% (95% CI: 53,85% - 81,93%), die sich aus der Prävalenz von 48,24% (95% CI: 40,52% - 56,01%) ergab. Die PLR und die NLR betrugen 2,47 (95% CI: 1,25 - 4,87) und 0,81 (95% CI: 0,70 - 0,95). Ein Chi-Quadrat Test ergab statistisch signifikante Ergebnisse innerhalb der ROP Kohorte (χ^2: FG=1; N = 170; Chi^2 Statistik = 7,55; P $<$ 0,006).

	alle Gruppen	ROD	CHR	ROP
über alle Zeitpunkte	1,16	0,90	1,49	1,10
IV3	1,12	1,13	1,14	1,08
IV6	1,18	1,08	1,46	1,00
T1	1,29	1,05	1,40	1,42
IV12	1,24	1,25	1,36	1,10
IV15	1,48	1,63	1,48	1,32
T2	1,30	1,25	1,56	1,09
IV27	1,35	1,34	1,36	1,35
IV36	0,99	1,00	0,65	1,32

Tabelle 4.23: Die Tabelle zeigt die RMSE Ergebnisse des Regressions GAF S Modells. Die Zeilen zeigen die Zeitabschnitte, während die Spalten die Gruppen der Teilnehmenden definieren. Der RMSE über alle Zeitpunkte und Gruppen war: 1,16.

	alle Gruppen	ROD	CHR	ROP
über alle Zeitpunkte	57,62%	55,54%	60,38%	58,34%
IV3	68,12%	68,05%	65,00%	70,83%
IV6	64,57%	65,00%	60,91%	66,67%
T1	56,67%	53,97%	61,52%	56,87%
IV12	52,04%	48,88%	56,67%	51,43%
IV15	51,85%	48,88%	63,64%	45,45%
T2	50,61%	50,15%	52,50%	46,43%
IV27	49,45%	53,33%	48,33%	40,00%
IV36	50,00%	50,00%	NaN	50,00%

Tabelle 4.24: Die Tabelle zeigt die balancierte Korrektheit (balanced accuracy) Ergebnisse der GAF S Prädiktionen. Es wurde Post-Hoc auf die Regessionsergebnisse der Grenzwert von 65 angewendet. Die BAC über alle Zeitpunkte betrug 57,62%

Kapitel 5

Diskussion

Im Rahmen dieser Arbeit kann gezeigt werden, dass sowohl mit dem GF als auch mit dem GAF reliable Vorhersagen getroffen werden können. Diese übersteigen mit dem hier präsentierten LSTM Modell auch die Vorhersagequalität anderer, bereits an ähnlichen Datensätzen, genutzter Algorithmen [107, 65]. So war es bisher durch den SVM (eng: support vector machine) [1] Algorithmus möglich gewesen bei einem Modell, welches mit GF S Daten zum Einschluss den GF S in 9 Monaten bei ROD Patienten vorhersagten konnte, eine BAC von 66.1% zu erreichen. Im Rahmen der gleichen Studie wurden auch die GF S Werte von CHR Patienten vorhergesagt, welche eine BAC von 76,9% aufwiesen [107]. Bei unserem Modell konnte über alle Zeitpunkte der ROD Patienten in GF S Klassifikationsanalyse eine BAC von 88,01% und bei den CHR Patienten eine BAC von 82,48% erreicht werden. Bei Betrachtung der T1 Ergebnissen, welche einem Neunmonats-Follow-up entsprechen, waren die BAC Werte mit 91,58% bei den ROD Patienten und 87,50% bei den CHR Patienten ebenfalls höher als in der Vergleichsstudie (siehe Tabelle 4.3). Auch das rekursive Modell, welches ausschließlich Daten, die zum Einschluss erhoben wurden, für dessen Prädiktionen nutzt, zeigt mit einer BAC von 75,41 % bei den ROD Patienten zu T1 und bei einer BAC von 63,33% bei den CHR Patienten, jeweils eine bessere Vorhersagegüte (siehe Tabelle 4.7). Der GF S erzielte im Rahmen unserer Auswertung die beste Vorhersagegenauigkeit, ist jedoch, anders als der GAF, klinisch weniger etabliert.
Mit dem GAF D/I, konnten bei einer BAC von 83,53% (siehe Tabelle 4.14) zuverlässig vorausgesagt werden, ob der Patient in den nächsten drei Monaten einen GAF D/I über oder unter 65 aufweisen wird. Im Rahmen einer wissenschaftlichen Arbeit von Koutsouleris et al. aus dem Jahr 2017 konnte der GAF nach 12 Monaten bei Patienten, die ihre erste Psychose erlitten haben, mit einer BAC von 71,1% vorhergesagt werden. Die Arbeit verwendete zahlreiche weitere Parameter, wie etwa das Geschlecht, Alter, sozioökonomischer Status und den

[1] Durch das zusätzliche Einführen von Vektoren in einem Vektorraum sog. „Kerneltrick", kann die Separierbarkeit eines Datensatzes verbessert werden. Mit Hife eines SVMs können sowohl Regressionen als auch 2-Gruppen Klassifizierungen durchgeführt werden [172]

Bildungsgrad [65]. Die BAC nach 12 Monaten betrug bei den ROP Patienten im GAF D/I LSTM Klassifikationsmodell 88,24%. Die GAF S Klassifikation konnte nur einen BAC Wert von 67,14% erzielen. Die GAF S Regressionsergebnisse erreichten mit einer BAC von 75,71% (siehe Abbildung 4.22) jedoch auch genauere Vorhersagen. Traditionell galt der GAF als ein Score, dessen prädiktive Fertigkeiten beschränkt sind [46]. Diese Arbeit, aber auch andere Arbeiten weisen auf Gegenteiliges hin [173]. Die vorgestellte Arbeit von Koutsouleris et al. aus dem Jahr 2017 verwendete den allgemeinen GAF Wert und unterschied nicht zwischen GAF S und GAF D/I. Im Rahmen dieser Arbeit wiesen die Vorhersagen, welche den GAF S als Grundlage nutzten eine deutlich schlechtere Vorhersagequalität auf, was darauf hinweist, dass der GAF S inhärent eine geringere prädiktive Qualität aufweist. Bei der Beurteilung eines einzelnen GAF Wertes, werden zu Beurteilung auch die subjektiven Symptome der jeweiligen Patienten automatisch Hinzugezogen. So könnte die prädiktive Qualität des GAF beeinträchtigt werden. Im Ergebnisteil war deutlich zu erkennen, dass das reine GAF D/I Modell mit einer Gesamt BAC von 83,53% (siehe Tabelle 4.14) bessere Ergebnisse lieferte als das GAF S Modell, welches eine Gesamt BAC von 78,28 % aufweist (siehe Tabelle 4.20). Tatsächlich konnte mit dem GF S die höchste BAC erreicht werden, was sich auch mit der Arbeit von Koutsouleris et al. deckt [107]. Der GF S ist ein Maß mit dem nicht die subjektiv empfundenen Symptome bewertet werden, sondern der Umgang des Patienten mit sozialen Situationen und dessen Qualität seiner sozialen Kontakte [73]. Stellen wir nun die jeweiligen Klassifikationsergebnisse gegenüber (siehe Tabelle Anhang B.3.1), zeigt sich deutlich, dass im Rahmen der GF S Analysen bei den ROD und bei den ROP Patienten die besten BAC Werte erzielt werden konnten. Die beste Performance bei den CHR Patienten findet sich bei den GAF D/I Analysen.
Die hohe Güte der GF S Prädiktionen bei den ROD Patienten deckt sich mit vorangegangen Publikationen, in denen die Verbesserung der sozialen Funktionsfähigkeit mit einer Linderung der depressiven Symptome verbunden war [174, 175]. Ebenso ist die soziale Funktionsfähigkeit mit dem funktionalen Outcome bei Patienten mit einer Erkrankung aus dem schizophrenen Formenkreis korreliert [176, 177]. Bei CHR Patienten war etwa Arbeitslosigkeit, niedriger Bildungsgrad, sowie eine niedrige soziale Funktionsfähigkeit bezeichnend. Da für die Bewertung des GAF D/I sowohl die Funktionsfähigkeit im Alltag und Beruf, als auch die sozialen Beeinträchtigungen bewertet werden, könnte dies die hohe BAC bei den CHR Patienten erklären [178]. Eine Übersicht der Klassifikationsergebnisse kann im Anhang unter A.3.1 eingesehen werden.
Durch die LSTM Regression wird versucht den nächsten, zukünftigen Wert so gut wie möglich anzunähern. Die Abweichung von diesem Zielwert kann in Form des RMSE gemessen werden, der im besten Fall so niedrig wie möglich sein sollte, da dies auf eine geringe Abweichung hinweisen würde. Im Vergleich wird deutlich, das der GF S die genausten Regressionsergebnisse liefert, während der GAF S über alle Gruppen die schlechteste Performance zeigt. Übereinstimmend mit dem Klassifikationsmodell kann auch das GF S Modell die besten Ergebnisse für die ROP und ROD Populationen liefern. Mit einem Wert von 0,69 und

0,72 findet sich hier jeweils der niedrigste RMSE über alle Modelle. Anders als zuvor liefert nicht mehr der GAF D/I die besten Ergebnisse für die CHR Population, sondern ebenfalls der GF S.
Der GF R konnte über alle Gruppen einen RMSE von 0,76 erzeugen und wies somit einen geringfügig höheren Wert als das GF S Modell auf. Mit einer AUC von 0,91 übertraf der GF R den GF S sogar, der eine AUC von 0,90 aufwies. Eine AUC von 0,90 oder größer, spricht für ein exzellentes Modell [149]. Die Regressionsergebnisse, sowie die AUC-Werte der jeweiligen Modelle können im Anhang unter B.3.2 in der Genüberstellung betrachtet werden.

5.1 Klinische Implikationen der Ergebnisse

Der PRONIA Datensatz bestand ausschließlich aus ersterkrankten Patienten, welche entweder an einer Erkrankung aus dem schizophrenen Formenkreis erkrankt waren, sich erstmals mit einer depressiven Episode vorstellig machten, oder die CHR Kriterien (siehe Material und Methoden 3.2.1) erfüllten [57]. In unserer Arbeit zeigte der GF und im Speziellen der GF S die beste Prädizierbarkeit. Auch bei dem rekursiven Modell konnte das GF S Modell noch einen Korrelationskoeffizienten $S\rho$ von 0,57 erreichen. Sowohl bei dem rekursiven GF R Modell als auch bei dem GAF D/I betrug dieser 0,51. Beim GAF S konnte für $S\rho$ ein Wert von 0,47 ermittelt werden. Der GF mit den Subdomänen „*Role*ünd „*Social Functioning*" wurde von Cornblatt et al. etabliert, um CHR Patienten mit einem erhöhten Transitionsrisiko besser detektieren zu können [179].
Unsere Ergebnisse weisen darauf hin, dass der GF nicht nur für CHR Patienten eine geeignete Metrik ist, sondern wie der GAF auch für andere Erkrankungsbilder zur Beurteilung des Funktionsniveaus herangezogen werden kann. Besonders zu beleuchten sind in diesem Zusammenhang die exzellenten Vorhersagen, die mit dem GF S Modell auch bei ROP und den ROD Patienten erzeugt werden konnten. Da es auch bei dem rekursiven Modell zu einer moderaten Korrelation kam, weist dies darauf hin, dass die bei Einschluss ermittelten Werte bereits einen deutlichen Einfluss auf den weiteren Krankheitsverlauf der Patienten haben. Ein verringertes soziales Funktionsniveau konnte als eigene, unabhängige Trajektorie identifiziert werden, die unabhängig von Positivsymptomen den Krankheitsverlauf der CHR Patienten maßgeblich beeinflusst [180]. In einem Review zur Social Media Nutzung von Probanden konnte gezeigt werden, dass positive Interaktionen, soziale Unterstützung und soziale Verbundenheit durchweg mit einem niedrigeren Niveau von Depressionen und Ängsten verbunden war, während negative Interaktionen und soziale Vergleiche mit einem höheren Niveau von Depressionen und Ängsten korrelierte [181]. Ein chronischer Verlauf psychiatrischer Erkrankungen war mit einem verringerten sozialen Funktionsniveau und kleineren Familienverbünden assoziiert[182]. Zudem war Psychotherapie bei schizophrenen Patienten mit der Verbesserung des sozialen Funktionsniveau assoziiert, während eine intensive

Pharmakotherapie über mehrere Monate mit einer Reduktion des sozialen Funktionsniveaus in Verbindung gebracht wurde [177]. Dies scheint sich auch in den Daten (Tabelle 4.4) der GF S Analyse widerzuspiegeln. Bis T1 sind die RMSE Werte mit einem Durchschnittswert von 0,81 deutlich höher, als im späteren Verlauf. Von IV12 bis IV36 betrug der RMSE im Schnitt nur noch 0,61. Dies deutet darauf hin, dass sich vor allem in den ersten Monaten der Erkrankung eine deutliche Variabilität des sozialen Funktionsniveaus zeigt. Bei zuletzt einem RMSE von 0,55 bei IV36, scheint das soziale Funktionsniveau einem festen Muster zu folgen. Wissenschaftliche Arbeiten, die interventionell das soziale Funktionsniveau von Patienten gezielt verbessern wollten, konnten im Rahmen einer Literaturrecherche nicht gefunden werden.
Insgesamt scheint die Thematik des Sozialen- , aber auch des Rollenfunktionsniveaus, also der Erfüllung beruflicher und familiärer Aufgaben, erst in den letzten Jahren mehr an Bedeutung zu gewinnen. Auffällig ist, dass es bisher keine einheitliche Nomenklatur gibt.
Gibt es alternative Erklärungsansätze, die die Prädiktionsgüte des GF S erklären könnten? Da es sich bei dem PRONIA Datensatz um eine Kohorte Ersterkrankter handelte, waren diese mit einem Durchschnittalter von 24,96 Jahren verhältnismäßig junge Patienten. Der jüngste Patient innerhalb des Datensatzes war 15 Jahre alt. Auch dies könnte die gute Prädizierbarkeit des GF S im Vergleich zum GAF D/I und zum GF R erklären.
Viele, jedoch nicht alle, unserer Probanden waren Studierende oder Schülerinnen und Schüler. Die Bewertung des Rollenfunktionsniveaus bezog sich daher bei diesen Probanden fast ausschließlich auf deren schulische oder universitäre Leistung. Da vor allem Schulen viel Struktur vorgeben und Schule selbst ein protektiver Faktor ist, der bei einem starken Klassenverbund das Risiko an einer psychiatrischen Erkrankung zu leiden deutlich verringert, könnte durch die unterschiedlichen Lebenssituationen und somit Verpflichtungen der Patienten eine einheitliche Bewertung des GAF D/I, sowie des GF R deutlich erschwert worden sein [183, 184]. Zudem scheint bei Schülern mit psychiatrischen Erkrankungen die Funktionsfähigkeit im Alltag vom Geschlecht abhängig zu sein [185]. In einer bekannten Studie von Fergusson et al. konnte herausgefunden werden, dass Patienten, die bereits im Alter zwischen 14 und 16 Jahren depressiv waren, ein mehr als dreimal so hohes Risiko für spätere Depressionen und ein mehr als doppelt so hohes Risiko für das spätere Auftreten von Angststörungen aufwiesen. Diese Ergebnisse waren unabhängig von sozialen Hintergründen, familiären Umständen, individuellen Merkmalen und anderen Komorbiditäten [186]. Durch die schützenden Strukturen der Schule oder Familie, sowie einem gesicherten Wohnort, den Jugendliche häufig aufweisen, ist es wahrscheinlich, dass die funktionalen Beschwerden aller Probanden über die GF S Achse des GF besser abgebildet werden konnten. Folglich könnte die Hinzuziehung des Geschlechts und Alters die Vorhersagequalität aller Modelle zusätzlich verbessern. In einem Review konnte festgestellt werden, dass Depressionen zu Einschränkungen im Bereich Familie, Ehe, Beruf, aber auch der körperlichen Funktionsfähigkeit und der allgemeinen Zufriedenheit führen [174]. Dies erzeugt verständlicherweise subjektiven Leidensdruck, der vor allem durch den GAF S erfasst wird. Inner-

