Zürcher Hochschule
für Angewandte Wissenschaften

School of
Engineering

IDP Institut für Datenanalyse
und Prozessdesign

Bachelorarbeit
Wirtschaftsingenieurwesen

Verwendung von 3D CNNs zur Prognose von Therapieergebnissen bei Schlaganfall-Patienten mithilfe von TOF-MRA Bildern

Autorin	Michèle Wieland
Hauptbetreuung	Beate Sick
Nebenbetreuung	Helmut Grabner

Datum	11.06.2021

Zürcher Hochschule
für Angewandte Wissenschaften

School of Engineering

Erklärung betreffend das selbstständige Verfassen einer Bachelorarbeit an der School of Engineering

Mit der Abgabe dieser Bachelorarbeit versichert der/die Studierende, dass er/sie die Arbeit selbständig und ohne fremde Hilfe verfasst hat. (Bei Gruppenarbeiten gelten die Leistungen der übrigen Gruppenmitglieder nicht als fremde Hilfe.)
Der/die unterzeichnende Studierende erklärt, dass alle zitierten Quellen (auch Internetseiten) im Text oder Anhang korrekt nachgewiesen sind, d.h. dass die Bachelorarbeit keine Plagiate enthält, also keine Teile, die teilweise oder vollständig aus einem fremden Text oder einer fremden Arbeit unter Vorgabe der eigenen Urheberschaft bzw. ohne Quellenangabe übernommen worden sind.

Bei Verfehlungen aller Art treten die Paragraphen 39 und 40 (Unredlichkeit und Verfahren bei Unredlichkeit) der ZHAW Prüfungsordnung sowie die Bestimmungen der Disziplinarmassnahmen der Hochschulordnung in Kraft.

Ort, Datum:	**Name Studierende:**
Zürich, 11.06.2021	Michèle Wieland M. Wieland

Zusammenfassung

Eine verlässliche und effiziente Bildgebung spielt bei der Behandlung von Schlaganfällen eine entscheidende Rolle. Mithilfe der Computertomographie (CT) und der Magnetresonanztomographie (MRT) können Schlaganfälle detektiert werden. Ergänzend wird die Magnetresonanzangiographie (MRA) eingesetzt, um unter anderem Ursachen zu identifizieren und über passende Therapiemassnahmen zu entscheiden. Zur Unterstützung von medizinischen Fachpersonen bei der Bildanalyse können automatische Bildverarbeitungsmethoden wie zum Beispiel Convolutional Neural Networks (CNNs) eingesetzt werden. In dieser Arbeit wird die Verwendung von 3D CNNs zur Detektion von Schlaganfällen und zur Prognose des funktionellen Outcomes (Therapieergebnis) untersucht. Als Trainingsdaten für die 3D CNNs dienen Time-of-Flight-MRA (TOF-MRA) Bilder von Patienten, welche einen ischämischen Schlaganfall oder eine transitorische ischämische Attacke (TIA) erlitten haben. Aufgrund der beschränkten Verfügbarkeit von Trainingsdaten werden unter anderem Methoden wie Data Augmentation und Dropout eingesetzt. Die Vorhersagegüte der 3D CNNs wird mithilfe einer 5-fachen Kreuzvalidierung evaluiert. Zusätzlich wird die Methode des Deep Ensemblings eingesetzt, um zum einen die Vorhersagen zu verbessern und zum anderen Informationen zur Verlässlichkeit der Vorhersagen zu erhalten. Das Modell zur Detektion von Schlaganfällen erreicht einen AUC von 0.597 [0.543, 0.651]. Die TOF-MRA Bilder werden in der klinischen Praxis nicht zur Detektion von Schlaganfällen verwendet, was ein Grund dafür sein kann, dass das Modell keine besseren Vorhersagen erzielt. Das Modell zur Prognose des funktionellen Outcomes erreicht eine AUC von 0.740 [0.683, 0.796]. Ähnliche Studien zur Prognose des funktionellen Outcomes erreichen vergleichbare Resultate (Zihni et al. 2020, Hilbert et al. 2019, Bacchi et al. 2020). Es ist somit anzunehmen, dass die TOF-MRA Bilder Informationen zum Verlauf der Erkrankung liefern. Um die Modell-Leistung weiter zu verbessern, sollten zum einen mehr Trainingsdaten gesammelt werden und zum anderen wäre das Hinzufügen von klinischen Daten eine vielversprechende Möglichkeit.

Abstract

Reliable and efficient medical imaging plays a crucial role in stroke care. Strokes can be detected using computed tomography (CT) and magnetic resonance tomography (MRT). Complementary, magnetic resonance angiography (MRA) is used to identify causes and to decide on suitable therapies for treatment. Automatic image processing methods such as Convolutional Neural Networks (CNNs) can be used to support physicians in image analysis. In this work, the use of 3D CNNs for stroke detection and functional outcome prognosis are investigated. Time-of-flight MRA (TOF-MRA) images of patients who have suffered an ischemic stroke, or a transient ischemic attack (TIA) serve as training data for the 3D CNNs. Due to limited availability of training data, methods such as data augmentation and dropout are used. The predictive quality of the 3D CNNs is evaluated using 5-fold cross-validation. In addition, deep ensembling is used to improve predictions and to provide information on the reliability of predictions. The model for stroke detection achieves an AUC of 0.597 [0.543, 0.651]. TOF-MRA images are not used for stroke detection in clinical practice, which may be a reason why the model does not yield better results. The model for functional outcome prediction achieves an AUC of 0.740 [0.683, 0.796]. Similar studies on the prediction of functional outcome achieve comparable results (Zihni et al. 2020, Hilbert et al. 2019, Bacchi et al. 2020). It can be assumed that TOF-MRA images provide information on the development of the disease. To further improve model performance, more training data should be collected and adding clinical data could be a promising option.

Inhaltsverzeichnis

1 Einleitung

Gemäss der Weltgesundheitsorganisation war im Jahr 2019 der Schlaganfall die zweithäufigste Todesursache weltweit und der Grund für über 6 Millionen Todesfälle (World Health Organization 2020). Wenn ein Schlaganfall aufgrund des Verschlusses eines hirnversorgenden Gefässes auftritt, spricht man von einem ischämischen Schlaganfall. Ist eine intrazerebrale Blutung die zugrundeliegende Ursache, so bezeichnet man dies als hämorrhagischen Schlaganfall. 80-85% aller Schlaganfälle sind ischämischer Art (Schwert & Bäzner 2018).

Ein ischämischer Schlaganfall wird auch Hirninfarkt genannt und beschreibt eine akute fokale Minderdurchblutung des Gehirns. Die Minderdurchblutung tritt aufgrund eines Verschlusses einer hirnversorgenden Arterie auf. Dies hat zur Folge, dass nachgeschaltete Hirnabschnitte nicht mehr genügend durchblutet werden und es dadurch zu einer strukturellen Hirnschädigung kommt. Leidet der Patient an vorübergehenden, flüchtigen, fokal-neurologischen Symptomen, welche Minuten bis wenige Stunden, aber nicht mehr als 24 Stunden dauern, spricht man von einer transitorischen ischämischen Attacke (TIA). Die Symptome einer TIA sind vollständig reversibel, jedoch ist das Risiko für einen Schlaganfall mit bleibenden Folgen in den Stunden und Tagen nach einer TIA erhöht (Cassier-Woidasky 2014, 30-31).

In der Akutphase des Schlaganfalls ist die Zeit ein wichtiger Faktor. Je länger der Patient nicht therapeutisch behandelt wird, desto mehr Hirngewebe wird geschädigt (Kessler 2014). Als Basis für eine erfolgreiche Therapie ist eine schnelle und verlässliche Bildgebung notwendig. Als Verfahren zur Bildgebung stehen die Computertomographie (CT) und die Magnetresonanztomographie (MRT) zur Verfügung. Dabei ist die MRT der CT in der Erkennung von Schlaganfällen überlegen. Die MRT wird häufig mit einer Magnetresonanzangiographie (MRA) kombiniert. Die MRA ermöglicht die Darstellung der hirnversorgenden Gefässe und hilft, die Ursachen der Akutsymptomatik wie zum Beispiel Gefässverschlüsse zu finden. Zudem lässt sich mit der MRA eine Indikation für die Therapiemassnahmen stellen. Eine Variante ist die Time-of-Flight-MRA (TOF-MRA) (Ringelstein & Forsting 2017).

Das Analysieren von MRA-Bildern erfordert ausgebildete Experten und ist zeitaufwändig. Um Experten bei der Analyse der Bilder zu unterstützen, könnten Convolutional Neural Networks (CNNs) eingesetzt werden (Singh et al. 2020, 1-2). CNNs sind in der Lage, Features in Bildern selbstständig zu erlernen (Hu et al. 2020). Damit müssen Features aus unstrukturierten Bilddaten nicht manuell extrahiert werden, wie dies bei herkömmlichen Machine Learning Methoden der Fall ist (Dürr et al. 2020, 22). Die Grundlagen zu CNNs sind in *Deep Learning with Python* zu finden (Chollet 2017).

