Aus dem Institut für Radiologie
der Medizinischen Fakultät Charité – Universitätsmedizin Berlin

DISSERTATION

Klassifikation von computertomographischen Befundtexten des Thorax anhand von Deep Learning

Classification of Computed Tomography Findings of the Chest Based on Deep Learning

zur Erlangung des akademischen Grades
Doctor medicinae (Dr. med.)

vorgelegt der Medizinischen Fakultät
Charité – Universitätsmedizin Berlin

von

Lina Xu

aus Berlin

Datum der Promotion: 23.03.2024

Inhaltsverzeichnis

Abkürzungsverzeichnis

2D	Zweidimensional
AUC	Area Under the Curve
AWD-LSTM	Average-Stochastic Gradient Descent Weight-Dropped LSTM
BERT	Bidirectional Encoder Representations from Transformers
CT	Computertomographie
DistilBERT	Distilled Bidirectional Encoder Representations from Transformers
F1	F1-Wert
FN	False Negative/Falsch Negativ
FP	False Positive/Falsch Positiv
GB	Gigabyte
HU	Hounsfield Unit
κ	Kappa
KI	Konfidenz-Intervall
KNN	Künstliche Neuronale Netze/Künstliche Neuronale Netzwerke
LAE	Lungenarterienembolie
LSTM	Long Short-Term-Memory
MRT	Magnetresonanztomographie
NLP	Natural Language Processing
PPV	Positiver prädiktiver Wert
R	Recall/Sensitivität
RIS	Radiologisches Informationssystem
RNN	Recurrent Neural Network/Rekurrentes Neuronales Netzwerk
ROC	Receiver Operating Characteristic
SCLC	Small Cell Lung Cancer/Kleinzelliges Bronchialkarzinom
Seq2seq	Sequence-to-Sequence
TN	True Negative/Richtig Negativ
TP	True Positive/Richtig Positiv
ULMFiT	Universal Language Model Fine Tuning
VRAM	Video Random Access Memory

Abbildungsverzeichnis

Tabellenverzeichnis

Zusammenfassung

Hintergrund: Die Computertomographie des Thorax ist eine häufige und bedeutsame Untersuchung der Radiologie. Die Ergebnisse einer CT-Untersuchung werden in einem Befundtext dargestellt, welcher jedoch keiner festen Struktur folgt, und bislang gibt es keine Kategorisierung von Befundtexten, obwohl dies den klinischen Alltag erheblich erleichtern würde. Um strukturierte Daten aus Befundtexten der CT des Thorax zu extrahieren, wurden drei verschiedene Deep-Learning-Modelle für das Natural Language Processing (NLP) entwickelt.

Methoden: Ein annotierter Datensatz bestehend aus 5.950 Befundtexten der CT-Diagnostik des Thorax (inklusive CT-Untersuchungen zur Lungenarterienembolie) wurde für das Training dreier Deep-Learning-Modelle erstellt und die Befundtexte auf das Auftreten 21 verschiedener Befunde untersucht. Für die Klassifikation der Befundtexte mittels Natural Language Processing wurden zum einen ein AWD-LSTM sowie zwei Transformer-Architekturen (BERT und DistilBERT) verwendet. Im Anschluss wurde die Klassifikationsleistungsfähigkeit der Modelle mithilfe der Metriken Genauigkeit, Sensitivität, positivem prädiktiven Wert, F1-Wert sowie AUC beurteilt.

Ergebnisse: Alle drei Modelle erzielten hohe Metriken, welche zwischen den verschiedenen Befunden variierten. Die Genauigkeit erreichte bei allen Befunden >0,96 für das AWD-LSTM, >0,89 für BERT und >0,87 für DistilBERT. Dabei stiegen die Parameter mit zunehmender Prävalenz des jeweiligen Befundes.

Schlussfolgerung: Mithilfe dreier Deep-Learning-Modelle (AWD-LSTM, BERT, DistilBERT) konnten auf Basis eines verhältnismäßig geringen Datensatzes an Texten verschiedene computergestützte Klassifikationssysteme von Befundtexten der CT des Thorax entwickelt werden, welche in der Lage waren, selbstständig die Befunde zu identifizieren. Die Modelle können nun auf sämtliche Befundtexte der CT-Bildgebung des Thorax angewendet und die extrahierten Labels für weiterführende Aufgaben genutzt werden.

Abstract

Background: Computed tomography of the chest is a common and very important examination in radiology. The results of a CT examination are presented in a report text that does not follow a fixed structure and so far, there is no categorization of the findings, although this would make clinical practice easier. In order to extract structured data from diagnostic texts of the chest CT, three different deep learning models for natural language processing (NLP) were developed.

Methods: An annotated data set consisting of 5,950 report texts from CT chest examinations (including CT examinations of pulmonary artery embolism) was created for the training of three deep learning models, and the report texts were screened for the occurrence of 21 different findings. An LSTM and two transformer architectures (BERT and DistilBERT) were used to classify the reports using natural language processing. The classification performance of the models was then assessed using the metrics accuracy, sensitivity, precision, F1 value and AUC.

Results: All three models were able to achieve high metrics, which varied between the different findings. The accuracy for every finding reached >0.96 for the LSTM, >0.89 for BERT and >0.87 for DistilBERT. During the process the parameters increased with higher prevalence of the respective finding.

Conclusion: With the help of three deep learning models (AWD-LSTM, BERT, DistilBERT) and based on a relatively small dataset of reports, various computer-aided classification systems of chest CT reports could be developed, which were able to identify the findings independently. The models may now be applied to all reports of chest CTs, and the extracted labels can be used for further tasks.

1. Einleitung

1.1 Computertomographie in der Bildgebung des Thorax

Bildgebende Verfahren haben einen hohen diagnostischen Stellenwert in der Medizin und werden zur Beantwortung zahlreicher Fragestellungen herangezogen. Die Anzahl und Komplexität der Untersuchungen haben hierbei über die Jahre hinweg beständig zugenommen. Die hohe diagnostische Aussagekraft der Computertomographie (CT) bei diversen Fragestellungen hat dazu geführt, dass sie in den letzten Jahren mit steigender Häufigkeit eingesetzt wurde. So stieg beispielsweise die mittlere Anzahl an CT-Untersuchungen pro Einwohner zwischen 2007 und 2014 um etwa 40% an (1).

Ein modernes CT-Gerät dritter Generation besteht aus einer Röntgenröhre, die einen schmalen Fächerstrahl erzeugt, sowie mehreren Detektoren, welche die nach dem Durchdringen des Körpers abgeschwächte Röntgenstrahlung erfassen. Bei der Untersuchung rotiert die Röntgenquelle um das zu untersuchende Körperteil, wodurch es möglich wird, die Organe und Gewebe in der Körperschicht als dreidimensionale Objekte aufzunehmen. Bei der Rotation der Röntgenröhre um die Patientin oder den Patienten wird kontinuierlich Strahlung ausgesendet, welche teilweise durch die untersuchte Person absorbiert wird. Die der Strahlenquelle gegenüberliegende Detektorreihe registriert anschließend ein lineares Absorptionsprofil des untersuchten Körperteils, aus welchem durch die gefilterte Rückprojektion ein dreidimensionales Bild des Körpers errechnet werden kann. Dabei werden die entstehenden dreidimensionalen Bildpunkte als Voxel bezeichnet (2, 3).

Aus diesem Volumendatensatz der untersuchten Körperregion können anschließend multiplanare Rekonstruktionen errechnet werden. Standardmäßig sind dies Rekonstruktionen als axiale, sagittale und koronare 2D-Schichten der untersuchten Körperregion in unterschiedlicher Schichtdicke. Zudem kann in der CT für jedes Voxel direkt der pixelspezifische Schwächungskoeffizient errechnet werden, dessen lineare Transformation als Hounsfield Unit (HU) bezeichnet wird. Der Schwächungsgrad wird durch einen Grauton visualisiert. Es entstehen auf diese Weise zahlreiche Schwarz-Weiß-Bilder in verschiedenen Graustufen, wobei ein Gewebe mit einer höheren Absorption hyperdens dargestellt wird. Ein Voxel aus Luft hat eine HU von -1000, ein Voxel Wasser eine HU von 0 und ein Voxel aus dichtem, kortikalem Knochen eine HU zwischen +900 bis +2000. Da das menschliche Auge nicht in der Lage ist, mehrere

tausend verschiedene Graustufen zeitgleich zu differenzieren, ist es durch die Bildwiedergabe-Software möglich, eine Fensterung in relevante HU-Bereiche durchzuführen. Somit können beispielsweise gezielt das Lungen-, Weichteil- oder Knochenfenster eingestellt werden (2, 3).

Bei der CT handelt es sich um eine Technik, die der herkömmlichen Röntgen-Technik ähnelt. Bei der klassischen Röntgen-Diagnostik wird das darzustellende Objekt von einer Röntgenquelle durchleuchtet und anschließend auf einem Röntgenfilm abgebildet, wodurch eine Projektion eines dreidimensionalen Körpers auf eine Fläche entsteht. Das führt dazu, dass sich in Strahlrichtung hintereinander liegende Bildteile des durchleuchteten Körpers zwangsläufig überlagern, woraus teilweise eine eingeschränkte Aussagekraft resultiert (4).

Der große Vorteil der CT gegenüber der konventionellen Röntgen-Aufnahme ist daher, dass sie eine genauere Darstellung der Strukturen mit einer besseren Auflösung und einer präziseren anatomischen Lokalisation ermöglicht. Zudem lassen sich mithilfe der CT auch Weichteilstrukturen genauer darstellen, wenngleich hier die Magnetresonanz-tomographie (MRT) überlegen ist. Allerdings ist letztere, bedingt durch die lange Dauer der Untersuchung, nicht immer zugänglich bzw. geeignet. Zudem kann die MRT insbesondere die Lunge aufgrund der relativen Abwesenheit von Wasserstoffmolekülen in der eingeatmeten Luft nicht gut darstellen.

Die computertomographische Untersuchung des Thorax ist ein zentraler Bestandteil der Diagnostik bei unterschiedlichen Indikationen. Einen besonders hohen Stellenwert hat die CT des Thorax in der Tumordiagnostik (5). Die CT wird zum einen routinemäßig bei Bronchialkarzinomen eingesetzt, um die Anzahl, Größenausdehnung sowie lymphogene Metastasierung der Tumore zu bestimmen. In diesem Fall wird es als diagnostisches Mittel zur Ersteinschätzung und Verlaufskontrolle verwendet. Es kann sowohl das Ansprechen auf eine Therapie evaluieren als auch als Kontrolle bei abgeschlossener Therapie zur frühzeitigen Erkennung eines Rezidivs dienen. Zum anderen wird die CT häufig im Rahmen des Stagings anderer Tumorentitäten, wie Lymphomen oder Ösophaguskarzinomen, oder bei der Frage nach pulmonalen Metastasen verwendet (6). Insgesamt lassen sich sämtliche Strukturen des Mediastinums und der Lunge anatomisch genau darstellen und die Dichtemessungen ermöglichen in vielen Fällen eine weitere Einschätzung der Gewebszusammensetzung. Da die CT auch die knöchernen

Strukturen der Wirbelsäule und Rippen abbildet, lassen sich in der Tumordiagnostik ebenso osteoplastische bzw. osteolytische Metastasen in Wirbelsäule und Rippen erfassen. Darüber hinaus wird in klinischen Studien die Low-Dose-CT auch als Screening-Methode für Risikogruppen zur frühzeitigen Detektion eines Bronchialkarzinoms untersucht (7).

Weitere Indikationen sind die Frage nach Infiltraten, vor allem bei komplizierteren Pneumonien, sowie nach Lungengerüstveränderungen wie bei einem Lungenemphysem oder einer Lungenfibrose. Hierbei geht der CT häufig eine konventionelle Röntgenaufnahme voraus, auf der es einen auffälligen Befund gab, der weiter spezifiziert werden soll. Zudem kann die CT in der traumatologischen Diagnostik angewendet werden, um Verletzungen von Herz, Blutgefäßen, Lunge, Rippen und Wirbelsäule zu beurteilen (8).

Eine häufige Anwendung der CT des Thorax stellt zudem die Bildgebung der Gefäße (CT-Angiographie) der Lunge bei der Frage nach einer Lungenarterienembolie dar (9). Bei der CT-Angiographie wird während der Untersuchung ein iodhaltiges Kontrastmittel intravenös verabreicht, welches zu einer temporär höheren Strahlenabsorption des Blutes führt. Hierdurch lassen sich die Intravasalräume hochgenau abgrenzen und intravasale Fremdkörper, wie z.B. Thromben, genauestens darstellen. Durch die hohe Ortsauflösung erlaubt die CT bei dieser Art der Untersuchung auch die Beurteilung kleinster Gefäßabschnitte, sodass darüber hinaus mit dieser Methode auch Gefäßanomalien oder Aneurysmen dargestellt werden können (6).

1.2 Befundstruktur in der Radiologie

1.2.1 Allgemeiner Aufbau eines Befundtextes

Der radiologische Befundtext ist das wichtigste Mittel zur Kommunikation zwischen dem radiologisch ärztlichen Personal und den zuweisenden Ärztinnen und Ärzten. Wenn die diagnostische Bildgebung bei der Untersuchung einer Patientin oder eines Patienten erfolgt ist, so wird in der Regel ein radiologischer Befundtext verfasst, um dem zuweisenden Personal die Ergebnisse mitzuteilen (10). Die radiologische Berichterstattung unterliegt hierbei einem stetigen Wandel, der durch die fortschreitende Entwicklung neuer technischer Innovationen geprägt ist (11). Der Freitextbefund setzte sich über lange Zeit durch und wird mittlerweile überwiegend mithilfe eines Diktiergerätes verfasst, woraufhin der Befund anschließend entweder durch eine Schreibkraft, welche den Text manuell überträgt, oder mit Hilfe automatischer Spracherkennung erstellt wird.

Darüber hinaus existieren mittlerweile auch neue Entwicklungsansätze wie z.B. die standardisierte Nomenklatur RadLex (12).

Trotz Bemühungen, eine universelle Befundstruktur einzuführen, bestehen für den radiologischen Befundtext zu diesem Zeitpunkt keine universellen Regeln, sodass dessen Aufbau von Befund zu Befund variieren kann. Dennoch gibt es in der Regel Richtlinien, die den Radiologinnen und Radiologen Hinweise zum Inhalt und Format des Befundtextes geben, wie beispielsweise die Leitlinie der European Society of Radiology (10).

Insgesamt lässt sich die Struktur eines radiologischen Befundtextes trotz fehlender Standardisierung zumeist in mehrere Bereiche unterteilen, die im Folgenden näher beleuchtet werden.

Klinik, Fragestellung, rechtfertigende Indikation

Hier wird beschrieben, welche medizinische Fragestellung der diagnostischen Untersuchung zugrunde liegt. Es werden die aktuelle klinische Entscheidung der Patientin oder des Patienten sowie relevante Details der medizinischen Vorgeschichte geschildert, welche die radiologische Untersuchung erforderlich machen. Das können beispielsweise wichtige Vorerkrankungen oder auch aktuelle Laborwerte sein. Der Abschnitt sollte so kompakt wie möglich sein, aber dennoch alle relevanten Informationen enthalten. Dieser Abschnitt ist der einzige im radiologischen Befund, der durch das anfordernde medizinisch ärztliche Personal erstellt und durch das radiologische Personal nur in den Befundtext übernommen wird.

Allerdings muss vor jeder computertomographischen Untersuchung eine Prüfung der rechtfertigenden Indikation erfolgen. Dabei entscheidet eine Ärztin oder ein Arzt mit der erforderlichen Fachkunde im Strahlenschutz darüber, ob und in welcher Form die Untersuchung an der Patientin oder dem Patienten angewendet werden kann. Bei der Stellung einer rechtfertigenden Indikation muss daher festgestellt werden, dass der gesundheitliche Nutzen der Verwendung ionisierender Strahlung verglichen mit dem Strahlungsrisiko überwiegt (vgl. § 83 Absatz 3 Strahlenschutzgesetz).

Aufklärung und Einwilligung

Hierbei handelt es sich um einen kurzen Textblock, der bei der Gabe von Kontrastmittel eingefügt werden kann. Bei einer CT mit Kontrastmittel ist es notwendig, Patientinnen

und Patienten über den Untersuchungsablauf inklusive möglicher Risiken der Untersuchung aufzuklären. Daher werden die Erhebung der Risikoanamnese, die erfolgte mündliche und schriftliche Aufklärung über die Kontrastmittelapplikation sowie die vorhandene schriftliche Einwilligung der Patientin oder des Patienten dokumentiert.

Technik

Dieser Abschnitt beschreibt die verschiedenen technischen Details der Untersuchung, wobei Elemente der Untersuchung, welche vom Standard abweichen, wie z.B. zusätzlich durchgeführte Sequenzen, explizit erwähnt werden sollten.

Im Falle der CT wird dabei zunächst beschrieben, welche Art der CT angewendet wurde (native Aufnahme, Spiral-CT, Mehrzeilenspiraltechnik), welches Kontrastmittel in welcher Dosis verwendet wurde und in welcher Kontrastmittelphase (arteriell, venös, Spätaufnahme) die Bildakquise erfolgte. Außerdem werden die anschließend durchgeführten Rekonstruktionen sowie die Schichtdicke beschrieben.

Verschiedene Kontrastmittelsequenzen eignen sich unterschiedlich gut für die Darstellung verschiedener Gewebe. Durch den Technik-Abschnitt können fachkundige Leserinnen und Leser des Befundes somit Limitationen der Untersuchungstechnik frühzeitig erkennen und die Ergebnisse des Befundes hierdurch besser interpretieren bzw. eine angemessene Ergänzungsuntersuchung anordnen.

Befund

Im Befund werden die wesentlichen, auf den radiologischen Bildern erkannten Strukturen beschrieben, wobei der Text stets deskriptiv und ohne Wertung verfasst werden sollte.

Der Abschnitt sollte eine systematische und verständliche Darstellung aller Pathologien beinhalten. Die Beobachtungen werden unter Verwendung radiologischer Terminologie dargestellt und bestenfalls so präzise wie möglich beschrieben, wobei unspezifische Begriffe zu vermeiden sind. Bei der Beschreibung der Pathologie ist auf eine möglichst spezifische Charakterisierung zu achten. Im Falle der CT ist es hier beispielsweise wichtig, auf die Gewebsdichte (Densität) sowie auf gegebene spezifische Eigenschaften wie Kalzifikationen oder Kavitationen einzugehen. Zudem ist es von großer Bedeutung, die anatomische Lokalisation der pathologischen Struktur sowie deren Bezug zu benachbarten Strukturen klar darzustellen. Hier ist es für das zuweisende ärztliche Personal hilfreich, die relevanten Bilder, welche die Pathologie am besten präsentieren,

zu benennen (z.B. „Serie 2 Bild 10“). So kann der Befund besser nachvollzogen werden. Auch relevante negative Befunde sollten beschrieben werden. Wichtig ist darüber hinaus, alle neuaufgetretenen Strukturen zu erwähnen und zu analysieren. Wenn Voraufnahmen vorhanden sind, sollte der aktuelle Befund mit den vergangenen Untersuchungen verglichen und dabei das Datum der Voraufnahme gelistet werden. Sind keine Voraufnahmen vorhanden, ist es sinnvoll auch dies explizit zu erwähnen (10).

Beurteilung

In der Beurteilung werden die wesentlichen Befunde zusammengefasst und interpretiert, wobei alle Details der Bildgebung mit den relevanten klinischen Informationen und Laborwerten verknüpft werden, um einen Gesamteindruck zu formulieren. Das Ziel ist, eine möglichst präzise Diagnose oder gegebenenfalls verschiedene Differential-diagnosen zu stellen, wobei diese in absteigender Wahrscheinlichkeit genannt werden sollten. In diesen Fällen können die Hinweise, welche für bzw. gegen die entsprechende Differentialdiagnose sprechen, erklärt werden. Die Beurteilung sollte Bezug auf die initiale Fragestellung nehmen und diese möglichst beantworten. Neuaufgetretene Befunde sind als relevant oder nicht-signifikant einzustufen und klinisch relevante pathologische Befunde klar darzustellen (10).

Besondere Anmerkung/Hinweise

Hier werden optional spezielle Informationen geschildert. Beispielsweise werden in diesem Abschnitt notwendige Prämedikationen wie Sedativa oder weitere Medikamente, welche durch Grunderkrankungen von Patientinnen und Patienten notwendig sind, beschrieben. Ein Beispiel wären hierbei Personen mit einer latenten Hyperthyreose, die sich aufgrund des Einsatzes von jodhaltigem Kontrastmittel einer weiteren medikamentösen Behandlung unterziehen sollten. Zudem wird gelegentlich erläutert, warum eine Untersuchung lediglich mit einer bestimmten Methodik durchgeführt werden konnte (z.B. „Aufgrund fehlender Schilddrüsen- und Nierenwerte erfolgte nur eine native Aufnhame.“). Auch allergische Reaktionen auf das Kontrastmittel sollten selbstverständlich dokumentiert werden.

Darüber hinaus wird gegebenenfalls beschrieben, ob und mit wem die Ergebnisse der Untersuchung kommuniziert wurden (z.B. „Eine telefonische Befundübermittlung an die Station ist erfolgt“). Sollte eine Wiedervorstellung der Patientin oder des Patienten erforderlich sein, wird diese mit Zeitangabe notiert.

Abweichungen

In einzelnen Fällen wird von der standardmäßigen Befundstruktur abgewichen, was insbesondere bei Notfällen oder im Bereitschafts-/Nachtdienst vorkommt. In diesen Fällen folgen auf die klinische Fragestellung lediglich Befund und Beurteilung, welche dann auch häufig komprimierter formuliert werden, sodass zumeist nur auf die Kernpunkte und Pathologien eingegangen wird. Dabei wird der Schwerpunkt auf die Beantwortung der Fragestellung gelegt.

1.2.2 Befundtext einer Computertomographie des Thorax

Befundtexte der CT des Thorax folgen wie auch alle anderen CT-Untersuchungen der oben erwähnten Befundstruktur. Die genaue Struktur von Befund und Beurteilung bei der Thorax-Computertomographie variiert hierbei zwischen den radiologischen Instituten und Kliniken, aber es gibt einige Kernpunkte, die im Wesentlichen thematisiert werden:

Herz: Ein zentraler Kernpunkt des CT-Thorax-Befundtextes sind das Herz sowie das Perikard. Es wird auf die Größe des Herzens, das Vorhandensein und gegebenenfalls das Ausmaß eines Perikardergusses, die Koronararterien sowie die Diameter der großen abführenden Gefäße (Aorta, Truncus pulmonalis) eingegangen.

Lymphknoten: Ein weiterer Kernbestandteil sind die abgebildeten Lymphknoten, welche sich typischerweise in drei Hauptgruppen unterteilen lassen – die mediastinalen, hilären und axillären Lymphknoten, bei denen jeweils die Anzahl, Größe und Konfiguration beschrieben werden können. In der Beurteilung kann bei einer Lymphadenopathie eingeschätzt werden, ob diese eher reaktiv oder malignitätssuspekt bedingt ist.

Mediastinum: In Bezug auf das Mediastinum wird auf Auffälligkeiten bezüglich der Schilddrüse und Halsweichteile, des Thymus, der Trachea sowie des Ösophagus geachtet.