halb der berichteten Symptomschwere und deren Verläufe konnten von Seiten des Modells teils nur ungenügende Vorhersagen gemacht werden. So betrug bei dem rekursiven GAF S Modell die Genauigkeit lediglich 55,04% und die BAC 57,62%. Dieser Unterschied könnten darin begründet sein, dass sich der GAF D/I, aber auch der GF, anders als der GAF S an objektivierbareren Kriterien bedient. So wird für den Bereich von 41-50 beim GAF D/I verlangt, dass jegliche schwere Einschränkung im Bereich Schule, Arbeit oder Sozialleben automatisch zu einer Kategorisierung in diesen Bereich führt. Demgegenüber verlangt die Beschreibung des GAF S, dass eine Ziffer zwischen 41 und 50 gewählt werden soll, wenn die Person an „schweren Symptomen“ leidet. Als Beispiele werden häufige Diebstähle, schwere Zwangsrituale und Suizidgedanken genannt [187]. Symptome können intraindividuell, etwa auf Grund der inhärenten Persönlichkeit der jeweiligen Patienten, sowie auf Grund ihrer psychiatrischen Erkrankungen, sehr unterschiedlich geschildert werden. Ob ein Patient fähig ist zu arbeiten, ist ein objektiveres Maß und kann bei berechtigtem Zweifel leichter fremdanamnestisch erhoben werden. Zudem können Symptome in viel kürzeren Intervallen schwanken als das Maß der Beeinträchtigung [188]. Somit ist in einer GAF S Erhebung, die alle drei Monate stattfindet, mit einem deutlich größeren Hintergrundrauschen zu rechnen, was zuverlässige Vorhersagen deutlich erschwert. Die divergierenden Ergebnisse bezüglich der Prädizierbarkeit des GAF D/I und des GAF S könnten die Widersprüche in der aktuellen wissenschaftlichen Literatur bezüglich des GAFs [46] begründen. Zudem könnte die Verwendung von diagnosespezifischen Symptomen und Funktionskriterien bei der Bewertung den Informationsgehalt und somit die Prädizierbarkeit des GAF zusätzlich verbessern [189]. Die Unterscheidung zwischen GAF S und GAF D/I reflektiert unterschiedliche Aspekte klinischer Beeinträchtigung. Der GAF D/I nimmt häufig niedrigere Werte an als der GAF S. Der GAF als globaler Indikator ist ein gutes Maß, um die Belastung durch Krankheitssymptome, sowie eine Einschränkung des sozialen und beruflichen Funktionsniveaus einzuschätzen. In etwa 10% der Fälle unterscheiden sich der GAF S und der GAF D/I deutlich. Typischerweise wird dann zu der Bewertung des GAF der niedrigere Wert aus den beiden Domänen herangezogen [116].

5.1.1 Der GAF im klinischen Alltag

Inzwischen gehört der GAF zur psychiatrischen Routine in den Krankenhäusern. Der GAF erwies sich im klinischen Setting als nützlich bei der Kategorisierung von Patienten und bei der Erkennung von Veränderungen zwischen Aufnahme und Entlassung [190]. Im DSM-IV ist der GAF in der Achse V enthalten [191] und unterstreicht die Bedeutung, welche dieser bereits gewonnen hat. Dennoch gibt es auch zahlreiche Daten, die an der Aussagekraft des GAFs zweifeln lassen. In einer Studie in welcher, der GAF von Klinikern im Rahmen der stationären Routine in einem Krankenhaus erhoben und im Anschluss mit den Bewertungen von Wissenschaftlern verglichen wurde, war die Interraterkorrelation gering (ICC = 0,39)[72]. Es gibt viele Ursachen, die den GAF im klinischen Routinealltag beeinträchtigen

könnte. In Forschungsprojekten werden Rater üblicherweise trainiert, um die Skalen und Fragebögen exakt und gemäß der Empfehlungen ausfüllen zu können, während dies im klinischen Alltag nicht gegeben ist. Ein Rater, der im Rahmen einer Studie tätig ist, weiß, dass seine erhobenen Werte dokumentiert und gesammelt werden. An den erhobenen Datensätzen hat dieser ein direktes Interesse, da es bei einem sauer erhobenen Datensatz deutlich wahrscheinlicher sein wird, dass aus der Arbeit eine Publikation resultiert. Bei größeren Projekten erfolgen meistens Qualitätskontrollen des Datensatzes. Ungenau erhobene Datensätze könnten im Rahmen einer Kontrolle auf die jeweilige Person zurückgeführt werden, während dies im klinischen Setting nicht unbedingt der Fall ist[192]. GAF Ratings werden meistens zur Verlaufsdokumentation des Funktionsniveaus bei Aufnahme, sowie bei Entlassung erhoben, dabei könnte die behandelnde Ärztin oder der behandelnde Arzt geneigt sein eine Verbesserung des Niveaus einzutragen, obwohl dies nicht unbedingt gegeben ist [187]. Inzwischen ist die Verordnung kassenärztlicher Leistungen an GAF-Werte geknüpft. So muss je nach Indikation ein GAF$\leq$ 50 oder $\leq$ 40 vorliegen, damit Ärzte psychiatrische häusliche Krankenpflege verordnen können [193]. Da die häusliche Krankenpflege als konkretes Beispiel von dem GAF-Wert bei Entlassung abhängt, kann auch dies das Rating maßgeblich beeinflussen. Um die Interraterkorrelation auch im Klinkalltag zu verbessern, sind Weiterbildungen und objektive, strukturierte Bewertungen sehr wichtig, da so nachweislich die Verzerrung von Daten reduziert werden kann [194, 195]. Um den GAF und somit auch den Vorhersagealgorithmus im stationären und ambulanten Setting sinnvoll nutzen zu können, müssen die Anwender vorher ausreichend geschult worden sein, um reliable Prädiktionen erhalten zu können. Die Bewertung, sowohl des GAF S, als auch des GAF D/I bietet einige Vorteile, da die Art und Weise, wie sich die Skalen im Rahmen der Therapie verändern, unterschiedlich ist. Patienten die bei Einschluss keine Unterschiede im Bezug auf ihre GAF S und ihre GAF D/I Bewertung aufzuweisen hatten, zeigten häufig bei Entlassung deutliche Unterschiede. Bei diesen Patientin würde die Verwendung eines einzelnen GAF Wertes die Zustandsverbesserung verschleiern, da für die Bewertung des GAF jeweils die Achse mit dem niedrigsten Wert herangezogen werden würde [116]. Um die Prädiktionsfähigkeit des GAF zu verbessern, empfehlen wir, basierend auf den Erkenntnissen dieser Arbeit, die zukünftige Bewertung des GAF S und des GAF D/I als individuelle Metriken.

5.1.2 Der Einfluss der Pharmakotherapie

Die Pharmakotherapie bildet eine wichtige Säule in der psychiatrischen Behandlung von Patienten. Manche Patienten wünschen die Pharmakotherapie zur Besserung ihrer Beschwerden, während andere Medikamente ablehnen und diese im Verlauf absetzen. Absetzphänomene führen häufig zu einer Verschlechterung des psychopathologischen Zustandes der betreffenden Personen und somit auch zu einer Beeinträchtigung des Funktionsniveaus [196]. Im Rahmen unseres Modells wurden keine Daten zur Medikamenteneinnahme mitaufgenommen, dennoch konnten sehr gute Vorhersagen erreicht werden.

Verschiedene Effekte könnten diese Diskrepanz erklären. So könnten Patienten, die nicht compliant bezüglich ihrer Medikation waren, auch die Studie vorzeitig abgebrochen haben. Somit wäre es möglich, dass die damit verbundenen Einbrüche des Funktionsniveaus innerhalb des PRONIA Datensatzes nicht abgebildet wurden [197]. Auf Grund der klinikinternen, sowie der Abläufe im Rahmen der PRONIA Studie konnten die Patienten nicht am ersten Tag bei Aufnahme direkt eingeschlossen werden. Erst nach ausführlichem Screening der Patienten und nachdem diese schriftlich der Teilnahme an der Studie zustimmten, begannen die Studienteams mit dem Einschluss der Probanden [57]. Dies führte dazu, dass Patienten bei Einschluss häufig schon seit mehreren Tagen mit einer initalen Medikation anbehandelt worden waren. Bei Einschluss wurden bei den GF und GAF Bewertungen jeweils auch retrospektive Werte ermittelt (siehe Material und Methoden 3.7.3), somit könnte das Ansprechen auf eine bereits eingeleitete Therapie in den Funktionsniveaus bei Einschluss bereits abgebildet sein. Für Replikationsstudien sollten diese Effekte in Erwägung gezogen werden. Im Rahmen der PRONIA Studie wurde die Medikation der jeweiligen Patienten anamnestisch erhoben.
Es gibt Hinweise, dass das soziale Umfeld einen größeren Effekt auf das langfristige Funktionsniveau haben könnte, als die Pharmakotherapie [198, 177].

5.1.3 Remission und Recovery

Im Rahmen der Behandlung von Depressionen konnte bereits früh beobachtet werden, dass eine vorhandene Residualsymptomatik nach erfolgter Behandlung zu einem erhöhten Risiko führt erneut an einer Depression zu erkranken. Es zeigte sich, dass die Symptomfreiheit ein besserer Bewertungsstandard für Studien war als die ledigliche Besserung der Symptome [199]. Diese Erkenntnisse führten auch bei anderen Erkrankungen wie etwa der Schizophrenie dazu, dass man zunächst eine vollständige Genesung, also eine *Recovery* anstrebte. Auf Grund der klinischen Erkenntnisse, stellte sich jedoch heraus, dass nach einer psychotischen Episode zumeist Krankheitsresiduen übrig blieben und eine vollständige Genesung bei Patienten mit einer Schizophrenie häufig nicht möglich ist. So kam es im Verlauf zu einem Paradigmenwechsel in der Schizophreniebehandlung [200].
Statt einer *Recovery*, wird die *Remission* der Symptome angestrebt, welche durch die Bestimmung des Funktionsniveaus gemessen wird. Demnach sollte das Ansprechen auf die Therapie mit Hilfe von Parametern bestimmt werden, die das Funktionsniveau messen [201]. Dies führte dazu, dass in den letzten Jahren der Schwerpunkt zunehmend auf klinische Behandlungsergebnisse, sowie auf die funktionelle Besserung, gelegt wurde. In einer bekannten wissenschaftlichen Arbeit von Andreasen et al. konnte herausgefunden werden, dass neurokognitive Defizite, wie die Verarbeitungsgeschwindigkeit mit der Sozialen- und der Rollenfunktionsfähigkeit assoziiert war. Diese Erkenntnisse wiesen zudem darauf hin, dass eine Beeinträchtigung der Funktionsfähigkeit bereits vor dem Beginn einer Psychose vorhanden ist [202]. Über die Funktionsfähigkeit der Patienten kann in der Folge eine Verschlechte-

rung oder die Verbesserung des Krankheitszustandes abgeleitet werden. In der Literatur wurden verschiedene Metriken diskutiert, um die Beeinträchtigung des Funktionsniveaus, sowie *Remission* oder *Recovery* zu messen [203]. Insbesondere durch die hier präsentierten GF und die GAF D/I Modelle können genaue Vorhersagen zum Funktionsniveau und somit der Krankheitslast in drei Monaten bestimmt werden. In einem Systematischen Review von Mausbach et al. schlussfolgerte dieser, dass es aktuell keinen „Goldstandard“ gäbe, um eine reliable Vorhersage des Funktionsniveaus zu treffen. Insbesondere wurde der eindimensionale GAF als unzureichend für die Messung der Funktionsfähigkeit befunden [204]. Wir schlagen im Rahmen dieser Diskussion den GF oder den GAF D/I als neue Metrik vor, die zur Bewertung des Funktionsniveaus herangezogen werden kann.

5.1.4 Das aufstrebende Gebiet der „Precision Psychiatry”

Aufbauend auf dem Konzept der „*Präzisionsmedizin*” (eng: precision medicine), bildete sich in der jüngsten, wissenschaftlichen Literatur der Begriff der „Precision Psychiatry” heraus [205]. Das zu Grunde liegende Konzept der Präzisionsmedizin ist, für den jeweiligen Patienten gemäß seiner klinischen Daten, zugeschnittene Behandlungskonzepte anbieten zu können. Fortschritte in den Bereichen Genetik, Bildgebung und Datenverarbeitung lassen diese Zukunftsvision sukzessive zu klinischen Realität werden [206]. Die Psychiatrie im Speziellen konnte von diesen Fortschritten bisher nicht ausreichend profitieren [207].
Dennoch konnten in einigen Bereichen der psychiatrischen Forschung deutliche Fortschritte erzielt werden. Die Präzisionspsychiatrie läutete unter anderem die kritische Beleuchtung und Veränderung alter Behandlungskonzepte und Theorien zur Krankheitsentstehung ein. Die Methoden der Präzisionspsychiatrie ermöglichen einen neuen Zugang um durch Phänotypisierung, etwa der Gehirnfunktion und dessen Physiologie, neue Erklärungsmodelle zur Pathophysiologie zu entwickeln [208]. Mit Maschinellem Lernen können hoch-dimensionale Datensätze präzise ausgewertet werden, was einerseits neue neurobiologische Erkenntnisse generiert, anderseits die Entwicklung individueller Biomarker ermöglicht, sowie die Vorhersage von Ansprechen auf Therapie und Prognose verbessern kann [205]. Neuere Daten suggerieren, dass durch Clusteralgorithmen gefundene Subgruppen innerhalb einer psychiatrischen Patientenkohorte, Behandlungsoutcomes besser vorhergesagt werden können, als durch die DSM oder ICD - 10 Diagnosen [209]. Da wir im Rahmen dieser Arbeit maschinelles Lernen nutzen, um individuelle, funktionelle Outcomes zu prädizieren, schlagen wir in die Kerbe der „Precision Psychiatry” und legen den ersten Stein in Richtung klinischer Anwendung, dieses aufstrebenden Konzeptes. Wir bieten neue Perspektiven für eine integrative und evidenzbasierte Medizin. Die Vorhersage des Funktionsniveaus mittels LSTM Netzen kann ein neues Zeitalter der personalisierten Psychiatrie einläuten.

5.2 Die Entwicklung eines Modells für die klinische Anwendung

Laut Fusar-Poli et al. erweist sich bei der Entwicklung neuer Prädiktionsmodelle der Transfer in die Klinik am kompliziertesten. Aktuell werden zahlreiche Modelle entwickelt, die die klinische Prognose verbessern sollen. Da deren Nutzung häufig sehr komplex ist und teils sehr teure oder aufwändige Anwendungen genutzt werden, kommt es meistens nicht zu einer tatsächlichen klinischen Anwendung [66]. Da sich die jeweiligen hier vorgestellten Modelle nur jeweils einem Fragebogen bedienen, gestaltet sich der Transfer in die Klinik verhältnismäßig einfach. Es sind keine zusätzlichen, teuren und zeitaufwändigen MRT Aufnahmen oder neurokognitiven Testungen nötig, die ebenfalls viel Zeit kosten und speziell geschultes Personal in Anspruch nehmen. Insbesondere der GAF wird bereits regelmäßig genutzt. Der Score findet viel Anwendung in der Klinik und wird teils auch explizit in Arztbriefen angeführt. Ein GAF Score kann auch telefonisch erhoben werden[210]. Es gibt Belege dafür, dass die von geschultem Personal der Psychiatrie durchgeführten GAF-Ratings eine zufriedenstellende Zuverlässigkeit aufweisen [211, 212]. Generell würden wir auf Basis unserer gewonnenen Erkenntnisse die Nutzung des GAF D/I oder des GF vorziehen.