Aufgrund der sinkenden Kosten für Rechenleistung und der Verfügbarkeit von rechenstarken Grafikprozessoren (GPUs) ist es in den letzten Jahren möglich geworden, medizinische 3D Bilder mit 3D CNNs zu analysieren (Singh et al. 2020, 2). Bei der Anwendung von 3D CNNs müssen die Schichten der medizinischen 3D Bilder nicht einzeln betrachtet werden, wie dies bei der Verwendung von 2D CNNs der Fall wäre (Nie et al. 2016). Bei den 3D CNNs geht somit der 3D Kontext der ursprünglichen Bilder nicht verloren (Huang et al. 2017). Zudem wird pro 3D Bild, das heisst pro Patienten, nur ein Bildlabel benötigt. Würde man die Schichten einzeln analysieren, so wäre für jede 2D Bildschicht ein separates Label notwendig. Mit der Verwendung von 3D CNNs geht jedoch einher, dass viel weniger Trainingsbilder zur Verfügung stehen. Diese Tatsache stellt die grösste Herausforderung für das Training von 3D CNNs dar (Singh et al. 2020, 2).

In dieser Arbeit wird vorgeschlagen, 3D CNNs einzusetzen, um 3D TOF-MRA Bilder von ischämischen Schlaganfall- und TIA-Patienten zu analysieren. Die 3D CNNs werden verwendet, um

a) zu unterscheiden, ob der Patient einen ischämischen Schlaganfall oder eine TIA erlitten hat.

b) den funktionellen Outcome (Therapieergebnis) des Patienten drei Monate nach der Behandlung vorherzusagen.

Der Aufbau des 3D CNNs ist im Grundsatz gleich wie dieser eines 2D CNNs mit dem Unterschied, dass wo nötig 3D anstelle von 2D Layern verwendet werden. Die 3D Convolutional Layer bestehen aus 3D Kernel, welche die Feature-Extraktion aus dem Volumen ermöglichen (Hesamian et al. 2019). Um genügend Daten für das Training zu haben, wird 3D Data Augmentation eingesetzt und um ein Overfitting auf den Trainingsdaten zu vermeiden wird unter anderem Dropout zur Regularisierung verwendet (Huang et al. 2017).

1.1 Verwandte Arbeiten

In den letzten Jahren sind Deep Learning Methoden für die Verarbeitung von medizinischen Bildern populärer geworden. Seit dem Jahr 2015 lässt sich ein exponentieller Anstieg der Publikationen mit Deep Learning Methoden im medizinischen Bereich verzeichnen. Seit dem Jahr 2017 steigt zudem die Anzahl Publikationen im Bereich 3D Deep Learning für medizinische Zwecke (Singh et al. 2020, 2-3), jedoch ist 3D Deep Learning vergleichsweise viel weniger verbreitet (Merino et al. 2021). 3D Deep Learning Modelle wurden in weiteren Studien bereits erfolgreich verwendet, um zum Beispiel Alzheimer (Korolev et al. 2017), Schizophrenie (Hu et al. 2020) oder Hirntumore (Amin et al. 2020) in 3D MRT-Bildern des Gehirns zu erkennen. Viele solcher Studien kämpfen mit der Herausforderung, dass im Vergleich zu Deep Learning aus anderen Anwendungsgebieten nur wenige Daten vorliegen (Singh et al. 2020, 2). Dafür verantwortlich ist unter anderem die Tatsache, dass medizinische Bilder nur mit Einwilligung des Patienten für Forschungszwecke zur Verfügung stehen. Zusätzlich erfordert das Klassifizieren von medizinischen Bildern Fachwissen, ist aufwändig, zeitintensiv und somit teuer (Vu et al. 2020).

Wenn wenige Daten vorliegen, hat sich Transfer Learning im Bereich der 2D Bildanalyse etabliert (Chollet 2017, 143). Da aber die Forschung und Anwendungen im Bereich 3D Deep Learning noch nicht so weit fortgeschritten ist wie im Bereich des 2D Deep Learnings, fehlt es an vortrainierten 3D CNNs, welche für Transfer Learning benutzt werden können. Einige Forscher bedienen sich deshalb an vortrainierten 2D CNNs und erweitern deren Architekturen und Gewichte in die dritte Dimension. Für die Transformation verwenden zum Beispiel Merino et al. erfolgreiche 2D Netzwerken wie das ResNet oder das EfficientNet B0. Die transformierten Gewichte werden als initiale Gewichte für das Training des neuen 3D CNNs eingesetzt (Merino et al. 2021, Carreira & Zisserman 2018).

Herkömmliche Machine Learning Methoden, bei denen Features zuerst manuell aus den Bildern extrahiert werden müssen, bieten eine Alternative zu Deep Learning Methoden, die direkt die Bilder als Eingabe verwenden (Dürr et al. 2020, 22). Einige Studien haben die Anwendung solcher manuellen Feature-Extraktionen für die Analyse von MRT-Bildern getestet. Dabei wurden Methoden wie Support Vector Machine und Random Forest verwendet, um mit den manuell extrahierten Features eine Klassifikations-Vorhersage zu machen. In den betrachteten Studien konnte gezeigt werden, dass mit den 3D Deep Learning Methoden in diesen spezifischen Anwendungen bessere Resultate als mit den herkömmlichen Machine Learning Methoden erzielt werden können (Hu et al. 2020, Rachmadi et al. 2017, Vu et al. 2020).

2 Material und Methoden

2.1 Datensatz

Der Datensatz besteht aus TOF-MRA Bildern von 508 Patienten, welche entweder einen ischämischen Schlaganfall oder eine TIA erlitten haben. Die Daten wurden am Departement für Neurologie des Universitätsspitals Zürich gesammelt. Die Bilder wurden bis zu drei Tage nach der Behandlung der Patienten aufgenommen.

Pro Patienten liegt eine Bildserie von 2D Grauwertbildern vor. Die Bildserien bestehen aus einer unterschiedlichen Anzahl Bildern. Das Minimum ist 47 und das Maximum 264 Bilder. Am häufigsten (in 60% der Fälle) liegen 200 Bilder pro Patienten vor. Über alle Patienten hinweg besteht der Datensatz aus 96'718 2D Bildern. Die einzelnen Bilder der Bildserien haben unterschiedliche Formate, wobei das Format 640 x 560 Pixel am häufigsten vorkommt (in 85% der Fälle). Wenn die einzelnen Bilder einer Bildserie aufeinandergestapelt werden, dann ergibt dies ein 3D Grauwertbild. Die Bilder liegen im DICOM-Format vor. Über alle Bilder ist der minimale Signalwert bei 0 und der maximale Signalwert bei 2'393. Die Abbildung 2.1 zeigt vier Bilder eines Patienten mit den dazugehörenden Histogrammen der Signalwerte.

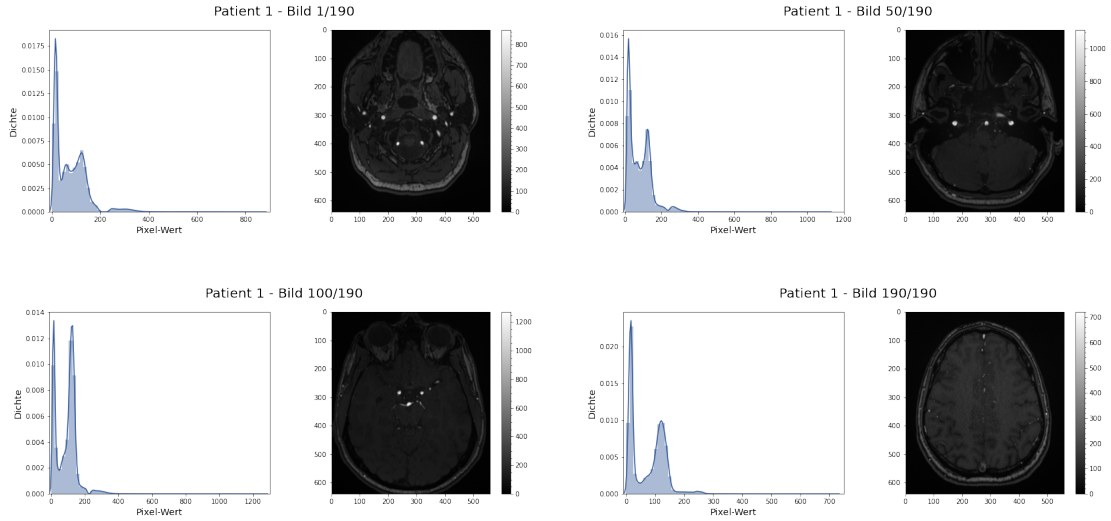

Abbildung 2.1: **Vier Bilder eines Patienten im Original-Format und die dazugehörenden Histogramme der Signalwerte:** Es ist zu erkennen, dass sich der maximale Signalwert je nach Bild unterscheidet.

Von den 508 Patienten haben 351 Patienten einen ischämischen Schlaganfall und 157 Patienten eine TIA erlitten. Diese Information liegt als Label pro Patienten vor (0 für TIA

und 1 für ischämischer Schlaganfall). Für 478 Patienten ist zusätzlich der mRS (modified Ranking Scale) drei Monate nach der Behandlung vorhanden. Der mRS ist ein ordinaler Score, der von Medizinern häufig zur Bewertung der Genesung nach einem Schlaganfall eingesetzt wird und eine Abschätzung des funktionellen Outcomes ermöglicht (Schmitt 2016). In der Tabelle 2.1 sind die Grade des mRS und deren Bedeutungen erläutert (Banks & Marotta 2007).