Lungenarterienembolie und Aortendissektion: Bei einer CT-Angiographie wird häufig die pulmonalarterielle Strombahn verfolgt und auf mögliche Kontrastmittelaussparungen bis auf die Ebene der subsegmentalen Lungenarterien untersucht. Auch die Aorta kann durch die CT-Angiographie mit Hinblick auf Aneurysmen oder Dissektionen betrachtet werden.

Belüftung: Dieser Abschnitt geht auf die Ventilation der Lunge und dementsprechend auf die Darstellung von Belüftungsstörungen wie Dystelektasen oder Atelektasen ein.

Pleura: Die Pleura ist ein weiterer wichtiger Kernpunkt des Befundes, wobei diesbezüglich sowohl auf die Pleura selbst mit etwaigen Verdickungen, Plaques oder Karzinosen als auch auf den Pleuraspalt eingegangen wird. In Bezug auf den Pleuraspalt werden Lokalisation und Ausdehnung eines Pleuraergusses oder Pneumothorax beschrieben. In der Beurteilung kann eine Empfehlung zur Punktion bzw. zum Anlegen einer Thoraxdrainage gestellt werden.

Rundherde: Hierunter werden Noduli der Lunge hinsichtlich ihrer Lokalisation und Größe sowie bei Voraufnahmen mit Dynamik dargeboten. Es kann anschließend die Dignität beurteilt werden, also ob diese beispielsweise Granulomen entsprechen, postspezifisch oder auch malignitätssuspekt sind. Bei Tumoren kann der Verlauf beobachtet und das Ansprechen auf die Therapie evaluiert werden.

Infiltrate: In diesem Abschnitt werden die Ausdehnung sowie Konfiguration vorhandener Infiltrate geschildert. In der Beurteilung kann zwischen typischen und atypischen Infiltraten unterschieden werden und in Kombination mit der Klinik der Patientin oder des Patienten ein Rückschluss auf das Erregerspektrum gezogen werden.

Lungenstruktur: Im Befund wird ebenfalls auf die Dichte sowie die Zeichnung des Lungenparenchyms und ein mögliches Emphysem sowie etwaige Fibrosen oder Bronchiektasen eingegangen. Deren Progredienz oder Regredienz kann im Verlauf beurteilt werden.

Rippen- und Wirbelsäule: Hier werden in Bezug auf die ossären Strukturen Knochendichte sowie mögliche Frakturen beschrieben. Zudem wird auch auf die Bandscheiben und ligamentären Strukturen eingegangen. Es ist beispielsweise eine Beurteilung darüber möglich, ob die Knochendichte und gegebenenfalls die degenerativen Veränderungen altersentsprechend sind.

Weichteilmantel: Hier können Auffälligkeiten wie ein Weichteilemphysem, Gynäkomastie, Kachexie oder Anasarka dargestellt werden.

Oberbauch/Abdomen: Bei einer CT des Thorax werden häufig die Organe des Oberbauchs mitangeschnitten, sodass auch Auffälligkeiten der dargestellten Strukturen von Leber, Pankreas, Milz und Nieren sowie derer Gefäße erkannt werden können.

Fremdmaterialien: Neben der Beschreibung der untersuchten Körperregion werden auch nicht körpereigene Materialien beschrieben, z.B. Thoraxdrainagen, zentrale

Venenkatheter, Portsysteme, Magensonden, Endotrachealtuben oder Herzschrittmacher. In der Beurteilung sollten Fehllagen erwähnt und es kann eine Empfehlung zur Repositionierung gestellt werden.

1.3 Einführung in die radiologische Terminologie

Ein radiologischer Befundtext wird in der Regel in Freisprache verfasst, unterscheidet sich aber dennoch in verschiedenen Gesichtspunkten von einem Alltagstext. Das liegt größtenteils an der Verwendung radiologischer Terminologie und medizinischer Fachsprache, welche einige Besonderheiten aufweisen.

Zunächst fällt auf, dass ein deutscher Befundtext auch viele fremdsprachige Wörter enthält. Das hat überwiegend historische Gründe, da der Beginn der medizinischen Fachsprache auf rund 500 v.Chr. zu datieren ist. Die ältesten schriftlichen medizinischen Dokumente der westlichen Welt gehen hierbei auf Hippokrates zurück, der diese auf Altgriechisch verfasste. Aus diesem Grund leiten sich viele der heute gebräuchlichen medizinischen Termini, sowohl was Symptome als auch was anatomische Strukturen anbelangt, aus dieser Sprache ab (Beispiele: *Dyspnoea* – heute Fachbegriff Dyspnoe für Luftnot, *Pyloros* – heute Pylorus als gebräuchlichen Begriff für eine anatomische Struktur des Magens) (13).

Seit etwa 100 n.Chr. wurde die lateinische Sprache zunehmend wichtiger für die Medizin, sodass sich die Bezeichnungen der meisten anatomischen Leitstrukturen aus dieser Sprache ableiten (Beispiele: *Costa* = Rippe, *Cor* = Herz). Heutzutage werden alle Artikel der wichtigsten medizinischen Fachzeitschriften auf Englisch verfasst, weshalb es einige englische Begriffe gibt wie z.B. *Bypass Operation, Base Excess* oder *Coiling,* die auch in nicht-englischsprachigen Ländern verwendet werden (13). Dies bezieht sich auch insbesondere auf spezifische radiologische Termini wie *Tree-in-Bud-Sign* oder *Honeycombing*. Des Weiteren werden häufig im medizinischen Bereich geläufige Abkürzungen verwendet wie beispielsweise *LAE* oder *SCLC*. Diese leiten sich teilweise von dem deutschen (Lungenarterienembolie – *LAE*) oder aber von dem englischen (Small Cell Lung Cancer – *SCLC*) Fachbegriff ab.

Eine Besonderheit von Befundtexten speziell im Bereich der Radiologie ist, dass sie sehr bildlich geschrieben sind. So werden zum einen Helligkeitsstufen beschrieben, wodurch je nach Untersuchungsart beispielsweise eine Aussage zur Densität (Röntgenverfahren wie bei einer CT), Intensität (MRT) oder Echogenität (Ultraschall) getroffen werden kann.

Zum anderen werden häufig bildhafte Adjektive wie „konfluierend“, „fleckförmig“ oder „infiltrierend“ verwendet, um Strukturen zu charakterisieren.

1.4 Möglichkeiten im Natural Language Processing

1.4.1 Informationsextraktion durch Natural Language Processing

Im Fachbereich der Radiologie entsteht täglich eine große Anzahl an digitalen Befundtexten. Die Verwendung der darin enthaltenen Daten würde viele Vorteile für die Gesundheitsforschung oder das Dokumentationswesen mit sich bringen, allerdings müssen diese Daten erst herausgefiltert werden. Obwohl die Befundtexte für die Dokumentation der diagnostischen Bildgebung gespeichert werden, wird für die volle Ausschöpfung deren Potentials eine Form der automatisierten Informationsextraktion durch Nutzung von Methoden des Natural Language Processing (NLP) benötigt (14).

Natural Language Processing bzw. natürliche Sprachprozessierung beschäftigt sich mit der Interaktion zwischen Computern und der menschlichen Sprache. Es werden verschiedene Methoden aus den Sprachwissenschaften, der modernen Informatik sowie der künstlichen Intelligenz kombiniert, um große Datenmengen der natürlichen Sprache zu analysieren. Ziel ist es, Informationen aus Freitexten in ein standardisiertes, strukturiertes Format zu überführen und damit in maschinenlesbare Informationen umzuwandeln. Das ist die Voraussetzung, um den Befundtext anschließend für verschiedene Analysen nutzen zu können. Mögliche Anwendungsbereiche sind hierbei u.a. die Beurteilung der Befundkonsistenz, die Erhebung epidemiologischer Daten wie beispielsweise der Inzidenz von Pathologien sowie die Feststellung der Ursachen von Krankheiten oder Nebenwirkungen von Therapien (15, 16).

NLP muss die Sprache erkennen, analysieren und den Sinn erfassen, um den Inhalt für die weitere Verwendung korrekt aufzubereiten. Dafür ist neben dem Verständnis der einzelnen Textbestandteile auch die Erfassung von Zusammenhängen innerhalb der Texte notwendig. Die Komplexität und Vielschichtigkeit der menschlichen Sprache stellt insgesamt eine große Herausforderung für das NLP dar (17).

Der Ablauf des NLP folgt dabei meistens einem gewissen Grundschema. Der erste Schritt ist die Informationsextraktion. Hierbei werden Freitexte zunächst in Abschnitte segmentiert, in einzelne Sätze geteilt und es erfolgt eine Tokenisierung des Textes. Anschließend kann noch eine Überprüfung der Rechtschreibung, der Syntax oder der

Semantik vorgenommen werden. Man erhält eine Sammlung von Informationen, welche im nächsten Schritt prozessiert und für eine spezielle Aufgabe angewendet werden. Im Anschluss kann die Performance anhand eines annotierten Datensatzes evaluiert werden, um im letzten Schritt den Algorithmus für seinen Verwendungszweck zu implementieren (18). Eine zentrale Rolle im Ablauf des NLP bildet dabei die Erstellung eines annotierten bzw. gelabelten Datensatzes durch eine menschliche Fachkraft. Die Annotation stellt eine deskriptive oder analytische Anmerkung dar, die sich auf den eigentlichen Text bezieht.

Die Methodik des NLP entwickelt sich stetig weiter und es gibt mittlerweile viele verschiedenen Formen der Prozessierung. Da Computer anders als Menschen nicht auf Erfahrungen zum besseren Verständnis von Sprache zurückgreifen können, werden zunehmend Algorithmen und Techniken der künstlichen Intelligenz und insbesondere des Deep Learning angewendet.

Im Folgenden wird deshalb zunächst die klassische regelbasierte Methodik des NLP erläutert und anschließend auf zwei Methoden des Deep Learning eingegangen: die rekurrenten neuronalen Netzwerke (RNN) sowie die relativ neue Transformer-Architektur.

1.4.2 Regelbasierte Methoden des NLP

Regelbasierte Algorithmen stellen die frühesten Methoden des NLP dar. Bei dieser Form werden Regeln erstellt, welche auf linguistischen Strukturen basieren, die der menschlichen Grammatik ähneln. Das impliziert typischerweise, dass ein Mensch in den Prozess der schrittweisen Entwicklung und Verbesserung des Programms involviert ist, welcher in der Lage ist, mitzuverfolgen, ob das System die Aufgaben durchführen kann, und gegebenenfalls Fehler auszubessern. Die grammatikalischen Regeln können flexibel erstellt und aktualisiert werden, sodass kein aufwändiges Training benötigt wird. Allerdings erfordert diese Art des NLP die manuelle Eingabe jeder einzelnen Regel durch erfahrene Linguistinnen und Linguisten oder Ingenieurinnen und Ingenieure. So werden hierbei Techniken wie „Word Stemming“ und „Lemmatization“ angewendet, was bedeutet, dass Worte gruppiert werden können, wenn sie denselben Wortstamm besitzen. Das „Stop Word Removal“ ermöglicht darüber hinaus, dass relevante Wörter von irrelevanten Füllwörtern unterschieden werden können (19).

1.4.3 Rekurrente Neuronale Netzwerke (RNN)

Den regelbasierten Algorithmen gegenüber stehen Methoden des Deep Learning wie die rekurrenten neuronalen Netzwerke (RNN). Deep Learning stellt einen Teilbereich des maschinellen Lernens dar, bei dem künstliche neuronale Netze (KNN) mit mehreren Zwischenschichten (Hidden Layers) zwischen Eingabe- und Ausgabeschicht eingesetzt werden, wodurch sich eine umfangreiche innere Struktur ausbildet. Die Besonderheit des Deep Learning ist, dass es der Maschine ermöglicht, eine Technik selbstständig zu entwickeln und ohne menschliche Anweisung ihre Fähigkeiten zu verbessern. Das wird erreicht, indem aus vorhandenen Informationen und Daten Muster extrahiert und erlernt werden. Die gewonnenen Erkenntnisse lassen sich wiederum in Korrelation zu anderen Daten setzen und können somit in einem weiteren Kontext verknüpft werden. Anschließend ist die Maschine in der Lage, Entscheidungen auf Grundlage der Verknüpfungen zu treffen. Im Gegensatz zu den regelbasierten Algorithmen, bei denen der Mensch jede Regel selbst erstellen muss, sucht sich der Computer beim maschinellen Lernen die Regeln selbst (20).

Der Aufbau der KNN orientiert sich hierbei am Aufbau des menschlichen Gehirns, wobei die Neuronen verschiedene Verknüpfungen besitzen und somit netzartig miteinander verbunden sind. Die erste Schicht des KNNs verarbeitet eine Eingabe von Rohdaten und wird als sichtbare Eingabeschicht bezeichnet, da die Dateneingabe Variablen enthält, die der Beobachtung zugänglich sind. Im Anschluss überträgt die erste Schicht die Informationen an die zweite Schicht, welche wiederum die Informationen verarbeitet und an die nächste Schicht weitergibt. Nach diesem Prinzip werden die Daten weiter prozessiert und es entstehen zahlreiche versteckte Schichten (Hidden Layers). Innerhalb der zwischen Ein- und Ausgabeschicht entstehenden versteckten Schichten erhalten die Informationsverknüpfungen durch ständiges Hinterfragen der Entscheidungen und Konzepte bestimmte Gewichtungen, die sich durch Bestätigung bzw. Revision der Entscheidungen verändern. Es entstehen zunehmend weitere Zwischenschichten und Verknüpfungen, und schließlich wird das Ergebnis in der sichtbaren letzten Schicht, der Ausgabeschicht, dargeboten (21).

Klassische Feed-Forward-Netzwerke funktionieren nur in eine Richtung, nämlich von der Eingabe- zur Ausgabeschicht. Die Besonderheit von RNN innerhalb der KNN besteht darin, dass im Gegensatz zu Feed-Forward-Netzwerken Neuronen einer Schicht des

RNNs auch zu sich selbst oder Neuronen vorausgegangener Schichten kommunizieren können. Bei der Verarbeitung sequenzieller Eingaben ist es hierdurch möglich, Informationen vorausgegangener Eingaben einer Sequenz bei dem Prozessieren späterer Eingaben der Sequenz zu nutzen, ähnlich einer Art inneren Kurzzeitgedächtnisses. Diese Eigenschaft macht RNN besonders nützlich für die Anwendung auf sequenziellen Daten wie Texten oder Zeitserien (21).

Ein klassisches Beispiel für die Anwendung von RNN sind die Sequence-to-Sequence-Modelle (seq2seq), welche eine bestimmte Sequenz in eine andere konvertieren und beispielsweise für die Übersetzung eines Satzes in eine andere Sprache verwendet werden können. Dieses Modell findet in vielen Bereichen Anwendung, wie beispielsweise der maschinellen Übersetzung, der Textzusammenfassung oder auch der Spracherkennung (22).

Hierbei gibt es den Encoder und den Decoder, welche beide RNN sind. Die Besonderheit von RNN gegenüber klassischen KNN ist, dass sie jedes Wort des Satzes als einen eigenständigen Input betrachten, der zu einer bestimmten Zeit auftritt. Im zu prozessierenden Satz wird jedes Wort in einen Input-Vektor umgewandelt. Die erste versteckte Schicht nimmt als Input den Vektor, der das erste Wort im Satz codiert. Der Output wird auch als Kontext-Vektor bezeichnet und dient hierbei als Input für die zweite versteckte Schicht, welche diesen gemeinsam mit dem Input des Vektors, der das zweite Wort repräsentiert, miteinander verrechnet und anschließend an die dritte versteckte Schicht übermittelt. Diese folgt demselben Prinzip und verknüpft die Information mit dem Input des dritten Wortes. In jedem Zeitschritt im Encoder nimmt also das RNN einen Wortvektor der Input-Sequenz sowie eine Information vom vorherigen Zeitschritt, wobei dieses bei jedem Schritt aktualisiert wird. Der Kontext-Vektor wird schließlich an den Decoder weitergegeben, welcher die Zielsequenz generiert (23).

Eine Verbesserung der RNN stellen die Modelle mit LSTM dar. Der Term LSTM steht hierbei für Long Short-Term-Memory und bedeutet übersetzt langes Kurzzeitgedächtnis. Bei herkömmlichen RNN besteht das Problem, dass bei der Verarbeitung einer Sequenz Informationen aus dem Anfang der Sequenz mit Voranschreiten der Verarbeitung der Sequenz zunehmend weniger genutzt werden können und somit vergessen werden. Durch das Anwenden verschiedener Informations-Gates im LSTM-Modul ist es hingegen

möglich, Informationen länger zu konservieren; sie funktionieren wie eine Art längeres Kurzzeitgedächtnis (24).

1.4.4 Transformer

Eine neue Architektur innerhalb der KNN stellen die Transformer dar. Transformer setzen sich aus einer Encoding- und einer Decoding-Komponente zusammen, wobei beide Komponenten aus Stacks von Encodern bzw. Decodern bestehen.

Die Encoder besitzen hierbei alle dieselbe Struktur, welche sich in zwei Teilschichten unterteilen lässt: eine Self-Attention-Schicht und ein Feed-Forward-Netzwerk. Der Input des Encoders durchläuft hierbei zuerst die Self-Attention-Schicht, welche ihm dabei hilft, die um das betreffende Wort herumliegenden Wörter zu analysieren und wichtigen Wörtern der Sequenz eine stärkere Gewichtung (Aufmerksamkeit) zukommen zu lassen. Der Output der Self-Attention-Schicht wird anschließend dem Feed-Forward-Netzwerk zugeführt, wobei dasselbe Netzwerk unabhängig auf jede Position angewendet wird. Hierbei wird innerhalb eines Satzes jedes Wort in einen Vektor umgewandelt, welcher anschließend als Input dient (25).

Der Decoder ist ähnlich aufgebaut. Er enthält dieselben beiden Schichten, wobei sich zwischen den beiden zusätzlich noch eine Attention-Schicht befindet, die dem Decoder dabei hilft, sich auf die relevanten Teile des Input-Satzes zu fokussieren (25).

Die Besonderheit der Modellarchitektur des Transformers ist, dass er einen großen Schwerpunkt auf den Attention-Mechanismus legt, um globale Abhängigkeiten zwischen Input und Output herzustellen. Bei diesem Aufmerksamkeitsmechanismus werden zwei Sätze in eine Matrix umgewandelt; dabei formen die Wörter des einen Satzes die Reihen und die des anderen die Spalten. Auf diese Art und Weise werden relevante Kontexte zwischen den beiden Sätzen hergestellt. Das funktioniert auch, wenn man zweimal denselben Satz verwendet, um zu verstehen, wie einzelne Satzteile in Bezug zueinanderstehen. Dies wird auch als Self-Attention bezeichnet (25).

Der wesentliche Unterschied zu RNN besteht daher darin, dass der Transformer nicht wie die RNN bei der Darstellung von Input und Output auf die Wiederholung von Sequenzen innerhalb der versteckten Schichten ausgerichtet ist, sondern ausschließlich den Aufmerksamkeitsmechanismus verwendet, um Abhängigkeiten herzustellen. Hinzu kommt, dass der Transformer auf diese Weise nicht nur Verknüpfungen zwischen

benachbarten Wörtern erstellt, sondern auch Abhängigkeiten von weiter entfernt voneinander vorkommenden Worten erkennt. Transformer sind zudem in der Lage, gezielt Aufmerksamkeit auf spezielle Wörter zu richten, unabhängig von der Position des Wortes im Text. Dies steht ebenfalls im Gegensatz zu RNN, bei welchen die Verbindung von Wörtern am Anfang einer Textsequenz zu Wörtern am Ende einer Textsequenz nur schwach ist. Zudem können hierdurch Wörter gezielter in Verbindung gebracht werden. Fängt ein Folgesatz beispielsweise mit einem Personalpronomen an, kann dies zu dem Subjekt des vorangegangenen Satzes in Beziehung gesetzt werden (25).

Durch den Verzicht auf die wiederholten Sequenz-Prozessierungen ermöglicht der Transformer somit eine Parallelisierung verschiedener Aufgaben, welche den RNN nicht möglich ist (25).

Mittlerweile existieren verschiedene auf der Transformer-Architektur entwickelte Modelle wie beispielsweise das von Google entwickelte Modell BERT (Bidirectional Encoder Representations from Transformers) (26).

1.5 Herausforderungen des NLP in medizinischen Texten

NLP wird mittlerweile in unterschiedlichen Lebensbereichen auf vielfältige Arten genutzt. Sehr bekannte Beispiele sind unter anderem die automatisierte Übersetzung oder die Spracherkennung. Auch im medizinischen Sektor wird NLP zunehmend eingesetzt. Hierbei kann die Anwendung von NLP auf medizinische Texte eine große Bereicherung für die Verbesserung des Gesundheitssystems darstellen. Allerdings bestehen diverse Herausforderungen, wobei die Hauptschwierigkeit aus der hohen Komplexität der medizinischen Sprache resultiert. Während medizinisches Personal ein natürliches Sprachverständnis besitzt, fehlt dieses den Computern, was sich in einigen grundsätzlichen und zu lösenden Problemen präsentiert (20).

Zunächst werden medizinische bzw. radiologische Texte in Fachsprache geschrieben, sodass für das korrekte Verständnis der Fachbegriffe und der Interpretation der Texte eine Kenntnis der Terminologie benötigt wird (18, 27).

Die fehlende Standardisierung des Vokabulars stellt eine große Hürde für die automatisierte Datenextraktion dar, da beispielsweise für ein bestimmtes Phänomen viele verschiedene Synonyme verwendet werden können. Des Weiteren entsteht das Problem der sogenannten „Word-Sense-Disambiguation“, was das Beinhalten verschiedener

Bedeutungen für ein und dasselbe Wort umschreibt. In der Medizin/Radiologie betrifft dies häufig Akronyme. So steht im Englischen die Abkürzung CT beispielsweise sowohl für das Wort „Chest Tube“ (Thoraxdrainage) als auch für den Begriff „Computed Tomography“ (Computertomographie) (19).

Auch die Semantik stellt eine Schwierigkeit dar, da Computern das natürliche Wort- und Textverständnis des Menschen fehlt. So muss NLP in der Lage sein, einige linguistische Komponenten erfassen zu können. Darunter fallen beispielsweise negative Aussagen oder bedingte Aussagen. Diese können häufig zu Missinterpretationen führen, wie die folgenden Beispiele zeigen:

„Die Patientin verneint jegliche Müdigkeit, Appetitlosigkeit, Kopfschmerzen oder Übelkeit.“

Bei dieser Aussage muss das NLP in der Lage sein, die Verneinung mit allen vier Symptomen in Verbindung zu bringen.

„Es gab keinen Hinweis auf einen Bandscheibenprolaps.“

Das NLP muss in diesem Fall das Wort „keinen“ korrekt auf das Wort „Hinweis“ und nicht auf „Bandscheibenprolaps“ beziehen.

Insgesamt besitzen verschiedene medizinische Texte häufig eine hohe Variabilität in Länge, Aufbau und Schreibstil, was es schwieriger für NLP-Modelle macht, sie korrekt zu erfassen. Klinische Daten wie radiologische Befunde erfordern zudem die Erkennung von Interaktionen weiter voneinander entfernter Worte innerhalb des Textes. Das stellt eine große Herausforderung für traditionelle NLP-Techniken dar, die lediglich den Zusammenhang von nahestehenden Wörtern gut erkennen, während die Beziehung mit zunehmendem Abstand der Wörter abnimmt.