5.2.1 Anwendung des LSTM Modells in der Praxis

Diese aber auch zukünftige Entwicklungen, können die Arbeit der Ärzte und Psychologen in Zukunft erheblich erleichtern, können deren Präsenz aber nicht ersetzen [213]. Potenzielle Schäden, die Risikokategorisierung und Risikoskalen den Patienten zufügen können wären etwa, dass Patienten, die als „Hochrisikopatienten“ bezeichnet werden, unnötig restriktiver behandelt werden könnten. Somit könnte die Stigmatisierung der Betroffenen verstärkt werden [214]. In der klinischen Praxis könnte die Risikobewertung Ärzten und Pflegepersonal falsche Sicherheit geben [215]. Auf Grund dieser Problematik sind weitere Studien nötig, die das Verhalten der Anwender untersucht und inwiefern die Prädiktionen dadurch verändert werden [216]. Der scharfe klinische Blick muss trotz Prädiktionsalgorithmen im Allgemeinen erhalten bleiben. Ein Neuronales Netz kann nur wiedererkennen, was es im Rahmen seines Trainings erlernte. Prädiktionen, die mit Daten, die etwa ganz andere Patientengruppen betreffen, generiert werden, sind deutlich ungenauer [74]. Unsere Anwendung kann, Ärzte und Therapeuten in Ihrer Arbeit unterstützen, ist aber auf deren korrekte Einschätzung angewiesen. Nur wenn sich der Anwender bei einem Follow-up nach drei Monaten durch eine falsche Zeitprädiktion nicht beirren lässt und eine erneute, unverfälschte GAF Bewertung in den Algorithmus einspeist, ermöglicht dem LSTM die Fehlprädiktion zu korrigieren und im nächsten Abschnitt eine korrektere Prognose für das kommende Intervall abzugeben.

5.2.2 Voraussetzungen zur Anwendung im ambulanten Setting

Die aktuellen Daten wurden in 10 verschiedenen Standorten aus fünf verschiedenen Ländern generiert. Dabei waren jeweils Krankenhäuser beteiligt, also fand die Datengenerierung hauptsächlich im stationären Umfeld statt. Da sich eher kränkere Patienten im stationären Rahmen bewegen, könnte der aktuelle Algorithmus im ambulanten Setting dazu tendieren eher zu schlechte Prognosen vorherzusagen. Um dies zu zeigen oder zu widerlegen wären jedoch weitere Daten aus dem ambulanten Rahmen nötig [217].
Folglich müsste eine erneute Follow-up Studie gemäß dem PRONIA Schema im ambulanten Rahmen erfolgen, um die Anwendbarkeit des Algorithmus zu belegen. Solche Daten existieren aktuell noch nicht [65]. Im Verlauf sind Replikationsstudien in einem ambulanten Setting notwendig.

5.2.3 Mögliche Ansätze für die klinische Anwendung

Für die tatsächliche klinische Anwendung schlagen wir einen kombinierten Ansatz vor. Ein möglicher Weg, um dieses Modell klinisch Umzusetzen wäre als Experten-Anwendung: Sucht ein Patient erstmals das stationäre Setting auf, kann von dem ausgebildeten Behandler der GF, sowie der GAF ermittelt werden. Mit Hilfe des rekursiven Modells erhält dieser Vorhersagen bis IV36. Bereits Bei T0 scheint der GF S viel Information bezüglich der weiteren Entwicklung des sozialen Funktionsniveaus zu bergen, sodass bereits bei Einschluss der Patienten Trends zu erkennen sind. Bei weiterer regelmäßigen Nutzung der Anwendung, werden die zusätzlichen Werte für die Vorhersagen herangezogen und die Vorhersagen aktualisiert. Um die besten Ergebnisse zu generieren, müssten pro Patient Datenbanken angelegt werden, um die vorhanden Funktionsniveau Daten zu speichern. Dabei ist wichtig die Daten der Patienten so zu pseudonymisieren, dass Dritte die Daten nicht auf die jeweilige Person zurückführen könnten. [218].
In der Literatur gibt es auch Überlegungen zu App-basierten Anwendungen, die sich der Patient auf sein Mobiltelefon herunter laden kann. Auch hier könnte über die Nutzung einer zweiten Anwendung, die nur ein Behandler nutzen kann, eine Bewertung des Funktionsniveaus erfolgen [180]. Jedoch stellt sich auch hier die Frage, ob die Daten im Anschluss auf den Mobiltelefonen oder auf einem Server gespeichert werden sollten. Bei einer lokalen Sicherung auf dem Handy würde ein Verlust der Hardware auch den Verlust der Verlaufsdaten bedeuten. Die Wahrung des Datenschutzes und der Datensicherheit, sollte für die Entwicklung eines Proptotypen in Erwägung gezogen werden [219].
Für die Anwendung des Algorithmus in der Klinik scheint die LSTM Regression die zu bevorzugende Modalität zu sein. Einerseits können so Vorhersagen bis IV36 erfolgen, andererseits kann bei Verwendung anderer Grenzwerte (siehe 5.3.3) die Performance der LSTM Regression zusätzlich verbessert werden.

5.3 Methodische Überlegungen

5.3.1 Die Größe des Datensatzes

Mit steigender Probandenzahl ist auch in unserem Modell mit einer weiteren Steigerung der BAC zu rechnen. Ebenso wird sich der RMSE bei den Regressionsmodellen deutlich verbessern [220]. Üblicherweise arbeiten Neuronale Netze mit sehr großen Datensätzen, deren unteren Grenze etwa 1000 Stichproben sind. Da wir mit etwa 400 Datensätzen (ja nach Konstellation etwas abweichend) unsere Netze trainiert haben, befinden wir uns mit unserem Datensatz tief im unteren Spektrum. Bei einem weiteren Trainieren des Algorithmus mit weiteren Daten, ist mit einer Verbesserung der Prädiktionen zu rechnen.
Die einzelnen Zeitintervalle waren im Rahmen des Projektes nicht immer drei Monate von einander entfernt. Vor allem T1 Untersuchungen (Untersuchung nach 9 Monaten nach Einschluss) wurden teils sehr spät nachgeholt. Patienten, die kleine Intervalluntersuchungen aufsuchten, nahmen die Termine pünktlicher wahr.

5.3.2 Replizierbarkeit und Reproduzierbarkeit der Ergebnisse

Obwohl viele Machine Learning-Studien von erheblichen Vorteilen gegenüber anderen State-of-the-Art-Modellen hinsichtlich Effektivität berichten, lassen sie oft zwei Faktoren außer Acht:

1. Replizierbarkeit - ob das berichtete experimentelle Ergebnis mit hoher Wahrscheinlichkeit mit demselben Modell und denselben Daten annähernd reproduziert werden kann.

2. Reproduzierbarkeit - ob ein berichtetes experimentellen Ergebnisse durch neue Experimente mit demselben experimentellen Protokoll und Modell, aber unterschiedlichen realen Daten reproduziert werden können.

Ein Blick auf die LSTM Trainingsgraphen, wie etwa in Abbildung 4.4, lässt erkennen, dass die Genauigkeiten der Test- und Trainingsdaten sich häufig überlappen. Dies ist ein Zeichen dafür, dass während des Trainings die Minibatches sehr unterschiedlich zusammengesetzt sind. Es ist eine gute Generalisierbarbeit unserer LSTM Netze gegeben, da aus den Validierungsdaten, welche jeweils aus unbekannten Daten der Universitätsklinik Köln bestand, gute BAC Werte berechnet werden konnten [221]. Somit kann von einer gegebenen Replizierbarkeit ausgegangen werden[222].
In dieser Doktorarbeit wurde jeweils der erste erfolgte Durchlauf vermerkt. Es erfolgten keine mehrmaligen Durchläufe, um das bestmögliche Ergebnis zu generieren. Bei erneuten Durchläufen könnten die Ergebnisse je nach Zusammensetzung der Trainings- / Test-

batches teils besser oder schlechter ausfallen. Es ist zu erwarten, dass bei größeren Trainingsdatensätze die Ergebnisse stabiler sein werden. Es wurde eine feste Zahl von 300 Trainings-Epochen gewählt, um die aktuellen Ergebnisse zu erzielen. Zur Kontrolle eines möglichen Overfittings wurde einerseits mit Regularisierungsmethoden und andererseits mit einer Leave-Site Out Validierung gearbeitet [223, 224]. Im Rahmen dieser Arbeit konnte gezeigt werden, dass der GAF, sowie der GF mit Hilfe von LSTM Netzen vorhergesagt werden kann. Durch weitere Experimente und zusätzliche Daten können die Prädiktionen verbessert und die Reproduzierbarkeit erhöht werden [225]. Um die Ergebnisse so reproduzierbar wie möglich zu machen, wurden die Hyperparameter für jedes verwendete Neuronale Netz im Zuge dieser Arbeit genannt. Der endgültige Nachweis der Replizierbarkeit kann erst durch Wiederholungsstudien an gänzlich anderen Daten erfolgen.

5.3.3 Alternative Grenzwerte

Bei einer GF S Regression werden keine ordinalen, sondern metrische Daten ausgegeben. Es ist ein GF Wert von zum Beispiel *6,8* als Regressionsergebnis möglich. Unter Betrachtung der GF S ROC Kurve, welche in Abbildung 4.6 zu erkennen ist, kann vermutet werden, dass ein Grenzwert von 6,5 bessere Ergebnisse liefert als der in der Literatur bisher verwendete Wert von 7. Bei einer erneuten Berechnung der BAC mit einem Grenzwert von 6,5 ergibt sich eine BAC 87,22% (siehe Anhang B.1) und übertrifft alle bisherigen Modelle. Der Negativ Prädiktive Wert betrug bei dem niedrigeren Grenzwert 91,38%.

5.4 Weitere, zukünftige Fragen

In dieser Arbeit haben wir nur an Hand von Daten gearbeitet, die sich dynamisch von Zeitabschnitt zu Zeitabschnitt verändert haben. Für weitere Experimente wäre es sehr interessant statische Daten, wie etwa das Geschlecht oder die Erhebung der CTQ (Childhood Trauma Questionnaire) zu erfassen. [2]Dieser wäre ähnlich leicht zu erheben wie der GAF, da der Patient lediglich einen Zettel ausfüllen müsste, könnte jedoch die BAC erhöhen. Ein alternativer Modellansatz wäre wie Esteban et al. in einer Publikation aus 2016 beschreibt[228]. Durch die Kombination von dynamischen und statischen Daten, könnte man zudem genetische Marker, sowie MRT Daten in einen Algorithmus einfließen lassen und somit Sensitivität und Spezifität weiter erhöhen [229]. Allerdings müsste bei einem Patienten, der den Vorhersagealgorithmus nutzen möchte in der Folge ein MRT erfolgen, sowie genetische Analysen

[2]Dieser Fragebogen besteht aus 28-Elementen, die in fünf Subgruppen unterteilt werden können: Emotionale, körperliche und sexuelle Misshandlung, sowie emotionale und körperliche Vernachlässigung[226].Die Items sind auf einer fünfstufigen Likert-Skala zu beantworten, die von „überhaupt nicht" (1) bis „sehr häufig" (5) reicht; somit reflektieren höhere Werte ein größeres Ausmaß an Misshandlungen [227].

durchgeführt werden, was den Zugang im ambulanten Setting, sowie den Transfer in die Klinik deutlich erschweren würde.

Anhang A

Ausgewählte Fragebögen

A.1 Globale Beurteilung der Funktionsfähigkeit (GAF)

Bei der Beurteilung des GAF werden psychologische, soziale und berufliche Funktionen auf einem hypothetischen Kontinuum der psychischen Gesundheit/Krankheit berücksichtigt. Einschränkungen der Funktionsfähigkeit auf Grund körperlicher Leiden oder (umweltbedingten) Einschränkungen werden nicht berücksichtigt.

A.1.1 Globale Beurteilung der Funktionsfähigkeit: Symptome (GAF S)

Beurteilen Sie den Grad der psychischen Belastung durch Krankheitssymptome, welcher die Lebenssituation innerhalb des betreffenden Zeitraumes am besten darstellt. Verwenden Sie gegebenenfalls Zwischencodes (z. B. 45, 68, 72), wenn dies angemessen erscheint.

91-100	Keine Symptome
81-90	Fehlende oder minimale Symptome (z.B. leichte Angst vor einer Prüfung)
71-80	Wenn Symptome vorhanden sind, handelt es sich um vorübergehende und erwartbare Reaktionen auf psychosoziale Stressoren (z.B. Konzentrationsschwierigkeiten nach einem Familienstreit)
61-70	Einige leichte Symptome (z.B. depressive Stimmung und leichte Schlaflosigkeit)
51-60	Moderate Symptome (z.B. flacher Affekt und umständliches Sprechen, gelegentliche Panikattacken)

41-50	Schwere Symptome (z.B. Selbstmordgedanken, schwere Zwangsrituale, häufige Ladendiebstähle)
31-40	Eine gewisse Beeinträchtigung der Realitätsprüfung oder der Kommunikation (z.B. die Sprache ist manchmal unlogisch, unverständlich oder sachfremd)
21-30	Das Verhalten wird in erheblichem Maße durch Wahnvorstellungen oder Halluzinationen beeinflusst ODER es besteht eine schwerwiegende Beeinträchtigung der Kommunikation oder des Urteilsvermögens (z.B. manchmal inkohärent, grob unangemessenes Verhalten, starkes Eingenommensein von Selbstmordgedanken)
11-20	Das Vorhandensein einer Gefahr, sich selbst oder andere zu verletzen (z.B. Selbstmordversuche ohne konkrete Todesabsichten; häufige Gewalttätigkeit; manische Erregung) ODER grobe Beeinträchtigung der Kommunikation (z.B. weitgehend inkohärent oder mutistisch).
1-10	Anhaltende Gefahr, sich selbst oder andere schwer zu verletzen (z.B. wiederholte Gewalt) ODER schwere Suizidversuche mit eindeutiger Todesabsicht.
0	Unzureichende Informationen

Schätzen Sie den jeweils höchsten Wert des betreffenden Zeitraumes der GAF-Skala SYMPTOME ab:

Während des gesamten Lebens: _____
Innerhalb des letzten Jahres: _____
Innerhalb des letzten Monats: _____

A.1.2 Globale Beurteilung der Funktionsfähigkeit: Einschränkungen im Alltag (GAF D/I)

Beurteilen Sie die Beeinträchtigung der Funktionsfähigkeit, welcher die Lebenssituation innerhalb des betreffenden Zeitraumes am besten darstellt. Verwenden Sie gegebenenfalls Zwischencodes (z. B. 45, 68, 72), wenn dies angemessen erscheint. Beeinträchtigungen der Funktionsfähigkeit aufgrund körperlicher Einschränkungen sind nicht einzubeziehen.