Tabelle 2.1: Erläuterung der sechs Stufen des mRS

mRS	Beschreibung
0	Keine Symptome
1	Keine wesentlichen Beeinträchtigungen trotz der Symptome; in der Lage allen üblichen Pflichten und Aktivitäten nachzugehen
2	Leichte Beeinträchtigungen; nicht in der Lage, alle bisherigen Tätigkeiten auszuführen, aber in der Lage, sich ohne Unterstützung um die eigenen Angelegenheiten zu kümmern
3	Moderate Beeinträchtigungen; benötigt etwas Hilfe, kann aber ohne Unterstützung gehen
4	Moderat schwere Beeinträchtigungen; nicht in der Lage, ohne Unterstützung zu gehen; nicht in der Lage, eigene Bedürfnisse ohne Hilfe zu erfüllen
5	Schwere Beeinträchtigungen; bettlägerig, inkontinent und benötigt konstante Pflege und Betreuung
6	Tod

Um aus dem 6-stufigen ordinalen mRS Score einen binären Score zu machen, ist es üblich bei mRS $= 2$ den Schwellwert zu setzen (Zihni et al. 2020). Werte von 0 bis 2 bedeuten einen guten funktionellen Outcome und Werte von 3 bis 6 einen schlechten funktionellen Outcome.

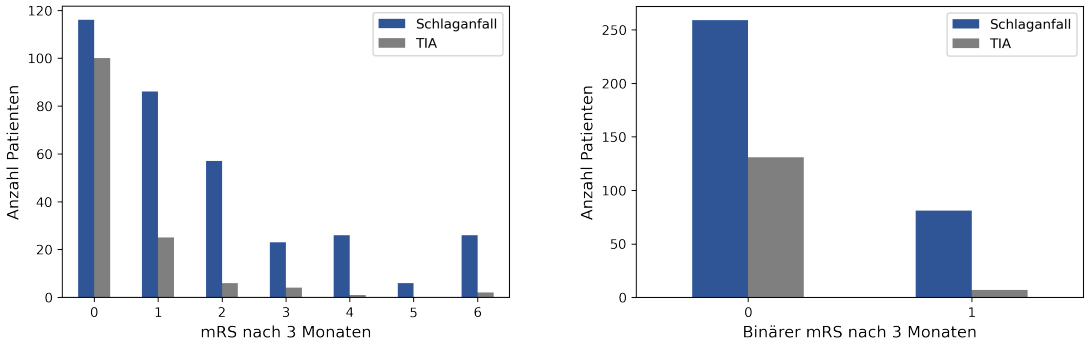

Abbildung 2.2: **Anzahl Patienten pro mRS-Grad:** Die linke Abbildung zeigt die Anzahl Patienten pro mRS-Grad und wie viele Patienten pro mRS-Grad einen Schlaganfall respektive eine TIA erlitten haben. Die rechte Abbildung zeigt die Anzahl Patienten für den binären Score, wobei 0 für einen guten funktionellen Outcome und 1 für einen schlechten funktionellen Outcome steht.

Nach der Zusammenfassung beinhaltet der Datensatz 390 Patienten mit einem guten funktionellen Outcome und 88 Patienten mit einem schlechten funktionellen Outcome.

2.2 Datenaufbereitung

Wie im Kapitel 2.1 erwähnt, kann sich die Grösse der TOF-MRA Bilder zwischen den Patienten unterscheiden. Für den Input des 3D CNNs müssen die Bilder ein einheitliches Format aufweisen und aufgrund der beschränkten Rechenkapazität verkleinert werden (Singh et al. 2020, 17). Dafür wird die lineare Spline-Interpolation verwendet. Bei der linearen Spline-Interpolation definiert man ein gewünschtes Zielformat für die Bilder und berechnet dann für jede Dimension den entsprechenden Zoom-Faktor, um das Zielformat zu erhalten. Nun zoomt man in jeder Dimension mit dem berechneten Faktor entlang der entsprechenden Achse (Zunair et al. 2020, 3-4). Während der Experimente wurden die Bildformate 128 x 112 x 40 und 64 x 56 x 20 getestet. Dabei konnten mit dem Bildformat 128 x 112 x 40 bessere Ergebnisse erzielt werden.

Die ursprünglichen Signalwerte in den DICOM-Bildern befinden sich je nach Patienten in einem unterschiedlichen Bereich. Das heisst, die maximalen Signalwerte unterscheiden sich je nach Patient stark. Unterschiedliche Bereiche der Signalwerte und grosse Werte erschweren das Training des CNNs. Aus diesem Grund werden die Bilder zuerst normalisiert. Die hier gewählte Methode zur Normalisierung besteht aus zwei Schritten. Im ersten Schritt werden die Werte pro 3D Bild auf das Intervall zwischen 0 und 255 skaliert. Im zweiten Schritt wird dann der Mittelwert der Signalwerte des 3D Bilds auf 0 und die Standardabweichung auf 1 gebracht (Chollet 2017, 101).

Als Input für ein 3D CNN dienen 5D Tensoren in der Form (Batchgrösse, Höhe, Breite, Tiefe, Farbkanäle). Um die hier vorliegenden Bilder in ein 3D CNN zu geben, müssen diese um eine Dimension erweitert werden. Da die TOF-MRA Bilder Grauwertbilder sind, haben diese nur einen Farbkanal (Chollet 2017, 37). Nach der Dimensionserweiterung liegen somit Bilder im Format 128 x 112 x 40 x 1 vor.

2.3 Modell-Architekturen

2.3.1 3D CNN Modell

Typische CNNs für die Klassifikation von Bildern bestehen aus zwei Teilen. Der erste «Convolutional Teil» ist für die Feature-Extraktion aus den Bildern verantwortlich und der zweite «Fully Connected Teil» liefert im Falle eines Klassifikationsproblems die Vorhersagen für die Klassen (Chollet 2017, 120-123). Für die Feature-Extraktion nutzen CNNs dank ihrer Architektur ein hierarchisches Vorgehen. Während die ersten Layer des CNNs, allgemeine Features wie Ecken oder Kanten erlernen, können höhere Layers domänenspezifische Features extrahieren (Swati et al. 2019, 35). Im ersten Teil werden Convolutional und MaxPooling Layers verwendet. Im zweiten Teil werden Fully Connected Layers für die Klassifikations-Vorhersage eingesetzt (Chollet 2017, 120-123).

Während 2D CNNs räumlich lokale Merkmale auf Bildern erkennen können, ermöglichen 3D CNNs, Merkmale aus dem ganzen 3D Volumen zu extrahieren und diese zu lernen (Singh et al. 2020, 4). Gegenüber 2D CNNs enthalten 3D CNNs unter anderem 3D

anstelle von 2D Convolutional und MaxPooling Layers. Die 3D Convolutional Layers bestehen aus 3D Kernels, welche über die 3D Inputdaten gefahren werden und so 3D Feature Maps erzeugen. Die 3D Eigenschaft der Kernel führt zu einem dreidimensionalen rezeptiven Feld und ermöglicht so die Informationsgewinnung aus dem ganzen Volumen (Gessert et al. 2018). Als Nachteil von 3D CNNs ist zu nennen, dass die Architektur einfach gehalten werden muss und die Input-Bilder reduziert werden müssen, da der Rechenaufwand ansonsten zu hoch wäre und die Rechenkapazitäten beschränkt sind (Hu et al. 2020).

In dieser Arbeit werden zwei 3D CNNs erstellt. Das Modell zur «Schlaganfall Detektion» hat zum Ziel, anhand der 3D TOF-MRA Bilder vorherzusagen, ob der Patient einen ischämischen Schlaganfall oder eine TIA erlitten hat. Das Modell zur «Outcome Prognose» sagt voraus, wie der funktionelle Outcome des Patienten drei Monate nach der Behandlung aussieht. Die Modelle wurden in Keras mit Tensorflow Backend implementiert (Chollet 2017, 62). Die verwendete Tensorflow Version ist die Version 2.4.0. Der Code für die verwendeten Modelle und Methoden steht auf GitHub zur Verfügung: `https://github.com/wielandmichele/TOF-MRA-images`.

Modell «Schlaganfall Detektion»

Die für das erste Modell verwendete Architektur besteht aus fünf 3D Convolutional Layers mit Kernelgrösse 3 x 3 x 3, auf welche jeweils ein MaxPooling 3D Layer mit Pool-Size 2 x 2 x 2 und ein Batch-Normalization Layer folgen. Der vierdimensionale Output des letzten 3D Convolutional Layers wird mithilfe eines Flatten-Layers in einen eindimensionalen Vektor transformiert. Nach dem Flatten-Layer folgen ein Dense-Layer sowie ein Dropout-Layer. In den Convolutional-Layers und dem Dense-Layer wird die nicht lineare Aktivierungsfunktion ReLU (Rectified Linear Unit) verwendet. Im Dense-Layer wird zusätzlich eine L2-Regulierung für die Gewichte eingesetzt. Der Output-Layer ist ein Dense-Layer mit zwei Outputs und Softmax-Aktivierungsfunktion. Somit liefert der letzte Layer die Klassifikations-Vorhersagen für die beiden Klassen. Als Loss-Funktion für das Training wurde die Categorical Crossentropy und als Optimizer Adam verwendet. Die initiale Lernrate beträgt 0.00001 und die Batchgrösse 2. Die Abbildung 2.3 zeigt eine Visualisierung der Modell-Architektur.