Viele NLP-Modelle, wie regelbasierte Algorithmen oder RNN, wurden bereits auf radiologische Befundtexte angewendet (18, 28). Eine Schwierigkeit in der Entwicklung der Modelle besteht allerdings darin, dass ein sehr zeitaufwendiges Labeln der Befundtexte durch Expertinnen und Experten erforderlich ist.

Eine weitere Hürde stellt der eingeschränkte Zugang zu den Befundtexten dar. Befundtexte enthalten detaillierte Aussagen zum Gesundheitszustand von Patientinnen und Patienten sowie auch Angaben zu deren Lebensstil und unterliegen daher dem Datenschutzgesetz, um die Rechte der betroffenen Personen zu wahren. Da die

Patientendaten unter Datenschutz stehen, können sie nur in der Klinik ausgewertet werden. Es gibt daher nur wenige frei zugängliche Datensätze, welche darüber hinaus einem erhöhten Risiko für einen Selektions-Bias unterliegen, da sie unter Umständen nur eine gewisse Teilpopulation repräsentieren. Eine Hürde stellt auch die sogenannte „Sanitization“ dar, die als minimaler Standard zum Datenschutz gilt und bei dem sensitive Daten wie Namen oder geographische Lokalisationen ersetzt oder entfernt werden. Obwohl diese Maßnahme häufig notwendig zum Schutz der Patientinnen und Patienten ist, stellt sie gleichzeitig eine mögliche Gefährdung der Integrität und Nutzbarkeit der Daten dar (16).

Zudem werden für NLP relativ große Mengen an Daten benötigt, die eine bestimmte Pathologie beinhalten. Besonders schwierig ist es hierbei, an große Mengen von annotierten Daten zu gelangen, vor allem, wenn es sich um nicht-englischsprachige Berichte handelt (29). Aus diesem Grund müssen die Daten gegebenenfalls vor ihrer Verwendung noch manuell gelabelt werden. Das manuelle Annotieren der Texte sowie das Training des Computersystems können sich hierbei als äußerst zeitaufwendig erweisen.

1.6 Besonderheiten im Training von Deep-Learning-Modellen

Im Bereich des NLP sind in den letzten Jahren zahlreiche vielversprechende Techniken entwickelt worden. Allerdings stellt deren Anwendung auf medizinische Texte durch die eingeschränkte Verfügbarkeit und die medizinische Terminologie immer noch eine gewisse Herausforderung dar.

Aktuell zeichnet sich ein Trend im Training von Deep-Learning-NLP-Modellen ab, in welchem das Training in zwei Teilschritte untergliedert wird. Diese Trainingsart kann dabei sowohl für RNN als auch für Modelle mit Transformer-Architektur angewendet werden.

Der erste Schritt ist das nicht supervidierte **Prä-Training**, bei dem das Modell mit einer großen Datenmenge (mitunter mehrere Gigabytes) von nicht-gelabelten Texten trainiert wird, um eine Kenntnis der Textbestandteile und -struktur zu erlangen. Die Nutzung dieser großen Menge an Daten für das Prä-Training ermöglicht den Modellen bereits das Erlernen eines Sprachverständnisses (29). Das Prä-Training kann dabei mit unterschiedlichen Daten erfolgen. Während Repräsentationsmodelle für die Alltagssprache beispielsweise mithilfe vom englischsprachigen Wikipedia trainiert werden, ist

dieses Standard-Modell aufgrund der vielen in radiologischen Befundtexten enthaltenen Fachbegriffe, die oft falsch verstanden werden, für den medizinischen Gebrauch nur bedingt geeignet. Das Modell kann allerdings auch explizit mit domänenspezifischen Texten trainiert werden, wie z.B. bei BioBERT, welches mit Abstracts und Artikeln der Datenbank PubMed und damit mithilfe biomedizinischer Arbeiten trainiert wurde und somit ein medizinisches Textverständnis erhielt (30).

Im zweiten Schritt, der supervidierten **Feinjustierung (Fine Tuning)**, werden die vortrainierten Modelle mithilfe von Annotationen/Labels auf eine spezifische Aufgabe angepasst. Annotationen, also Anmerkungen, die von einer menschlichen Person zum Text gemacht wurden, können im medizinisch-radiologischen Kontext beispielsweise Befunde, die im Befundtext erwähnt wurden, oder bestimmte Befundeigenschaften sein.

Die Anzahl an verfügbaren gelabelten bzw. annotierten Daten ist meist sehr begrenzt (gar keine bis einige tausende Texte), während der Zugriff auf bis zu mehreren Billionen an frei verfügbaren nicht-gelabelten Texten oft möglich ist. Das im Prä-Training erlangte Textverständnis der Modelle führt im Endeffekt dazu, dass im Anschluss weniger gelabelte Daten benötigt werden. Es ergibt sich dadurch oft eine immense Zeitersparnis, da manuelles Labeln eines Textes äußerst zeitaufwendig ist. Mit dieser Art des Trainings können bessere Ergebnisse als mit konventionellen Trainingsmethoden erzielt und eine optimierte Form des NLP ermöglicht werden (29).

Eine Studie von Rivera Zavala et al. hat beispielsweise die Fähigkeiten zweier vortrainierten Modelle (eines LSTM- sowie eines BERT-Modells) untersucht, Negativierungen und Spekulationen in medizinischen Texten zu identifizieren. Sie zeigte, dass beide Modelle ähnlich gute Leistungen und höhere Genauigkeiten als vorangegangene Methoden, wie z.B. regelbasierte Algorithmen, erbringen konnten (31).

Im Bereich des NLP werden Deep-Learning-Modelle mittlerweile sehr häufig eingesetzt. Die Tendenz geht innerhalb der verschiedenen Modelle in den letzten Jahren oft zum vermehrten Einsatz des Transformers, sodass er seit seiner Entwicklung im Jahre 2017 auch im medizinischen und insbesondere auf dem radiologischen Sektor bereits in verschiedenen Fragestellungen Anwendung fand. Einige von zahlreich existierenden Beispielen hierfür sind die Feinjustierung für die Überprüfung der diagnostischen Sicherheit in MRT-Aufnahmen (32), die Textklassifikation von radiologischen Befunden für die Knie-Osteoarthritis (33) oder die Extraktion klinischer Informationen für Brustkrebs

(34). Insgesamt wird der Transformer in vielen Bereichen des NLP als State-of-the-Art bezeichnet, also als neuester Stand der Entwicklung (35).

1.7. Zielsetzung

Die CT des Thorax ist eine der am häufigsten durchgeführten und wichtigsten diagnostischen Verfahren der Radiologie, deren Befunde mithilfe eines Befundtextes dargestellt und kommuniziert werden. Da es für den Befundtext jedoch keine einheitlichen Regeln gibt, existiert bislang keine strukturierte Befundung. Es ist daher auch nicht möglich, die Befundtexte zu kategorisieren oder gezielt nach Befundtexten mit einer bestimmten Pathologie zu suchen.

Im Zentrum dieser Arbeit stand die Klassifikation von computertomographischen Befundtexten des Thorax anhand von Deep Learning. Mithilfe von Deep-Learning-Techniken wurden drei verschiedene NLP-Modelle (AWD-LSTM, BERT, DistilBERT) entwickelt, wofür verschiedene Schritte notwendig waren (siehe Abbildung 1).

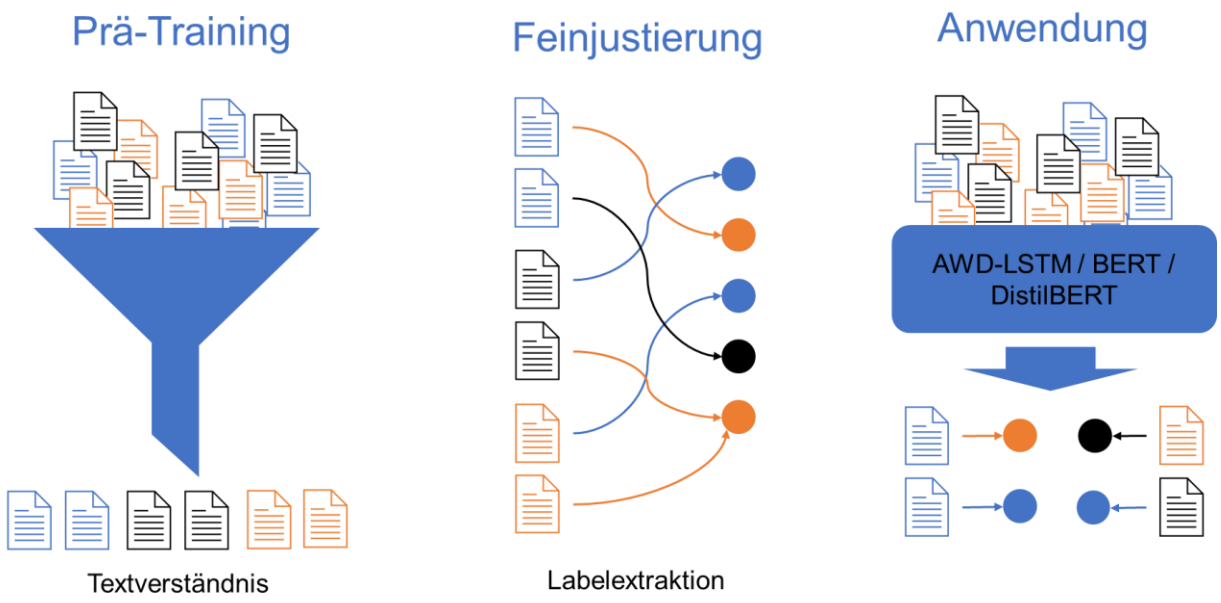

Abbildung 1: Entwicklungsmechanismen der Deep-Learning-Modelle

In dieser Grafik ist das Training der drei in dieser Arbeit verwendeten Modelle AWD-LSTM, BERT sowie DistilBERT dargestellt, welches einem gemeinsamen Grundprinzip folgt (Prä-Training, Feinjustierung, Anwendung). In der Abbildung stehen die Text-Symbole für Befundtexte und die Punkte für Label.

Es wurde zunächst ein Prä-Training auf einem großen Datensatz mit ungelabelten radiologischen Daten durchgeführt, damit die Modelle ein Textverständnis erlangten.

Anschließend erfolgte die Feinjustierung (Fine Tuning), bei dem die Modelle mithilfe eines Datensatzes annotierter Texte trainiert wurden, Label aus den Befundtexten zu

extrahieren. Dafür wurden Befundtexte der CT des Thorax sowie der CT zur Detektion der Lungenarterienembolie mittels Annotationen manuell strukturiert klassifiziert bzw. gelabelt.

Ziel war es, die Modelle zu trainieren, damit sie im Anschluss imstande sind, die wichtigsten Diagnosen der CT des Thorax in den Befundtexten selbstständig zu identifizieren, und auf alle weiteren Befundtexten angewendet werden können. Die Extraktion sämtlicher Labels ermöglicht es, alle Befundtexte nach bestimmten Eigenschaften zu filtern, und bietet darüber hinaus die Möglichkeit der Weiterverwendung für andere Aufgaben, wie beispielsweise der Entwicklung eines Bildklassifikations-modells.

2. Materialien und Methoden

2.1 Modellarchitekturen

Für die Klassifikation von Befundtexten der CT des Thorax wurden insgesamt drei verschiedene Modelle unterschiedlicher Komplexität trainiert und anschließend ihre Leistung evaluiert. Jedes der drei Deep-Learning-Modelle besaß hierbei eine individuelle Architektur, welche seine spätere Genauigkeit mitbestimmte.

Modell 1: AWD-LSTM

Das erste Modell war ein AWD-LSTM (Average-Stochastic Gradient Descent Weight-Dropped LSTM), welches mit der ULMFiT-Technik (siehe 2.2) trainiert wurde. Bei diesem Modell handelte es sich um ein RNN mit multiplen LSTM-Modulen. LSTM-Modelle unterscheiden sich von herkömmlichen RNN dadurch, dass sie Aktivierungen aus vorausgegangenen Phasen selektiv konservieren bzw. aktiv vergessen und sich somit an vorherige Aktivierungen erinnern können (36). Ihre Architektur besteht im Wesentlichen aus einem Eingangstor (Input Gate), einem Ausgangstor (Output Gate), einem Erinnerungs- und Vergesstor (Forget Gate) sowie dem Zellinneren mit seinen Verknüpfungen. Dieser Aufbau ermöglicht durch dieses gezielte Behalten von Informationen, tiefe Netzwerke mit multiplen LSTM-Modulen zu konstruieren ohne die Gefahr verschwindender Gradienten bei der Backpropagation. Zusätzlich wird bei dem AWD-LSTM-Modell noch die Regularisierungstechnik des Drop-Connect angewendet, bei welcher zufällige Gewichte einzelner Schichten während des Trainings nicht aktualisiert werden. Dies verleiht dem Modell zusätzliche stochastische Tiefe (37).

Modell 2: BERT

Bei den beiden anderen Modellen handelte es sich um zwei Formen des Transformers, welche die in 1.5 beschriebene Grundarchitektur mit Aufmerksamkeitsmechanismen besaßen. Das eine Modell war ein klassisches BERT-Modell.

BERT (Bidirectional Encoder Representations from Transformers) stellt ein tief bidirektionales Modell dar, was bedeutet, dass es die Bedeutung eines Wortes in einem Satz basierend auf den Worten davor und dahinter erkennen kann, also der Wortzusammenhang bei der Erstellung der Wortkodierung eine Rolle spielt. Das unterscheidet BERT von vorausgegangenen ähnlichen Modellen, welche zumeist

unidirektional oder nur marginal bidirektional blieben. BERT hat durch den Aufmerksamkeitsmechanismus die Fähigkeit, Wort-Interaktionen über eine lange Spannweite zu erkennen (29).

Das hier genutzte BERT-Modell bestand aus zwölf Schichten mit insgesamt 110 Millionen Parametern.

Modell 3: DistilBERT

Bei dem dritten KNN handelte es sich um DistilBERT, eine größen- und performanceoptimierte Variante des klassischen BERT-Modells (38).

DistilBERT ist ein Modell, welches grundsätzlich dieselbe Transformer-Architektur wie BERT besitzt, aber durch eine spezielle Kompressionstechnik, die sogenannte „Knowledge Distillation“, kleiner und somit besser anwendbar und auch weniger ressourcenintensiv während des Trainings ist. Bei dieser Technik wird ein kompakteres Modell, der sogenannte Student, darauf trainiert, das Verhalten eines größeren Modells, des sogenannten Lehrers, zu reproduzieren. Es besitzt eine Größe von sechs Schichten mit insgesamt 66 Millionen Parametern und ist dadurch im Vergleich zum klassischen BERT-Modell um 40% kleiner, 60% schneller und dabei mit einer Sprachverständnis-Kapazität von 97% nahezu genauso leistungsstark (38).

2.2 Textextraktion

2.2.1 Datensätze

Zu Beginn wurde ein Datensatz für das Training der Modelle benötigt. Hierzu wurden alle Befundtexte aus dem radiologischen Datenarchiv der Charité – Universitätsmedizin Berlin, die im Zeitraum vom 01.01.2009 bis zum 31.12.2018 (n = 4.790.287) verfasst wurden, extrahiert. Es handelte sich ausschließlich um Freitextbefunde. Diese beinhalteten dabei sämtliche Ergebnisse von Untersuchungen der Röntgen-Diagnostik, CT, MRT, Sonographie, interventionellen Radiologie, Bestrahlungstherapie und Nuklearmedizin.

Von allen Befundtexten wurden insgesamt n = 948.457 isoliert, da sie keine Befunde des Thorax enthielten, weil es sich beispielsweise um Konstanztestungen der Untersuchungsgeräte, abgesagte Untersuchungen oder importierte Bilder handelte.

Es ist nicht möglich, die An- oder Abwesenheit eines Befundes zu eruieren, wenn er im Befundtext nicht beschrieben wurde. Da die aus diesen Texten generierten Labels ansonsten ungenau wären und die weiterführenden Aufgaben beeinträchtigen würden, wurde beschlossen, die unvollständigen Befunde auszuschließen. Um diese Befunde zu identifizieren, wurden alle Befundtexte danach untersucht, wie häufig bestimmte Wort-Kombinationen auftauchten. Weil es sehr unwahrscheinlich ist, dass in mehreren Freitextbefunden der exakt gleiche Wortlaut auftritt, konnten automatisch ausgefüllte Texte, z.B. über das Nicht-Erscheinen einer Patientin oder eines Patienten, zuverlässig identifiziert und isoliert werden. Auch Texte mit einer Wortanzahl von weniger als 50 wurden ausgeschlossen, sodass n = 3.841.830 Befunde für das Prä-Training übrig blieben (415.702.033 Wörter, 3,36 GB). Nach dem sogenannten „White-Space Stripping“, bei dem alle Leerzeichen entfernt wurden, wurden die Befunde in eine einzige Textdatei übertragen.

2.2.2 Tokenisierung

Tokenisierung beschreibt die Unterteilung eines Textes in multiple wiederkehrende kleinere Einheiten (Tokens). Ein Token kann hierbei ein Buchstabe, ein Wortteil oder ein vollständiges Wort sein. In dieser Arbeit wurde eine Subwort-Tokenisierung verwendet, welche den Text in wiederkehrende Wortteile untergliederte. Vor der Tokenisierung wurde die Größe des Vokabulars (maximale Anzahl der letztendlichen Tokens) auf 30.000 festgelegt. Zusätzlich zu den 30.000 festgelegten Tokens wurden noch spezielle Tokens eingeführt, welche spezifisch für die jeweiligen Modelle waren. In dieser Arbeit wurden verschiedene Modelle und Techniken angewendet, welche eine spezifische Tokenisierung erforderten. Die Tokenisierung für das Universal Language Model Fine Tuning (ULMFiT), die für das LSTM-Modell verwendet wurde, unterschied sich demzufolge von den Tokens, welche für BERT benötigt wurden.

ULMFiT:

- xxbos: Markiert den Beginn einer Sequenz (BOS = Beginn of Stream).
- xxmaj: Indiziert, dass das nächste Token mit einem Großbuchstaben beginnt (alle Wörter wurden zuvor in Kleinbuchstaben konvertiert).
- xxup: Indiziert, dass das nächste Token nur aus Großbuchstaben besteht.
- xxunk: Repräsentiert ein unbekanntes Token (z.B. ein sehr seltenes Token).
- xxeos: Markiert das Ende der Sequenz (EOS = End of Stream).

BERT:

- [CLS]: Markiert den Beginn einer Sequenz.
- [SEP]: Markiert das Ende einer Sequenz.
- [PAD]: Wird genutzt, um Sequenzen unterschiedlicher Länge aufzufüllen.
- [UNK]: Repräsentiert ein unbekanntes (sehr seltenes) Token.
- [MASK]: Ersetzt beim unsupervidierten Prä-Training des Modells zufällig 15% der Wörter. Das ersetzte Wort muss durch das Modell vorhergesagt werden.

Die Tokenisierung wurde für ULMFiT mit der Python-Programmbibliothek „SpaCy“ (39) durchgeführt und für BERT erfolgte die Tokenisierung mit „Tokenizers“ (40), welches ebenfalls eine Python-Programmbibliothek ist. Es wurden Subwort-Tokenizer verwendet, welche Worte in ihre häufigsten Bestandteile unterteilten.

Beispielsweise konnte der Satz „Kein Nachweis einer Lungenarterienembolie bis auf Subsegmentebene“ durch den ULMFiT-Tokenizer in folgende Tokens unterteilt werden:

„xxbos“, „xxmaj“, „kein“, „xxmaj“, „nach“, „##weis“, „ein“, „##er“, „xxmaj“, „lungen“, „##arterien“, „##embolie“, „bis“, „auf“, „xxmaj“, „sub“, „##segmentebene“, „xxeos“

Der BERT-Tokenizer konnte folgende Unterteilung vornehmen:

„[CLS]“, „Kein“, „Nach“, „##weis“, „ein“, „##er“, „Lungen“, „##arterien“, „##embolie“, „bis“, „auf“, „Sub“, „##segmentebene“, „[SEP]“

Im Gegensatz zur ULMFiT-Technik unterschied der BERT-Tokenizer zwischen Groß- und Kleinschreibung, „Sub“ und „sub“ entsprachen somit verschiedenen Tokens. Dies wurde bewusst gewählt, da in der deutschen Sprache die Großschreibung von Wörtern teilweise entscheidend über die Bedeutung eines Wortes sein kann. Beim Tokenizer ULMFiT kennzeichnete das Spezial-Token „xxmaj“, dass das darauffolgende Wort großgeschrieben wurde. Die Mindesthäufigkeit eines Tokens wurde auf zehn festgelegt, sodass Tokens, welche seltener im Text auftraten, durch „[UNK]“ bzw. „xxunk“ ersetzt wurden. Da die Deep-Learning-Modelle keine Texte, sondern nur Zahlen verarbeiten können, wurden in einem letzten Schritt die Tokens durch Zahlen ersetzt, wobei jedes Token einer Ziffer zugeordnet war.

2.3 Prä-Training der Deep-Learning-Modelle

2.3.1 Prä-Training vom AWD-LSTM

Im Rahmen des ULMFiT wurde das AWD-LSTM auf die Folgewortprädiktion in einem Satz trainiert. Die Aufgabe des Netzwerkes bestand also darin, nach Vorgabe einer kurzen Sequenz von Tokens das nächste Token vorherzusagen. Hierfür wurden ausschließlich Befundtexte von CT-Thorax-Untersuchungen verwendet.

Die Sequenzlänge während des Trainings wurde auf 512 festgelegt und die Größe des Minibatches, einer Untergruppe von Daten, so angepasst, dass die gesamte Größe des Videoarbeitsspeichers der Grafikkarte (12 GB) genutzt wurde. Das Training erfolgte gestaffelt. Zuerst wurde die letzte Schicht des Netzwerkes für eine Epoche bei einer Lernrate von 0,02 trainiert, anschließend das gesamte Netzwerk für 25 Epochen mit einer Lernrate von 0,0001. Beim Prä-Training wurde auf die Metriken Trainingsloss und Validierungsloss geachtet (siehe Abbildung 2). Der Trainingsloss zeigte an, wie gut das Modell an die Trainingsdaten angepasst war, während der Validierungsloss indizierte, wie gut das Modell sich an neue Daten anpassen können würde. Der Loss spiegelte dabei die Fehleranfälligkeit des Modells wider, sodass ein niedrigerer Loss ein besseres Modell darstellte.