91-100	Überlegenes Auftreten in einem breiten Spektrum von Aktivitäten, die Probleme des Lebens scheinen nie außer Kontrolle zu geraten, wird von anderen wegen ihrer oder seiner vielen positiven Qualitäten aufgesucht
81-90	Gutes Funktionieren in allen Bereichen, interessiert und beteiligt an einem breiten Spektrum von Aktivitäten, sozial aktiv, im Allgemeinen zufrieden mit dem Leben, nicht mehr als alltägliche Probleme oder Sorgen (z.B. ein gelegentlicher Streit mit Familienmitgliedern)
71-80	Nicht mehr als eine leichte Beeinträchtigung des sozialen, beruflichen oder schulischen Funktionierens (z.B. temporärer Rückstand bei den Hausaufgaben oder Deadlines)
61-70	Einige wenige Schwierigkeiten im sozialen, beruflichen oder schulischen Bereich (z.B. gelegentliches Schwänzen oder Diebstahl innerhalb des Familienverbundes), aber im Allgemeinen funktioniert sie oder er ziemlich gut, hat einige bedeutsame zwischenmenschliche Beziehungen
51-60	Moderate Schwierigkeiten im sozialen, beruflichen oder schulischen Bereich (z.B. wenige Freunde, Konflikte mit Gleichaltrigen oder Arbeitskollegen)
41-50	Jede ernsthafte Beeinträchtigung des sozialen, beruflichen oder schulischen Lebens (z.B. keine Freunde, Unfähigkeit in einem Beschäftigungsverhältnis zu verbleiben)
31-40	Starke Beeinträchtigung in mehreren Bereichen, z.B. Arbeit oder Schule, Familienbeziehungen, Urteilsvermögen oder Denken (z.B. ein depressiver Mann meidet Freunde, vernachlässigt die Familie und ist arbeitsunfähig; Kind verprügelt häufig jüngere Kinder, ist zu Hause aufsässig und versagt in der Schule)
21-30	Unfähigkeit, in fast allen Bereichen (z.B. bleibt den ganzen Tag im Bett, keine Arbeit, kein Zuhause, keine Freunde)
11-20	Gelegentlich unterlässt er ein Mindestmaß an persönlicher Hygiene (z.B. Verschmieren von Fäkalien)
1-10	Anhaltende Unfähigkeit, ein Minimum an Körperpflege zu betreiben
0	Unzureichende Informationen

Schätzen Sie den jeweils höchsten Wert des betreffenden Zeitraumes der GAF-Skala EINSCHRÄNKUNGEN IM ALLTAG ab:

Während des gesamten Lebens: _____
Innerhalb des letzten Jahres: _____
Innerhalb des letzten Monats: _____

A.2 Globale Funktionsfähigkeit (GF)

A.2.1 Globale Funktionsfähigkeit: Social Scale (GF S)

Aktueller Wert:	
Schlechtester Wert letzten Jahres	
Bester Wert letzten Jahres	
Bester Wert innerhalb der gesamten Lebensspanne	

Die Bewertung basiert auf den Angaben der folgenden Personen (Mehrfachauswahl möglich):	
□	Proband
□	Familienmitglieder oder Freunde
□	Krankenhausmitarbeiter, Sozialpädagogen oder andere regelmäßiger Betreuer

Instruktionen: Bitte bewerten Sie den niedrigsten Grad der sozialen Funktionsfähigkeit des Patienten für den angegebenen Zeitraum, der das Funktionsniveau während dieses Zeitraums beschreibt, um die Kategorie „Schlechtester Wert letzten Jahres“ zu beurteilen. Für „Aktueller Wert“ geben Sie den am stärksten beeinträchtigten Grad der Funktionsfähigkeit in der letzten Woche an. Bewerten Sie die aktuelle Funktionsfähigkeit unabhängig von der Ätiologie der sozialen Probleme. Befindet sich der Patient jedoch in den Schulferien (z. B. Sommerferien) oder war er <2 Wochen im Krankenhaus und wird nun entlassen, ohne dass er zwischen Entlassung und Beurteilung des GF die Möglichkeit hatte, seine gewohnte soziale Rolle wahrzunehmen (z. B. Beurteilung des GF am Tag der Entlassung aus dem Krankenhaus), sollten Sie die Situation vor dem Ereignis beurteilen (z. B. Wenn sich die Person zum Zeitpunkt der Beurteilung im Krankenhaus befindet und dies <2 Wochen her ist, sollten Sie den GF S unmittelbar vor dem Krankenhausaufenthalt ermitteln). Wenn die Person aus nicht psychischen Gründen (z. B. komplizierte Grippe) länger als zwei Wochen im Krankenhaus war, erfolgt die Bewertung auch nach dem Grad der sozialen Funktionsfähigkeit unmittelbar vor dem Krankenhausaufenthalt (ebenso bei einem längeren Urlaub). Befindet sich eine Person in einer Übergangsphase (z. B. zwischen Schule und Studium oder zwischen zwei Arbeitsstellen), muss auch das Funktionsniveau unmittelbar vor dieser Situation berücksichtigt werden.
Hinweis: Der Schwerpunkt liegt auf sozialen Kontakten/Interaktionen mit Personen außer-

halb der eigenen Familie, es sei denn, diese sind die einzigen zwischenmenschlichen Kontakte, die eine Person hat (z. B. am unteren Ende der Skala). Zu beachten ist auch, dass die Bewertung der intimen Beziehungen der Beurteilung primärer Freundschaften untergeordnet ist und das Alter der Person berücksichtigt werden sollte. So ist bei älteren Personen davon auszugehen, dass sie intime Beziehungen haben, die Verabredungen, Zusammenleben mit einer anderen Person oder Heirat beinhalten, während bei jüngeren Personen lediglich romantische Interessen (d. h. Flirts oder Verknallt sein) oder enge Freundschaften zu erwarten sind.

Fragen zur Bewertung des GF S:

1. Erzählen Sie mir von Ihrem Sozialleben. Haben Sie Freunde?
2. Sind es lockere oder enge Beziehungen? Falls es sich um lockere Beziehungen handelt, sind es nur Schul- oder Arbeitsfreunde? Bei engen Freundschaften: Wie lange sind Sie schon enge Freunde?
3. Wie oft sehen Sie Ihre Freunde? Sehen Sie diese außerhalb der Arbeit/Schule? Wann haben Sie einen Ihrer Freunde das letzte Mal außerhalb der Arbeit/Schule gesehen? (Versuchen Sie, die tatsächliche Menge an sozialen Kontakten gegenüber der wahrgenommenen Menge an sozialen Kontakten zu bestimmen).
4. Initiieren Sie normalerweise den Kontakt oder die Aktivitäten mit Freunden oder rufen diese Sie normalerweise an oder laden Sie ein? Vermeiden Sie manchmal den Kontakt zu Freunden?
5. Hatten Sie jemals Probleme/ Streitigkeiten mit Freunden? Streitereien oder Auseinandersetzungen? Wie werden sie normalerweise gelöst?
6. Haben Sie einen Freund oder eine Freundin oder sind Sie an einem Freund der Freundin interessiert? (Nach Bedarf ändern, um altersgemäße intime Beziehungen zu beurteilen)
7. Verbringen Sie Zeit mit Familienmitgliedern (zu Hause)? Wie oft kommunizieren Sie mit ihnen? Meiden Sie manchmal den Kontakt zu Familienmitgliedern?

10	Herausragende soziale Kompetenz	Herausragendes Funktionieren in einem breiten Spektrum von sozialen und zwischenmenschlichen Aktivitäten. Sucht häufig andere Menschen auf und unterhält zahlreiche befriedigende zwischenmenschliche Beziehungen, darunter mehrere enge und gelegentliche Freunde. Wird von anderen auf Grund vieler positiven Eigenschaften aufgesucht. Altersgemäße Beteiligung an intimen Beziehungen (erforderlich).
9	Überdurchschnittliche soziale Kompetenz	Gutes Funktionieren in allen sozialen Bereichen und gute zwischenmenschliche Beziehungen. Interesse und Beteiligung an einem breiten Spektrum sozialer und zwischenmenschlicher Aktivitäten, einschließlich enger und lockerer Freundschaften. Altersgemäße Beteiligung an intimen Beziehungen (erforderlich). Nicht mehr als alltägliche zwischenmenschliche Probleme oder Sorgen (z.B. ein vereinzelter Streit mit dem Ehepartner, einer Freundin/einem Freund, Freunden, Arbeitskollegen oder Klassenkameraden). In der Lage, solche Konflikte angemessen zu lösen.
8	Gute soziale Kompetenz	Leichte vorübergehende Beeinträchtigung der sozialen Funktionsfähigkeit. Leichte soziale Beeinträchtigungen sind vorhanden, aber vorübergehende und erwartbare Reaktionen auf psychosoziale Stressfaktoren (z.B. nach kleineren Auseinandersetzungen mit dem Ehepartner, der Freundin/dem Freund, Freunden, Arbeitskollegen oder Klassenkameraden). Hat einige bedeutsame zwischenmenschliche Beziehungen zu Gleichaltrigen (gelegentliche und enge Freunde) und/oder altersgemäße intime Beziehungen. Seltene zwischenmenschliche Konflikte mit Gleichaltrigen.
7	Milde Beeinträchtigung sozialer und zwischenmenschlicher Beziehungen	Anhaltende leichte Schwierigkeiten beim sozialen Funktionieren. Leichte Beeinträchtigung, die NICHT nur eine erwartbare Reaktion auf psychosoziale Stressfaktoren ist (z.B. leichte Konflikte mit Gleichaltrigen, Arbeitskollegen oder Klassenkameraden; Schwierigkeiten, Konflikte angemessen zu lösen). Hat einige bedeutsame zwischenmenschliche Beziehungen zu Gleichaltrigen (gelegentlich und/oder enge Freunde). Gewisse Schwierigkeiten, altersgemäße Beziehungen zu entwickeln oder aufrechtzuerhalten intime Beziehungen aufzubauen oder aufrechtzuerhalten (z.B. mehrere kurzfristige Beziehungen)
6	Moderate Beeinträchtigung sozialer und zwischenmenschlicher Beziehungen	Mäßige Beeinträchtigung der sozialen Kompetenz. Mäßige Beeinträchtigung vorhanden (z.B. wenige enge Freunde; erhebliche, aber intermittierende Konflikte mit Gleichaltrigen, Arbeitskollegen oder Klassenkameraden). Mäßige Schwierigkeiten, altersgemäße intime Beziehungen aufzubauen (z.B. seltene Verabredungen). Gelegentlich sucht sie oder er andere auf. Reagiert, wenn sie oder er von anderen zur Teilnahme an einer Aktivität aufgefordert wird.

5	Deutliche Beeinträchtigung sozialer und zwischenmenschlicher Beziehungen	Deutliche Beeinträchtigung der sozialen Funktionfähigkeit. Keine engen Freunde oder Intimpartner, hat aber einige lockere soziale Kontakte (z.B. Bekannte, nur Schul-/Arbeitsfreunde). Sucht selten nach anderen. Gelegentlich kämpferisches oder verbal streitsüchtiges Verhalten gegenüber Gleichaltrigen. Beginnt sich von Familienmitgliedern zurückzuziehen (z. B. initiiert keine Gespräche mit der Familie, reagiert aber, wenn er angesprochen wird).
4	Schwere Beeinträchtigung sozialer und zwischenmenschlicher Beziehungen	Schwere Beeinträchtigung der sozialen Funktionsfähigkeit. Schwere Beeinträchtigung der Beziehungen zu Freunden oder Gleichaltrigen (z.B. sehr wenige oder keine Freunde, häufige Konflikte mit Freunden oder Vermeidungsverhalten gegenüber Freunden). Häufiges streitlustiges oder verbal argumentatives Verhalten mit Gleichaltrigen. Seltener Kontakt mit Familienmitgliedern (z.B. reagiert er manchmal nicht auf Familienmitglieder oder meidet sie).
3	Marginale soziale Kompetenz	Äußerst geringe Fähigkeit, in der Gesellschaft zu funktionieren oder zwischenmenschliche Beziehungen zu pflegen. Häufig allein und sozial isoliert. Schwere Beeinträchtigung der Beziehungen zu allen Gleichaltrigen, einschließlich Bekannten. Wenige Interaktionen mit Familienmitgliedern (z.B. oft allein im Zimmer). Schwerwiegende Beeinträchtigung der Kommunikation mit anderen (vermeidet z.B. die Teilnahme an den meisten sozialen Aktivitäten).
2	Unfähigkeit innerhalb der Gesellschaft zu funktionieren	Unfähig, sozial zu funktionieren oder zwischenmenschliche Beziehungen zu pflegen. Typischerweise allein und sozial isoliert. Verlässt nur selten das Haus. Geht nur selten ans Telefon oder an die Tür. Nimmt selten an Interaktionen mit anderen zu Hause oder in anderen Umgebungen teil (z.B. Arbeit, Schule).
1	Extreme soziale Isolation	Extreme soziale Isolation. Keinerlei soziale Kontakte, sowie kein Kontakt zu Familienmitgliedern. Verlässt das Haus nicht. Weigert sich, an das Telefon oder die Tür zu gehen.

A.2.2 Globale Funktionsfähigkeit: Role Scale (GF R)

Aktueller Wert:	
Schlechtester Wert letzten Jahres	
Bester Wert letzten Jahres	
Bester Wert innerhalb der gesamten Lebensspanne	

Die Bewertung basiert auf den Angaben der folgenden Personen (Mehrfachauswahl möglich):	
□	Proband
□	Familienmitglieder oder Freunde
□	Krankenhausmitarbeiter, Sozialpädagogen oder andere regelmäßiger Betreuer

Instruktionen: Bitte bewerten Sie den niedrigsten Grad der Funktionsfähigkeit des Patienten innerhalb der Bereiche Beruf, Ausbildung und / oder Hausarbeit für den angegebenen Zeitraum, der das Funktionsniveau innerhalb dieses Zeitraums beschreibt, um die Kategorie „Schlechtester Wert letzten Jahres“ zu beurteilen. Für „Aktueller Wert“ geben Sie den am stärksten beeinträchtigten Grad der Funktionsfähigkeit in der letzten Woche an. Bewerten Sie die aktuelle Funktionsfähigkeit unabhängig von der Ätiologie der Probleme. Befindet sich der Patient jedoch in den Schulferien (z. B. Sommerferien) oder war er <2 Wochen im Krankenhaus und wird nun entlassen, ohne dass er zwischen Entlassung und Beurteilung des GF die Möglichkeit hatte, seine gewohnte berufliche/schulische Rolle wahrzunehmen (z. B. Beurteilung des GF am Tag der Entlassung aus dem Krankenhaus), sollten Sie die Situation vor dem Ereignis beurteilen (z. B. Wenn sich die Person zum Zeitpunkt der Beurteilung im Krankenhaus befindet und dies <2 Wochen her ist, sollten Sie den GF R unmittelbar vor dem Krankenhausaufenthalt ermitteln). Wenn die Person aus nicht psychischen Gründen (z. B. komplizierte Grippe) länger als zwei Wochen im Krankenhaus war, erfolgt die Bewertung auch nach dem Grad der sozialen Funktionsfähigkeit unmittelbar vor dem Krankenhausaufenthalt (ebenso bei einem längeren Urlaub). Befindet sich eine Person in einer Übergangsphase (z.B. zwischen Schule und Studium oder zwei Arbeitsstellen), muss auch das Funktionsniveau unmittelbar vor dieser Situation berücksichtigt werden.
Bestimmen und bewerten Sie den GF R für die primäre Rollenverteilung (Arbeit, Schule oder Zuhause) anhand der Fragen unten. Wenn die Testperson in verschiedenen Aufgabenbereichen involviert ist, ist die Gesamtdauer, die mit den rollenspezifischen Aktivitäten verbracht wird (d. h. Teilzeitschule plus Teilzeitarbeit entspricht Vollzeitrollenstatus) entscheidend.

Hinweis: Diese Skala hebt den Grad der Unterstützung hervor, der in der Umgebung der Person bereitgestellt wird, und die Leistung der Person, die diese Unterstützung erhält. Der Begriff „unabhängig“, wie er in diesem Instrument verwendet wird, impliziert, dass eine Person auf einem „altersgemäßen Niveau“ ohne die Hilfe von externen Unterstützungen oder Anpassungen funktioniert. Beispiele für ein unabhängiges Funktionieren sind (1) ein altersgemäßes Funktionieren in einer Regelschule, ohne dass zusätzliche Hilfe, spezielle Klassen oder besondere Vorkehrungen für Tests erforderlich sind, (2) wettbewerbsfähige Vollzeitbeschäftigung ohne zusätzliche Anleitung, Unterstützung, Job-Coaching oder andere Formen besonderer Hilfe; und (3) Vollzeit-Hausfrau, die für die Erledigung von Aufgaben und Aktivitäten im Haushalt für eine Familie ohne zusätzliche Anleitung, Unterstützung oder Aufsicht verantwortlich ist.