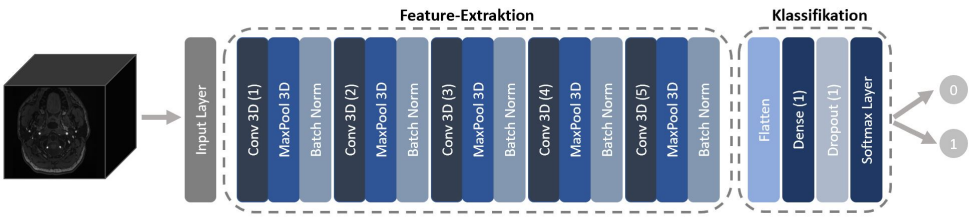

Abbildung 2.3: Schematische Darstellung der Modell-Architektur für das Modell zur Schlaganfall Detektion (in Anlehnung an (Zunair et al. 2020, 6))

Modell «Outcome Prognose»

Für die Outcome Prognose werden fünf Modelle mit verschiedenen Hyperparametern verwendet (siehe Kapitel 2.4). Im ersten Teil (Feature-Extraktion) unterscheiden sich

die Modelle nur bezüglich der Anzahl Filter in den Convolutional Layers vom ersten Modell. Der Klassifikationsblock beinhaltet je nach Modell einen oder zwei Dense- und Dropout-Layer. Als Output-Layer dient wiederum ein Dense-Layer mit zwei Outputs und Softmax-Aktivierungsfunktion. Als Loss-Funktion für das Training wurde ebenfalls die Categorical Crossentropy und als Optimizer Adam verwendet. Die initiale Lernrate beträgt 0.00001 und die Batchgrösse 4. Die Abbildung 2.4 visualisiert die verwendete Architektur.

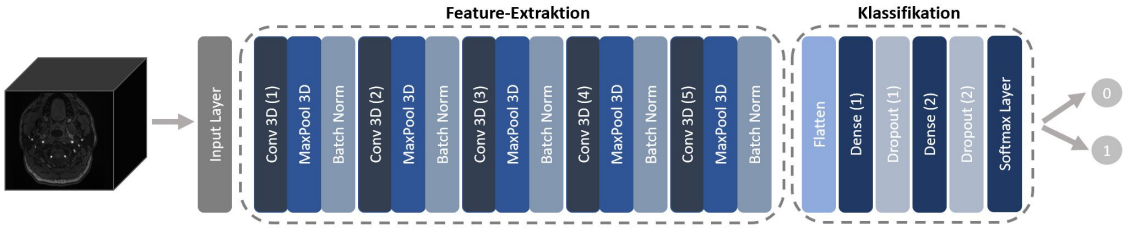

Abbildung 2.4: Schematische Darstellung der Modell-Architektur für das Modell zur Outcome Prognose (in Anlehnung an (Zunair et al. 2020, 6))

2.3.2 Transfer Learning Modell

Wenn Deep Learning mit einem kleinen Datensatz praktiziert wird, dann liegt es nahe, Transfer Learning zu nutzen (Chollet 2017, 37). Wie im Kapitel 1.1 erwähnt, gibt es im Bereich von 3D Deep Learning erst wenige vortrainierte Netzwerke. Aus diesem Grund werden Gewichte von vortrainierten 2D CNNs in die dritte Dimension erweitert. Eine dafür verwendete Methode ist, die zweidimensionalen Gewichte der 2D CNNs zu extrudieren. Dabei werden die zweidimensionalen Gewichte entlang der z-Achse kopiert, sodass man dreidimensionale Gewichte erhält (Merino et al. 2021).

In dieser Arbeit wurde das von João Carreira und Andrew Zisserman entwickelte Inflated 3D ConvNet getestet (Carreira & Zisserman 2018). Das Inflated 3D ConvNet nutzt die Architektur und die Gewichte des Inception-V1 CNNs und erweitert diese in die dritte Dimension (Stolz 2020). Zusätzlich wird das transformierte Netzwerk noch auf dem Kinetics-400 Datensatz (bestehend aus 240'000 Kurzvideos) trainiert (Carreira & Zisserman 2018).

Das vortrainierte Inflated 3D ConvNet wird mit den Gewichten, aber ohne den Klassifikationsblock geladen und dient als Basis für das eigene 3D CNN. Auf diese Basis wird ein eigener Klassifikationsblock gesetzt. Dieser besteht aus einem Flatten-Layer und einem Dense-Layer mit 32 Knoten. Darauf folgt der Output-Layer mit zwei Outputs und Softmax-Aktivierungsfunktion.

Als erstes wird nur der zusätzliche Klassifikationsblock trainiert und die restlichen Layer aus der Basis werden eingefroren. Anschliessend werden einige Layer aus der Basis gemeinsam mit dem Klassifikationsblock trainiert (Chollet 2017, 154). Es gab Hinweise dafür, dass bessere Resultate erzielt werden können, wenn man alle Layer mit dem Klassifikationsblock trainiert. Jedoch konnte mithilfe des Inflated 3D ConvNets die Modell-

Leistung nicht wie erhofft verbessert werden. Aus diesem Grund wurde dieser Ansatz nicht weiterverfolgt.

2.4 Optimierung der Hyperparameter

Um die Hyperparameter der beiden 3D CNNs zu bestimmen, wird eine Bayessche Optimierungsmethode eingesetzt. Bei der Bayesschen Optimierung handelt es sich um einen iterativen Algorithmus, bei dem die nächsten zu evaluierenden Hyperparameter basierend auf den zuvor erhaltenen Resultaten bestimmt werden. Die Bayessche Optimierung besteht aus zwei Hauptkomponenten: dem Surrogatmodell und der Akquisitionsfunktion. Das Surrogatmodell wird verwendet, um die Zielfunktion zu modellieren (Yang & Shami 2020, 303-304). Die iterative Maximierung der Akquisitionsfunktion hilft beim Finden des globalen Optimums, indem sowohl in noch nicht erkundeten Bereichen als auch in Bereichen, in denen basierend auf den erhaltenen Resultaten vermutlich das Optimum liegt, gesucht wird (Wang et al. 2017, 507). Als Surrogatmodell wird trees of Parzen estimators verwendet.

Für das Modell zur Schlaganfall Detektion erfolgt die Optimierung auf einem der fünf Validierungsteile (siehe Kapitel 2.6). Dabei wurde die in Tabelle 2.2 ersichtliche Auswahl an Hyperparametern vorgegeben. In der Spalte «Gewählter Wert» ist ersichtlich, welche Hyperparameter am Ende für das Modell verwendet wurden.

Tabelle 2.2: Zur Auswahl gestellte Hyperparameter für das Modell «Schlaganfall Detektion»

Hyperparameter	Auswahl an Parametern	Gewählter Wert
Anzahl Filter im Conv 3D (1)	8, 16, 32	8
Anzahl Filter im Conv 3D (2)	8, 16, 32	16
Anzahl Filter im Conv 3D (3)	16, 32, 64	32
Anzahl Filter im Conv 3D (4)	16, 32, 64	32
Anzahl Filter im Conv 3D (5)	32, 64, 128	64
Anzahl Knoten im Dense (1)	32, 64, 128	64
Dropbout-Rate im Dropout (1)	Gleichverteilung zw. 0 und 0.6	0.01

Die Anzahl Filter in den Convolutional 3D Layern entspricht der Anzahl 3D Feature Maps, welche pro 3D Convolutional Layer erzeugt werden (Vu et al. 2020). Das heisst, dass im ersten 3D Convolutional Layer 8 3D Feature-Maps, im zweiten 16 3D Feature-Maps erstellt werden usw. Bei der optimierten Modell-Architektur nimmt somit die Anzahl Feature Maps entlang der 3D Convolutional Layer zu, während die Grösse der Feature Maps abnimmt. Dieses Schema ist in fast allen CNNs zu beobachten (Chollet 2017, 134). Die gefundene Dropout-Rate ist tiefer als erwartet, da normalerweise eine Rate zwischen 0.2 und 0.5 verwendet wird. Eine Dropout-Rate von 0.01 bedeutet, dass 1% aller Features auf null gesetzt werden (Chollet 2017, 109).

Für das Modell zur Outcome Prognose erfolgt die Optimierung auf allen fünf Validierungsteilen einzeln (siehe Kapitel 2.6), woraus fünf Modelle mit verschiedenen Hyperparametern resultieren. Für die Optimierung wurden die in Tabelle 2.3 definierten Hyperparameter vorgegeben.