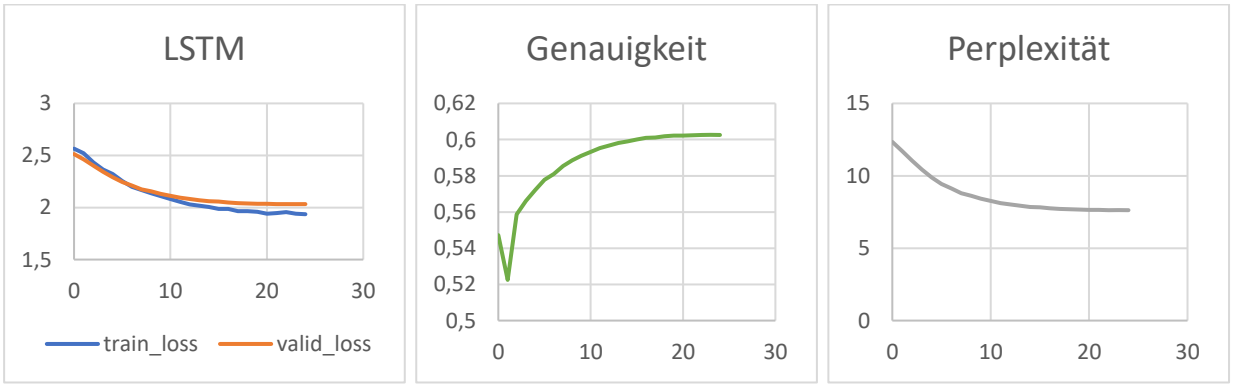

Abbildung 2: Prä-Training vom AWD-LSTM

Die Abbildung stellt verschiedene Werte dar, die während des Prä-Trainings des AWD-LSTM erhoben wurden. Es werden Trainings- (blaue Kurve) und Validierungsloss (orange Kurve) auf der linken Grafik, die steigende Genauigkeit (grüne Kurve) auf der mittleren Grafik sowie die abnehmende Perplexität (graue Kurve) auf der rechten Grafik im Verhältnis zur Trainingsepisode abgebildet.

Zur Evaluation des Trainings wurde die Perplexität als Metrik genutzt, wobei eine niedrigere Perplexität für eine niedrigere Entropie und eine bessere Genauigkeit des Modells bei der Vorhersage der individuellen Wörter sprach. Der beste Perplexitätswert wurde nach 24 Epochen erreicht und betrug 7,63. Dieses Modell wurde für den nächsten

Trainingsschritt gespeichert. Das Prä-Training erfolgte unter Nutzung der Programmbibliotheken „PyTorch“ (41) und „Fastai“ (42) und dauerte durchschnittlich 27 Sekunden pro Epoche.

2.3.2 Prä-Training von BERT und DistilBERT

Beide in dieser Arbeit genutzten Transformer-Modelle wurden identisch vortrainiert. Die Aufgaben des Modells für das Training umfassten hierbei die Vorhersage maskierter Wörter (Masked Word Prediction) und die Prädiktion eines Folgesatzes. Durch die daraus entwickelte Fähigkeit zur Folgesatzprädiktion konnte das Modell somit feststellen, ob und inwiefern zwei Sätze miteinander in Verbindung standen (43).

Das Training der Modelle erfolgte unter Nutzung der Programmbibliotheken „PyTorch“ (41) und „Transformers“ (40). Als Corpus für das Training dienten die gesammelten Befundtexte der Radiologie im Zeitraum vom 01.01.2009 bis 31.12.2018. Nach Anonymisierung der Befundtexte wurden alle Texte in einer einzigen großen Datei zusammengeführt. Die Erstellung der Wortmaskierung oder das Vorbereiten der Sätze für die Satzprädiktion erfolgte direkt während des Trainingsvorgangs. Die Anzahl maskierter Wörter wurde auf 15% festgelegt. Die Größe des Minibatches betrug 32 GB und das Modell wurde für eine Epoche trainiert. Ein Abspeichern der Modellgewichte erfolgte alle 25.000 Schritte, wobei ein Schritt einem Minibatch entsprach.

2.4 Annotation von Befundtexten

Für die spätere Feinjustierung und die Evaluation der Leistung der Modelle wurde ein Datensatz aus gelabelten Befundtexten benötigt. Hierzu wurden insgesamt 5.950 Befundtexte von computertomographischen Aufnahmen des Thorax aus dem radiologischen Informationssystem (RIS) extrahiert und manuell annotiert. Die Annotationen wurden mithilfe eines eigens zu diesem Zweck programmierten Hilfsprogramms (Annotator) durchgeführt, welches in Abbildung 3 dargestellt ist. Die Befundtexte wurden auf das Auftreten von insgesamt 21 Befunden hin untersucht: bronchiale Auffälligkeit, Emphysem, Fibrose, Frakturen, Fremdmaterialien, Herzgröße (vergrößertes Herz), Hiatushernie, kardiale Stauung, Knochentumoren, Lungenarterienembolie, Lungentrauma, Lungenrundherd, Lymphadenopathie, mediastinaler Tumor, Perikarderguss, Pleuraerguss, pleurale Auffälligkeit, Pneumonie, Pneumothorax, Ventilationsstörung und Weichteilemphysem. Hierbei wurden passend zu jedem Befundtext die darin beschriebenen Pathologien annotiert.

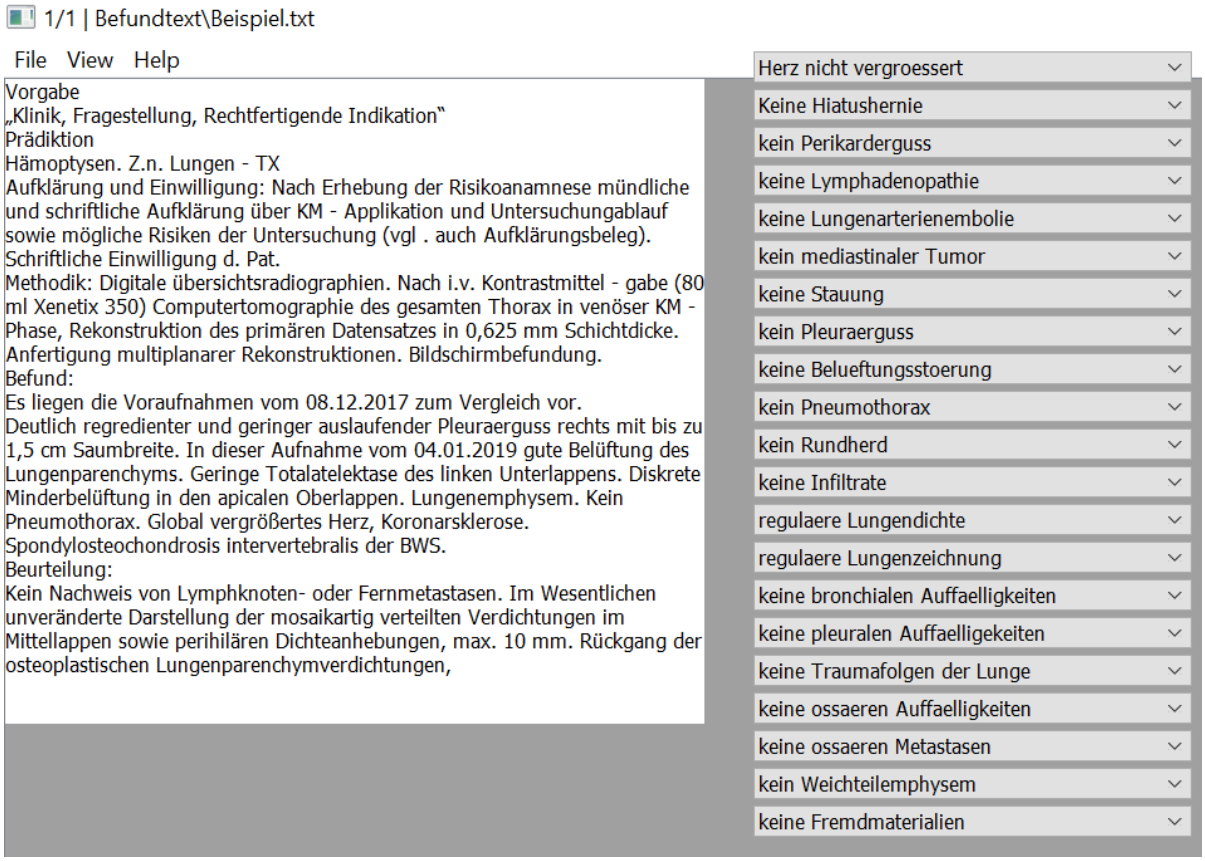

Abbildung 3: Screenshot des Annotators

Der Annotator ist das Programm, mit dessen Hilfe die Annotationen durchgeführt wurden. Um Patientendaten zu schützen, handelt es sich bei dem im Bild links dargestellten Text nicht um einen realen Befundtext, sondern um einen Befund, der durch das AWD-LSTM generiert wurde und nur zur Veranschaulichung dient. Auf der rechten Seite sind die 21 verschiedenen Labelkategorien abgebildet.

Der Gesamtdatensatz wurde schrittweise in verschiedene Datensätze unterteilt (siehe Abbildung 4). Im Vorfeld wurden vom gesamten Datensatz 1.000 Befundtexte isoliert und daraus ein Testdatensatz erstellt. Dieser war für die spätere Evaluation der Modelle von großer Bedeutung. Durch die Erhebung der Performance-Metriken nicht nur auf den Validierungsdatensatz, sondern auch auf den Testdatensatz, kann verhindert werden, dass von einer falsch hohen Genauigkeit bzw. einem falsch niedrigen Generalisierungsfehler ausgegangen wird. Diese vermeintlich guten Vorhersage-ergebnisse würden darauf basieren, dass das Modell den Datensatz, auf dessen Grundlage es im vorherigen Training entwickelt wurde, sehr exakt abbildet und daher verhältnismäßig gut vorhersagen kann. In solch einem Fall spräche man von einem Overfitting, also einer Überanpassung des Modells (44). Dieses würde man daran erkennen, dass die Ergebnisse des Testdatensatzes deutlich unter denen des Validierungsdatensatzes lägen.

Um Befundtexte ohne ausreichende Informationen für die Klassifikation mittels der definierten Labels (beispielsweise über das Nicht-Erscheinen einer Patientin oder eines Patienten oder abgesagte Untersuchungen) als inkomplett zu markieren, wurden alle Texte, die weniger als 50 Zeichen beinhielten, gekennzeichnet. Wie auch im Prä-Training wurden diese Befunde ausgeschlossen, um in den nachfolgenden Aufgaben keine Ungenauigkeiten zu erzeugen.

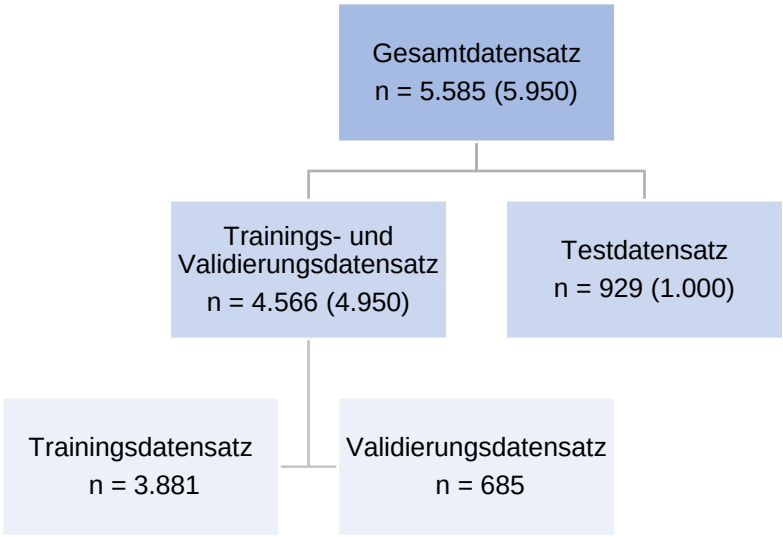

Abbildung 4: Unterteilung des Datensatzes

Das Diagramm stellt die Vorgehensweise bei der Unterteilung des gesamten Datensatzes in einen Trainingsdatensatz, einen Validierungsdatensatz sowie einen Testdatensatz dar. Es wurden dabei in Klammern die absolute Anzahl der Texte vor dem Ausschluss unvollständiger Befunde angegeben.

Das betraf im Trainings- und Validierungsdatensatz n = 384 Befunde, sodass der finale Datensatz hier aus n = 4.566 Befunden bestand. Dieser Datensatz setzte sich aus 4.216 Befundtexten der Untersuchung CT Thorax und 350 Befundtexten, die aus einer computertomographischen LAE-Untersuchung, also einer CT-Angiographie, stammten, zusammen. Der Datensatz wurde anschließend im Verhältnis 85:15 in einen Trainingsdatensatz (n = 3.881) und Validierungsdatensatz (n = 685) unterteilt.

Die Texte des Testdatensatzes (n = 1.000) wurden nach dem Training der Modelle aus dem radiologischen Informationssystem (RIS) extrahiert und von zwei Annotatorinnen gelabelt, um das Interrater-Agreement zu erheben. Im Falle einer großen Diskrepanz zwischen den zwei Annotatorinnen erfolgte eine Überprüfung durch einen weiteren radiologischen Fachexperten. Zudem wurden die Texte durch eine der Annotatorinnen nach einem längeren Zeitintervall erneut gelabelt, um das Intrarater-Agreement zu erfassen.

Auch im Testdatensatz wurden inkomplette Texte markiert und ausgeschlossen, sodass n = 929 Texte für die Evaluation der Modelle übrig blieben. Davon stammten 852 von der Untersuchung CT Thorax und 77 wurden im Rahmen einer CT zur Lungenarterienembolie erstellt.

2.5 Feinjustierung der Deep-Learning-Modelle

Für die Feinjustierung (Fine Tuning) der Sprachmodelle wurden die in 2.3 vortrainierten Sprachmodelle geladen und die letzte Schicht durch eine zusätzliche Klassifikationsschicht ergänzt. Die Modelle wurden mithilfe der annotierten Texte des Trainings- und Validierungsdatensatzes auf ihre spezifische Aufgabe angepasst: die Klassifikation der Befundtexte von computertomographischen Aufnahmen des Thorax.

2.5.1 Feinjustierung vom AWD-LSTM

Die Feinjustierung unterteilte sich in zwei Phasen. In der ersten Phase erfolgte nur eine Anpassung der Gewichte der neu hinzugefügten Klassifikationsschicht; die Gewichte des vortrainierten Encoders wurden konstant gehalten. Phase 1 umfasste eine Epoche mit einer Lernrate von 0,001. In Phase 2 konnten alle Gewichte des Netzwerks angepasst werden und sie umfasste 100 Epochen mit einer Lernrate von 0,0001. Erneut wurde der Validierungsloss überwacht und das Modell, das den geringsten Loss zeigte, zwischengespeichert. Dieser Loss wurde nach 52 Trainingsepochen erreicht, wobei eine Epoche durchschnittlich 17 Sekunden lang war. Anschließend stieg der Validierungsloss wieder, was für ein beginnendes Overfitting des Modells sprach (siehe Abbildung 5).

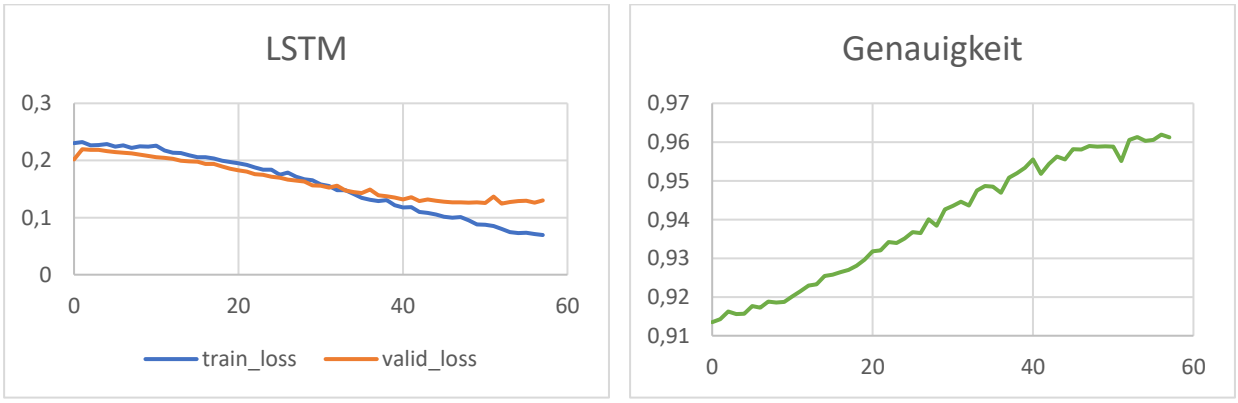

Abbildung 5: Feinjustierung vom AWD-LSTM

Das Diagramm stellt verschiedene erhobene Parameter während der Feinjustierung des AWD-LSTM dar. Es werden abnehmende Werte von Trainings- (blaue Kurve) und Validierungsloss (orange Kurve) auf der linken Grafik sowie die die steigende Genauigkeit (grüne Kurve) auf der rechten Grafik im Verhältnis zur Trainingsepisode abgebildet.

2.5.2 Feinjustierung von BERT und DistilBERT

Die Feinjustierung erfolgte für BERT und DistilBERT unter Beobachtung von Trainings- und Validierungsloss mit einer uniformen Sequenzlänge von 512 Tokens, einer Batchgröße von 16, einer Lernrate von 0,0001 und einem Strafterm von 0,001.

Beim BERT-Modell wurde der beste Wert für den Validierungsloss nach der achten Epoche erreicht (siehe Abbildung 6), wobei das Training durchschnittlich 44 Sekunden pro Epoche dauerte.

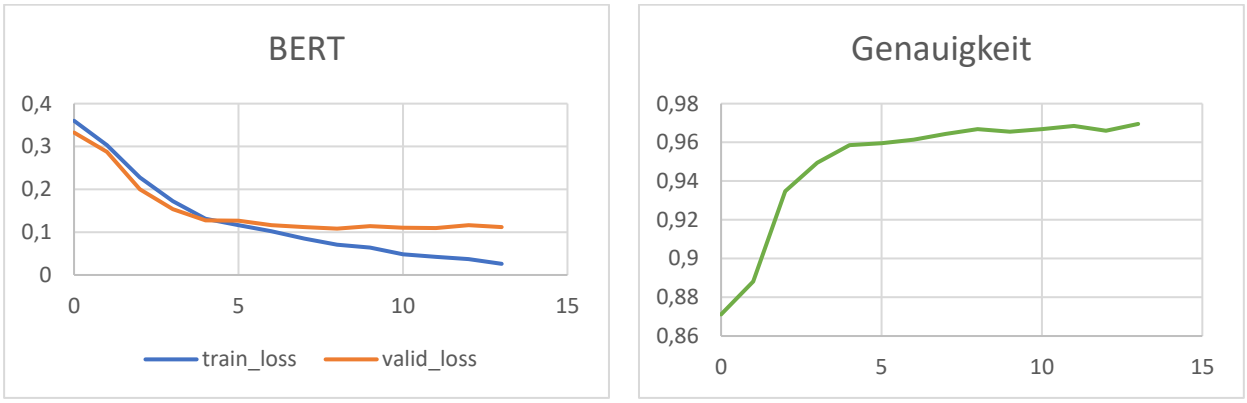

Abbildung 6: Feinjustierung vom BERT-Modell

Das Diagramm stellt verschiedene erhobene Parameter während der Feinjustierung des BERT-Modells dar. Es werden abnehmende Werte von Trainings- (blaue Kurve) und Validierungsloss (orange Kurve) auf der linken Grafik sowie die die steigende Genauigkeit (grüne Kurve) auf der rechten Grafik im Verhältnis zur Trainingsepisode abgebildet.

Die Epochendauer des DistilBERT-Modells betrug durchschnittlich 24 Sekunden. Bei diesem Modell wurde nach der neunten Trainingsepoche keine Abnahme des Validierungslosses mehr beobachtet und das Training beendet (siehe Abbildung 7).

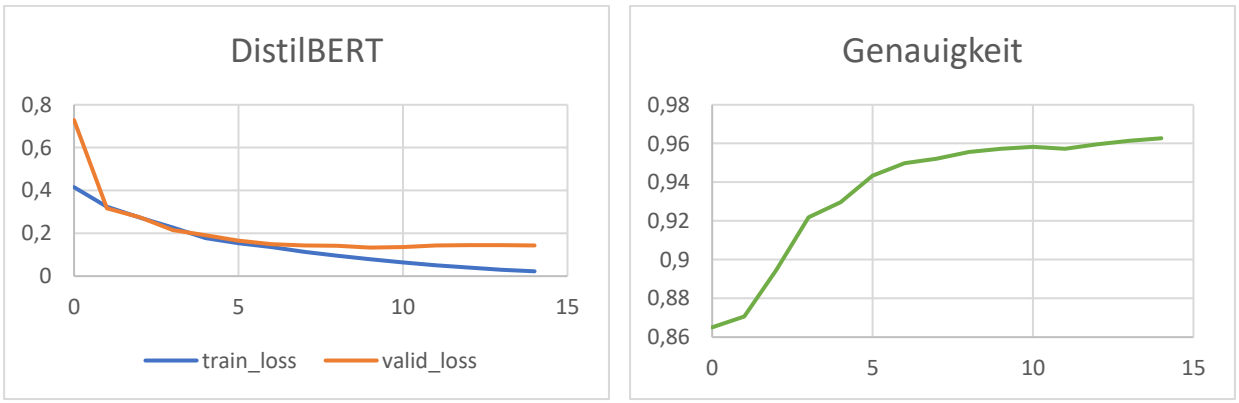

Abbildung 7: Feinjustierung vom DistilBERT-Modell

Das Diagramm stellt verschiedene erhobene Parameter während der Feinjustierung des DistilBERT-Modells dar. Es werden abnehmende Werte von Trainings- (blaue Kurve) und Validierungsloss (orange Kurve) auf der linken Grafik sowie die steigende Genauigkeit (grüne Kurve) auf der rechten Grafik im Verhältnis zur Trainingsepisode abgebildet.

2.6. Metriken

Um die Leistung des AWD-LSTM, des BERT-Modells sowie das DistilBERT-Modells zu ermitteln, erhielten die Modelle dieselbe Aufgabe wie die menschlichen Annotatorinnen bei ihrem manuellen Labeling der Befundtexte. Sie sollten somit die Labels vergeben bzw. klassifizieren, welcher der 21 möglichen Befunde vorlag.

Zur besseren Evaluation der drei Deep-Learning-Modelle wurden deren Leistungen anhand eines Validierungsdatensatzes sowie eines vom Trainings- und Validierungs-datensatz unabhängigen Testdatensatzes bewertet. Die Werte beider Datensätze wurden darüber hinaus verglichen, um mögliche Generalisierungsfehler sowie Überanpassungen zu identifizieren.

Die Prädiktionen der Modelle konnten Werte zwischen minus unendlich und unendlich annehmen, welche mit Hilfe einer Softmax-Funktion in Werte zwischen 0 und 1 skaliert wurden. Da es sich um binäre Klassifikationsmodelle handelte, wurden die Vorhersagen in zwei Kategorien unterteilt. Die Prädiktionen wurden somit entweder in 1 (Befund wurde als vorhanden vorhergesagt) oder 0 (Befund wurde als nicht vorhanden vorhergesagt) transformiert, wobei ein Schwellenwert von 0,5 gewählt wurde. Das bedeutet, dass Werte, welche größer als 0,5 waren, indizierten, dass der Befund vorhanden war. Dieser Schwellenwert wurde über das gesamte Training hinweg konstant gehalten.

Zur Evaluation der jeweiligen Leistung des Modells sowie zum Vergleich der Modelle untereinander wurden die Werte Genauigkeit (Accuracy), Sensitivität (Recall, R), positiver prädiktiver Wert (PPV), F1-Wert sowie die Area under the curve (AUC) erhoben. Die Metriken wurden ermittelt, indem die Vorhersagen der Deep-Learning-Modelle mit den vergebenen Labels durch die Annotatorinnen verglichen wurden.