1. Wie verbringen Sie Ihre Zeit während des Tages?
2. Wenn Sie derzeit arbeiten:
   (a) Wo arbeiten Sie? Was sind Ihre beruflichen Aufgaben?
   (b) Wie viele Stunden pro Woche arbeiten Sie?
   (c) Wie lange arbeiten Sie schon in Ihrem derzeitigen Beruf? Gab es in letzter Zeit irgendwelche Veränderungen in Ihrem Job (z. B. Verlust des Arbeitsplatzes, Beendigung des Arbeitsverhältnisses, Änderung der Position oder der Arbeitsbelastung)?
   (d) Benötigen Sie bei der Arbeit normalerweise Unterstützung oder regelmäßige Aufsicht? Wie oft brauchen Sie zusätzliche Hilfe? Gibt es Aufgaben, die Sie nicht allein bewältigen können?
   (e) Haben Sie manchmal Schwierigkeiten, auf dem Laufenden zu bleiben? Sind Sie in der Lage, einen Rückstand aufzuholen, wenn Sie in Verzug geraten?
   (f) Haben Sie (positive oder negative) Kommentare oder anderweitige Rückmeldungen zu Ihrer Leistung erhalten? Haben andere Sie auf Dinge hingewiesen, die Sie gut oder schlecht gemacht haben?
3. Wenn Sie derzeit eine Schule besuchen:
   (a) Welche Art von Schule besuchen Sie? (allgemeinbildende Schule, nicht-öffentliche Schule, Wohnheim/Krankenhaus)
   (b) Waren Sie schon einmal an einer Sonderschule oder einer anderen nicht allgemeinbildenden Schule eingeschrieben?
   (c) Wie lange sind Sie schon an dieser Schule? Hatten Sie in letzter Zeit irgendwelche Änderungen in Ihrer Schul-Einstufung?

(d) Benötigen Sie zusätzliche Hilfe oder Unterstützung, um den aktuellen Schulstoff zu bewältigen? Bekommen Sie Nachhilfe oder zusätzliche Hilfe während des Unterrichts oder nach der Schule? Bekommen Sie zusätzliche Zeit, um Tests zu schreiben oder können Sie das Klassenzimmer verlassen, um an einem ruhigen Ort Tests zu schreiben?

(e) Haben Sie Schwierigkeiten, mit Ihren Kursarbeiten Schritt zu halten? Können Sie den Rückstand aufholen, wenn ein solcher anfällt?

(f) Wie sind Ihre Noten (beste und schlechteste)? Ist ihre Versetzung gefährdet?

4. Wenn Sie Hausfrau oder Hausmann sind:

(a) Welche Aufgaben erledigen Sie im Haushalt oder für die Familie?

(b) Wie lange sind Sie schon für den Haushalt zuständig?

(c) Wie viele Stunden pro Woche verbringen Sie mit Aufgaben im Haushalt?

(d) Sind Sie in der Lage, mit den Anforderungen Ihres Haushalts Schritt zu halten? Geraten Sie jemals in Rückstand? Wenn ja, sind Sie in der Lage, den Rückstand aufzuholen, oder sind Sie auf die Hilfe anderer angewiesen? Vermeiden Sie irgendwelche Aufgaben? Benötigen Sie für bestimmte Aufgaben im Haushalt regelmäßig Unterstützung oder Aufsicht?

10	Herausragende Rollenfunktion	Erfüllt selbstständig Aufgaben mit überdurchschnittlichem Ergebnis in anspruchsvollen Funktionen (Vollzeit oder gleichwertig). Erhält nur überdurchschnittliche Leistungsbewertungen bei kompetetiven Arbeitsstellen. Erhält fast ausschließlich Einsen in der Regelschule. Plant, organisiert und erledigt alle Aufgaben im Haushalt mit Leichtigkeit.
9	Überduchschnittliche Rollenfunktion	Arbeitet selbstständig und sehr gut in anspruchsvollen Funktionen (erfordert Vollzeitbeschäftigung oder eine gleichwertige Tätigkeit). Selten abwesend oder nicht arbeitsfähig. Erzielt gute bis überdurchschnittliche Leistungen bei kompetetiven Arbeitsstellen. Erzielt in allen Kursen der Regelschule entweder Einsen oder Zweien. Plant, organisiert und erledigt alle hauswirtschaftlichen Aufgaben.

8	Gute Rollenfunktion	Erfüllt selbstständig Aufgaben in anspruchsvollen Funktionen (erfordert Vollzeitbeschäftigung oder eine gleichwertige Tätigkeit). Gelegentlicher Rückstand bei Aufgaben, holt diesen aber immer wieder auf. Erzielt eine zufriedenstellende Leistungsbeurteilung bei einem kompetitiven Praktikumsplatz. Erzielt in der Regelschule eine Drei und bessere Leistungen. Gelegentliche Schwierigkeiten bei der Erledigung oder Organisation von Hausarbeiten. Oder Erzielt überdurchschnittliche Leistungen mit minimaler Unterstützung (z. B. Nachhilfeunterricht; reduzierte akademische Kurse an der 4-Jahres-Universität; Besuch der Volkshochschule; möglicherweise zusätzliche Unterstützung bei der Arbeit weniger als 1-2 Mal pro Woche). Erhält gute Arbeits-/Schulbeurteilungen, erledigt alle Aufgaben mit diesem Maß an Unterstützung.
7	Milde Beeinträchtigung der Rollenfunktionsfähigkeit	Geringfügige Beeinträchtigung der Funktionsfähigkeit bei anspruchsvollen Aufgaben. Häufig mit Aufgaben im Rückstand oder unfähig, diese zu erfüllen. Erhält häufig schlechte Leistungsbeurteilungen in kompetitiven Arbeitsplätzen oder Vieren oder besser in der Regelschule. Häufige Schwierigkeiten bei der Erledigung oder Organisation von Aufgaben im Haushalt. Erzielt bessere Leistungen mit minimaler Unterstützung (z. B., erhält zusätzliche Anleitung/Unterstützung bei der Arbeit 1-2x pro Woche), erhält dann eine Drei oder besser, zufriedenstellende Arbeits-/Schulbeurteilungen und erledigt die meisten Aufgaben im Haushalt mit diesem Maß an Unterstützung.
6	Moderate Beeinträchtigung der Rollenfunktionsfähigkeit	Beeinträchtigung der Funktionsfähigkeit bei selbstständig ausgeübten Tätigkeiten. Kann gelegentlich eine sechs in normalen Kursen erhalten, anhaltend schlechte Leistungsbewertungen bei kompetetiven Arbeitsplätzen, Arbeitsplatzwechsel aufgrund schlechter Leistungen, anhaltende Schwierigkeiten bei der Erledigung oder Organisation von Hausarbeiten. Oder Benötigt teilweise Unterstützung (einige Hilfsmittel- oder Sonderschulkurse; erhält 2+ Mal/Woche Anleitung/Unterstützung bei der Arbeit). Benötigt möglicherweise weniger anspruchsvolle Vollzeit- oder Teilzeitjobs und/oder eine gewisse Beaufsichtigung in der häuslichen Umgebung, funktioniert aber gut oder angemessen, wenn die Unterstützung angeboten wird (mit Unterstützung zufriedenstellende Beurteilungen am Arbeitsplatz oder gute Noten in der Schule).

5	Deutliche Beeinträchtigung der Rollenfunktionsfähigkeit	Schwere Beeinträchtigung der Funktionsfähigkeit bei selbstständig ausgeübten Tätigkeiten. Durchfallen in mehreren Kursen der Regelschule, möglicherweise Verlust des Arbeitsplatzes (z.B. Bewährung am Arbeitsplatz) oder nicht in der Lage, die meisten Aufgaben im Haushalt selbstständig zu erledigen. Oder in vollständig sonderpädagogischen Klassen, benötigt weniger anspruchsvolle berufliche/alltägliche Unterstützung oder Anleitung, benötigt möglicherweise berufliche Rehabilitation und/oder etwas Betreuung im häuslichen Umfeld, hält aber überdurchschnittliche Leistungen mit Unterstützung aufrecht - erhält Einsen und Zweien, gute Beurteilungen in Arbeit/Schule, erledigt alle Aufgaben.
4	Schwere Beeinträchtigung der Rollenfunktionsfähigkeit	Sehr schwere Beeinträchtigung der Funktionsfähigkeit bei selbstständig ausgeübten Tätigkeiten. Nur Sechsen in der Regelschule oder Schulabbruch. Kann eigenständig keine Arbeit finden oder halten. (z.B. kürzlich entlassen, mehrere kurzfristige Jobs in der jüngeren Vergangenheit, aber aktiv auf Arbeitssuche oder neue Arbeit in naher Zukunft erwartet), oder nicht in der Lage, praktisch alle Aufgaben im Haushalt selbständig zu erledigen. Oder ausreichende bis gute Funktionsfähigkeit mit großer Unterstützung. Benötigt eine supportive Arbeitsumgebung, vollständig sonderpädagogischen Unterricht, eine nicht-öffentliche oder psychiatrische Schule, Heimunterricht zum Zweck einer unterstützenden schulischen Umgebung und/oder eine unterstützte häusliche Umgebung, ABER funktioniert mit dieser Unterstützung angemessen (erhält zufriedenstellende Leistungsbewertungen am Arbeitsplatz oder gute Noten).
3	Marginale Rollenfunktionsfähigkeit	Beeinträchtigte Funktion trotz großer Unterstützung. Benötigt ein supportives Arbeitsumfeld, vollständig sonderpädagogischen Unterricht, eine nicht-öffentliche oder psychiatrische Schule, Heimunterricht mit dem Ziel eines supportiven schulischen Umfelds und/oder supportiven häuslichen Umfelds, ABER trotz dieser Unterstützung schlechte Funktionsfähigkeit (ist ständig im Rückstand bei Aufgaben, ist häufig nicht in der Lage, Leistungen zu erbringen, erhält schlechte Leistungsbewertungen am Arbeitsplatz oder fällt in der Schule durch).
2	Unfähigkeit eigenständig eine gesellschaftliche Rolle einzunehmen	Behindert, nimmt aber an strukturierten Aktivitäten teil. Invalidität oder ein gleichwertiger, nicht selbstständiger Status. Arbeitet nicht gegen Bezahlung, besucht keinen Unterricht oder lebt nicht autonom. Verbringt 5 oder mehr Stunden pro Woche mit strukturierten, rollenbezogenen Aktivitäten (z. B. aktive Arbeitssuche/Bewerbung, stationäre Behandlung, Freiwilligenarbeit, Nachhilfe, geschützte Arbeitsprogramme).
1	Extreme Dysfunktion	Schwerbehindert. Invalidität oder gleichwertiger unselbstständiger Status. Arbeitet nicht gegen Bezahlung (z. B. längere Arbeitslosigkeit oder nie erwerbstätig UND nicht auf der Suche nach Arbeit), besucht keinen Unterricht oder lebt nicht autonom. Verbringt weniger als 5 Stunden pro Woche mit strukturierten, rollenbezogenen Aktivitäten.

Anhang B

Weitere Ergebnisse

B.1 GF S Ergebnisse mit einem Grenzwert von 6,5

Nach Abbildung 4.6 kann abgeleitet werden, dass der Grenzwert <6,5 (statt <7) ein geeigneterer Grenzwert für die GF S Regressionsanalyse sein könnte. Probatorisch wurde daher auch mit dem Grenzwert von 6,5 eine Analyse durchgeführt. Hierbei zeigt sich eine Sensitivität von 82,84% (95% CI: 76,96% - 87,75%), eine Spezifität von 91,60% (95% CI: 88,47% - 94,12%), eine PLR von 9,87 (95% CI: 7,11-13,69) und eine NLR von 0,19 (95% CI: 0,14 - 0,25). Insgesamt ergibt sich somit eine BAC von 87,22% und eine Genauigkeit von 88,67%. Der PPW beträgt 83,25% (95% CI: 78,17% - 87,34%) und der NPW 91,38% (95% CI: 88,67% - 93,49%).

B.2 GF R Ergebnisse mit einem Grenzwert von 6,5

Wie auch bei Punkt A.1 wurde bei der GF R Regression nachträglich mit unterschiedlichen Grenzwerten experimentiert. Es wurde ebenfalls ein Grenzwert von 6,5 genutzt, um hohe und niedrige GF Werte voneinander zu unterscheiden. So konnte man nachträglich die BAC und andere Gütekriterien, welche typischerweise für die Klassifikation herangezogen werden, berechnen. Hierbei ergibt sich eine BAC von 86,41% und eine Genauigkeit von 86,21%. Die Sensitivität beträgt 87.22% (95% CI: 82.17% - 91.27%) und die Spezifität 85,60% (95% CI: 81,68% - 88.97%). Der NPW ist mit einem Wert von 91,85% (95% CI: 88,90% - 94,08%) sehr hoch. Der PPW beträgt 78,26% (95% CI: 73,72% - 82,21%).

B.3 Zusammenfassende Tabellen der Ergebnisse

B.3.1 Klassifikationsergebnisse

	alle Gruppen	ROD	CHR	ROP
GF S	85,04%	88,01%	82,48%	82,40%
GAF D/I	83,53%	84,51%	84,68%	76,50%
GF R	82,75%	85,26%	81,58%	78,91%
GAF S	78,28%	84,10%	74,75%	72,15%

Tabelle B.1: Die Tabelle zeigt die balancierte Korrektheit (balanced accuracy) der verschiedenen Klassifikationsergebnisse im Vergleich.

B.3.2 Regressionsergebnisse

	alle Gruppen	ROD	CHR	ROP	AUC	$S\rho$
GF S	0,72	0,69	0,73	0,72	0,90	0,81
GF R	0,76	0,68	0,74	0,76	0,91	0,81
GAF D/I	0,89	0,89	0,89	0,89	0,90	0,81
GAF S	0,85	0,88	0,87	0,85	0,88	0,75

Tabelle B.2: Die Tabelle zeigt die RMSE Werte der verschiedenen Regressionsergebnisse, sowie Spearmans ρ und die AUC im Vergleich.

Literaturverzeichnis

[1] J. McGrath, S. Saha, D. Chant, and J. Welham, "A concise overview of incidence prevalence and mortality," *Epidemiol Rev*, vol. 30, no. 1, pp. 67–76, 2008.

[2] L. Löhrs and A. Hasan, "Risikofaktoren für die Entstehung und den Verlauf der Schizophrenie," *Fortschritte der Neurologie· Psychiatrie*, vol. 87, no. 02, pp. 133–143, 2019.

[3] P. C. Fletcher and C. D. Frith, "Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia," *Nature Reviews Neuroscience*, vol. 10, no. 1, pp. 48–58, 2009.

[4] H.-J. Möller, G. Laux, and A. Deister, *Psychiatrie, Psychosomatik und Psychotherapie*. Thieme Stuttgart, 2020.

[5] T. J. Crow, "Molecular pathology of schizophrenia: more than one disease process?" *British medical journal*, vol. 280, no. 6207, p. 66, 1980.

[6] N. C. Andreasen and S. Olsen, "Negative v positive schizophrenia: Definition and validation," *Archives of general psychiatry*, vol. 39, no. 7, pp. 789–794, 1982.

[7] Arbeitsgemeindschaft für Methodik und Dokumentation in der Psychiatrie, *Das AMDP-System*. hogrefe, 2018.

[8] K. Jaspers, *Allgemeine Psychopathologie*. Springer-Verlag, 1965.

[9] S. Arndt, R. J. Alliger, and N. C. Andreasen, "The distinction of positive and negative symptoms," *The British Journal of Psychiatry*, vol. 158, no. 3, pp. 317–322, 1991.

[10] A. Wolkin, M. Sanfilipo, A. P. Wolf, B. Angrist, J. D. Brodie, and J. Rotrosen, "Negative symptoms and hypofrontality in chronic schizophrenia," *Archives of general psychiatry*, vol. 49, no. 12, pp. 959–965, 1992.

[11] N. C. Andreasen, "Affective flattening and the criteria for schizophrenia." *The American Journal of Psychiatry*, 1979.

[12] M. Kiang, B. K. Christensen, G. Remington, and S. Kapur, "Apathy in schizophrenia: clinical correlates and association with functional outcome," *Schizophrenia research*, vol. 63, no. 1-2, pp. 79–88, 2003.