Tabelle 2.3: Zur Auswahl gestellte Hyperparameter für das Modell «Outcome Prognose»

Hyperparameter	Auswahl an Parametern
Anzahl Filter im Conv 3D (1)	8, 16, 32
Anzahl Filter im Conv 3D (2)	16, 32, 64
Anzahl Filter im Conv 3D (3)	16, 32, 64
Anzahl Filter im Conv 3D (4)	32, 64 , 128
Anzahl Filter im Conv 3D (5)	32, 64, 128
Anzahl Knoten im Dense (1)	32, 64, 128
Dropbout-Rate im Dropout (1)	Gleichverteilung zw. 0 und 0.6
Verwendung Dense-Layer (2)	Ja, Nein
Anzahl Knoten im Dense (2)	32, 64, 128
Dropbout-Rate im Dropout (2)	Gleichverteilung zw. 0 und 0.6

Mithilfe der Optimierung wurden die folgenden Hyperparameter für die fünf Modelle gewählt:

Tabelle 2.4: Gewählte Hyperparameter für die fünf Modelle

Hyperparameter	Modell 1	Modell 2	Modell 3	Modell 4	Modell 5
Anzahl Filter im Conv 3D (1)	32	16	16	32	32
Anzahl Filter im Conv 3D (2)	32	16	32	32	32
Anzahl Filter im Conv 3D (3)	32	32	32	32	32
Anzahl Filter im Conv 3D (4)	64	32	64	64	64
Anzahl Filter im Conv 3D (5)	64	32	64	64	64
Anzahl Knoten im Dense (1)	128	64	64	32	128
Dropbout-Rate im Dropout (1)	0.3	0.1	0.3	0.1	0.35
Verwendung Dense-Layer (2)	Ja	Nein	Ja	Nein	Nein
Anzahl Knoten im Dense (2)	128	-	64	-	-
Dropbout-Rate im Dropout (2)	0.3	-	0.3	-	-

In allen fünf Modellen steigt die Anzahl 3D Feature Maps in den Convolutional Layers an. Im Modell 1 sowie im Modell 3 werden zwei Dense Layer und in den restlichen Modellen wird ein Dense Layer im Klassifikationsblock verwendet. Die Dropout-Rate befindet sich je nach Modell zwischen 0.1 und 0.35.

2.5 Data Augmentation

Die Idee von Data Augmentation ist, aus den verfügbaren Trainingsdaten mehr Daten zu generieren, sodass das Modell mehr Aspekte der Daten sieht und so besser generalisieren kann (Chollet 2017, 139). In dieser Arbeit werden mittels randomisierter Transformationen neue 3D Bilder generiert. Die verwendeten Transformationen sind Zufälliges Drehen, Zoomen, Verschieben, Spiegeln und ein zufälliger Gauss-Filter (Dörn 2017). In der Tabelle 2.5 werden die verwendeten Funktionen genauer erläutert.

Tabelle 2.5: Verwendete Funktionen für 3D Data Augmentation

Funktion	Beschrieb	Parameter
Zufälliges Drehen	Das 3D-Bild wird zufällig in der xy-, xz- oder yz-Ebene gedreht.	Min. und max. Drehung in der xy-Ebene: -20, 20 Min. und max. Drehung in der xz-Ebene: -5, 5 Min. und max. Drehung in der yz-Ebene: -5, 5
Zufälliges Zoomen	Das 3D-Bild wird zufällig gezoomt.	Min. Zoom: 0.8 Max. Zoom: 1.3
Zufälliges Verschieben	Das 3D-Bild wird zufällig in x- und y-Richtung verschoben.	Min. und max. Verschiebung in x-Richtung: -20, 20 Min. und max. Verschiebung in y-Richtung: -20, 20
Zufälliges Spiegeln	Das 3D-Bild wird zufällig vertikal gespiegelt.	
Zufälliger Gauss-Filter	Rauschreduzierung und Bildglättung mittels Gauss-Filter.	Max. Standardabweichung des Gauss Kernels: 0.2

Für das Training werden auf jedes Bild zufällig zwei verschiedene Data Augmentation Funktionen aus der Tabelle 2.5 angewendet.

2.6 Training und Testing

Um alle Daten nutzen zu können, wird eine 5-fache Kreuzvalidierung verwendet. Dabei wird der Datensatz zufällig in fünf Teile zerlegt. In jedem der fünf Teile entspricht die Verteilung der Klassen etwa der gleichen wie im ganzen Datensatz. Im ersten Trainingsdurchgang wird der erste der fünf Teile zu Testzwecken, der zweite zur Validierung und die restlichen drei zu Trainingswecken eingesetzt. Dieser Vorgang wiederholt sich fünf Mal, sodass jeder Teil einmal zu Testzwecken verwendet wird. Für das Modell zur Schlaganfall Detektion bestehen die Testsätze jeweils aus 102 oder 101 Patienten und für das Modell zur Outcome Prognose aus 96 oder 95 Patienten. Die Abbildung 2.5 zeigt eine Visualisierung der Prozedur.

Abbildung 2.5: Schematische Darstellung der 5-fachen Kreuzvalidierung

In jedem der fünf Durchgänge wird das Modell für 150 Epochen trainiert. Für die Vorhersage werden die Gewichte der Epoche mit dem tiefsten Loss auf dem Validierungsdatensatz verwendet. Damit kann einem Overfitting auf den Trainingsdaten entgegengewirkt werden (Chollet 2017, 249-250). Da es sich um unbalancierte Datensätze handelt, werden in jeder Epoche gleich viele Bilder von der positiven Klasse wie von der negativen Klasse für das Training verwendet. Auf diese Weise findet das Training auf einem balancierten Datensatz statt (Tensorflow 2021).

Da die Resultate der 3D CNNs aufgrund der zufälligen Initialisierung zufälligen Schwankungen ausgesetzt sind, wird von der Methode des Modell Ensemblings Gebrauch gemacht. Die Idee von Modell Ensembling ist, Vorhersagen von Modellen, die sich mit verschiedenen Initialisierungsgewichten ergeben, zu kombinieren, um bessere Resultate zu erzielen (Lakshminarayanan et al. 2017, Chollet 2017, 264 - 266). In diesem Fall wird mit einer Ensemble-Grösse von 5 gearbeitet, d.h. die 5-fache Kreuzvalidierung wird fünf Mal wiederholt, sodass man für alle Patienten fünf Vorhersagen erhält. Anschliessend werden die Mittelwerte und die Varianzen aus diesen fünf vorhergesagten Wahrscheinlichkeiten berechnet. Die gemittelten Wahrscheinlichkeiten ergeben die Vorhersage des Ensemble Modells und die erhaltenen Varianzen liefern eine Indikation für die Unsicherheit des Modells bei den Vorhersagen.

2.7 Modell-Bewertung

Wie in Kapitel 2.3.1 erwähnt, wird die Categorical Crossentropy als Loss-Funktion zur Modelloptimierung eingesetzt. Im Falle eines Klassifikationsproblems ist die Categorical Crossentropy das Gleiche wie der negative Log Likelihood (NLL) (Dürr et al. 2020, 104-105). Neben dem Einsatz als Loss-Funktion eignet sich diese Metrik auch, um die Leistung eines probabilistischen Modells zu beurteilen und liefert eine Bewertung der Vorhersageunsicherheit (Ovadia et al. 2019). Die Crossentropy ist im Falle einer binären Klassifikation mit zwei Outputs wie folgt gegeben:

$$Crossentropy = -\frac{1}{n}\Big(\sum_{j\,mit\,y_j=0} log(p_0(x_j)) + \sum_{j\,mit\,y_j=1} log(p_1(x_j))\Big) \tag{2.1}$$

Wobei n die Anzahl Daten aus beiden Klassen, y_j die Klassenlabels, p_0 die vorhergesagten Wahrscheinlichkeiten für die wahre Klasse 0 und p_1 die vorhergesagten Wahrscheinlich-

keiten für die wahre Klasse 1 sind (Dürr et al. 2020, 104-105). Verwendet man das untrainierte CNN Modell für die Vorhersage, so klassifiziert es zufällig alle als Klasse 0 oder alle als Klasse 1. Angenommen das Modell klassifiziert alle als Klasse 0, so gehören alle Patienten mit einer Wahrscheinlichkeit von knapp über 0.5 der Klasse 0 an. Das Gleiche gilt, wenn alle als Klasse 1 klassifiziert werden. Setzt man diese Wahrscheinlichkeiten in die Gleichung 2.1 ein, so ergibt dies ein Wert von ungefähr 0.69 für das untrainierte Modell.

Um die Resultate der Modelle zu beurteilen, werden primär der Test-NLL, als Proper Score, und der Test-AUC verwendet. Der AUC ist definiert als die Fläche unter der ROC Kurve und kann Werte zwischen 0 und 1 annehmen, wobei ein Klassifikator erst ab einem AUC grösser als 0.5 nützlich ist (Held & Bové 2014, 308). Für den Test-AUC wird jeweils das 95% Konfidenzintervall mit der Methode von Delong berechnet (Robin et al. 2011). Neben dem NLL und dem AUC werden die Accuracy, Sensitivität und Spezifität auf dem Testset ausgewiesen. Für diese Metriken wird jeweils das 95% Wilson Konfidenzintervall berechnet (Brown et al. 2001, 107-108).

3 Resultate

Die Experimente wurden mithilfe der beiden in Kapitel 2.3.1 vorgestellten 3D CNNs durchgeführt. Das Modell «Schlaganfall Detektion» hat zum Ziel, zwischen TIA- und Schlaganfall-Patienten zu unterscheiden, während das Modell «Outcome Prognose» den funktionellen Outcome der Patienten drei Monate nach der Behandlung voraussagt. Die Resultate für die beiden Modelle sind in den folgenden Kapiteln zu finden.