Die **Genauigkeit**, welche im Englischen als Accuracy bezeichnet wird, beschreibt den Anteil der korrekt klassifizierten Variablen. Sie umfasst somit sowohl die richtig positiven (true positive/TP) als auch die richtig negativen (true negative/TN) Befunde, welche ins Verhältnis zur Gesamtheit der Befunde, welche auch die falsch positiven (false positive/FP) und falsch negativen (false negative/FN) Befunde beinhaltet, gesetzt werden. In dieser Arbeit sagt sie also aus, in wie vielen Befundtexten der jeweilige Befund korrekt als vorhanden bzw. nicht-vorhanden klassifiziert wurde.

$$Genauigkeit = \frac{TP+TN}{TP+TN+FP+FN}$$

Die **Sensitivität** stellt die Anzahl der korrekt als positiv klassifizierten Variablen geteilt durch die Anzahl aller positiven Variablen dar. In dieser Arbeit beschreibt sie, in welchem Prozentsatz der Fälle die vorliegenden Befunde tatsächlich vom Modell als vorhanden identifiziert wurden.

$$Sensitivität = \frac{TP}{TP+FN}$$

Der **positive prädiktive Wert** (PPV) bezeichnet die Wahrscheinlichkeit, dass eine Variable bei einem positiven Testergebnis tatsächlich positiv ist. Das beschreibt in dieser Studie die Wahrscheinlichkeit, dass ein Befund, den das Modell als vorhanden vorhersagt, auch tatsächlich ein positiver Befund ist.

$$PPV = \frac{TP}{TP+FP}$$

Der **F1-Wert** ist ein Maß, welches die Sensitivität und den positiven prädiktiven Wert kombiniert, und in dem beide Werte mithilfe des harmonischen Mittels gleich gewichtet sind. Im Gegensatz zur Genauigkeit fließen in diesen Wert daher keine richtig negativen Befunde ein.

$$F1\text{-}Wert = 2\ x\ \frac{Sensitivität\ x\ PPV}{Sensitivität+PPV}$$

Die **AUC** (Area under the Curve) wird mithilfe der Operationscharakteristik, der sogenannten Receiver Operating Characteristic (ROC), erhoben. Zur Erstellung dieser Kurve wird die Sensitivität/Richtig-Positiv-Rate (y-Achse) im Verhältnis zur Falsch-Positiv-Rate (x-Achse) gesetzt. Der Vorteil der AUC ist, dass sie Informationen über die Performance eines Modells bietet, unabhängig vom gewählten Schwellenwert.

Zusätzlich wurden das **Interrater-** und **Intrarater-Agreement** erhoben. Hierzu wurde Cohen‘s Kappa verwendet und ein 95-prozentiges Konfidenzintervall erhoben, indem ein Bootstrapping mit jeweils 1.000 Wiederholungen durchgeführt wurde. Bootstrapping stellt eine Form des Resamplings (Stichprobenwiederholung) dar. Hier werden zufällige multiple Datenpunkte aus den ursprünglichen Daten gezogen und so verschiedene artifizielle Datensätze erstellt, auf denen die statistischen Parameter erneut errechnet werden. Die entstehende statistische Verteilung der Parameter hat nun im Anschluss eine Aussagekraft über die Verteilung der Grundgesamtheit, aus der die Stichprobe stammt.

3. Ergebnisse

3.1 Prävalenz der verschiedenen Befunde

3.1.1 Trainings- und Validierungsdatensatz

Für den Trainings- und Validierungsdatensatz wurden insgesamt n = 4.950 Texte annotiert. Nach dem Ausschluss von Befundtexten ohne ausreichende Informationen für die Klassifikation mittels der definierten Labels (n = 384) blieben für das Training der Modelle während der Feinjustierung n = 4.566 Texte übrig. Die Annotation umfasste eine binäre Klassifikation 21 verschiedener Befunde mit je zwei Labels, die das Vorhandensein bzw. die Abwesenheit des jeweiligen Befundes widerspiegelten, sodass insgesamt 42 unterschiedliche Labels vergeben werden konnten. Eine detaillierte Übersicht der Verteilung verschiedener Befunde ist in Tabelle 1 gegeben.

Der häufigste pathologische Befund in den annotierten Texten war ein Lungenrundherd mit einer Prävalenz von 42,1% (n = 1.924), gefolgt von einer Lymphadenopathie mit einer Prävalenz von 41,8% (n = 1.910) und Störung der Ventilation mit einer Prävalenz von 32,7% (n = 1.492). Ein Pleuraerguss trat in 26,6% der Befunde auf (n = 1.215) und eine Pneumonie war in 25,3% der Fälle zu finden (n = 1.155) und kam somit auch relativ oft vor. Bei diesen Pathologien handelte es sich insgesamt um Befunde, die durch die hohe Sensitivität der CT in vielen Untersuchungen detektiert werden konnten. Beispielsweise wurde das Label Lungenrundherd auch bereits bei kleineren unspezifischen Noduli oder Granulomen vergeben, welche bei vielen Patientinnen und Patienten zu finden sind.

Der mit Abstand seltenste Befund war das Lungentrauma, welches lediglich in n = 15 Befundtexten annotiert wurde, was einer Prävalenz von 0,3% entsprach. Deutlich häufiger, aber dennoch relativ selten waren die Hiatushernie (n = 119), die pleurale Auffälligkeit (n = 126), Knochentumoren (n = 154) sowie mediastinale Tumoren (n = 156). Diese Befunde wurden alle weniger als 200 Male annotiert und die Prävalenz betrug in allen diesen Fällen unter 4%. Die geringe Prävalenz der Labels war hierbei am ehesten auf die ebenfalls geringe Prävalenz der Grundkrankheiten im Patientenkollektiv der Radiologie der Charité zurückzuführen.

Darüber hinaus wurden in 25,2% der Befunde Fremdmaterialien (n = 1.150) annotiert, wobei das Label nur dann vergeben wurde, wenn es sich um therapeutische körperfremde Materialien handelte, die eine bestimmte Größe hatten und entfernt werden

könnten, wie etwa Thoraxdrainagen, Magensonden oder zentrale Venenkatheter, während etwa Nahtmaterial oder Clips nicht dazu gezählt wurden.

Die durchschnittliche Prävalenz eines Befundes betrug 14,0% und pro Befundtext kamen im Schnitt 3,0 Befunde vor.

3.1.2 Testdatensatz

Für den Testdatensatz wurden insgesamt n = 1.000 Texte annotiert. Nach dem Ausschluss von Befundtexten ohne ausreichende Informationen für die Klassifikation mittels der definierten Labels (n = 71) blieben für die Evaluation n = 929 Befunde übrig. Wie auch beim Trainings- und Validierungsdatensatz umfasste die Annotation 21 verschiedene Befunde mit je zwei Labels, die das Vorhandensein bzw. die Abwesenheit des jeweiligen Befundes widerspiegelten, sodass auch hier insgesamt 42 unterschiedliche Labels vergeben werden konnten.

Insgesamt verhielten sich die Häufigkeiten im Großen und Ganzen kongruent zu denen des Trainings- und Validierungsdatensatzes.

So waren im Testdatensatz die zwei häufigsten Befunde ebenfalls die Lymphadenopathie (n = 389), welche mit einer Prävalenz von 41,9% vorkam, sowie der Lungenrundherd (n = 380), welcher in 40,9% der Texte annotiert wurde. Auch andere häufige Befunde des Trainings- und Validierungsdatensatzes traten im Testdatensatz mit ähnlichen Prävalenzen auf. Beispiele wären die Störung der Ventilation mit 31,1% (n = 289), der Pleuraerguss mit 25,5% (n = 238) oder die Pneumonie mit 25,1% (n = 233).

Das Lungentrauma hingegen war, ebenfalls konkordant zum Trainings- und Validierungsdatensatz, der mit Abstand seltenste Befund und wurde lediglich in fünf Texten bzw. rund 0,5% aller Texte annotiert. Ebenso wie im Trainings- und Validierungsdatensatz waren die Befunde Hiatushernie (n = 25), Knochentumoren (n = 24), mediastinale Tumoren (n = 43) sowie pleurale Auffälligkeit (n = 28) deutlich häufiger als das Lungentrauma, die Labels wurden aber dennoch alle seltener als 50 Male vergeben und die Prävalenz betrug stets unter 5%. Beim Testdatensatz zeigte zusätzlich auch das Weichteilemphysem (n = 23) eine Prävalenz von nur 2,5% und war somit etwas seltener als im Trainings- und Validierungsdatensatz.

Es wurden in 22,4% der Befunde Fremdmaterialien (n = 208) annotiert, die durchschnittliche Prävalenz eines Befundes betrug 13,6% und pro Befundtext kamen im Schnitt 2,8 Befunde vor.

Tabelle 1: Prävalenz der Befunde

Darstellung der Prävalenzen sowie der absoluten Verteilungen der jeweiligen Label innerhalb des Trainings- und Validierungsdatensatzes sowie des Testdatensatzes. Die Zahlen wurden in Bezug auf die Datensatzmenge nach Ausschluss der inkompletten Texte gesetzt und die Prävalenz wurde jeweils auf eine Nachkommastelle gerundet. Es war möglich, multiple Labels pro Befundtext zu vergeben.

Befund	Prävalenz Trainings- und Validierungsdatensatz		Prävalenz Testdatensatz	
Bronchiale Auffälligkeit	9,8%	(n = 447)	9,8%	(n = 91)
Emphysem	17,8%	(n = 812)	17,2%	(n = 160)
Fibrose	8,8%	(n = 401)	8,2%	(n = 76)
Frakturen	12,5%	(n = 571)	13,8%	(n = 128)
Fremdmaterialien	25,2%	(n = 1.150)	22,4%	(n = 208)
Herzgröße	12,4%	(n = 564)	10,3%	(n = 96)
Hiatushernie	2,6%	(n = 119)	2,7%	(n = 25)
Kardiale Stauung	4,2%	(n = 193)	4,7%	(n = 44)
Knochentumoren	3,4%	(n = 154)	7,3%	(n = 24)
Lungenarterienembolie	4,9%	(n = 222)	5,5%	(n = 51)
Lungenrundherd	42,1%	(n = 1.924)	40,9%	(n = 380)
Lungentrauma	0,3%	(n = 15)	0,5%	(n = 5)
Lymphadenopathie	41,8%	(n = 1.910)	41,9%	(n = 389)
Mediastinaler Tumor	3,3%	(n = 156)	4,6%	(n = 43)
Perikarderguss	7,5%	(n = 343)	6,8%	(n = 63)
Pleuraerguss	26,6%	(n = 1.215)	25,6%	(n = 238)
Pleurale Auffälligkeit	2,8%	(n = 126)	3,0%	(n = 28)
Pneumonie	25,3%	(n = 1.155)	25,1%	(n = 233)
Pneumothorax	6,3%	(n = 287)	5,6%	(n = 52)
Ventilationsstörung	32,7%	(n = 1.492)	31,1%	(n = 289)
Weichteilemphysem	4,7%	(n = 215)	2,5%	(n = 23)

3.2 Ergebnisse des AWD-LSTM

3.2.1 Ergebnisse der Textgenerierung

Durch das Prä-Training des AWD-LSTM auf die Prädiktion des nächsten Wortes einer Wortsequenz war es möglich, Texte zu generieren. Das Vortraining erfolgte auf allen vorhandenen Befundtexten einer CT des Thorax, und anschließend war das Modell zur Generierung künstlich erzeugter CT-Texte fähig. Nach Vorgabe einer kleinen Anfangssequenz erzeugte das Modell eine vorgegebene Anzahl an Wörtern. Es zeigte sich, dass das AWD-LSTM in der Lage war, sinnvoll klingende Sätze und Satzfragmente zu bilden. Jedoch wurde kein längerfristiger Zusammenhang zwischen Sätzen am Anfang und am Ende des Textes hergestellt. Das Modell war also imstande, die Struktur der Sprache zu kopieren, ein intelligentes, menschenähnliches Sprachverständnis lag allerdings nicht vor, sodass die Gesamttexte keinen logischen Inhalt besaßen. Hierzu sind in Tabelle 2 einige Beispiele gegeben.

Tabelle 2: Drei mit dem AWD-LSTM-Modell generierte Befundtexte

Es werden exemplarisch drei Befundtexte, die nach dem Prä-Training vom AWD-LSTM generiert werden konnten, dargestellt. Die Texte wurden nachträglich formatiert, der Wortlaut jedoch nicht verändert.

Vorgabe 1	„Klinik, Fragestellung, Rechtfertigende Indikation“
Prädiktion 1	Hämoptysen. Z.n. Lungen – TX **Aufklärung und Einwilligung**: Nach Erhebung der Risikoanamnese mündliche und schriftliche Aufklärung über KM – Applikation und Untersuchungablauf sowie mögliche Risiken der Untersuchung (vgl. auch Aufklärungsbeleg). Schriftliche Einwilligung d. Pat. **Methodik:** Digitale übersichtsradiographien. Nach i.v. Kontrastmittelgabe (80 ml Xenetix 350) Computertomographie des gesamten Thorax in venöser KM – Phase, Rekonstruktion des primären Datensatzes in 0,625 mm Schichtdicke. Anfertigung multiplanarer Rekonstruktionen. Bildschirmbefundung. **Befund**: Es liegen die Voraufnahmen vom 08.12.2017 zum Vergleich vor. Deutlich regredienter und geringer auslaufender Pleuraerguss rechts mit bis zu 1,5 cm Saumbreite. In dieser Aufnahme vom 04.01.2019 gute Belüftung des Lungenparenchyms. Geringe Totalatelektase des linken Unterlappens. Diskrete Minderbelüftung in den apicalen Oberlappen. Lungenemphysem. Kein Pneumothorax. Global vergrößertes Herz, Koronarsklerose. Spondylosteochondrosis intervertebralis der BWS.

	Beurteilung: Kein Nachweis von Lymphknoten- oder Fernmetastasen. Im Wesentlichen unveränderte Darstellung der mosaikartig verteilten Verdichtungen im Mittellappen sowie perihilären Dichteanhebungen, max. 10 mm. Rückgang der osteoplastischen Lungenparenchymverdichtungen,
Vorgabe 2	„Frage nach Lungenarterienembolie“
Prädiktion 2	Infiltrat, Sarkoidose, Sarkoidose Veränderungen. **Aufklärung und Einwilligung:** Nach Erhebung der Risikoanamnese mündliche und schriftliche Aufklärung über die Kontrastmittelapplikation und den Untersuchungsablauf sowie mögliche Risiken der Untersuchung (vergleiche auch Aufklärungsbeleg). Schriftliche Einwilligung d. Pat. **Untersuchungstechnik**: Bolusförmige intravenöse Applikation von 80 ml Xenetix 350 und kontinuierliche axiale Schichtung durch den Thorax in Mehrzeilenspiraltechnik. Multiplanare Rekonstruktionen. DLP 236. **Befund und Beurteilung:** Keine Voraufnahmen. • Glatt berandete Formation mit Verkalkungen in Segment Iva, SE 8 / 28, S3 / IM91, multiplen Makroverkalkungen links lateral mit möglicher Kavernenbildung bei inhomogener Kontrastierung und dorsobasalen Kontrastierung. Verdacht auf Minderperfusion im oberen Mediastinum, Se 4 / 33 DD nachrangig Artefakt. • CT – grafisch kein Anhalt für ein Malignom. Einzelne kleine bipulmonale Mikronoduli. • Geringer dorsobasaler Pericarderguss mit maximal 8 mm Saumbreite. Rechter Ventrikel regelrecht. Kein Hinweis auf pneumonische Infiltrate/ malignitätssuspekte pulmonale Einblutung. Kein Pleuraerguss. Kein Pneumothorax. Akzentuierte axilläre Lymphknoten beidseits. Kein Nachweis malignitätssuspekter ossärer Destruktionen
Vorgabe 3	„Keine Voraufnahmen zum Vergleich vorliegend“
Prädiktion 3	• Verdichtete Resektionshöhle großer TEP rechts. PTCD bei Metallartefakt und Dichtewerte I88 Serie 4 Image 227, Serie 5 Bild 99. • Schmaler Erguss links mit einer Saumbreite von 32 mm. • sd – Pleuraerguss rechts mit einer Saumbreite von ca. 20 mm. Kein Perikarderguss. • Kein Nachweis einer hilären, mediastinalen und axillären Lymphadenopathie. • Leberzysten, teils zystisch, DD a.e. reaktiv.

	CT Thorax vom 12.02.2010: **Klinik**: Pleuraerguss links bei ausgeprägter Struma nodosa. Lymphknoten rechts hilär. Supraaortale Gefäße normkalibrig, normkalibrig, regelrecht kontrastiert. Lymphknoten nativ. Hili cervical nativ nicht ausreichend beurteilbar. Nachweis kleinerer Lymphknoten retrosternal rechts paratracheal. Verdacht auf Nebennierenmetastase links bei Peritonealkarzinose. Ausgedehnte Raumforderung der rechten Nebenniere. Nebennierenmetastase links und Nebennierenmetastase rechts. Lebermetastasen. Rechte Nebenniere schlank. Raumforderung Parenchym links medial subkutan, parakardial bis zum Prominenz reichend. Raumforderung Segment 3 links 9 cm Durchmesser. Links dorsal der Vena brachiocephalica sowie der Arteria pulmonalis sinistra – DD Tumorinfiltration. In den mitabgebildeten Oberbauchanteilen bekannte osteolytische bzw. osteolytische Metastasen in der gesamten abgebildeten

3.2.2 Ergebnisse der Textklassifikation

Anschließend konnte das AWD-LSTM zur Textklassifikation verwendet werden, wobei es in Bezug auf die meisten Befunde eine sehr gute Performance erreichte. Eine genaue Darstellung der ermittelten Metriken aller 21 Befunde ist in Tabelle 3 dargestellt.

Der Befund, für den mit 0,999 die höchste Genauigkeit bestimmt werden konnte, war das Lungentrauma. Die AUC betrug ebenfalls 0,999, allerdings wurde aufgrund der Sensitivität, die 1,000 betrug, und einem positiven prädiktiven Wert von 0,500 lediglich ein F1-Wert von 0,667 erreicht. Da das Lungentrauma allerdings der mit Abstand seltenste Befund war, waren diese Werte nur bedingt zur Beurteilung der Leistung des AWD-LSTM geeignet.

Die zweithöchste Genauigkeit hingegen wurde mit 0,996 für den Pleuraerguss erzielt, der mit einer Prävalenz von 26,6% in mehr als jedem vierten Befund und somit recht häufig vorkam. Für den Pleuraerguss war anders als beim Lungentrauma neben der AUC, welche 0,994 betrug, durch eine hohe Sensitivität (R = 0,992) sowie sehr guten positiven prädiktiven Wert (PPV = 0,992) auch der F1-Wert (F1 = 0,992) beachtlich. Diese Werte ergaben eine äußerst hohe Klassifikationsleistung und spiegelten zusammen die beste Performance des AWD-LSTM wider.

Neben dem Pleuraerguss konnte auch das Vorkommen von Fremdmaterialien vom AWD-LSTM sehr zuverlässig erkannt werden. So wurden für das Modell bei diesem Befund eine Genauigkeit von 0,990, eine Sensitivität von 0,991, ein positiver prädiktiver Wert von 0,970, ein F1-Wert von 0,980 und eine AUC von 0,990 ermittelt. Darüber hinaus wurden auch für die Befunde Emphysem, Herzgröße und Lymphadenopathie in allen Metriken Werte über 0,950 erreicht, sodass das AWD-LSTM auch diese Befunde sehr sicher kategorisieren konnte.

Insgesamt betrug die Genauigkeit des AWD-LSTM für sämtliche Befunde über 0,960, sodass das Modell daher in mindestens 96 von 100 Vorhersagen richtig lag. Der niedrigste Wert in dieser Hinsicht wurde beim Weichteilemphysem mit 0,968 ermittelt. Dieser Befund erzielte auch für die anderen Metriken jeweils keine sehr hohen Ergebnisse (R = 0,512, PPV = 0,733, F1 = 0,603, AUC = 0,751).

Das AWD-LSTM tat sich bei der Klassifikation einiger anderer Befunde ebenfalls schwer. So lag etwa der F1-Wert für die pleurale Auffälligkeit lediglich bei 0,091, was aus der geringen Sensitivität (R = 0,048) resultierte. Obgleich die Werte für die Genauigkeit, welche 0,978 betrug, und den positiven prädiktiven Wert (PPV = 1,000) sehr gut waren, ergab auch die AUC nur 0,523.

Insgesamt erlangte das Modell für 13 der 21 Befunde eine Sensitivität über 0,900 und konnte daher den Großteil der Befunde sehr zielsicher erkennen. Für drei Befunde war die Sensitivität allerdings nur unter 0,300. Das betraf neben der pleuralen Auffälligkeit die Hiatushernie (R = 0,120) und den mediastinalen Tumor (R = 0,267), welche aufgrund der niedrigen Sensitivitäten auch nur geringe F1-Werte erreichten (F1 = 0,214 bzw. 0,421).

Der positive prädiktive Wert war bei 18 von 21 Befunden über 0,930. Somit hatte das AWD-LSTM insgesamt eine sehr hohe Präzision und lag bei der positiven Klassifikation fast immer richtig. Der F1-Wert, welcher Sensitivität und PPV kombiniert, erreichte bei zwölf Befunden Werte über 0,900.

Für 14 von 21 Befunden konnten zudem AUC-Werte über 0,900 bestimmt werden, was somit eine recht hohe Entscheidungskraft des Klassifikationssystems abbildete.

Tabelle 3: Klassifikationsleistung des AWD-LSTM (Validierungsdatensatz)

Es wurden die Metriken Genauigkeit, Sensitivität, positiver prädiktiver Wert (PPV), F1-Wert sowie Area Under the Curve (AUC) erhoben, um die Performance des Modells in Bezug auf den Validierungsdatensatz zu evaluieren. Alle Metriken wurden jeweils auf drei Nachkommastellen gerundet.

	Genauigkeit	Sensitivität	PPV	F1-Wert	AUC
Bronchiale Auffälligkeit	0,974	0,878	0,859	0,868	0,931
Emphysem	0,987	0,962	0,972	0,967	0,977
Fibrose	0,972	0,750	0,957	0,840	0,873
Frakturen	0,988	0,943	0,966	0,954	0,969
Fremdmaterialien	0,990	0,991	0,970	0,980	0,990
Herzgröße	0,990	0,955	0,964	0,960	0,975
Hiatushernie	0,976	0,120	1,000	0,214	0,560
Kardiale Stauung	0,991	0,800	0,966	0,875	0,899
Knochentumoren	0,988	0,686	1,000	0,814	0,843
Lungenarterienembolie	0,996	0,933	0,977	0,955	0,966
Lungenrundherd	0,974	0,988	0,954	0,971	0,975
Lungentrauma	0,999	1,000	0,500	0,667	0,999
Lymphadenopathie	0,977	0,978	0,965	0,972	0,977
Mediastinaler Tumor	0,976	0,267	1,000	0,421	0,633
Perikarderguss	0,993	0,968	0,938	0,953	0,982
Pleuraerguss	0,996	0,992	0,992	0,992	0,994
Pleurale Auffälligkeit	0,978	0,048	1,000	0,091	0,524
Pneumonie	0,979	0,954	0,966	0,960	0,971
Pneumothorax	0,996	0,982	0,947	0,964	0,989
Ventilationsstörung	0,977	0,949	0,983	0,966	0,970
Weichteilemphysem	0,968	0,512	0,733	0,603	0,751

3.3 Ergebnisse der Transformer-Modelle

Durch eine andere Strategie des Prä-Trainings der verwendeten Transformer-Modelle (Masked Word Prediction) eigneten sich diese im Gegensatz zum AWD-LSTM nicht für die Generierung von Texten.