[13] P. Harvey, J. Pruessner, Y. Czechowska, and M. Lepage, "Individual differences in trait anhedonia: a structural and functional magnetic resonance imaging study in non-clinical subjects," *Molecular psychiatry*, vol. 12, no. 8, pp. 767–775, 2007.

[14] E. M. Joyce, S. Collinson, and P. Crichton, "Verbal fluency in schizophrenia: relationship with executive function, semantic memory and clinical alogia," *Psychological medicine*, vol. 26, no. 1, pp. 39–49, 1996.

[15] C. A. Wilson and J. I. Koenig, "Social interaction and social withdrawal in rodents as readouts for investigating the negative symptoms of schizophrenia," *European Neuropsychopharmacology*, vol. 24, no. 5, pp. 759–773, 2014.

[16] B. J. Freedman, "The subjective experience of perceptual and cognitive disturbances in schizophrenia: A review of autobiographical accounts," *Archives of general psychiatry*, vol. 30, no. 3, pp. 333–340, 1974.

[17] J. E. Cooper, *Taschenführer zur ICD-10-Klassifikation psychischer Störungen*, 2012.

[18] A. S. Cohen, J. E. McGovern, T. J. Dinzeo, and M. A. Covington, "Speech deficits in serious mental illness: A cognitive resource issue?" *Schizophrenia Research*, vol. 160, no. 1, pp. 173–179, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S092099641400601X

[19] F. G. Guggenheim and H. M. Babigian, "Catatonic schizophrenia: Epidemiology and clinical course: A 7-year register study of 798 cases." *Journal of Nervous and Mental Disease*, 1974.

[20] M. T. Tsuang and G. Winokur, "Criteria for subtyping schizophrenia: Clinical differentiation of hebephrenic and paranoid schizophrenia," *Archives of General Psychiatry*, vol. 31, no. 1, pp. 43–47, 1974.

[21] H. S. Sullivan, "The onset of schizophrenia," *American Journal of Psychiatry*, vol. 84, no. 1, pp. 105–134, 1927.

[22] W. Mayer-Gross *et al.*, *Handbuch der Geisteskrankheiten*, 1932, vol. 9.

[23] S. J. Keith and S. M. Matthews, "The diagnosis of schizophrenia: a review of onset and duration issues," *Schizophrenia bulletin*, vol. 17, no. 1, pp. 51–68, 1991.

[24] J. Klosterkötter, M. Hellmich, E. M. Steinmeyer, and F. Schultze-Lutter, "Diagnosing schizophrenia in the initial prodromal phase," *Archives of general psychiatry*, vol. 58, no. 2, pp. 158–164, 2001.

[25] R. A. Berk, "An introduction to sample selection bias in sociological data," *American sociological review*, pp. 386–398, 1983.

[26] P. Fusar-Poli, S. Borgwardt, A. Bechdolf, J. Addington, A. Riecher-Rössler, F. Schultze-Lutter, M. Keshavan, S. Wood, S. Ruhrmann, L. J. Seidman *et al.*, "The psychosis high-risk state: a comprehensive state-of-the-art review," *JAMA psychiatry*, vol. 70, no. 1, pp. 107–120, 2013.

[27] G. Huber and G. Gross, "The concept of basic symptoms in schizophrenic and schizoaffective psychoses." *Recenti progressi in medicina*, vol. 80, no. 12, pp. 646–652, 1989.

[28] F. Schultze-Lutter, "Subjective symptoms of schizophrenia in research and the clinic: the basic symptom concept," *Schizophrenia bulletin*, vol. 35, no. 1, pp. 5–8, 2009.

[29] F. Schultze-Lutter, M. Debbané, A. Theodoridou, S. J. Wood, A. Raballo, C. Michel, S. J. Schmidt, J. Kindler, S. Ruhrmann, and P. J. Uhlhaas, "Revisiting the basic symptom concept: toward translating risk symptoms for psychosis into neurobiological targets," *Frontiers in Psychiatry*, vol. 7, p. 9, 2016.

[30] P. Fusar-Poli and F. Schultze-Lutter, "Predicting the onset of psychosis in patients at clinical high risk: practical guide to probabilistic prognostic reasoning," *Evidence-based mental health*, vol. 19, no. 1, pp. 10–15, 2016.

[31] T. J. Miller, T. H. McGlashan, J. L. Rosen, K. Cadenhead, J. Ventura, W. McFarlane, D. O. Perkins, G. D. Pearlson, and S. W. Woods, "Prodromal assessment with the structured interview for prodromal syndromes and the scale of prodromal symptoms: Predictive validity, interrater reliability, and training to reliability," *Schizophrenia Bulletin*, vol. 29, no. 4, pp. 703–715, 2003. [Online]. Available: http://dx.doi.org/10.1093/oxfordjournals.schbul.a007040

[32] J. B. Williams and M. First, "Diagnostic and statistical manual of mental disorders," in *Encyclopedia of social work*, 2013.

[33] A. R. Yung, P. D. McGorry, C. A. McFarlane, H. J. Jackson, G. C. Patton, and A. Rakkar, "Monitoring and care of young people at incipient risk of psychosis," *Schizophrenia bulletin*, vol. 22, no. 2, pp. 283–303, 1996.

[34] P. Fusar-Poli, M. Cappucciati, A. De Micheli, G. Rutigliano, I. Bonoldi, S. Tognin, V. Ramella-Cravaro, A. Castagnini, and P. McGuire, "Diagnostic and prognostic significance of brief limited intermittent psychotic symptoms (blips) in individuals at ultra high risk," *Schizophrenia bulletin*, vol. 43, no. 1, pp. 48–56, 2017.

[35] T. D. Cannon, T. G. Van Erp, and D. C. Glahn, "Elucidating continuities and discontinuities between schizotypy and schizophrenia in the nervous system," *Schizophrenia research*, vol. 54, no. 1-2, pp. 151–156, 2002.

[36] T. D. Cannon, K. Cadenhead, B. Cornblatt, S. W. Woods, J. Addington, E. Walker, L. J. Seidman, D. Perkins, M. Tsuang, T. McGlashan *et al.*, "Prediction of psychosis in youth at high clinical risk: a multisite longitudinal study in north america," *Archives of general psychiatry*, vol. 65, no. 1, pp. 28–37, 2008.

[37] J. van Os and S. Guloksuz, "A critique of the "ultra-high risk" and "transition" paradigm," *World Psychiatry*, vol. 16, no. 2, pp. 200–206, 2017.

[38] H. Häfner, K. Maurer, and W. An Der Heiden, "ABC Schizophrenia study: an overview of results since 1996," *Social psychiatry and psychiatric epidemiology*, vol. 48, no. 7, pp. 1021–1031, 2013.

[39] H. Häfner and W. an der Heiden, "The course of schizophrenia in the light of modern follow-up studies: the ABC and WHO studies," *European archives of psychiatry and clinical neuroscience*, vol. 249, no. 4, pp. S14–S26, 1999.

[40] S.-H. Hosseini and M. K. Yousefi, "Quality of life and GAF in schizophrenia correlation between quality of life and global functioning in schizophrenia," *Iranian journal of psychiatry and behavioral sciences*, vol. 5, no. 2, p. 120, 2011.

[41] G. S. Malhi and J. J. Mann, "Depression." *Lancet*, vol. 392, no. 10161, pp. 2299–2312, 2018.

[42] M. Zimmerman, M. A. Posternak, and I. Chelminski, "Defining remission on the Montgomery-Asberg depression rating scale," *The Journal of clinical psychiatry*, vol. 65, no. 2, pp. 163–168, 2004.

[43] S. Tashiro, S. Hosoda, and R. Kawahara, "Naikan therapy for prolonged depression: Psychological changes and long-term efficacy of intensive Naikan therapy," *Seishin shinkeigaku zasshi*, vol. 106, no. 4, pp. 431–457, 2004.

[44] M. L. Chatterton, E. Stockings, M. Berk, J. J. Barendregt, R. Carter, and C. Mihalopoulos, "Psychosocial therapies for the adjunctive treatment of bipolar disorder in adults: network meta-analysis," *The British journal of psychiatry*, vol. 210, no. 5, pp. 333–341, 2017.

[45] L. Luborsky, L. Diguer, J. Cacciola, J. P. Barber, K. Moras, K. Schmidt, and R. J. DeRubeis, "Factors in outcomes of short-term dynamic psychotherapy for chronic vs. nonchronic major depression," *The Journal of Psychotherapy Practice and Research*, vol. 5, no. 2, p. 152, 1996.

[46] R. H. Moos, L. McCoy, and B. S. Moos, "Global assessment of functioning (GAF) ratings: Determinants and role as predictors of one-year treatment outcomes," *Journal of clinical psychology*, vol. 56, no. 4, pp. 449–461, 2000.

[47] P. Fusar-Poli, M. Cappucciati, S. Borgwardt, S. W. Woods, J. Addington, B. Nelson, D. H. Nieman, D. R. Stahl, G. Rutigliano, A. Riecher-Rössler *et al.*, "Heterogeneity of psychosis risk within individuals at clinical high risk: a meta-analytical stratification," *JAMA psychiatry*, vol. 73, no. 2, pp. 113–120, 2016.

[48] N. A. Christakis, *Death foretold: prophecy and prognosis in medical care*. University of Chicago Press, 2001.

[49] J. D. Childs and J. A. Cleland, "Development and application of clinical prediction rules to improve decision making in physical therapist practice," *Physical Therapy*, vol. 86, no. 1, pp. 122–131, 2006.

[50] L. Flyckt, M. Mattsson, G. Edman, R. Carlsson, and J. Cullberg, "Predicting 5-year outcome in first-episode psychosis: construction of a prognostic rating scale," *Journal of Clinical Psychiatry*, vol. 67, no. 6, pp. 916–924, 2006.

[51] M. M. Hunink, M. C. Weinstein, E. Wittenberg, M. F. Drummond, J. S. Pliskin, J. B. Wong, and P. P. Glasziou, *Decision making in health and medicine: integrating evidence and values*. Cambridge University Press, 2014.

[52] A. R. Yung and P. D. McGorry, "The prodromal phase of first-episode psychosis: past and current conceptualizations," *Schizophrenia bulletin*, vol. 22, no. 2, pp. 353–370, 1996.

[53] W. H. Organization *et al.*, *The global burden of disease: 2004 update*. World Health Organization, 2008.

[54] A. Gustavsson, M. Svensson, F. Jacobi, C. Allgulander, J. Alonso, E. Beghi, R. Dodel, M. Ekman, C. Faravelli, L. Fratiglioni *et al.*, "Cost of disorders of the brain in europe 2010," *European neuropsychopharmacology*, vol. 21, no. 10, pp. 718–779, 2011.

[55] J. Leal, R. Luengo-Fernández, A. Gray, S. Petersen, and M. Rayner, "Economic burden of cardiovascular diseases in the enlarged european union," *European heart journal*, vol. 27, no. 13, pp. 1610–1619, 2006.

[56] E. C. Harris and B. Barraclough, "Suicide as an outcome for mental disorders. a meta-analysis," *British journal of psychiatry*, vol. 170, no. 3, pp. 205–228, 1997.

[57] N. Koutsouleris, "Application for ethical approval of the medical faculty of the Ludwig-Maximilian-University," 2015.

[58] I. Kooyman, K. Dean, S. Harvey, and E. Walsh, "Outcomes of public concern in schizophrenia," *The British Journal of Psychiatry*, vol. 191, no. S50, pp. s29–s36, 2007.

[59] A. Barbato, "Psychiatry in transition: outcomes of mental health policy shift in Italy," *Australian and New Zealand Journal of Psychiatry*, vol. 32, no. 5, pp. 673–679, 1998.

[60] T. Wykes, C. Reeder, S. Landau, B. Everitt, M. Knapp, A. Patel, and R. Romeo, "Cognitive remediation therapy in schizophrenia: randomised controlled trial," *The British journal of psychiatry*, vol. 190, no. 5, pp. 421–427, 2007.

[61] C. R. Bowie, M. Grossman, M. Gupta, L. K. Oyewumi, and P. D. Harvey, "Cognitive remediation in schizophrenia: efficacy and effectiveness in patients with early versus long-term course of illness," *Early Intervention in Psychiatry*, vol. 8, no. 1, pp. 32–38, 2014.

[62] P. D. McGorry, A. R. Yung, A. Bechdolf, and P. Amminger, "Back to the future: predicting and reshaping the course of psychotic disorder," *Archives of General Psychiatry*, vol. 65, no. 1, pp. 25–27, 2008.

[63] B. A. Cornblatt, R. E. Carrión, A. Auther, D. McLaughlin, R. H. Olsen, M. John, and C. U. Correll, "Psychosis prevention: a modified clinical high risk perspective from the recognition and prevention (rap) program," *American Journal of Psychiatry*, vol. 172, no. 10, pp. 986–994, 2015.

[64] R. E. Carrión, D. J. Walder, A. M. Auther, D. McLaughlin, H. O. Zyla, S. Adelsheim, R. Calkins, C. S. Carter, B. McFarland, R. Melton *et al.*, "From the psychosis prodrome to the first-episode of psychosis: No evidence of a cognitive decline," *Journal of psychiatric research*, vol. 96, pp. 231–238, 2018.

[65] N. Koutsouleris, R. S. Kahn, A. M. Chekroud, S. Leucht, P. Falkai, T. Wobrock, E. M. Derks, W. W. Fleischhacker, and A. Hasan, "Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: a machine learning approach," *The Lancet Psychiatry*, vol. 3, no. 10, pp. 935–946, 2016.

[66] P. Fusar-Poli, Z. Hijazi, D. Stahl, and E. W. Steyerberg, "The science of prognosis in psychiatry: a review," *JAMA psychiatry*, vol. 75, no. 12, pp. 1289–1297, 2018.

[67] B.-A. Morel, *Traité des maladies mentales*. V. Masson, 1860.

[68] M. Bleuler and R. Bleuler, *Dementia praecox oder die Gruppe der Schizophrenien: Eugen Bleuler*. Cambridge University Press, 1986, vol. 149, no. 5.

[69] J. Hoenig, "The Concept of Schizophrenia Kraepelin–Bleuler–Schneider," *The British Journal of Psychiatry*, vol. 142, no. 6, pp. 547–556, 1983.

[70] S. H. Jones, G. Thornicroft, M. Coffey, and G. Dunn, "A brief mental health outcome scale: Reliability and validity of the Global Assessment of Functioning (GAF)," *The British Journal of Psychiatry*, vol. 166, no. 5, pp. 654–659, 1995.

[71] A. R. Yung, L. J. Phillips, P. D. Mcgorry, C. A. McFarlane, S. Francey, S. Harrigan, G. C. Patton, and H. J. Jackson, "Prediction of psychosis: a step towards indicated prevention of schizophrenia," *The British Journal of Psychiatry*, vol. 172, no. S33, pp. 14–20, 1998.

[72] T. Vatnaland, J. Vatnaland, S. Friis, and S. Opjordsmoen, "Are GAF scores reliable in routine clinical use?" *Acta Psychiatrica Scandinavica*, vol. 115, no. 4, pp. 326–330, 2007.

[73] B. A. Cornblatt, A. M. Auther, T. Niendam, C. W. Smith, J. Zinberg, C. E. Bearden, and T. D. Cannon, "Preliminary findings for two new measures of social and role functioning in the prodromal phase of schizophrenia," *Schizophrenia bulletin*, vol. 33, no. 3, pp. 688–702, 2007.

[74] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85 – 117, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0893608014002135

[75] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, 1943.

[76] D. O. Hebb, "The organization of behavior. a neuropsychological theory," 1949.

[77] H. Sompolinsky, "The theory of neural networks: The hebb rule and beyond," in *Heidelberg colloquium on glassy dynamics*. Springer, 1987, pp. 485–527.

[78] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain." *Psychological review*, vol. 65, no. 6, p. 386, 1958.