3.1 Ergebnisse Modell «Schlaganfall Detektion»

Mit 5-facher Kreuzvalidierung wurden die Performance Metriken des Ensembling Modells bestimmt. Dabei wurde ein AUC von 0.597 [0.543, 0.651], eine Accuracy von 0.656 [0.613, 0.696] und ein NLL von 0.644 erreicht. Die Klasse Schlaganfall sagt das Modell in 79.5% aller Fälle richtig vorher, während nur 34.4% aller TIA-Patienten erkannt werden. Die Klasse Schlaganfall kann somit deutlich besser erkannt werden als die Klasse TIA. Eine Ursache dafür könnte sein, dass für die Klasse TIA weniger Trainingsdaten zur Verfügung stehen. Die Performance Metriken und Konfidenzintervalle sind in der Tabelle 3.1 aufgeführt.

Tabelle 3.1: Test Performace Metriken für das Modell «Schlaganfall Detektion»

Metrik	Wert	95% Konfidenzintervall
AUC	0.597	von 0.543 bis 0.651
Accuracy	0.656	von 0.613 bis 0.696
Sensitivität	0.795	von 0.750 bis 0.834
Spezifität	0.344	von 0.274 bis 0.421
NLL	0.644	

Mithilfe des Ensembling-Modells erhält man für jede Vorhersage eine Varianz. Zur Beschreibung der Verteilung der Varianzen dienen als Lagemasse der Mittelwert sowie der Median und als Streuungsmass der IQR (interquartile range). An den berechneten Werten in Tabelle 3.2 und der Verteilung der Varianzen in Abbildung 3.1 lässt sich erkennen, dass falsche Klassifizierungen eine höhere Varianz und eine höhere Streuung aufweisen. Die höhere Varianz kann als grössere Unsicherheit des Modells bei der Vorhersage interpretiert werden. Bei der Klasse Schlaganfall ist die Varianz bei den falschen Klassifizierungen im Mittelwert und Median deutlich höher als diese der richtigen Klassifizierungen. Dies lässt sich auch an den Histogrammen in der zweiten Zeile der Abbildung 3.1 erkennen. Bei der Klasse TIA scheint man keine Aussage treffen zu können. Dies könnte daran

liegen, dass diese Klasse grundsätzlich schlechter erkannt wird und auch weniger Daten für diese Klasse vorliegen.

Tabelle 3.2: Varianzen der Vorhersagen für das Modell «Schlaganfall Detektion»

Klassifikation	Mittelwert	Median	IQR
Richtige Klassifizierungen (n = 333)	0.031	0.025	0.029
Falsche Klassifizierungen (n = 175)	0.037	0.029	0.038
Richtige Klassifizierungen Klasse Schlaganfall (n = 279)	0.031	0.026	0.029
Falsche Klassifizierungen Klasse Schlaganfall (n = 72)	0.052	0.046	0.055
Richtige Klassifizierungen Klasse TIA (n = 54)	0.026	0.020	0.026
Falsche Klassifizierungen Klasse TIA (n = 103)	0.027	0.022	0.023

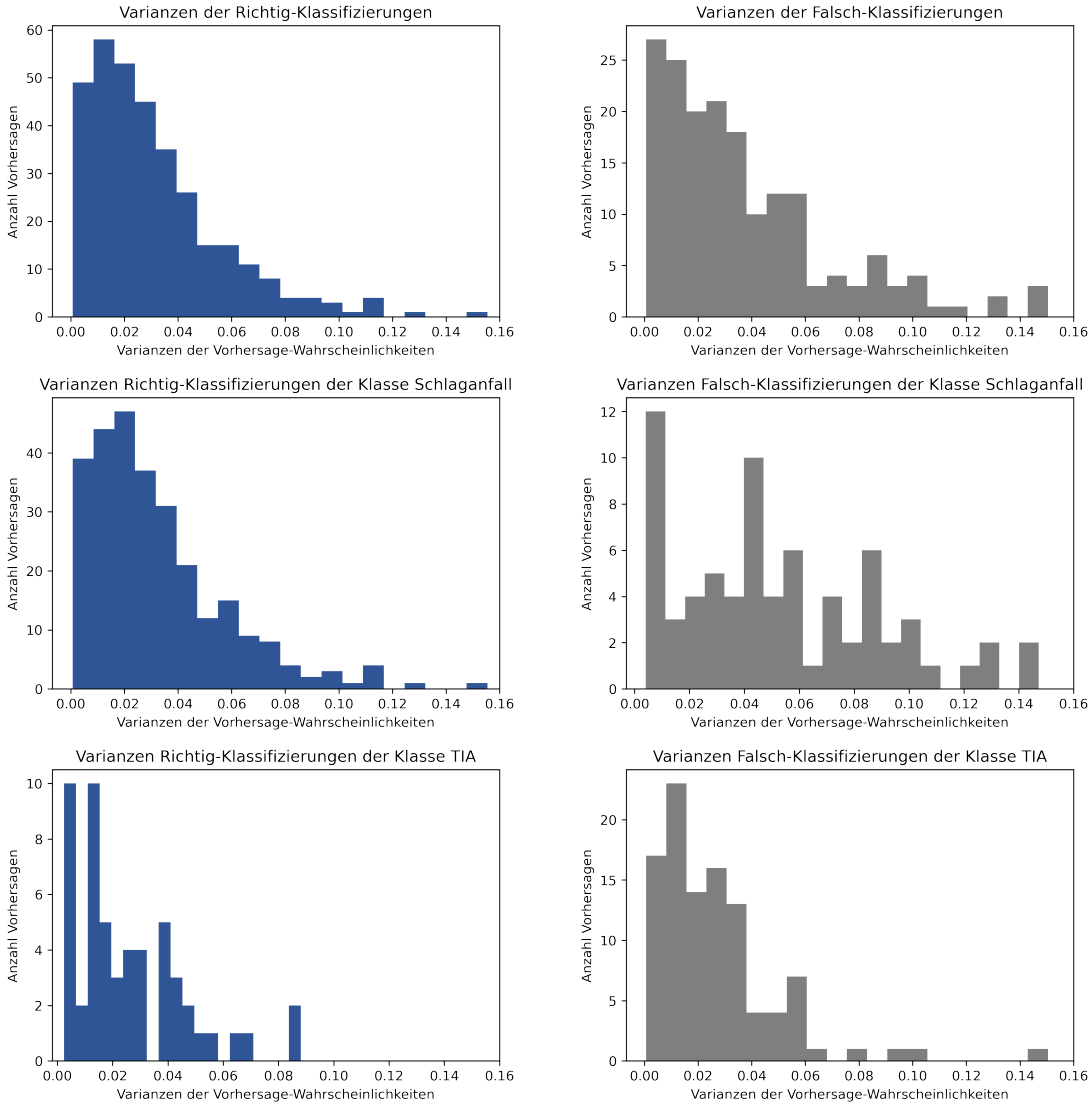

Abbildung 3.1: **Verteilung der Varianzen für das Modell «Schlaganfall Detektion»:** In der ersten Zeile sind die Varianzen über alle richtigen und falschen Klassifizierungen ersichtlich. In der zweiten Zeile werden die Varianzen für die Klasse Schlaganfall und in der dritten Zeile für die Klasse TIA gezeigt.

Als Vergleichswert zur Beurteilung der Modell-Leistung dient die grobe Abschätzung einer erfahrenen Neurologin. Gemäss ihrer Aussage würde sie 40-50% aller Schlaganfälle in den TOF-MRA Bildern der 508 Patienten erkennen. Um zudem die Klassifikations-Vorhersagen zu validieren, wurden der Neurologin einige Falsch-Klassifikationen präsentiert. Bei den vom Modell fälschlicherweise als Klasse TIA klassifizierten Bildern konnte die Neurologin bestätigen, dass auf diesen auch für sie kein Schlaganfall zu erkennen ist. Bei den fälschlicherweise als Schlaganfall klassifizierten Bildern, konnte sich die Neurologin nicht erklären, weshalb diese vom Modell der Klasse TIA zugeordnet wurden. Durch die Betrachtung der Aussagen der Neurologin ist anzunehmen, dass nicht auf allen TOF-MRA Bildern ein Schlaganfall bzw. eine TIA zu erkennen ist. Dies könnte ein Grund dafür sein, dass das Modell Schwierigkeiten hat die Bilder akkurat zu klassifizieren.

3.2 Ergebnisse Modell «Outcome Prognose»

Für das Modell zur Outcome Prognose wurden ebenfalls mit 5-facher Kreuzvalidierung die Performance Metriken des Ensembling Modells bestimmt. Das Modell klassifiziert 83.1% aller Patienten mit gutem funktionellen Outcome und 42.0% mit schlechtem funktionellen Outcome richtig. Die Klasse 0 (= guter funktioneller Outcome) kann somit deutlich besser vorhergesagt werden als die Klasse 1 (= schlechter funktioneller Outcome). Es ist zu vermuten, dass das Ungleichgewicht der beiden Klassen im Datensatz für dieses Resultat mitverantwortlich ist. Gesamthaft erreicht das Modell einen AUC von 0.740 [0.683, 0.796] eine Accuracy von 0.755 [0.715, 0.792] und einen NLL von 0.522. Die Konfidenzintervall für die Metriken sind in der Tabelle 3.3 ersichtlich. Durch den Einsatz des Modell Ensemblings konnten die Accuracy und der NLL verbessert werden.