3.3.1 BERT: Ergebnisse der Textklassifikation

Ähnlich wie das AWD-LSTM konnte auch das BERT-Modell in Bezug auf die meisten annotierten Befunde sehr gute Leistungen erbringen. So war bei diesem Modell die Genauigkeit ebenso wie beim AWD-LSTM konstant über 0,950 (siehe Tabelle 4).

Durchgehend besonders gute Werte konnte das BERT-Modell beim Erkennen eines Emphysems erzielen. Hier konnte nicht nur für die Genauigkeit mit 0,994 ein hoher Wert ermittelt werden, sondern auch für die anderen Metriken (R = 0,969, F1 = 0,984, AUC = 0,984). Insbesondere der positive prädiktive Wert von 1,000 war hierbei hervorzuheben, welcher ausgedrückt hat, dass das Modell, wenn es einen Befund als vorhanden erkannt hat, stets richtig lag. Für das Emphysem, das somit die beste Leistung des Transformers darstellte, konnten daher insgesamt äußerst hohe Ergebnisse erreicht werden.

Auch die Frage, ob eine Störung der Ventilation vorlag, konnte das Modell sehr zuverlässig beurteilen und daher eine Genauigkeit von 0,985, eine Sensitivität von 0,971, einen positiven prädiktiven Wert von 0,984, einen F1-Wert von 0,977 sowie eine AUC von 0,981 erzielen. Darüber hinaus wurden auch für die Befunde Fremdmaterialien, Lungenrundherd, Lymphadenopathie und Perikarderguss durchgängig sehr gute Werte von über 0,950 in Bezug auf alle Metriken erreicht.

Die größten Schwierigkeiten hatte das BERT-Modell beim Befund Lungentrauma. Die AUC lag nur bei 0,500 und Sensitivität, positiver prädiktiver Wert und F1-Wert betrugen jeweils 0,000, was bedeutet, dass das BERT-Modell diesen Befund generell nie erkannt hat. Da das Lungentrauma im Datensatz allerdings eine äußerst niedrige Prävalenz besaß, wurde aufgrund der vielen richtig negativen Befunde dennoch mit 0,999 wie auch beim AWD-LSTM der höchste Genauigkeitswert erreicht.

Neben dem Lungentrauma, welches nie als positiv erkannt werden konnte, gab es noch einige andere Befunde, die das BERT-Modell nicht gut detektieren bzw. vorhersagen konnte. So lag beispielsweise der F1-Wert für die pleurale Auffälligkeit lediglich bei 0,083, was durch eine niedrige Sensitivität von 0,043 bedingt war. Zwar war der positive prädiktive Wert gleich 1,000 und zeigte, dass das Modell bei der Vorhersage eines Befundes als vorhanden stets richtig lag, allerdings spiegelte die niedrige Sensitivität wider, dass nur in sehr wenigen Fällen das Modell den Befund überhaupt erkannt hat. Dementsprechend war auch die AUC nur bei 0,522.

Für einige andere Befunde konnte das Modell ebenfalls nur recht niedrige Werte erzielen. Das betraf vor allem den mediastinalen Tumor (R = 0,071, PPV = 0,750, F1 = 0,130, AUC = 0,535), Knochentumoren (R = 0,214, PPV = 0,667, F1 = 0,324, AUC = 0,605) und das Weichteilemphysem (R = 0,442, PPV = 0,792, F1 = 0,567, AUC = 0,718). Zu beachten ist allerdings, dass diese drei Befunde ähnlich wie das Lungentrauma eine sehr geringe

Prävalenz aufwiesen und somit aufgrund der vielen richtig Negativen eine gute Genauigkeit erreichten.

Insgesamt konnte bei dem BERT-Modell im Validierungsdatensatz für 14 der 21 Befunde eine Sensitivität über 0,900 errechnet werden und auch der positive prädiktive Wert lag für 14 der 21 Befunde über 0,900. Ebenso erreichte auch das kombinierte Maß der beiden Metriken, der F1-Wert, in 14 von 21 Befunden mehr als 0,900. Zudem erzielten 16 von 21 Befunden Werte über 0,900 bei der AUC.

Tabelle 4: Klassifikationsleistung des BERT-Modells (Validierungsdatensatz)

Es wurden die Metriken Genauigkeit, Sensitivität, positiver prädiktiver Wert (PPV), F1-Wert sowie Area Under the Curve (AUC) erhoben, um die Performance des Modells in Bezug auf den Validierungsdatensatz zu evaluieren. Alle Metriken wurden jeweils auf drei Nachkommastellen gerundet.

	Genauigkeit	Sensitivität	PPV	F1-Wert	AUC
Bronchiale Auffälligkeit	0,982	0,970	0,881	0,923	0,977
Emphysem	0,994	0,969	1,000	0,984	0,984
Fibrose	0,990	0,935	0,947	0,941	0,965
Frakturen	0,990	0,983	0,944	0,963	0,987
Fremdmaterialien	0,985	0,974	0,966	0,970	0,981
Herzgröße	0,989	0,990	0,919	0,953	0,990
Hiatushernie	0,997	0,950	0,905	0,927	0,974
Kardiale Stauung	0,985	0,838	0,795	0,816	0,914
Knochentumoren	0,972	0,214	0,667	0,324	0,605
Lungenarterienembolie	0,985	0,897	0,778	0,833	0,943
Lungenrundherd	0,962	0,962	0,953	0,957	0,962
Lungentrauma	0,999	0,000	0,000	0,000	0,500
Lymphadenopathie	0,960	0,958	0,953	0,956	0,960
Mediastinaler Tumor	0,956	0,071	0,750	0,130	0,535
Perikarderguss	0,994	0,954	0,969	0,961	0,976
Pleuraerguss	0,980	0,988	0,943	0,965	0,983
Pleurale Auffälligkeit	0,976	0,043	1,000	0,083	0,522
Pneumonie	0,970	0,962	0,926	0,943	0,967
Pneumothorax	0,993	0,929	0,963	0,945	0,963
Ventilationsstörung	0,985	0,971	0,984	0,977	0,981
Weichteilemphysem	0,968	0,442	0,792	0,567	0,718

3.3.2 DistilBERT: Ergebnisse der Textklassifikation

Die Metriken des DistilBERT-Modells waren insgesamt etwas niedriger als bei den anderen beiden Modellen. Wie in Tabelle 5 zu sehen ist, konnten dennoch für viele Befunde gute Werte zustande gebracht werden.

Ein Befund, bei dem DistilBERT eine sehr gute Leistung erbringen konnte, war der Pleuraerguss. Hier lag der Wert für die Genauigkeit bei 0,962, der F1-Wert ergab 0,936, bedingt durch die Sensitivität (R = 0,961) sowie den positiven prädiktiven Wert (PPV = 0,911), und die AUC betrug 0,962.

DistilBERT war auch beim Kategorisieren der Befunde Emphysem, Fremdmaterialien, Lymphadenopathie sowie Ventilation recht zuverlässig. Für die Genauigkeit und AUC wurden bei diesen Befunden stets Werte über 0,900 bestimmt und auch der aussagekräftige F1-Wert lag konstant bei etwa 0,900.

Die Genauigkeit betrug insgesamt durchgängig mehr als 0,900. Die einzige Ausnahme bildete hierbei der Lungenrundherd, der einen Wert von lediglich 0,879 besaß. Wie eine AUC von 0,877 zeigte, waren die Metriken von diesem Befund zwar etwas niedriger (R = 0,866, PPV = 0,853, F1 = 0,859), aber recht ausbalanciert.

Genau wie BERT konnte allerdings auch DistilBERT einige Befunde nur sehr schlecht klassifizieren. So wurden die Befunde Hiatushernie, Lungentrauma sowie mediastinaler Tumor nie korrekt erkannt, weshalb die Sensitivität, der positive prädiktive Wert sowie der F1-Wert dementsprechend bei 0,000 lagen. Da es sich allerdings um seltene Befunde handelte, wurde für die Genauigkeit dennoch stets ein Wert über 0,960 ermittelt. Die AUC hingegen betrug nur 0,499 bzw. 0,500.

Auch für einige andere Befunde lagen die Werte im unteren Bereich. So konnte beispielsweise bei dem Befund pleurale Auffälligkeit eine Sensitivität von lediglich 0,033 erreicht werden. Trotz eines positiven prädiktiven Wertes von 1,000 betrugen daher der F1-Wert und die AUC ebenfalls nur 0,065 bzw. 0,517.

Zusätzlich konnten bei den Befunden Fibrose, kardiale Stauung, Knochentumoren, Lungenarterienembolie und Weichteilemphysem nur F1-Werte unter 0,700 errechnet werden. Das lag vor allem an einer jeweils niedrigen Sensitivität kleiner als 0,600. Auch die AUC-Werte lagen für diese Befunde nur zwischen 0,500 und 0,800. Für die restlichen

Befunde konnten hingegen F1-Werte von mindestens 0,800 sowie AUC-Werte von minimal 0,850 und Sensitivitätswerte über 0,700 gemessen werden.

Die Sensitivitäten waren insgesamt geringer als die der beiden anderen Modelle. So wurde lediglich für zwei Befunde eine Sensitivität von über 0,900 erzielt und bei neun weiteren Befunden erreichte die Sensitivität immerhin 0,750. Auch bei den anderen Metriken waren die Werte der Befunde in der Gesamtbetrachtung geringfügig niedriger als bei dem AWD-LSTM und dem BERT-Modell. So konnte bei dem positiven prädiktiven Wert der Wert 0,900 immerhin bei acht Befunden erbracht werden und weitere sechs Befunde erzielten über 0,800. Der F1-Wert überschritt den Wert 0,900 nur bei drei Befunden und die AUC betrug in nur fünf von 21 Befunden über 0,900.

Tabelle 5: Klassifikationsleistung des DistilBERT-Modells (Validierungsdatensatz)

Es wurden die Metriken Genauigkeit, Sensitivität, positiver prädiktiver Wert (PPV), F1-Wert sowie Area Under the Curve (AUC) erhoben, um die Performance des Modells in Bezug auf den Validierungsdatensatz zu evaluieren. Alle Metriken wurden jeweils auf drei Nachkommastellen gerundet.

	Genauigkeit	Sensitivität	PPV	F1-Wert	AUC
Bronchiale Auffälligkeit	0,971	0,776	0,938	0,849	0,885
Emphysem	0,967	0,888	0,922	0,904	0,936
Fibrose	0,939	0,559	0,788	0,654	0,771
Frakturen	0,956	0,791	0,887	0,836	0,887
Fremdmaterialien	0,945	0,927	0,867	0,896	0,939
Herzgröße	0,959	0,737	0,923	0,820	0,864
Hiatushernie	0,981	0,000	0,000	0,000	0,500
Kardiale Stauung	0,968	0,477	0,778	0,592	0,735
Knochentumoren	0,977	0,292	0,636	0,400	0,644
Lungenarterienembolie	0,968	0,519	0,871	0,651	0,757
Lungenrundherd	0,879	0,866	0,853	0,859	0,877
Lungentrauma	0,999	0,000	0,000	0,000	0,500
Lymphadenopathie	0,916	0,889	0,909	0,899	0,912
Mediastinaler Tumor	0,965	0,000	0,000	0,000	0,499
Perikarderguss	0,982	0,794	0,964	0,871	0,896
Pleuraerguss	0,962	0,961	0,911	0,936	0,962
Pleurale Auffälligkeit	0,968	0,033	1,000	0,065	0,517
Pneumonie	0,915	0,818	0,844	0,831	0,883
Pneumothorax	0,978	0,797	0,879	0,836	0,894
Ventilationsstörung	0,947	0,896	0,940	0,917	0,934
Weichteilemphysem	0,965	0,476	0,667	0,556	0,732

3.4 Ergebnisse des Testdatensatzes

3.4.1 Ergebnisse des AWD-LSTM

Das AWD-LSTM konnte im Testdatensatz, welcher ausschließlich Daten enthielt, die dem Modell nicht vom Training bekannt waren, fast durchgängig eine äußerst gute Leistung erzielen. In Tabelle 6 ist eine detaillierte Darstellung aller erhobenen Metriken gegeben.

Die zwei Befunde, für die mit jeweils 0,997 die höchsten Genauigkeiten bestimmt wurden, waren zum einen das Lungentrauma und zum anderen die Lungenarterienembolie. Die Lungenarterienembolie erzielte auch bei den anderen Metriken gute Werte und erreichte eine Sensitivität von 0,973, einen positiven prädiktiven Wert von 0,973, einen F1-Wert von 0,972 und eine AUC von 0,986. Das Lungentrauma hingegen erreichte für die anderen Metriken mit einer Sensitivität von 0,200, einem PPV sowie einer AUC von 0,600 und einem F1-Wert von 0,300 nur niedrige Werte, wobei zu beachten ist, dass es sich um den Befund mit der deutlich geringsten Prävalenz handelte.

Der hohe Einfluss der Prävalenz zeichnete sich auch in den übrigen Ergebnissen des AWD-LSTM ab. So konnten Befunde mit einer hohen Prävalenz wie der Pleuraerguss (Genauigkeit = 0,993, R = 0,984, PPV = 0,990, F1 = 0,987, AUC = 0,990) oder auch die Lymphadenopathie (Genauigkeit = 0,976, R = 0,981, PPV = 0,961, F1 = 0,971, AUC = 0,976) durchgehend sehr hohe Werte erzielen.

Demgegenüber standen beispielsweise die Befunde Hiatushernie (Genauigkeit = 0,975, R = 0,051, PPV = 1,000, F1 = 0,097, AUC = 0,525) und pleurale Auffälligkeit (Genauigkeit = 0,974, R = 0,048, PPV = 1,000, F1 = 0,091, AUC = 0,524), welche eine sehr geringe Prävalenz in der Gesamtmenge der Befunde hatten und dementsprechend nur niedrige Werte erreichen konnten. Insbesondere die Sensitivitäten waren für diese Befunde sehr gering und führten daher auch zu niedrigen F1-Werten. Die Metriken zeigten, dass das Modell bei einer positiven Klassifikation zwar immer korrekt lag, diese Befunde aber nur in den wenigsten Fällen überhaupt erkannt hat.

Insgesamt ergaben alle Genauigkeitswerte mehr als 0,960 und in Bezug auf den F1-Wert erreichten alle Befunde, welche in mehr als 0,5% der Befundtexte annotiert wurden, Werte über 0,800.

Bei zwölf der 21 Befunde wurde eine Sensitivität über 0,900 erreicht und auch der positive prädiktive Wert war sehr gut. So lag das Modell bei 18 von 21 Befunden in mindestens neun von zehn Fällen richtig, wenn es einen Befund positiv vorhersagte. Die AUC betrug in 14 von 21 Fällen über 0,900.

Insgesamt konnten für das AWD-LSTM im Testdatensatz ähnlich gute Metriken wie im Validierungsdatensatz ermittelt werden. Es war somit nicht übermäßig an den Trainings- und Validierungsdatensatz angepasst und erbrachte auch bei Befunden, die nicht vom Training des Modells bekannt waren, gute Leistungen.

Tabelle 6: Klassifikationsleistung des AWD-LSTM (Testdatensatz)

Es wurden die Metriken Genauigkeit, Sensitivität, positiver prädiktiver Wert (PPV), F1-Wert sowie Area Under the Curve (AUC) erhoben, um die Performance des Modells in Bezug auf den Validierungsdatensatz zu evaluieren. Alle Metriken wurden jeweils auf drei Nachkommastellen gerundet.

	Genauigkeit	Sensitivität	PPV	F1-Wert	AUC
Bronchiale Auffälligkeit	0,975	0,867	0,879	0,873	0,927
Emphysem	0,989	0,969	0,971	0,970	0,981
Fibrose	0,978	0,812	0,928	0,866	0,903
Frakturen	0,989	0,944	0,969	0,956	0,970
Fremdmaterialien	0,985	0,981	0,960	0,970	0,983
Herzgröße	0,989	0,962	0,950	0,956	0,978
Hiatushernie	0,975	0,051	1,000	0,097	0,525
Kardiale Stauung	0,986	0,710	0,951	0,813	0,854
Knochentumoren	0,987	0,654	0,952	0,775	0,826
Lungenarterienembolie	0,997	0,973	0,973	0,972	0,986
Lungenrundherd	0,970	0,975	0,955	0,965	0,971
Lungentrauma	0,997	0,200	0,600	0,300	0,600
Lymphadenopathie	0,976	0,981	0,961	0,971	0,976
Mediastinaler Tumor	0,976	0,321	0,926	0,476	0,660
Perikarderguss	0,994	0,955	0,967	0,961	0,976
Pleuraerguss	0,993	0,984	0,990	0,987	0,990
Pleurale Auffälligkeit	0,974	0,048	1,000	0,091	0,524
Pneumonie	0,982	0,958	0,969	0,964	0,974
Pneumothorax	0,991	0,947	0,909	0,928	0,971
Ventilationsstörung	0,977	0,958	0,972	0,965	0,972
Weichteilemphysem	0,967	0,512	0,703	0,592	0,751

3.4.2 Ergebnisse des BERT-Modells

Auch für das BERT-Modell konnten im Testdatensatz bei den meisten Befunden gute Ergebnisse bestimmt werden, wie in Tabelle 7 zu sehen ist.

Der Befund Pleuraerguss erreichte bei allen Parametern hohe Werte (Genauigkeit = 0,976, R = 0,979, PPV = 0,932, AUC = 0,978) und insbesondere beim F1-Wert wurde mit 0,955 das höchste Ergebnis ermittelt. Auch die Befunde Emphysem (Genauigkeit = 0,976, R = 0,944, PPV = 0,921, F1 = 0,932, AUC = 0,963) und Ventilationsstörung (Genauigkeit = 0,972, R = 0,965, PPV = 0,945, F1 = 0,955, AUC = 0,970) erbrachten konstant gute Metriken.

Es zeigte sich allerdings erneut, dass Befunde mit einer sehr geringen Prävalenz zwar mitunter die höchsten Genauigkeiten erzielten, aber insbesondere beim F1-Wert schlecht abschnitten. So erreichten die Befunde Knochentumoren, Lungentrauma, mediastinaler Tumor und pleurale Auffälligkeit zwar alle eine Genauigkeit über 0,950, allerdings betrugen die AUC-Werte nur zwischen 0,500 und 0,600 und sowohl die Sensitivität als auch der F1-Wert waren stets kleiner als 0,200, was zeigte, dass diese seltenen Befunde kaum vom Modell erkannt wurden.

Insgesamt wurde für sämtliche Befunde eine Genauigkeit über 0,890 errechnet und von drei Befunden abgesehen erzielten alle Befunde eine Genauigkeit von mindestens 0,950. Die F1-Werte lagen hingegen nur bei vier Befunden über 0,900, was durch die Sensitivitäten und positiven prädiktiven Werte bedingt wurde, die etwas niedriger als noch im Validierungsdatensatz waren.

Für das BERT-Modell wurde bei sechs Befunden eine Sensitivität von größer als 0,900 bestimmt und bei acht weiteren Befunden eine über 0,800. Der positive prädiktive Wert war bei vier Befunden über 0,900 und bei weiteren neun Befunden über 0,800. Die AUC lag in elf von 21 Fällen über 0,900.

Insgesamt waren die Metriken des BERT-Modells im Testdatensatz geringfügig niedriger als im Validierungsdatensatz, sodass unter Umständen ein leichtes Overfitting vorgelegen haben könnte.

Tabelle 7: Klassifikationsleistung des BERT-Modells (Testdatensatz)

Es wurden die Metriken Genauigkeit, Sensitivität, positiver prädiktiver Wert (PPV), F1-Wert sowie Area Under the Curve (AUC) erhoben, um die Performance des Modells in Bezug auf den Validierungsdatensatz zu evaluieren. Alle Metriken wurden jeweils auf drei Nachkommastellen gerundet.

	Genauigkeit	Sensitivität	PPV	F1-Wert	AUC
Bronchiale Auffälligkeit	0,962	0,846	0,786	0,815	0,912
Emphysem	0,976	0,944	0,921	0,932	0,963
Fibrose	0,975	0,829	0,863	0,847	0,909
Frakturen	0,969	0,875	0,896	0,885	0,929
Fremdmaterialien	0,950	0,885	0,893	0,889	0,927
Herzgröße	0,954	0,906	0,719	0,802	0,933
Hiatushernie	0,991	0,840	0,840	0,840	0,918
Kardiale Stauung	0,972	0,636	0,718	0,675	0,812
Knochentumoren	0,975	0,083	0,667	0,148	0,541
Lungenarterienembolie	0,982	0,804	0,854	0,828	0,898
Lungenrundherd	0,895	0,905	0,847	0,875	0,896
Lungentrauma	0,995	0,000	0,000	0,000	0,500
Lymphadenopathie	0,915	0,910	0,889	0,900	0,914
Mediastinaler Tumor	0,954	0,000	0,000	0,000	0,500
Perikarderguss	0,980	0,810	0,879	0,843	0,901
Pleuraerguss	0,976	0,979	0,932	0,955	0,978
Pleurale Auffälligkeit	0,973	0,107	1,000	0,194	0,554
Pneumonie	0,906	0,863	0,785	0,822	0,892
Pneumothorax	0,977	0,750	0,830	0,788	0,870
Ventilationsstörung	0,972	0,965	0,945	0,955	0,970
Weichteilemphysem	0,982	0,565	0,650	0,605	0,779

3.4.3 Ergebnisse des DistilBERT-Modells

Wie die detaillierte Darstellung der Metriken in Tabelle 8 zeigt, entsprach die Performance des DistilBERT-Modells im Testdatensatz zum größten Teil der des BERT-Modells.

So war auch bei diesem Modell der Pleuraerguss der Befund mit dem höchsten F1-Wert, der bei 0,949 lag. Auch in Bezug auf die anderen Beurteilungsparameter ergab die Evaluation hohe Werte mit einer Genauigkeit von 0,973, einer Sensitivität von 0,971, einem PPV von 0,928 und einer AUC von 0,972.

Wie bei dem BERT-Modell waren die Ergebnisse der Befunde Emphysem (Genauigkeit = 0,976, R = 0,938, PPV = 0,926, F1 = 0,932, AUC = 0,961) sowie Ventilationsstörung (Genauigkeit = 0,945, R = 0,913, PPV = 0,909, F1 = 0,911, AUC = 0,936) ebenfalls konstant über 0,900.

Ebenso konkordant zum BERT-Modell hatte auch das DistilBERT-Modell Schwierigkeiten, die Befunde Knochentumoren, Lungentrauma, mediastinaler Tumor sowie pleurale Auffälligkeit zu erkennen. Diese Befunde besaßen alle trotz einer Genauigkeit über 0,950 jeweils eine AUC von 0,500 bis 0,600 sowie einen F1-Wert unter 0,300, was dem Grund geschuldet ist, dass diese Befunde kaum detektiert wurden, wie die geringen Sensitivitäten von stets unter 0,200 zeigten.