[79] A. Steger, *Diskrete Strukturen: Band 1: Kombinatorik, Graphentheorie, Algebra*. Springer-Verlag, 2007.

[80] M. Minsky and S. Papert, "Perceptron expanded edition," 1969.

[81] M. T. Hagan, H. B. Demuth, M. H. Beale, and O. De Jesús, *Neural network design*. Pws Pub. Boston, 1996, vol. 20.

[82] D. E. Rumelhart and J. L. McClelland, "Parallel distributed processing: explorations in the microstructure of cognition. volume 1. foundations," 1986.

[83] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural networks*, vol. 2, no. 5, pp. 359–366, 1989.

[84] M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken, "Multilayer feedforward networks with a nonpolynomial activation function can approximate any function," *Neural networks*, vol. 6, no. 6, pp. 861–867, 1993.

[85] R. J. Williams and D. Zipser, "A learning algorithm for continually running fully recurrent neural networks," *Neural computation*, vol. 1, no. 2, pp. 270–280, 1989.

[86] D. E. Rumelhart, G. E. Hinton, R. J. Williams *et al.*, "Learning representations by back-propagating errors," *Cognitive modeling*, vol. 5, no. 3, p. 1, 1988.

[87] P. J. Werbos *et al.*, "Backpropagation through time: what it does and how to do it," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.

[88] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, no. 02, pp. 107–116, 1998.

[89] A. Shewalkar, D. Nyavanandi, and S. A. Ludwig, "Performance evaluation of deep neural networks applied to speech recognition: Rnn, lstm and gru," *Journal of Artificial Intelligence and Soft Computing Research*, vol. 9, no. 4, pp. 235–245, 2019.

[90] S. Hochreiter and J. Schmidhuber, "LSTM can solve hard long time lag problems," in *Advances in neural information processing systems*, 1997, pp. 473–479.

[91] Hochreiter, Sepp and Schmidhuber, Jürgen, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[92] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," 1999.

[93] F. A. Gers and J. Schmidhuber, "Recurrent nets that time and count," in *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, vol. 3. IEEE, 2000, pp. 189–194.

[94] I. Goodfellow, Y. Bengio, and A. Courville, "Deep learning," *http://www.deeplearningbook.org*, 2016.

[95] S. Gheisari, S. Shariflou, J. Phu, P. J. Kennedy, A. Agar, M. Kalloniatis, and S. M. Golzan, "A combined convolutional and recurrent neural network for enhanced glaucoma detection," *Scientific reports*, vol. 11, no. 1, pp. 1–11, 2021.

[96] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural networks for machine learning*, vol. 4, no. 2, pp. 26–31, 2012.

[97] D. Sussillo and L. Abbott, "Random walk initialization for training very deep feedforward networks," *arXiv preprint arXiv:1412.6558*, 2014.

[98] R. Pascanu, C. Gulcehre, K. Cho, and Y. Bengio, "How to construct deep recurrent neural networks," *arXiv preprint arXiv:1312.6026*, 2013.

[99] K. Doya, "Universality of fully connected recurrent neural networks," *Dept. of Biology, UCSD, Tech. Rep*, 1993.

[100] Y. Bengio, N. Boulanger-Lewandowski, and R. Pascanu, "Advances in optimizing recurrent networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 8624–8628.

[101] J. Li, Y. Zhong, J. Han, G. Ouyang, X. Li, and H. Liu, "Classifying ASD children with LSTM based on raw videos," *Neurocomputing*, vol. 390, pp. 226–238, 2020.

[102] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 6645–6649.

[103] T. Lang and M. Rettenmeier, "Understanding consumer behavior with recurrent neural networks," in *Workshop on Machine Learning Methods for Recommender Systems*, 2017.

[104] M. R. Hoehe and D. J. Morris-Rosendahl, "The role of genetics and genomics in clinical psychiatry," *Dialogues in clinical neuroscience*, vol. 20, no. 3, p. 169, 2018.

[105] S. A. Hunter and S. M. Lawrie, "Imaging and genetic biomarkers predicting transition to psychosis," in *Biomarkers in Psychiatry*. Springer, 2018, pp. 353–388.

[106] D. B. Dwyer, P. Falkai, and N. Koutsouleris, "Machine learning approaches for clinical psychology and psychiatry," *Annual review of clinical psychology*, vol. 14, pp. 91–118, 2018.

[107] N. Koutsouleris, L. Kambeitz-Ilankovic, S. Ruhrmann, M. Rosen, A. Ruef, D. B. Dwyer, M. Paolini, K. Chisholm, J. Kambeitz, T. Haidl, A. Schmidt, J. Gillam, F. Schultze-Lutter, P. Falkai, M. Reiser, A. Riecher-Rössler, R. Upthegrove, J. Hietala, R. K. R. Salokangas, C. Pantelis, E. Meisenzahl, S. J. Wood, D. Beque, P. Brambilla, S. Borgwardt, and for the PRONIA Consortium, "Prediction models of functional outcomes for individuals in the clinical high-risk state for psychosis or with recent-onset depression: A multimodal, multisite machine learning analysis," *JAMA Psychiatry*, vol. 75, no. 11, pp. 1156–1172, 11 2018. [Online]. Available: https://doi.org/10.1001/jamapsychiatry.2018.2165

[108] Verschiendene Autoren, " Datensicherung für alle – von Laptop bis Supercomputer," https://doku.lrz.de/pages/viewpage.action?pageId=30084073, 2020, [Online; accessed 06-Januar-2020].

[109] D. Dwyer, "Main page," http://mitnvpfs1.srv.med.uni-muenchen.de/mediawiki/index.php?title=Main_Page, 2019, [Online; accessed 06-Januar-2020].

[110] MATLAB, "version 9.5.0.944444 (r2018b)," 2018.

[111] A. Hasan, T. Wobrock, I. Großimlinghaus, J. Zielasek, B. Janssen, D. Reich-Erkelenz, I. Kopp, W. Gaebel, and P. Falkai, "Die Aktualisierung der DGPPN S3-Leitlinie Schizophrenie-aktueller Stand," *Die Psychiatrie*, vol. 12, no. 01, pp. 19–27, 2015.

[112] B. Nelson, K. Yuen, and A. R. Yung, "Ultra high risk (UHR) for psychosis criteria: are there different levels of risk for transition to psychosis?" *Schizophrenia research*, vol. 125, no. 1, pp. 62–68, 2011.

[113] G. Pedersen and S. Karterud, "The symptom and function dimensions of the Global Assessment of Functioning (GAF) scale," *Comprehensive psychiatry*, vol. 53, no. 3, pp. 292–298, 2012.

[114] J. Asendorpf and H. G. Wallbott, "Maße der Beobachterübereinstimmung: ein systematischer Vergleich," *Zeitschrift für Sozialpsychologie*, vol. 10, no. 3, pp. 243–252, 1979.

[115] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: uses in assessing rater reliability." *Psychological bulletin*, vol. 86, no. 2, p. 420, 1979.

[116] M. Startup, M. C. Jackson, and S. Bendix, "The concurrent validity of the Global Assessment of Functioning (GAF)," *British Journal of Clinical Psychology*, vol. 41, no. 4, pp. 417–422, 2002.

[117] R. E. Carrión, A. M. Auther, D. McLaughlin, R. Olsen, J. Addington, C. E. Bearden, K. S. Cadenhead, T. D. Cannon, D. H. Mathalon, T. H. McGlashan *et al.*, "The global functioning: social and role scales—further validation in a large sample of adolescents and young adults at clinical high risk for psychosis," *Schizophrenia bulletin*, vol. 45, no. 4, pp. 763–772, 2019.

[118] N. He, "Ethical considerations for clinical trials on medicinal products conducted with the paediatric population," *Eur J Health Law*, vol. 15, pp. 223–250, 2008.

[119] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker, and G. D. Tourassi, "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance," *Neural Networks*, vol. 21, no. 2, pp. 427 – 436, 2008, advances in Neural Networks Research: IJCNN '07. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0893608007002407

[120] B. Widrow and M. E. Hoff, "Adaptive switching circuits," Stanford Univ Ca Stanford Electronics Labs, Tech. Rep., 1960.

[121] Y. Ito, "Representation of functions by superpositions of a step or sigmoid function and their applications to neural network theory," *Neural Networks*, vol. 4, no. 3, pp. 385–394, 1991.

[122] B. L. Kalman and S. C. Kwasny, "Why tanh: choosing a sigmoidal function," in *[Proceedings 1992] IJCNN International Joint Conference on Neural Networks*, vol. 4. IEEE, 1992, pp. 578–581.

[123] D. F. Specht, "A general regression neural network," *IEEE transactions on neural networks*, vol. 2, no. 6, pp. 568–576, 1991.

[124] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," California Univ San Diego La Jolla Inst for Cognitive Science, Tech. Rep., 1985.

[125] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*. Springer, 2010, pp. 177–186.

[126] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner *et al.*, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[127] Verschiedene Autoren, "MATLAB Optimization Toolbox," 2018.

[128] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[129] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," 1999.

[130] M. Stone, "Cross-validation: A review," *Statistics: A Journal of Theoretical and Applied Statistics*, vol. 9, no. 1, pp. 127–139, 1978.

[131] A. Shabtai, Y. Elovici, and L. Rokach, *A survey of data leakage detection and prevention solutions*. Springer Science & Business Media, 2012.

[132] M. W. Browne, "Cross-validation methods," *Journal of mathematical psychology*, vol. 44, no. 1, pp. 108–132, 2000.

[133] J. D. Rodriguez, A. Perez, and J. A. Lozano, "Sensitivity analysis of k-fold cross validation in prediction error estimation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 3, pp. 569–575, 2009.

[134] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.

[135] D. R. Wilson and T. R. Martinez, "Instance pruning techniques," in *ICML*, vol. 97, no. 1997, 1997, pp. 400–411.

[136] J. Mockus, "On bayesian methods for seeking the extremum," pp. 400–404, 1975.

[137] The MathWorks, Inc., Verschiedene Autoren, "Bayesian optimization algorithm," https://de.mathworks.com/help/stats/bayesian-optimization-algorithm.html, 2020.

[138] K. B. Ensor and P. W. Glynn, "Stochastic optimization via grid search," *Lectures in Applied Mathematics-American Mathematical Society*, vol. 33, pp. 89–100, 1997.

[139] D. R. Jones, "A taxonomy of global optimization methods based on response surfaces," *Journal of global optimization*, vol. 21, no. 4, pp. 345–383, 2001.

[140] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," in *Advances in neural information processing systems*, 2012, pp. 2951–2959.

[141] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[142] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," pp. 3121–3124, 2010.

[143] J. Shreffler and M. R. Huecker, "Diagnostic testing accuracy: Sensitivity, specificity, predictive values and likelihood ratios," 2020.

[144] I. O. for Standardization, *ISO 5725-1: 1994: accuracy (trueness and precision) of measurement methods and results-part 1: general principles and definitions*. International Organization for Standardization, 1994.

[145] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature," *Geoscientific model development*, vol. 7, no. 3, pp. 1247–1250, 2014.

[146] T. Fawcett, "An introduction to ROC analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[147] I. The MathWorks. (2021) Trapezoidal numerical integration. Natick, Massachusetts, United States. [Online]. Available: https://de.mathworks.com/help/matlab/ref/trapz.html

[148] J. Huang and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," *IEEE Transactions on knowledge and Data Engineering*, vol. 17, no. 3, pp. 299–310, 2005.

[149] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[150] G. Modinos, M. J. Kempton, S. Tognin, M. Calem, L. Porffy, M. Antoniades, A. Mason, M. Azis, P. Allen, B. Nelson *et al.*, "Association of adverse outcomes with emotion processing and its neural substrate in individuals at clinical high risk for psychosis," *JAMA psychiatry*, vol. 77, no. 2, pp. 190–200, 2020.

[151] P. Allen, C. A. Chaddock, A. Egerton, O. D. Howes, G. Barker, I. Bonoldi, P. Fusar-Poli, R. Murray, and P. McGuire, "Functional outcome in people at high risk for psychosis predicted by thalamic glutamate levels and prefronto-striatal activation," *Schizophrenia bulletin*, vol. 41, no. 2, pp. 429–439, 2015.

[152] H. Rinne, "Taschenbuch der Statistik," 1997.

[153] R. Parikh, A. Mathai, S. Parikh, G. Chandra Sekhar, and R. Thomas, "Understanding and using sensitivity, specificity and predictive values," *Indian journal of ophthalmology*, vol. 56, no. 1, p. 45—50, 2008. [Online]. Available: https://europepmc.org/articles/PMC2636062

[154] J. J. Deeks and D. G. Altman, "Diagnostic tests 4: likelihood ratios," *Bmj*, vol. 329, no. 7458, pp. 168–169, 2004.

[155] D. Altman, "Diagnostic tests," *Statistics with confidence*, pp. 100–120, 2001.

[156] C. J. Clopper and E. S. Pearson, "The use of confidence or fiducial limits illustrated in the case of the binomial," *Biometrika*, pp. 404–413, 1934.

[157] N. D. Mercaldo, K. F. Lau, and X. H. Zhou, "Confidence intervals for predictive values with an emphasis to case–control studies," *Statistics in medicine*, vol. 26, no. 10, pp. 2170–2183, 2007.

[158] M. S. Ltd. (2021) Diagnostic test evaluation calculator. [Online]. Available: https://www.medcalc.org/calc/diagnostic_test.php

[159] M. L. McHugh, "The chi-square test of independence," *Biochemia medica*, vol. 23, no. 2, pp. 143–149, 2013.

[160] T. M. Franke, T. Ho, and C. A. Christie, "The chi-square test: Often used and more often misinterpreted," *American Journal of Evaluation*, vol. 33, no. 3, pp. 448–458, 2012.

[161] M. Raymond and F. Rousset, "An exact test for population differentiation," *Evolution*, pp. 1280–1283, 1995.

[162] R. A. Fisher, "On the interpretation of x^2 from contingency tables, and the calculation of p," *Journal of the Royal Statistical Society*, vol. 85, no. 1, pp. 87–94, 1922. [Online]. Available: http://www.jstor.org/stable/2340521

[163] J. Bortz, G. A. Lienert, and K. Boehnke, *Verteilungsfreie Methoden in der Biostatistik*. Springer-Verlag, 2008.

[164] J. D. Gibbons and S. Chakraborti, *Nonparametric statistical inference*. CRC press, 2020.

[165] I. The MathWorks, *Statistical Toolbox*, Natick, Massachusetts, United States, 2018. [Online]. Available: https://www.mathworks.com/help/stats/ranksum.html

[166] N. Cressie and H. Whitford, "How to use the two sample t-test," *Biometrical Journal*, vol. 28, no. 2, pp. 131–148, 1986.

[167] M. M. Mukaka, "A guide to appropriate use of correlation coefficient in medical research," *Malawi medical journal*, vol. 24, no. 3, pp. 69–71, 2012.

[168] C. Spearman, "The proof and measurement of association between two things." 1961.

[169] B. At, "An inventory for measuring depression," *Archives of General Psychiatry*, vol. 4, no. 6, pp. 561–571, 1961. [Online]. Available: +http://dx.doi.org/10.1001/archpsyc.1961.01710120031004

[170] A. Schaub, *Kognitiv-psychoedukative Therapie zur Bewältigung von Depressionen: Ein Therapiemanual*, 2013.

[171] S. R. Kay, A. Fiszbein, and L. A. Opler, "The positive and negative syndrome scale (PANSS) for schizophrenia," *Schizophrenia bulletin*, vol. 13, no. 2, pp. 261–276, 1987.

[172] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[173] C. G. Forero, E. Olariu, P. Álvarez, J.-I. Castro-Rodriguez, M. J. Blasco, G. Vilagut, V. Pérez, J. Alonso, I. Investigators *et al.*, "Change in functioning outcomes as a predictor of the course of depression: a 12-month longitudinal study," *Quality of Life Research*, vol. 27, no. 8, pp. 2045–2056, 2018.