Tabelle 3.3: Test Performance Metriken für das Modell «Outcome Prognose»

Metrik	Wert	95% Konfidenzintervall
AUC	0.740	von 0.683 bis 0.796
Accuracy	0.755	von 0.715 bis 0.792
Sensitivität	0.420	von 0.323 bis 0.525
Spezifität	0.831	von 0.790 bis 0.865
NLL	0.522	

Am Mittelwert und dem Median der Varianzen in Tabelle 3.4 lässt sich erkennen, dass richtige Klassifizierungen tiefere Varianzen aufweisen als falsche Klassifizierungen. Dem IQR in Tabelle 3.4 und den Histogrammen in Abbildung 3.2 ist zudem zu entnehmen, dass die Varianzen von falschen Klassifizierungen eine grössere Streuung aufweisen als diese der richtigen Klassifizierungen. Das Gleiche gilt für die Klasse 0 (= guter funktioneller Outcome). Für die Klasse 1 (= schlechter funktioneller Outcome) sind die Varianzen ähnlich verteilt und weisen ähnliche Mittelwerte auf. Da sich in dieser Klasse nur 88 Patienten befinden, lässt sich keine Aussage machen.

Tabelle 3.4: Varianzen der Vorhersagen für das Modell «Outcome Prognose»

Klassifikation	Mittelwert	Median	IQR
Richtige Klassifizierungen (n = 361)	0.014	0.005	0.018
Falsche Klassifizierungen (n = 117)	0.026	0.018	0.032
Richtige Klassifizierungen Klasse 0 (n = 324)	0.012	0.004	0.015
Falsche Klassifizierungen Klasse 0 (n = 66)	0.027	0.019	0.031
Richtige Klassifizierungen Klasse 1 (n = 37)	0.024	0.017	0.028
Falsche Klassifizierungen Klasse 1 (n = 51)	0.025	0.017	0.035

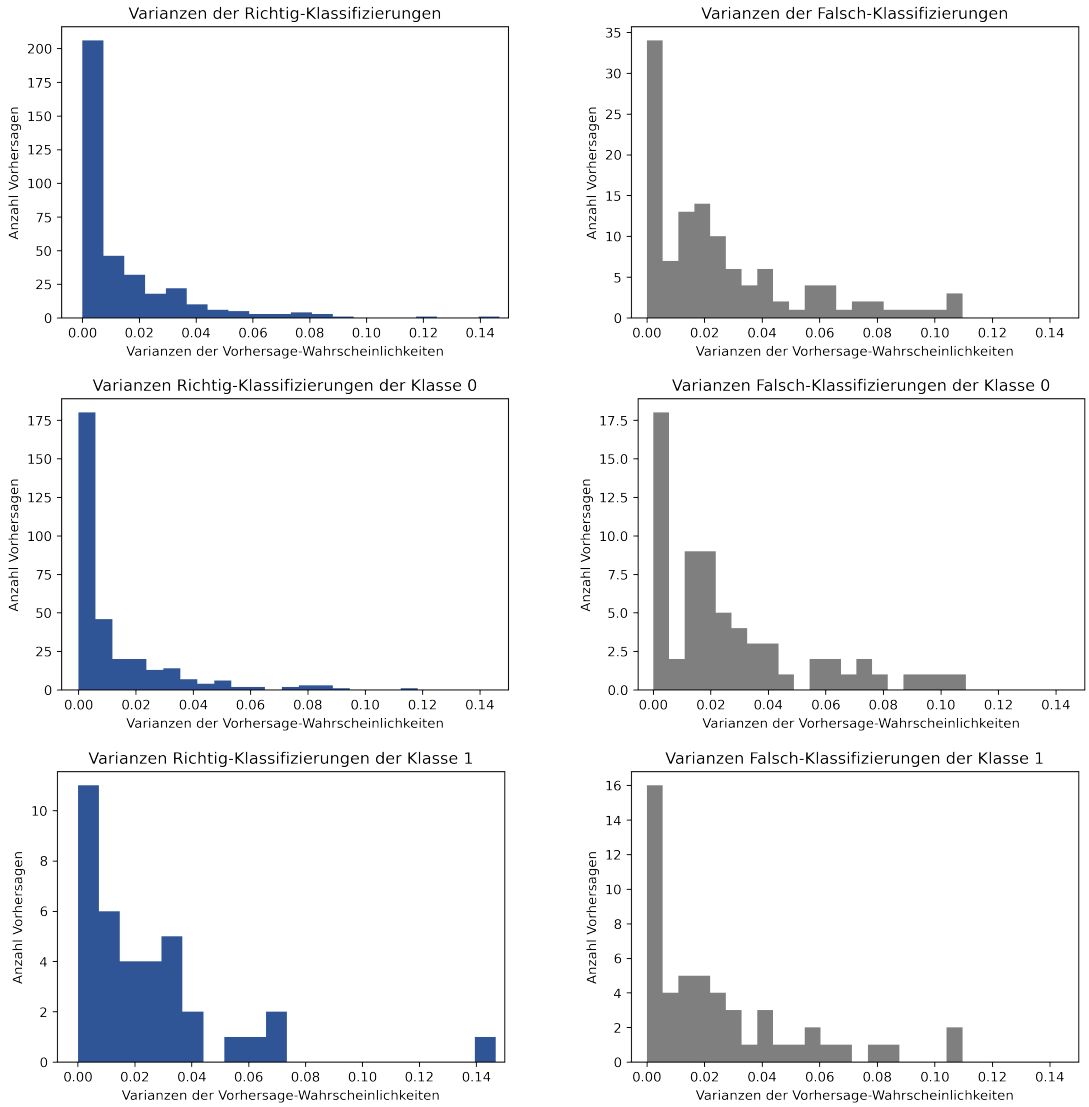

Abbildung 3.2: **Verteilung der Varianzen für das Modell «Outcome Prognose»:** In der ersten Zeile sind die Varianzen über alle richtigen und über alle falschen Klassifizierungen ersichtlich. In der zweiten Zeile werden die Varianzen für die Klasse 0 (=guter funktioneller Outcome) und in der dritten Zeile für die Klasse 1 (=schlechter funktioneller Outcome) dargestellt.

Um die Leistung des Modells zu beurteilen, dient ein Vergleich mit ähnlichen Studien, welche ebenfalls das Ziel haben, den mRS drei Monate nach der Behandlung vorherzusagen. Der Vergleich ist jedoch mit Vorsicht zu geniessen, da in den Studien verschiedene Patientenkollektive verwendet werden. Der Vergleich soll somit lediglich einen Anhaltspunkt zur Modellbewertung geben. Zihni et al. nutzen ein Deep Learning Modell bestehend aus einem 3D CNN zur Verarbeitung von TOF-MRA Bildern und einem MLP (Multilayer Perceptron) zur Verarbeitung von klinischen Daten. Der zur Verfügung stehende Datensatz beinhaltet 316 Patienten. Das 3D CNN allein erreicht einen AUC von 0.68. Kombiniert mit den klinischen Daten kann ein AUC von 0.76 erzielt werden. Ebenfalls ein kombiniertes Modell setzen Bacchi et al. ein, um die Outcome Prognose anhand von CT-Bildern und klinischen Daten von vorzunehmen. Der Datensatz besteht

aus 204 Patienten. Mit dem kombinierten Modell erzielen sie einen AUC von 0.75 und eine Accuracy von 0.74. Hilbert et al. nutzen RFNNs (Receptive Field Neural Networks), um den mRS basierend auf CT-Angiographien vorherzusagen. Dabei stehen Daten von 1'301 Patienten zur Verfügung. Mit dem besten Modell erreichen sie einen AUC von 0.71.

4 Diskussion und Konklusion

In dieser Arbeit wurde der Einsatz von 3D CNNs zur Detektion von Schlaganfällen und zur Prognose des funktionellen Outcomes untersucht. Die Vorteile des 3D CNN sind unter anderem, dass Features aus den TOF-MRA Bildern nicht manuell extrahiert werden müssen und pro Patienten nur ein Label notwendig ist (anstelle eines Labels für jedes Bild der Bildserie). Dies spart Zeit beim Labeln der Daten. Zu den Nachteilen zählt, dass mit den 3D Bildern viel weniger Daten zur Verfügung stehen als wenn man alle 2D Bilder der Bildserien für das Training nutzen könnte. Um einen 2D Ansatz zu testen, würde man jedoch ein Label für jedes 2D Bild benötigen und man müsste sich überlegen, wie man aus den Vorhersagen für die einzelnen 2D Bilder eine Vorhersage für das gesamte 3D Bild erhält. Die beschränkte Verfügbarkeit der 3D Bilder erschwert das Training und die Generalisierbarkeit der Modelle. Als weitere Nachteile sind zu nennen, dass die TOF-MRA Bilder in ihrer Grösse aufgrund der beschränkten Rechenkapazitäten stark reduziert werden müssen und es praktisch keine vortrainierten 3D Transfer Learning Modelle gibt. Der Transfer Learning Ansatz, der in dieser Arbeit getestet wurde, beinhaltet das Erweitern eines 2D Transfer Learning Modells in die dritte Dimension. Die von diesem Modell übernommenen Features sind den Features in den TOF-MRA Bildern vermutlich sind genügend ähnlich. Mit den selbst trainierten 3D CNNs konnten bessere Resultate erzielt werden.