Zusätzlich konnte auch der Befund Hiatushernie nie erkannt werden, was sich in den Werten einer Genauigkeit von 0,973, einem F1-Wert, einer Sensitivität sowie einem PPV von jeweils 0,000 und einer AUC = 0,500 bemerkbar machte. Das stellte somit einen Unterschied zum BERT-Modell dar, das bei diesem Befund recht gute Werte erreichen konnte (Genauigkeit = 0,991, R = 0,840, PPV = 0,840, F1 = 0,840, AUC = 0,918).

Insgesamt wurde für alle Befunde eine Genauigkeit von mindestens 0,880 errechnet. Wie auch im Validierungsdatensatz wurde für fünf von 21 Befunden eine AUC über 0,900 bestimmt. Die Sensitivität war bei vier Befunden über 0,900 und bei vier weiteren über 0,800. Bei neun Befunden wurden positive prädiktive Werte über 0,900 ermittelt, allerdings wurde diese Zahl in Bezug auf den F1-Wert nur von drei Befunden erzielt.

Im Testdatensatz erreichte das DistilBERT-Modell insgesamt im Vergleich mit den beiden anderen Modellen die erwartungsgemäß niedrigsten Metriken. Allerdings unterschieden sich die Werte nicht sonderlich vom Validierungsdatensatz, sodass keine Überanpassung des Modells vorlag.

Tabelle 8: Klassifikationsleistung des DistilBERT-Modells (Testdatensatz)

Es wurden die Metriken Genauigkeit, Sensitivität, positiver prädiktiver Wert (PPV), F1-Wert sowie Area Under the Curve (AUC) erhoben, um die Performance des Modells in Bezug auf den Validierungsdatensatz zu evaluieren. Alle Metriken wurden jeweils auf drei Nachkommastellen gerundet.

	Genauigkeit	Sensitivität	PPV	F1-Wert	AUC
Bronchiale Auffälligkeit	0,963	0,681	0,925	0,785	0,838
Emphysem	0,976	0,938	0,926	0,932	0,961
Fibrose	0,950	0,684	0,703	0,693	0,829
Frakturen	0,966	0,820	0,921	0,868	0,905
Fremdmaterialien	0,931	0,861	0,836	0,848	0,906
Herzgröße	0,968	0,781	0,893	0,833	0,885
Hiatushernie	0,973	0,000	0,000	0,000	0,500
Kardiale Stauung	0,977	0,545	0,960	0,696	0,772
Knochentumoren	0,974	0,167	0,500	0,250	0,581
Lungenarterienembolie	0,975	0,588	0,938	0,723	0,793
Lungenrundherd	0,899	0,900	0,859	0,879	0,899
Lungentrauma	0,995	0,000	0,000	0,000	0,500
Lymphadenopathie	0,889	0,830	0,897	0,862	0,881
Mediastinaler Tumor	0,955	0,023	1,000	0,045	0,512
Perikarderguss	0,973	0,698	0,880	0,779	0,846
Pleuraerguss	0,973	0,971	0,928	0,949	0,972
Pleurale Auffälligkeit	0,971	0,036	1,000	0,069	0,518
Pneumonie	0,926	0,841	0,860	0,850	0,898
Pneumothorax	0,977	0,750	0,830	0,788	0,870
Ventilationsstörung	0,945	0,913	0,909	0,911	0,936
Weichteilemphysem	0,982	0,565	0,650	0,605	0,779

3.4.4 Rater-Agreement

Das Interrater-Agreement lag für die verschiedenen Befunde zwischen 0,35 und 0,93, wobei es durchschnittlich 0,69 betrug (siehe Tabelle 9 für die Werte aller Befunde).

Das höchste Agreement wurde mit 0,93 für den Befund Lungenarterienembolie ermittelt. Auch die Befunde bronchiale Auffälligkeit, Emphysem, kardiale Stauung, Lungentrauma, Perikarderguss, Pleuraerguss und Ventilationsstörung erzielten Kappa-Werte über 0,8 und gaben somit eine fast perfekte Übereinstimmung wieder.

Die Befunde Fibrose, Frakturen Fremdmaterialien, Herzgröße, Knochentumoren, Lungenrundherd, Lymphadenopathie, Pneumonie, Pneumothorax sowie Weichteil-

emphysem erreichten Kappa-Werte zwischen 0,60 und 0,80 und damit eine starke Übereinstimmung.

Das niedrigste Agreement wurde für die Befunde Hiatushernie, mediastinaler Tumor und pleurale Auffälligkeit mit Werten von 0,56 bzw. 0,38 und 0,35 bestimmt, was eine eher moderate bis schwache Übereinstimmung widerspiegelte.

Das Intrarater-Agreement bewegte sich im Bereich zwischen 0,92 und 1,00 und betrug durchschnittlich 0,96. Das höchste Agreement von 1,00 wurde hierbei beim Befund Lungentrauma errechnet, welches eine vollkommene Übereinstimmung darstellte. Auch bei den restlichen Befunden wurden fast vollkommene Übereinstimmungen erzielt.

Tabelle 9: Interrater- und Intrarater-Agreement

*Es wurde für jeden der 21 verschiedenen Befunde jeweils Cohens Kappa (κ) sowie das 95-prozentige Konfidenzintervall (KI) erhoben, wobei die Werte auf zwei Nachkommastellen gerundet wurden. *Bei vollkommener Übereinstimmung war keine KI-Berechnung möglich.*

	Interrater-κ	Interrater-KI	Intrarater-κ	Intrarater-KI
Bronchiale Auffälligkeit	0,82	0,75 - 0,91	0,96	0,94 - 0,99
Emphysem	0,80	0,74 - 0,87	0,94	0,91 - 0,97
Fibrose	0,73	0,64 - 0.84	0,95	0,92 - 0,99
Frakturen	0,71	0,65 - 0,82	0,95	0,93 - 0,97
Fremdmaterialien	0,76	0,69 - 0,84	0,97	0,94 - 1,00
Herzgröße	0,74	0,64 - 0,87	0,94	0,90 - 0,97
Hiatushernie	0,56	0,31 - 0,88	0,96	0,92 - 1,00
Kardiale Stauung	0,82	0,67 - 0,91	0,98	0,95 - 1,00
Knochentumoren	0,68	0,52 - 0,83	0,94	0,88 - 1,00
Lungenarterienembolie	0,93	0,86 - 1,00	0,98	0,96 - 1,00
Lungenrundherd	0,69	0,64 - 0,75	0,95	0,93 - 0,97
Lungentrauma	0,86	0,71 - 1,00	1,00*	-
Lymphadenopathie	0,74	0,69 - 0,80	0,96	0,94 - 0,98
Mediastinaler Tumor	0,38	0,16 - 0,65	0,93	0,88 - 1,00
Perikarderguss	0,89	0,81 - 0,99	0,97	0,95 - 1,00
Pleuraerguss	0,83	0,78 - 0,90	0,96	0,94 - 0,98
Pleurale Auffälligkeit	0,35	0,21 - 0,51	0,96	0,93 - 1,00
Pneumonie	0,68	0,61 - 0,75	0,92	0,90 - 0,95
Pneumothorax	0,73	0,63 - 0,88	0,95	0,92 - 1,00
Ventilationsstörung	0,82	0,77 - 0,88	0,96	0,95 - 0,98
Weichteilemphysem	0,68	0,44 - 0,93	0,93	0,87 - 1,00

4. Diskussion

4.1 Zusammenfassung der Ergebnisse

4.1.1 Leistungen der drei Deep-Learning-Modelle

Mit dem Ziel computertomographische Befundtexte des Thorax anhand von Deep Learning zu klassifizieren, wurden in dieser Arbeit drei verschiedene Modelle des NLP entwickelt und trainiert: ein AWD-LSTM, das eine spezielle Form der RNN darstellt, ein BERT-Modell, welches sich den Transformern zuordnen lässt, sowie eine alltagstauglichere Form dessen, das DistilBERT-Modell.

Die Modelle erhielten jeweils eine spezielle Form des Trainings. Zunächst erfolgte ein Prä-Training, bei dem sie auf große Datenbanken von Texten vortrainiert wurden und Repräsentationen sowie Zusammenhänge erlernten. Sie erlangten ein Textverständnis für die radiologische Terminologie und konnten anschließend in der Feinjustierung auf eine spezielle Aufgabe angepasst werden, welche in dieser Arbeit die Klassifikation von computertomographischen Befundtexten des Thorax war.

Die Modelle konnten hinsichtlich der Labelextraktion bei vielen Befunden Ergebnisse erreichen, welche an die Genauigkeiten menschlicher Expertinnen und Experten heranreicht. Da die Modelle sinnvolle Repräsentationen aus den Texten generieren konnten, benötigten sie hierzu nur eine vergleichsweise niedrige Menge an manuell gelabelten Daten, waren anschließend in der Lage, aus den übrigen Texten die Labels zu extrahieren, und somit sehr viel zeiteffizienter, als es ein manuelles Labeling allein durch menschliche Hand wäre.

Alle drei Modelle konnten zur Textklassifikation angewendet werden. Insgesamt konnten sowohl im Validierungsdatensatz als auch im Testdatensatz für die meisten Befunde sehr gute Ergebnisse ermittelt werden.

Das AWD-LSTM konnte von den drei Modellen insgesamt die höchsten statistischen Metriken erzielen. Es erbrachte sowohl im Validierungsdatensatz als auch im Testdatensatz bei allen Befunden eine Genauigkeit von mindestens 0,960 und erreichte für alle Befunde, welche eine Prävalenz von über 0,5% hatten und somit etwas häufiger vorkamen, einen F1-Wert von mindestens 0,800.

Auch die Transformer-Modelle BERT und DistilBERT konnten zum größten Teil sehr gute Ergebnisse bei der Textklassifikation erbringen. Bei diesen beiden Modellen waren die Metriken jeweils etwas niedriger als beim AWD-LSTM, die Leistung war dennoch überwiegend sehr akkurat. So erzielte BERT durchgängig Genauigkeiten über 0,890 und DistilBERT über 0,870. Zudem konnten von den sehr seltenen Befunden abgesehen sowohl im Validierungs- als auch im Testdatensatz durchgehend ein F1-Wert von mindestens 0,600 erreicht werden.

Das DistilBERT-Modell lieferte hierbei aufgrund seiner kompakteren Architektur geringfügig niedrigere Messwerte als das BERT-Modell, wobei sich die Ergebnisse stets sehr kongruent verhielten. Befunde, die das BERT-Modell besonders gut erkennen konnte, erkannte auch das DistilBERT-Modell sehr gut. Bei Befunden hingegen, die schlecht vom BERT-Modell klassifiziert wurden, tat sich das DistilBERT-Modell ebenfalls schwer. Dieses Verhalten entsprach den Erwartungen, da das DistilBERT-Modell eine Architektur besitzt, die der des BERT-Modells zwar ähnelt, durch einen speziellen Komprimierungsmechanismus jedoch deutlich kleiner ist und somit ein minimales Maß an Genauigkeit verliert. Der Vorteil dieser Architektur ist jedoch der geringere Bedarf an Speicherplatz, Rechenressourcen und die schnellere Inferenz-Zeit (38). Insbesondere auf Geräten mit geringer Leistung (Edge-Devices wie z.B. Mobiltelefonen) überwiegen deshalb die Vorteile destillierter Modelle den geringen Verlust an Genauigkeit.

Die Metriken der drei Modelle unterschieden sich beim Testdatensatz nicht wesentlich vom Validierungsdatensatz, sodass keine übermäßige Anpassung der Modelle an die Befundtexte, mit denen sie trainiert wurden, vorlag. Kleine Differenzen gab es lediglich beim BERT-Modell, sodass dieses Modell unter Umständen geringfügig zu stark an den Trainings- und Validierungsdatensatz adaptiert war.

Das trainierte AWD-LSTM konnte darüber hinaus zur Textgenerierung verwendet werden. Es war in der Lage, Sätze und Satzfragmente zu generieren, die sinnvoll klangen, es wurde allerdings kein längerfristiger Zusammenhang zwischen Sätzen am Anfang und Ende des Textes hergestellt, sodass der Gesamttext keinen Sinn ergab. Das Modell war somit zwar imstande, die Struktur der Sprache zu kopieren, besaß aber nicht das intelligente Sprachverständnis eines Menschen.

Das Interrater-Agreement zeigte, dass die manuelle Klassifikation der Befunde zwischen den Annotatorinnen bei einigen Befunden fast identisch war, während sie bei anderen

Befunden mitunter stark schwankte. Diese Differenzen spiegelten somit die Realität des Klinikalltags wider, da es auch bei der radiologischen Befundung deutliche Unterschiede zwischen verschiedenen Ärzten gibt. So gibt es beispielsweise Grenzwerte, die unterschiedlich gesetzt werden, wie z.B. wann vergrößerte Lymphknoten auch als Lymphadenopathie bezeichnet werden oder ab wann ein Herz als vergrößert bezeichnet wird. Das führte somit teilweise zu Diskrepanzen beim Vergeben der Labels. Bei anderen Befunden hingegen, wie beispielsweise bei der Entscheidung, ob eine Lungenarterienembolie vorliegt, ist die Klassifikation auch in der Klinik eindeutiger.

4.1.2 Abhängigkeit der Metriken von der Prävalenz eines Befundes

Bei näherer Betrachtung der verschiedenen Metriken der Klassifikation ist einiges zu beachten.

Die Genauigkeit ist ein Wert, der sowohl die Befunde, welche korrekt als positive (vorhandene) als auch diejenigen, die korrekt als negative (nicht vorhandene) klassifiziert wurden, in die Kalkulation miteinbezieht. Es handelt sich somit um einen Wert, der stark von der Prävalenz einer Pathologie abhängt. Würde das Modell beispielsweise einen seltenen Befund, der eine Prävalenz von nur 0,5% besäße, stets als negativ klassifizieren, läge die Genauigkeit dennoch bei 0,995. Sie würde folglich eine gute Performance suggerieren, obwohl der Befund nie korrekt erkannt wurde. Bei einer Prävalenz von 40% hingegen läge die Genauigkeit lediglich bei 0,600, wenn der Befund durchgehend als negativ klassifiziert würde. Bei der Beurteilung der klinischen Einsatzfähigkeit ist somit die Aussagekraft der Genauigkeit etwas eingeschränkt. Während der Wert bei der Evaluation der Modelle in Bezug auf häufig vorkommende Befunde durchaus sinnvoll ist, ist seine Bewertung bei sehr seltenen Befunden zu vernachlässigen. Dennoch drückt die Genauigkeit die Gesamtleistung des jeweiligen Modells gut aus.

Der insgesamt aussagekräftigste Wert ist der F1-Wert, der sowohl die Sensitivität als auch den positiven prädiktiven Wert miteinbezieht. Da die richtig Negativen nicht miteinbezogen werden, handelt es sich somit um einen Wert, welcher nicht von der Prävalenz eines Befundes abhängig ist. Daher kann man mit dem F1-Wert sowohl bei häufigen als auch bei seltenen Befunden eine gute Aussage zur Leistung des Modells treffen.

Bei allen Modellen zeigte sich, dass Befunde mit einer hohen Prävalenz deutlich besser klassifiziert wurden als solche mit einer geringen Prävalenz. Insbesondere durch den Testdatensatz wurde erkennbar, dass mit steigender Prävalenz der Befunde auch die positiven prädiktiven Werte und insbesondere die Sensitivitäten der Befunde stiegen. Das führte wiederum zu einem besseren F1-Wert.

4.1.3 Beurteilung der klinischen Anwendbarkeit

Für den weiterführenden klinischen Einsatz lässt sich insgesamt sagen, dass die Modelle sehr zuverlässig angewendet werden könnten, um Befunde mit einer recht hohen Prävalenz zu klassifizieren.

Das DistilBERT-Modell als das insgesamt leistungsschwächste Modell kann die Befunde Emphysem, Pleuraerguss und Ventilationsstörung gut kategorisieren, wie ein F1-Wert von größer als 0,900 im Testdatensatz zeigt.

Das BERT-Modell ist zusätzlich zu diesen drei Befunden auch in der Lage, den Befund Lymphadenopathie sicher zu erkennen.

Das AWD-LSTM kann diese vier Befunde ebenfalls korrekt identifizieren und besitzt darüber hinaus im Testdatensatz auch F1-Werte über 0,900 bei den Befunden Frakturen, Fremdmaterialien, Herzgröße, Lungenarterienembolie, Lungenrundherd, Perikarderguss, Pneumonie sowie Pneumothorax.

Bei Befunden hingegen, die lediglich eine sehr geringe Prävalenz aufweisen, wären die Modelle insbesondere aufgrund einer niedrigen Sensitivität nur eingeschränkt in ihrem Nutzen. Die Befunde Hiatushernie, Lungentrauma, mediastinaler Tumor sowie pleurale Auffälligkeit konnten im Test-Datensatz bei keinem der Modelle F1-Werte über 0,500 erreichen und werden daher nicht zuverlässig kategorisiert. Das DistilBERT-Modell konnte zudem den Befund Knochentumoren nur mit geringer Genauigkeit klassifizieren.

4.1.4 Beurteilung der Ergebnisse

Insgesamt war das Ergebnis der Auswertung der Textklassifikation insofern zum Teil unerwartet, als dass nach dem aktuellen Stand der Forschung oft Modelle der Transformer-Architektur die höchsten Genauigkeiten erzielen und anderen Architekturen wie z.B. den LSTM-basierten Modellen überlegen sind.

Dass das AWD-LSTM in dieser Studie bessere Ergebnisse als die beiden BERT-Modelle erzielt hat, kann allerdings verschiedene Gründe haben. Zum einen wurde das AWD-LSTM ausschließlich auf Befundtexten der computertomographischen Untersuchung des Thorax vortrainiert und dann anschließend feinjustiert. Eventuell kam es auf diese Weise zu einer sehr starken Spezialisierung des Modells auf CT-Texte des Thorax, was als Erklärung für die sehr gute Performance dienen könnte. Die Transformer wurden hingegen auf allen radiologischen Texten des radiologischen Datenarchives, welche innerhalb des Zeitraums vom 01.01.2009 bis zum 31.12.2018 verfasst wurden, trainiert und waren dementsprechend weniger spezialisiert. Durch die angewendete Trainingsweise ergab sich für das AWD-LSTM deutlich weniger Varianz in den Trainingstexten, da Beschreibungen von Strukturen außerhalb des Thorax nicht im Datensatz für das Prä-Training vorkamen. Die gesamte Kapazität des Modells konnte hierdurch für das Erlernen der Struktur von CT-Thorax-Befundtexten verwendet werden. Die Transformer-Modelle mussten hingegen beim Prä-Training lernen, Text-repräsentationen für alle Arten von Befundtexten zu generieren. Es ist daher davon auszugehen, dass das BERT- und DistilBERT-Modell leicht auf andere Befundtexte wie z.B. von einer Kopf- oder Abdomen-Untersuchung anpassbar wären, während das AWD-LSTM aus dieser Arbeit solche Befunde nicht gut klassifizieren könnte.

Hinzu kommt, dass auch die Technisierung des AWD-LSTM effektiver war, da hier nur Vokabeln aus Befundtexten des Thorax genommen wurden, was die Genauigkeit im Verhältnis zu den Transformern, bei denen der Tokenizer mit allen Texten des ausgewählten Zeitraums erstellt wurde, noch zusätzlich gehoben hat.

Des Weiteren konnte festgestellt werden, dass das BERT-Modell im Testdatensatz geringfügig niedrigere Ergebnisse als im Validierungsdatensatz erbrachte und somit leicht überangepasst war. Durch diese geringfügige Überanpassung ging somit bei der Anwendung auf neue Texte ein gewisses Maß an Genauigkeit verloren.

Auch wenn in der Literatur meistens BERT-Modelle bessere Ergebnisse als RNN erzielten (45, 46), gab es auch andere Ergebnisse wie in der Studie von Rivera Zavala et al., in der ein LSTM-Modell und ein BERT-Modell insgesamt ähnliche Genauigkeiten erreichen. Die Überlegenheit eines der beiden Modelle unterschied sich dabei je nach Art des Textes, auf die sie angewendet wurden (31).

4.2 Stärken und Limitationen

In dieser Arbeit gab es einige allgemeine Stärken und Limitationen, die durch die Rahmenbedingungen des Studiendesigns bedingt wurden, sowie auch Vorteile und Einschränkungen, die in der Architektur der trainierten Modelle begründet waren.

Es hat sich sehr deutlich gezeigt, wie stark die Genauigkeiten der Modelle von der Prävalenz eines Befundes abhingen. So wurde stets eine gewisse Menge an Befundtexten benötigt, um eine Genauigkeit zu erreichen, welche das Modell klinisch einsetzbar machte.

Eine Stärke dieser Arbeit war die große Menge der annotierten Befundtexten. Somit war es nicht nur bei den sehr häufig vorkommenden Befunden wie z.B. dem Lungenrundherd, welcher eine Prävalenz von etwa 40% besaß, möglich, eine hohe Klassifikations-genauigkeit zu erreichen. Auch etwas seltenere Befunde wie die Lungenarterienembolie, die in den Befundtexten mit einer Prävalenz von rund 5% vorkam, erzielten sehr gute Metriken.

Dennoch lag eine Limitation dieser Arbeit darin begründet, dass bestimmte Befunde in computertomographischen Untersuchungen des Thorax generell eine nur äußerst geringe Prävalenz besaßen. Einige Befunde wie beispielsweise das Lungentrauma, welches Verletzungen wie eine Lungenkontusion oder Lungenlazeration umfasste, kamen im Patientenkollektiv nur in den seltensten Fällen vor. Das führte wiederum dazu, dass die Anzahl der Befunde, in denen sie annotiert werden konnte, selbst bei einer sehr hohen Menge an gelabelten Texten äußerst gering blieb. Somit war diese Menge zu klein, als dass die Modelle eine hohe Genauigkeit für das Erkennen dieser Befunde hätten erreichen können. Wie die Metriken der verschiedenen Befunde gezeigt haben, stiegen die Ergebnisse deutlich mit zunehmender Prävalenz eines Befundes. Um die Leistungen der Modelle in Bezug auf diese Befunde zu verbessern, müsste die Zahl annotierter Befundtexte massiv erhöht werden, da einige Pathologien beispielsweise nur etwa in einem von 200 Texten vorkommen.

Die Tatsache, dass noch immer alle Annotationen durch Fachkräfte manuell angefertigt werden müssen, stellt eine weitere Limitation der Deep-Learning-Modelle dar. Selbst wenn die Schnelligkeit der Annotatorinnen und Annotatoren mit steigender Befundzahl zunimmt, werden pro Befundtext noch einige Minuten benötigt. Ginge man von einem Zeitaufwand von zwei Minuten pro Befundtext aus, was bereits eine äußerst hohe

Geschwindigkeit bei der Annotation darstellte, würde die Annotation von 5000 Befundtexten noch immer über 160 Stunden benötigen. Die Genauigkeit der Modelle auf seltene Befunde zu verbessern, wäre somit mit hohem personellem Aufwand verbunden.