[174] R. M. Hirschfeld, S. A. Montgomery, M. B. Keller, S. Kasper, A. F. Schatzberg, H.-J. Moller, D. Healy, D. Baldwin, M. Humble, and M. Versiani, "Social functioning in depression: a review." *The Journal of clinical psychiatry*, vol. 61, no. 4, pp. 0–0, 2000.

[175] F. Renner, P. Cuijpers, and M. Huibers, "The effect of psychotherapy for depression on improvements in social functioning: a meta-analysis," *Psychological medicine*, vol. 44, no. 14, pp. 2913–2926, 2014.

[176] A.-K. J. Fett, W. Viechtbauer, D. L. Penn, J. van Os, L. Krabbendam *et al.*, "The relationship between neurocognition and social cognition with functional outcomes in schizophrenia: a meta-analysis," *Neuroscience & Biobehavioral Reviews*, vol. 35, no. 3, pp. 573–588, 2011.

[177] C. Barrowclough and N. Tarrier, "Social functioning in schizophrenic patients," *Social Psychiatry and Psychiatric Epidemiology*, vol. 25, no. 3, pp. 125–129, 1990.

[178] P. Fusar-Poli, M. Tantardini, S. De Simone, V. Ramella-Cravaro, D. Oliver, J. Kingdon, M. Kotlicka-Antczak, L. Valmaggia, J. Lee, M. Millan, and et al., "Deconstructing vulnerability for psychosis: Meta-analysis of environmental risk factors for psychosis in subjects at ultra high-risk," *European Psychiatry*, vol. 40, p. 65–75, 2017.

[179] B. A. Cornblatt, R. E. Carrión, J. Addington, L. Seidman, E. F. Walker, T. D. Cannon, K. S. Cadenhead, T. H. McGlashan, D. O. Perkins, M. T. Tsuang *et al.*, "Risk factors for psychosis: impaired social and role functioning," *Schizophrenia bulletin*, vol. 38, no. 6, pp. 1247–1257, 2011.

[180] R. E. Carrión, A. M. Auther, D. McLaughlin, J. Addington, C. E. Bearden, K. S. Cadenhead, T. D. Cannon, M. Keshavan, D. H. Mathalon, T. H. McGlashan *et al.*, "Social

decline in the psychosis prodrome: Predictor potential and heterogeneity of outcome," *Schizophrenia Research*, vol. 227, pp. 44–51, 2021.

[181] E. M. Seabrook, M. L. Kern, and N. S. Rickard, "Social networking sites, depression, and anxiety: a systematic review," *JMIR mental health*, vol. 3, no. 4, p. e5842, 2016.

[182] T. S. Brugha, J. Wing, C. Brewin, B. MacCarthy, and A. Lesage, "The relationship of social network deficits with deficits in social functioning in long-term psychiatric disorders," *Social Psychiatry and Psychiatric Epidemiology*, vol. 28, no. 5, pp. 218–224, 1993.

[183] T. W. Baskin, B. E. Wampold, S. M. Quintana, and R. D. Enright, "Belongingness as a protective factor against loneliness and potential depression in a multicultural middle school," *The Counseling Psychologist*, vol. 38, no. 5, pp. 626–651, 2010.

[184] M. d. C. García de Jesús and M. d. G. C. Ferriani, "School as a „protective factor" against drugs: perceptions of adolescents and teachers," *Revista latino-americana de enfermagem*, vol. 16, pp. 590–594, 2008.

[185] A. W. Riley, M. E. Ensminger, B. Green, and M. Kang, "Social role functioning by adolescents with psychiatric disorders," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 37, no. 6, pp. 620–628, 1998.

[186] D. M. Fergusson and L. J. Woodward, "Mental health, educational, and social role outcomes of adolescents with depression," *Archives of general psychiatry*, vol. 59, no. 3, pp. 225–231, 2002.

[187] H. L. Piersma and J. L. Boes, "The GAF and psychiatric outcome: a descriptive report," *Community Mental Health Journal*, vol. 33, no. 1, pp. 35–41, 1997.

[188] T. V. Lagerberg, O. A. Andreassen, P. A. Ringen, A. O. Berg, S. Larsson, I. Agartz, K. Sundet, and I. Melle, "Excessive substance use in bipolar disorder is associated with impaired functioning rather than clinical characteristics, a descriptive study," *BMC psychiatry*, vol. 10, no. 1, pp. 1–9, 2010.

[189] I. M. Aas, "Global Assessment of Functioning (GAF): properties and frontier of current knowledge," *Annals of general psychiatry*, vol. 9, no. 1, p. 20, 2010.

[190] B. D. Dufton and C. Siddique, "Measures in the day hospital: I. The Global Assessment of Functioning Scale." *International journal of partial hospitalization*, 1992.

[191] H. H. Goldman, A. E. Skodol, and T. R. Lave, "Revising axis V for DSM-IV: a review of measures of social functioning," *Am J Psychiatry*, vol. 149, p. 9, 1992.

[192] M. Ekström, "Do watching eyes affect charitable giving? evidence from a field experiment," *Experimental Economics*, vol. 15, no. 3, pp. 530–546, 2012.

[193] K. B. (KBV), "Neuerungen bei der psychiatrischen häuslichen krankenpflege," 2021, https://www.kbv.de/html/1150_39900.php/, Last accessed on 2021-05-10.

[194] R. Fernández-Ballesteros, "Psychological assessment: Future challenges and progresses." *European psychologist*, vol. 4, no. 4, p. 248, 1999.

[195] G. J. Meyer, S. E. Finn, L. D. Eyde, G. G. Kay, K. L. Moreland, R. R. Dies, E. J. Eisman, T. W. Kubiszyn, and G. M. Reed, "Psychological testing and psychological assessment: A review of evidence and issues." *American psychologist*, vol. 56, no. 2, p. 128, 2001.

[196] S. M. Sotsky, D. R. Glass, M. T. Shea, P. A. Pilkonis, F. Collins, I. Elkin, J. T. Watkins, S. D. Imber, W. R. Leber, J. Moyer *et al.*, "Patient predictors of response to psychotherapy and pharmacotherapy: Findings in the nimh treatment of depression collaborative research program," *Focus*, vol. 148, no. 2, pp. 997–290, 2006.

[197] A. Abdel-Baki, C. Ouellet-Plamondon, and A. Malla, "Pharmacotherapy challenges in patients with first-episode psychosis," *Journal of affective disorders*, vol. 138, pp. S3–S14, 2012.

[198] J. Addington, D. Piskulic, and C. Marshall, "Psychosocial treatments for schizophrenia," *Current Directions in Psychological Science*, vol. 19, no. 4, pp. 260–263, 2010.

[199] S. D. Hollon, M. E. Thase, and J. C. Markowitz, "Treatment and prevention of depression," *Psychological Science in the public interest*, vol. 3, no. 2, pp. 39–77, 2002.

[200] P. Buckley, G. Fenley, A. Mabe, and S. Peeples, "Recovery and schizophrenia," *Clinical Schizophrenia & Related Psychoses*, vol. 1, no. 1, pp. 96–100, 2007.

[201] A. Vita and S. Barlati, "Recovery from schizophrenia: is it possible?" *Current opinion in psychiatry*, vol. 31, no. 3, pp. 246–255, 2018.

[202] N. C. Andreasen, W. T. Carpenter Jr, J. M. Kane, R. A. Lasser, S. R. Marder, and D. R. Weinberger, "Remission in schizophrenia: proposed criteria and rationale for consensus," *American Journal of Psychiatry*, vol. 162, no. 3, pp. 441–449, 2005.

[203] R. M. Van Eck, T. J. Burger, A. Vellinga, F. Schirmbeck, and L. de Haan, "The relationship between clinical and personal recovery in patients with schizophrenia spectrum disorders: a systematic review and meta-analysis," *Schizophrenia Bulletin*, vol. 44, no. 3, pp. 631–642, 2018.

[204] B. T. Mausbach, R. Moore, C. Bowie, V. Cardenas, and T. L. Patterson, "A review of instruments for measuring functional recovery in those diagnosed with psychosis," *Schizophrenia bulletin*, vol. 35, no. 2, pp. 307–318, 2009.

[205] D. Popovic, K. Schiltz, P. Falkai, and N. Koutsouleris, "Precision psychiatry and the contribution of brain imaging and other biomarkers," *Fortschritte Der Neurologie-Psychiatrie*, vol. 88, no. 12, pp. 778–785, 2020.

[206] R. Hodson, "Precision medicine," *Nature*, vol. 537, no. 7619, pp. S49–S49, 2016.

[207] B. S. Fernandes, L. M. Williams, J. Steiner, M. Leboyer, A. F. Carvalho, and M. Berk, "The new field of 'precision psychiatry'," *BMC medicine*, vol. 15, no. 1, pp. 1–7, 2017.

[208] K. J. Friston, "Precision psychiatry," *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, vol. 2, no. 8, pp. 640–643, 2017.

[209] D. Bzdok and A. Meyer-Lindenberg, "Machine learning for precision psychiatry: opportunities and challenges," *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, vol. 3, no. 3, pp. 223–230, 2018.

[210] R. C. Hall, "Global Assessment of Functioning: a modified scale," *Psychosomatics*, vol. 36, no. 3, pp. 267–275, 1995.

[211] O. Sonesson, T. Tjus, and H. Arvidsson, "Reliability of a functioning scale (GAF) among psychiatric ward staff," *Nordic Psychology*, 2010.

[212] I. M. Aas, O. Sonesson, and S. Torp, "A qualitative study of clinicians experience with rating of the global assessment of functioning (gaf) scale," *Community Mental Health Journal*, vol. 54, no. 1, pp. 107–116, 2018.

[213] J. Goldhahn, V. Rampton, and G. A. Spinas, "Could artificial intelligence make doctors obsolete?" *Bmj*, vol. 363, p. k4563, 2018.

[214] R. Mulder, G. Newton-Howes, and J. W. Coid, "The futility of risk prediction in psychiatry," *The British Journal of Psychiatry*, vol. 209, no. 4, pp. 271–272, 2016.

[215] M. K. Chan, H. Bhatti, N. Meader, S. Stockton, J. Evans, R. C. O'Connor, N. Kapur, and T. Kendall, "Predicting suicide following self-harm: systematic review of risk factors and risk scales," *The British Journal of Psychiatry*, vol. 209, no. 4, pp. 277–283, 2016.

[216] D. Altman, Y. Vergouwe, and P. Royston, "Prognosis and prognostic research: application and impact of prognostic models in clinical practice," *BMJ*, vol. 338, p. b606, 2009.

[217] J. Heckman, "Varieties of selection bias," *The American Economic Review*, vol. 80, no. 2, pp. 313–318, 1990.

[218] A. Ferretti, M. Schneider, and A. Blasimme, "Machine learning in medicine: opening the new data protection black box," *Eur. Data Prot. L. Rev.*, vol. 4, p. 320, 2018.

[219] C. Behrendt, H. Pridöhl, K. Schaar, H. Federrath, and E. Debus, "Clinical registers in the twenty-first century: Balancing act between data protection and feasibility?" *Der Chirurg; Zeitschrift fur Alle Gebiete der Operativen Medizen*, vol. 88, no. 11, pp. 944–949, 2017.

[220] E. Byvatov, U. Fechner, J. Sadowski, and G. Schneider, "Comparison of support vector machine and artificial neural network systems for drug/nondrug classification," *Journal of chemical information and computer sciences*, vol. 43, no. 6, pp. 1882–1889, 2003.

[221] K. G. Mehrotra, C. K. Mohan, and S. Ranka, "Bounds on the number of samples needed for neural learning," 1980.

[222] C. Liu, C. Gao, X. Xia, D. Lo, J. Grundy, and X. Yang, "On the replicability and reproducibility of deep learning in software engineering," *arXiv preprint arXiv:2006.14244*, 2020.

[223] R. Caruana, S. Lawrence, and L. Giles, "Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping," *Advances in neural information processing systems*, pp. 402–408, 2001.

[224] L. Prechelt, "Early stopping-but when?" in *Neural Networks: Tricks of the trade*. Springer, 1998, pp. 55–69.

[225] J. B. Asendorpf, M. Conner, F. De Fruyt, J. De Houwer, J. J. Denissen, K. Fiedler, S. Fiedler, D. C. Funder, R. Kliegl, B. A. Nosek *et al.*, "Recommendations for increasing replicability in psychology," *European journal of personality*, vol. 27, no. 2, pp. 108–119, 2013.

[226] L. A. Fink, D. Bernstein, L. Handelsman, J. Foote, and M. Lovejoy, "Initial reliability and validity of the childhood trauma interview: a new multidimensional measure of childhood interpersonal trauma," *The American journal of psychiatry*, vol. 152, no. 9, p. 1329, 1995.

[227] K. Wingenfeld, C. Spitzer, C. Mensebach, H. J. Grabe, A. Hill, U. Gast, N. Schlosser, H. Höpp, T. Beblo, and M. Driessen, "The German version of the Childhood Trauma Questionnaire (CTQ): preliminary psychometric properties," *Psychotherapie, Psychosomatik, Medizinische Psychologie*, vol. 60, no. 11, pp. 442–450, 2010.

[228] C. Esteban, O. Staeck, S. Baier, Y. Yang, and V. Tresp, *Predicting clinical events by combining static and dynamic information using recurrent neural networks*, 2016.

[229] D. Bzdok and A. Meyer-Lindenberg, "Machine learning for precision psychiatry: Opportunities and challenges," *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, vol. 3, no. 3, pp. 223–230, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2451902217302069

Danksagung

Zuerst möchte ich mich bei meinem Doktorvater Nikolaos Koutsouleris bedanken, der mir die nötigen Ressourcen und Möglichkeiten zur Verfügung stellte, um Matlab lernen zu können und mit zahlreichen konstruktiven Verbesserungsvorschläge zu dieser Arbeit beitrug.
Zudem möchte ich mich bei meinem Betreuer David Popovic bedanken, der mich stets motivierte und oftmals mehr an das Projekt glaubte als ich selbst. Er war es, der mich auf den Themenkomplex „Machine Learning" in der Psychiatrie aufmerksam machte. Bei ihm muss ich mich auch für die kritische Auseinandersetzung mit dem Thema und für die Durchsicht meiner Arbeit herzlich bedanken. Die zahlreichen motivierenden Gespräche werden mir immer als bereichernder und konstruktiver Austausch in Erinnerung bleiben.
Aus dem restlichen Pronia-Team möchte ich mich gerne bei Dom, Marc und Anne bedanken, die ebenfalls durch interessante Anregungen und Hilfestellungen, sowie bei technischen Problemen stets hilfreich zur Seite standen.
Auch meiner Familie bin ich großen Dank schuldig. Vor allem meiner Mutter, die mich gegen Ende mit einem 3D Drucker motivierte diese Dissertation zu beenden. Ich möchte auch meinem Vater danken, ohne den mir mein Medizinstudium nicht möglich gewesen wäre, meinen Brüdern für die stetigen Ermunterungen. Zuletzt möchte ich mich bei meinem Partner, Moyuan, bedanken, der mich stets seelisch, technisch und mit leckerem Essen unterstützte.

Eidesstattliche Versicherung

Ich, *Julia Maria Eder*, erkläre hiermit an Eides statt, dass ich die vorliegende Dissertation mit dem Titel:

Die Prädiktion des klinischen Funktionsniveaus mit Hilfe von neuronalen Netzen

selbständig verfasst, mich außer der angegebenen keiner weiteren Hilfsmittel bedient und alle Erkenntnisse, die aus dem Schrifttum ganz oder annähernd übernommen sind, als solche kenntlich gemacht und nach ihrer Herkunft unter Bezeichnung der Fundstelle einzeln nachgewiesen habe.

Ich erkläre des Weiteren, dass die hier vorgelegte Dissertation nicht in gleicher oder in ähnlicher Form bei einer anderen Stelle zur Erlangung eines akademischen Grades eingereicht wurde.

München, 22.11.2022 Julia Eder

Ort, Datum Unterschrift