Das Modell zur Detektion von Schlaganfällen eignet sich nicht für den Gebrauch in der Praxis, da es keine verlässliche Vorhersage liefert. Es ist hier zu betonen, dass in der Praxis medizinische Fachpersonen Schlaganfälle nicht anhand von TOF-MRA Bildern detektieren, sondern CT- oder MRT-Modalitäten verwenden (Ringelstein & Forsting 2017, 1-2). Zudem kann auch eine erfahrene Neurologin, nicht auf allen TOF-MRA Bildern die Klasse Schlaganfall erkennen. Es erstaunt somit nicht, dass mit diesem Modell keine besseren Resultate erzielt werden können. Das Modell zur Prognose des funktionellen Outcomes erreicht vergleichbare Resultate wie Modelle in ähnlichen Studien (Zihni et al. 2020, Hilbert et al. 2019, Bacchi et al. 2020). Die 3D TOF-MRA Bildern beinhalten scheinbar Informationen, welche für die Outcome Prognose hilfreich sind. Dies ist nicht erstaunlich, da die TOF-MRA Bilder unter anderem in der klinischen Praxis verwendet werden, um Ursachen für die Symptomatik zu erkennen (Ringelstein & Forsting 2017, 1-2). So deuten gemäss einer erfahrenen Neurologin Gefässverschlüsse oder Engstellen in den TOF-MRA Bilder auf einen schweren Verlauf und auf ein höheres Rezidivrisiko hin. Um die Leistung des Modells zu verbessern, könnte es sinnvoll sein, mehr Trainingsdaten von der Klasse der Patienten mit schlechtem funktionellen Outcome zu sammeln, da diese zurzeit deutlich schlechter prognostiziert wird.

Durch die Methode des Deep Ensemblings konnten bei beiden Modellen stabilere Vorhersagen erzielt werden und sowohl eine bessere Accuracy als auch tieferer NLL erzielt

werden, wie man das Modell ohne Ensembling verwendet. Bei beiden Modellen ist eine höhere Vorhersageunsicherheit bei den falschen als bei den richtigen Klassifizierungen über beide Klassen zu erkennen. Gleiches gilt für die falschen und richtigen Klassifizierungen der jeweils grösseren Klasse. Für die jeweils kleinere Klasse kann aufgrund der wenigen Daten keine verlässliche Aussage gemacht werden. Die Informationen zur Zuverlässigkeit der Vorhersagen sind für die Anwendbarkeit der Modelle in der klinischen Praxis entscheidend (Herzog et al. 2020).

Abschliessend lässt sich sagen, dass es sich lohnen könnte das Modell zur Outcome Prognose weiterzuentwickeln. Dabei sollte das Modell mit weitere Bildmodalitäten und klinische Daten erweitert werden. So haben zum Beispiel Bacchi et al. und Zihni et al. gezeigt, dass mit dem Hinzufügen von klinischen Daten die Outcome Prognose der Modelle verbessert werden kann. Somit könnte dies auch für das hier entwickelte Modell eine vielversprechende Möglichkeit sein.

Literaturverzeichnis

Amin, A., Sharif, M., Gul, N., Yasmin, M. & Shad, S. A. (2020), 'Brain tumor classification based on DWT fusion of MRI sequences using convolutional neural network', *Pattern Recognition Letters* .

Bacchi, S., Zerner, T., Oakden-Rayner, L., Kleinig, T., Patel, S. & Jannes, J. (2020), 'Deep Learning in the Prediction of Ischaemic Stroke Thrombolysis Functional Outcomes: A Pilot Study', *Academic Radiology* .

Banks, J. L. & Marotta, C. A. (2007), 'Outcomes Validity and Reliability of the Modified Rankin Scale: Implications for Stroke Clinical Trials', *Stroke* .

Brown, L. D., Cai, T. T. & DasGupta, A. (2001), 'Interval Estimation for a Binomial Proportion', *Statistical Science* .

Carreira, J. & Zisserman, A. (2018), 'Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset'.

Cassier-Woidasky, A.-K. (2014), 'Pflege von Patienten mit Schlaganfall'.

Chollet, F. (2017), *Deep Learning with Python*, Manning Publications.

Dörn, S. (2017), *Bildverarbeitung*, Springer Berlin Heidelberg, pp. 307–357.

Dürr, O., Sick, B. & Murina, E. (2020), *Probabilistic Deep Learning: With Python, Keras and TensorFlow Probability*, Manning Publications.

Gessert, N., Schlüter, M. & Schlaefer, A. (2018), 'A deep learning approach for pose estimation from volumetric OCT data', *Medical Image Analysis* .

Held, L. & Bové, D. S. (2014), *Applied Statistical Inference*, Springer, Berlin, Heidelberg.

Herzog, L., Murina, E., Dürr, O., Wegener, S. & Sick, B. (2020), 'Integrating uncertainty in deep neural networks for MRI based stroke analysis', *Medical Image Analysis* .

Hesamian, M. H., Jia, W., He, X. & Kennedy, P. (2019), 'Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges', *Journal of Digital Imaging* .

Hilbert, A., Ramos, L., van Os, H., Olabarriaga, S., Tolhuisen, M., Wermer, M., Barros, R., van der Schaaf, I., Dippel, D., Roos, Y., van Zwam, W., Yoo, A., Emmer, B., Lycklama à Nijeholt, G., Zwinderman, A., Strijkers, G., Majoie, C. & Marquering, H. (2019), 'Data-efficient deep learning of radiological image data for outcome prediction after endovascular treatment of patients with acute ischemic stroke', *Computers in Biology and Medicine* .

Hu, M., Sim, K., Zhou, J. H., Jiang, X. & Guan, C. (2020), 'Brain MRI-based 3D Convolutional Neural Networks for Classification of Schizophrenia and Controls'.

Huang, X., Shan, J. & Vaidya, V. (2017), 'Lung nodule detection in CT using 3D convolutional neural networks', *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)* pp. 379–383.

Kessler, H. (2014), 'Das Kompetenzzentrum im regionalen Netzwerk', *UniversitätsSpital Zürich* .

Korolev, S., Safiullin, A., Belyaev, M. & Dodonova, Y. (2017), 'Residual and Plain Convolutional Neural Networks for 3D Brain MRI Classification'.

Lakshminarayanan, B., Pritzel, A. & Blundell, C. (2017), 'Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles'.

Merino, I., Azpiazu, J., Remazeilles, A. & Sierra, B. (2021), '3D Convolutional Neural Networks Initialized from Pretrained 2D Convolutional Neural Networks for Classification of Industrial Parts', *Sensors* .

Nie, D., Cao, X., Gao, Y., Wang, L. & Shen, D. (2016), 'Estimating CT Image from MRI Data Using 3D Fully Convolutional Networks', *Deep Learning and Data Labeling for Medical Applications* .

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B. & Snoek, J. (2019), 'Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift'.

Rachmadi, M. F., Valdés-Hernández, M. D. C., Agan, M. L. F. & Komura, T. (2017), 'Deep Learning vs. Conventional Machine Learning:Pilot Study of WMH Segmentation in Brain MRI with Absence or Mild Vascular Pathology', *Journal of Imaging* .

Ringelstein, A. & Forsting, M. (2017), 'Magnetresonanztomografie beim akuten ischämischen Schlaganfall'.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C. & Müller, M. (2011), 'pROC: an open-source package for R and S+ to analyze and compare ROC curves', *BMC Bioinformatics* .

Schmitt, M. (2016), Vergleich von zwei etablierten Schlaganfallklassifikationen in Bezug auf das funktionelle Outcome nach drei Monaten, PhD thesis, Ludwig-Maximilians-Universität München.

Schwert, H. & Bäzner, H. (2018), 'Ischämischer und hämorrhagischer Schlaganfall', *Klinikarzt* .

Singh, S. P., Wang, L., Gupta, S., Goli, H., Padmanabhan, P. & Gulyás, B. (2020), '3D Deep Learning on Medical Images: A Review', *Sensors (Basel, Switzerland)* .

Stolz, R. (2020), Classification of 3D MRI head scans using pretrained CNNs – comparison, optimization and evaluation of approaches, Master's thesis, Hochschule Landshut.

Swati, Z. N. K., Zhao, Q., Kabir, M., Ali, F., Ali, Z., Ahmed, S. & Lu, J. (2019), 'Brain tumor classification for MR images using transfer learning and fine-tuning', *Computerized Medical Imaging and Graphics* .

Tensorflow (2021), 'Classification on imbalanced data', *Tensorflow Tutorials* .

Vu, H., Kim, H.-C., Jung, M. & Lee, J.-H. (2020), 'fMRI volume classification using a 3D convolutional neural network robust to shifted and scaled neuronal activations', *NeuroImage* .

Wang, H., van Stein, B., Emmerich, M. & Back, T. (2017), 'A new acquisition function for Bayesian optimization based on the moment-generating function'.

World Health Organization (2020), 'Top 10 causes of death'.
URL: *https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death*

Yang, L. & Shami, A. (2020), 'On hyperparameter optimization of machine learning algorithms: Theory and practice', *Neurocomputing* .

Zihni, E., Madai, V., Khalil, A., Galinovic, I., Fiebach, J., Kelleher, J., Frey, D. & Livne, M. (2020), 'Multimodal Fusion Strategies for Outcome Prediction in Stroke', pp. 421–428.

Zunair, H., Rahman, A., Mohammed, N. & Cohen, J. P. (2020), 'Uniformizing Techniques to Process CT scans with 3D CNNs for Tuberculosis Prediction'.