Eine gemeinsame Limitation aller drei Modelle ist zudem die maximale Sequenzlänge von 512 Tokens, wobei ein Token ein Wort oder Wortteil beinhaltet, sodass Informationen aus längeren Texten verloren gehen können. Insbesondere bei langen Befundtexten kann dies zu einer Trunkierung der Texte führen und somit zu einem unvermeidbarem Informationsverlust. Zwar ist es hier möglich, Sliding-Window-Ansätze anzuwenden, in denen dem Modell multiple Ausschnitte des Textes gegeben werden und dann der Durchschnitt der Prädiktionen gebildet wird, die Nutzung höherer Sequenzlängen würde jedoch eine effektivere Maßnahme darstellen. Theoretisch wäre es möglich, die Sequenzlänge zu erhöhen, dies ginge jedoch mit einem gesteigerten Rechen- und Speicherbedarf der Modelle einher. Durch die Verwendung des Attention-Mechanismus erhöht sich beispielsweise der Speicherbedarf der Transformer quadratisch mit der Sequenzlänge. Folglich werden Sequenzen länger als 1024 Tokens nur sehr selten verwendet. In der vorgestellten Arbeit war es aufgrund des auf 11 GB limitierten VRAM (Video Random Access Memory) der Grafikkarten nicht praktikabel, die Sequenzlänge auf über 512 anzuheben. Neue Ansätze werden jedoch bereits entwickelt, um z.B. nur eine lineare Zunahme des Speicherbedarfs mit der Sequenzlänge zu erreichen (47).

Die zunehmende Sequenzlänge stellt auch für Netzwerke mit LSTM-Modulen ein Problem dar. Durch die intelligenten Gating-Mechanismen innerhalb dieser Modelle können Informationen zwar auch in langen Texten zueinander in Kontext gesetzt werden, je länger ein Text jedoch wird, desto stärker wirken Effekte wie verschwindende Gradienten auf die LSTM-Performance ein. Durch ihren Aufmerksamkeitsmechanismus stellt dies für Transformer-Modelle kein Problem dar.

Darüber hinaus gibt es noch für jedes der drei Modelle weitere individuelle Vorteile und Limitationen. So ist eine Stärke des in dieser Arbeit entwickelten AWD-LSTM gegenüber der BERT-Modelle die starke Spezialisierung, welche eine sehr hohe Genauigkeit ermöglicht, sowie die etwas einfachere Technisierung. Ein Nachteil des AWD-LSTM aus dieser Arbeit hingegen ist, dass es schlecht übertragen werden kann. Für die Durchführung einer anderen ähnlichen Aufgabe müsste das gesamte Training wiederholt werden.

Das stellt wiederum einen großen Vorteil der Transformer dar. Durch die Verwendung des Transfer-Lernprozesses ist es möglich, die BERT-Modelle lokal vorzutrainieren und dann öffentlich verfügbar zu machen, sodass andere eine Feinjustierung für spezifische Aufgaben vornehmen können, ohne das Prä-Training wiederholen zu müssen. Da die BERT-Modelle auf allen radiologischen Befunden trainiert wurden, sind sie besser als RNN-basierte Modelle dazu in der Lage, Repräsentationen von Befundtexten, welche sich nicht auf den Thorax beziehen, zu erstellen.

Es wäre auch möglich, ein LSTM-Modell ähnlich wie die BERT-Modelle auf allen radiologischen Befundtexten zu trainieren. In dieser Arbeit wurde sich jedoch dagegen entschieden. Das Vortraining auf sehr großen Datenmengen bedeutet einen immensen Rechenaufwand und um trotzdem umsetzbare Trainingszeiten zu erzielen, ist eine effiziente Parallelisierung nötig. Bei Modellen der Transformer-Architektur kann hierfür auf ausgereifte Programmbibliotheken zurückgegriffen werden wie z.B. „Transformers“ von Huggingface (40). Zusätzlich ermöglichen diese Programmbibliotheken eine Standardisierung der Modellarchitekturen und sichern so eine nachhaltige und langfristige Nutzbarkeit der Modelle. Bei LSTM-Modellen existieren keine derartig fortgeschrittenen Bibliotheken. Auch zeigt der aktuelle Trend in der NLP-Forschung einen starken Fokus auf Modelle der Transformer-Architektur. Insgesamt wurde deshalb von einem umfangreichen Vortraining des AWD-LSTM-Modells abgesehen.

Wie die Textgenerierung gezeigt hat, blieb das Erinnerungsvermögen des RNNs trotz der LSTM-Module begrenzt, sodass es kein menschenähnliches Sprachverständnis erlangen konnte. Auch bei den Transformer-Modellen war trotz der verbesserten Architektur mit dem Attention-Mechanismus die Erinnerungskapazität noch immer limitiert und auch das Textverständnis reichte nicht vollständig an das menschliche heran. So hat auch eine Studie von Ettinger et al. gezeigt, dass sich BERT-Modelle beispielsweise bei Negativierungen noch schwer taten (48).

Aufgrund von Einschränkungen des Gedächtnisses der verwendeten Hardware konnte für diese Arbeit nur das Basis-BERT-Modell (BERTbase) verwendet werden. In der Studie von Devlin et al. wurde ein größeres BERT-Modell (BERTlarge) verwendet, welches aus 24 Schichten bestand und noch genauer als BERT war (26). Die Anwendung eines größeren Modells hätte unter Umständen eine bessere Performance ermöglichen können als die Nutzung des Basis-Modells. Ein solches Modell erfordert allerdings eine

besondere Hardware in Form von Tensor Processing Units, welche z.B. via Google-Cloud verfügbar wären. Da sich der Cloud-Export sensitiver Patientendaten allerdings zumeist nicht mit den lokalen Datenschutzrichtlinien vereinbaren lässt, war der Einsatz dieses größeren Modells nicht möglich.

Der Einfluss der Größe eines Transformer-Modells wurde auch beim DistilBERT-Modell zum Ausdruck gebracht. DistilBERT hat gegenüber BERT den Vorteil, dass es kleiner und schneller ist und deshalb einfacher und kostengünstiger angewendet werden kann. Durch die im Training verwendete Kompressionstechnik ist es auf der anderen Seite geringfügig leistungsschwächer als das BERT-Modell. Die Wahl eines bestimmten Deep-Learning-Modells ist somit stets von seinem Verwendungszweck abhängig.

4.3 Vergleich mit ähnlichen Arbeiten

NLP wird seit Jahren in der Radiologie angewendet, um Informationen aus Befundtexten zu generieren, und die Techniken haben sich über die Zeit hinweg stetig verbessert. Während zu Beginn ausschließlich regelbasierte Algorithmen verwendet wurden, werden aktuell immer öfter auf Deep Learning basierende Techniken eingesetzt, da in diesem Bereich die Entwicklung neuer Modelle stetig vorangeht und diese eine effizientere Informationsextraktion ermöglichen. Verschiedene Deep-Learning-Methoden umfassen dabei häufig klassische RNN, verbesserte Formen davon mit LSTM-Modulen oder Transformer mit dem Attention-Mechanismus. Mithilfe dieser Modelle des NLP können aus radiologischen Befundtexten verschiedene Informationen gewonnen werden, mit deren Hilfe viele weitere Aufgaben durchgeführt werden können.

Das Ziel dieser Arbeit war es, unter Verwendung dreier Deep-Learning-Modelle Befundtexte der Computertomographie des Thorax zu klassifizieren und somit eine Kategorisierung von Befundtexten zu ermöglichen, welche in dieser Form bislang nicht existiert.

Im Bereich der Thorax-Diagnostik kamen auch in einigen anderen Studien bereits diverse NLP-Techniken zum Einsatz.

RNN fanden beispielsweise in der Arbeit von Cornegruta et al. ihre Anwendung, als ein LSTM-Modell verwendet wurde, um sowohl Pathologien als auch Negativierungen, welche folglich das Nicht-Vorhandensein eines Befundes implizieren, in Befundtexten des Röntgen-Thorax zu identifizieren (49). Auch Transformer-Modelle wurden in einigen

Studien verwendet. In der Arbeit von Moon et al. wurden z.B. verschiedene auf BERT basierende Modelle angewendet, um mithilfe eines annotierten Datensatzes von radiologischen Befundtexten sowie Röntgen-Bildern des Thorax unterschiedliche Aufgaben durchzuführen, wie etwa die Befundtextgenerierung, die Klassifikation von Diagnosen oder die Korrelation von den Bildern zu den Texten (50). Auch Smit et al. kreierten ein Transformer-Modell, CheXbert, um Befundtexte von Thorax-Untersuchungen zu klassifizieren und Labels für verschiedene Befunde zu extrahieren (51). Darüber hinaus wurden in einigen Studien auch Deep-Learning-Modelle unterschiedlicher Architektur miteinander verglichen, wie in der Studie von Datta et al., in der drei Modelle, darunter ein LSTM- und ein BERT-Modell, verwendet wurden, um Befundtexte des Röntgen-Thorax auszuwerten und Zusammenhänge zwischen den Befunden und ihrer anatomischen Lokalisation herzustellen (52).

Insgesamt handelte sich allerdings in der Thorax-Diagnostik bei den Befundtexten fast ausschließlich um Ergebnisse von Röntgen-Untersuchungen. Diese Modalität bietet aufgrund ihrer im Vergleich zur CT niedrigeren Komplexität nur eine begrenzte Anzahl möglicher gelabelter Befunde. Somit ist der Annotationsaufwand geringer und das Training der Modelle einfacher.

Es gibt bislang nur verhältnismäßig wenige Studien, die auch Befundtexte von computertomographischen Aufnahmen des Thorax in größerem Maße miteinbezogen haben. Eines der Beispiele dafür bildete eine Studie von Chen et al., in der Deep-Learning-Modelle und regelbasierte Methoden auf Befundtexte einer CT des Thorax angewendet wurden, um Texte mit einer Lungenarterienembolie zu identifizieren (53). Die Studie von Yuan et al. schloss ebenfalls CT-Befunde mit ein, aber auch hier wurde nur ein einziger Befund untersucht, in diesem Fall die Veränderung von pulmonalen Noduli (54). Ein weiteres Beispiel ist die Arbeit von Olthof et al., welche eine systematische multifaktorielle Analyse verschiedener Deep-Learning-Techniken im Bereich des NLP in der radiologischen Befundung darstellte (55). Es wurden vier verschiedene Modelle verwendet, darunter auch ein LSTM- sowie ein BERT-Modell, um u.a. Röntgen- und CT- Aufnahmen des Thorax zu klassifizieren. Die Studie bestätigte systematisch den Einfluss der Prävalenz auf die Genauigkeit eines Modells, was auch in dieser Arbeit deutlich zu bemerken war. Allerdings konzentrierte sich diese Arbeit ebenfalls nur auf einen einzigen Befund (pulmonale Infiltrate).

In der hier durchgeführten Arbeit wurde somit eine Klassifikation von Befundtexten der Computertomographie des Thorax durchgeführt, welche über den Umfang vorheriger ähnlicher Studien deutlich hinausgeht. Es wurde nicht nur eine einzige Pathologie herausgearbeitet, sondern insgesamt 21 verschiedene Befunde, welche sämtliche Hauptdiagnosen dieser Untersuchungsmodalität abdeckten. Andere Deep-Learning-Modelle, die im Bereich der CT-Thorax-Diagnostik eine solche Vielzahl verschiedener Befunde klassifizieren können, existieren bislang nicht in dieser Form.

Die durch die Klassifikation von radiologischen Befundtexten des Thorax gewonnenen Labels können auch für weiterführende Aufgaben weiterverwendet werden. So wurden in einigen Arbeiten die aus Texten extrahierten Labels zur Entwicklung eines Bild-klassifikationsmodells herangezogen (56, 57).

In vielen dieser Studien stellten sehr kurze oder inkomplette Texte eine große Herausforderung dar. Diese Befunde evaluierten beispielsweise nur eine spezifische Pathologie (z.B. „Kein Pleuraerguss. Ansonsten keine Befundwandel.“) oder erwähnten lediglich unveränderte Fremdmaterialien, ohne näher darauf einzugehen, um welche Fremdmaterialien es sich genau handelte. Diese Befunde stellten ein hohes Risiko für Ungenauigkeiten dar, wenn sie zur Labelextraktion verwendet wurden, da die Labels zwar die Befundtexte widerspiegelten, allerdings in Bezug auf das zugehörige Bild unvollständig waren (29). Eine hohe Datenqualität ist jedoch für die Entwicklung eines verlässlichen Machine-Learning-Algorithmus unerlässlich, sodass die Anwendung dieser Labels zu einer schlechteren Leistung des daraus trainierten Modells führen kann (58). Um solche Fehler zu vermeiden, wurden in dieser Arbeit im Voraus alle besonders kurzen bzw. inkomplette Befunde ausgeschlossen.

Die Entwicklung eines Bildklassifikationsmodells mithilfe von zuvor durch NLP gewonnene Labels war selbstverständlich nicht nur in der Thorax-Diagnostik möglich. Auf diese Weise wurde beispielsweise ebenso ein Modell für kraniale CT-Aufnahmen (59) oder Bilder der lumbalen Wirbelsäule entwickelt (60).

Auch insgesamt ist der Bereich des NLP äußerst vielseitig. So ist es zum einen möglich, Befundtexte jeder Untersuchungsmodalität auszuwerten. Neben der bereits erwähnten NLP-Applikation auf Röntgen- oder CT-Untersuchungen war in Studien auch die Anwendung auf Befundtexte der Mammographie (61), der MRT (62), der Szintigraphie (63) und des Ultraschalls (64) möglich.

Die untersuchte Körperregion variiert ebenfalls stark in den verschiedenen Arbeiten. In der Studie von Wang et. al stand beispielsweise die Klassifikation von proximalen Femurfrakturen im Vordergrund (65), während von Li et al. NLP-Modelle verwendet wurden, um Nierensteine in CT-Befunden zu identifizieren (66), und sich Senders et al. Patientinnen und Patienten mit Glioblastomen angesehen haben (67).

Hinzu kommt, dass sich die genutzten Methoden des NLP unterscheiden. Es gibt Arbeiten, die regelbasierte Methoden angewendet haben (62, 68, 69), und Studien, in denen Techniken des maschinellen Lernens verwendet wurden. Bei diesen Studien gibt es Arbeitsgruppen, die RNN, teilweise auch mit LSTM, eingesetzt haben (35, 49, 70), und andere, in denen Transformer wie BERT zur Anwendung kamen (33, 45, 71).

Aber auch das Ziel des NLP variiert in den Studien. Eine Übersicht über verschiedene Anwendungen von NLP im Bereich der Radiologie hat gezeigt, dass die Nutzung der Informationen aus radiologischen Befundtexten verschiedenen Zwecken dienen kann (72, 73).

Ein großer Bereich ist die Klassifikation von Befundtexten, bei der nach Texten mit einer bestimmten Pathologie, wie beispielsweise ischämischen Schlaganfällen (74) oder Hirnmetastasen (75), gesucht wird. In einigen Arbeiten wurden darüber hinaus auch die Eigenschaften bestimmter Pathologien näher charakterisiert. Beispiele hierfür sind Studien, in denen BI-RADS-Befunde aus Ultraschall-Befunden (76) oder Charakteristiken von hepatozellulären Karzinomen (71) herausgearbeitet wurden. Ein weiterer wichtiger Anwendungsbereich ist die diagnostische Überwachung, die häufig in der Onkologie angewendet wird und beispielsweise dazu dient, die Veränderung von Tumoren über die Zeit (77) oder das Ansprechen auf eine Therapie zu untersuchen (78).

Des Weiteren gibt es auch Arbeiten zur Sprachstruktur, welche versuchen, den Aufbau eines radiologischen Befundtextes zu optimieren, indem verschiedene Sprachaspekte der radiologischen Terminologie wie z.B. die Komplexität und Variabilität des Vokabulars analysiert werden (79). Darüber hinaus gibt es auch Publikationen zur Qualitätskontrolle mit dem Ziel, den radiologischen Alltag zu verbessern, indem Aspekte der Untersuchung, wie die Auswahl des Protokolls, sorgfältig geprüft werden (80). Auch die Compliance der Patientinnen und Patienten bei empfohlener Follow-Up-Bildgebung kann näher betrachtet werden (81). Der Einsatz von NLP muss somit nicht nur der Eruierung bestimmter Pathologien dienen. So ist beispielsweise auch die Identifikation bestimmter

weiterführender Therapieempfehlungen (82) oder die Evaluation von diagnostischen Unsicherheiten (83) möglich. Insgesamt kann NLP daher sehr vielseitig in der Radiologie eingesetzt werden.

4.4 Schlussfolgerungen und Ausblick

In dieser Arbeit war es möglich, die Labels aus allen computertomographischen Befundtexten des Thorax eines bestimmten Zeitraums vom radiologischen Informationssystem (RIS) der Charité zu extrahieren. Dies würde die Entwicklung eines Programms gestatten, das es ermöglicht, gezielt nach Befundtexten zu suchen. Die Befundtexte können mithilfe der Labels gefiltert werden, was die explizite Auswahl einer Pathologie möglich macht. Hierdurch könnten die Modelle dazu beitragen, eine strukturierte Befundung umzusetzen.

Ferner erlaubt das Deep Learning auch bei Studien, für die vorab Personen mit einer bestimmten Pathologie identifiziert werden müssen, eine schnellere und effizientere Identifikation derselben. Möchte man beispielsweise Daten zu Patientinnen und Patienten mit einer Pneumonie erheben, kann man nun beispielsweise alle computertomographischen Aufnahmen, die pneumonische Infiltrate zeigen, herausfiltern.

Darüber hinaus können diese Labels für diverse weiterführende Aufgaben angewendet werden. Zum einen ermöglichen die Labels das Training eines Bildklassifikationsmodells für computertomographische Aufnahmen des Thorax, welches erlauben würde, eine beträchtliche Anzahl an Bildern mit nur minimalem menschlichem Input zu klassifizieren. Manuelles Labeln ist stets eine sehr zeitaufwendige Arbeit, wobei die Klassifikation von CT-Thorax-Bildern noch erheblich länger dauern würde als die von Befundtexten. Durch die Korrelation der bereits gelabelten Befundtexte mit den zugehörigen computer-tomographischen Rekonstruktionsbildern kann somit dieser Schritt umgangen werden. Mit diesem Bildklassifikationsmodell kann anschließend ein System entwickelt werden, das in der Lage ist, gezielt nach Bildern mit einem bestimmten Befund zu suchen.

Da die Fähigkeiten und Genauigkeiten des radiologischen Personals mit steigender Erfahrung zunehmen, bieten somit verschiedene Techniken des maschinellen Lernens Möglichkeiten, um vor allem unerfahrene Radiologinnen und Radiologen zu unterstützen. So könnte ein Algorithmus beispielsweise Patientinnen und Patienten mit demselben Befund heraussuchen, was es ermöglichen würde, die Bilder zu vergleichen und eine bessere Diagnose zu stellen.

Die stetige Zunahme radiologischer Aufnahmen und die tägliche Entstehung neuer Bilder bergen noch sehr viel Potential für verschiedene Anwendungen. Um dieses Potential besser nutzen zu können, erfordert es eine größere Effizienz in der klinischen Routine, welche durch Methoden des Deep Learning erreicht werden kann.

So existiert beispielsweise an der Charité eine der größten medizinischen Datenbanken weltweit, sodass sich insbesondere für den Fachbereich der Radiologie noch viele verschiedene Möglichkeiten ergeben, in denen Modelle mit künstlicher Intelligenz weiterhelfen könnten.

Deep-Learning-Modelle im Bereich des NLP stellen dabei eine sehr effektive Methode zur Datenerhebung dar. So ist ein Modell zur Informationsextraktion schneller und deutlich effizienter, als alle Daten von Hand zu herauszuarbeiten, und kann nahezu dieselbe Genauigkeit wie menschliche Fachexpertinnen und Fachexperten erreichen. Außerdem ist es eine Aufgabe, die lediglich ein einziges Mal durchgeführt werden muss, woraufhin die Modelle für eine Vielzahl von Projekten in unterschiedlichen Anwendungsbereichen wie beispielsweise der Qualitätssicherung, der diagnostischen Überwachung oder der systematischen Analyse bestimmter Befunde wiederverwendet werden können.

Möglicherweise könnten Informationen verschiedener durchgeführter Projekte auch in eine multimodale Repräsentation für sämtliche Patientinnen und Patienten implementiert werden, sodass Ärztinnen und Ärzte aller Fachrichtungen einen schnellen Zugriff auf relevante Informationen der Modelle hätten, um bessere Diagnosen stellen zu können und Forschungsstudien effizienter zu gestalten.

Mit Hinblick auf die Zukunft bieten verschiedene Methoden des Deep Learning daher noch großes Potential, den klinischen Alltag in der Radiologie, aber auch in anderen Fachbereichen zu vereinfachen und zu verbessern.

Literaturverzeichnis

1. Nekolla EA, Schegerer AA, Griebel J, Brix G. Häufigkeit und Dosis diagnostischer und interventioneller Röntgenanwendungen. Der Radiologe. 2017;57(7):555-62.

2. Reiser M, Kuhn FP, Debus J. Computertomographie (CT). Duale Reihe Radiologie. Stuttgart: Georg Thieme Verlag KG; 2006. p. 79-83.

3. Goldman LW. Principles of CT and CT technology. J Nucl Med Technol. 2007;35(3):115-28; quiz 29-30.

4. Dössel O. Computer-Tomographie. Bildgebende Verfahren in der Medizin. Berlin, Heidelberg: Springer Verlag; 2016. p. 131-78.

5. Bhalla AS, Das A, Naranje P, Irodi A, Raj V, Goyal A. Imaging protocols for CT chest: A recommendation. Indian J Radiol Imaging. 2019;29(3):236-46.

6. Reiser M, Kuhn FP, Debus J. Thorax: Radiologische Methoden - Computertomographie (CT). Duale Reihe - Radiologie. Stuttgart: Thieme Verlag; 2006. p. 153-7.

7. van Beek E, Jr., Mirsadraee S, Murchison JT. Lung cancer screening: Computed tomography or chest radiographs? World J Radiol. 2015;7(8):189-93.

8. Ludes C, Schaal M, Labani A, Jeung MY, Roy C, Ohana M. [Ultra-low dose chest CT: The end of chest radiograph?]. Presse Med. 2016;45(3):291-301.

9. Grob D, Oostveen LJ, Prokop M, Schaefer-Prokop CM, Sechopoulos I, Brink M. Imaging of pulmonary perfusion using subtraction CT angiography is feasible in clinical practice. Eur Radiol. 2019;29(3):1408-14.

10. European Society of Radiology. Good practice for radiological reporting. Guidelines from the European Society of Radiology (ESR). Insights Imaging. 2011;2(2):93-6.

11. Glasser O. Röntgens vorläufige Mitteilung „Über eine neue Art von Strahlen“. Wilhelm Conrad Röntgen und die Geschichte der Röntgenstrahlen. Berlin, Heidelberg: Springer Berlin Heidelberg; 1959. p. 14-24.

12. Reiner BI, Knight N, Siegel EL. Radiology reporting, past, present, and future: the radiologist's perspective. J Am Coll Radiol. 2007;4(5):313-9.

13. Wulff HR. The language of medicine. J R Soc Med. 2004;97(4):187-8.

14. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? J Biomed Inform. 2009;42(5):760-72.

15. Jungmann F, Kuhn S, Tsaur I, Kämpgen B. Natural Language Processing in der Radiologie. Der Radiologe. 2019;59(9):828-32.

16. Šuster S, Tulkens S, Daelemans W. A Short Review of Ethical Challenges in Clinical Natural Language Processing. arxiv preprint. 2017; arXiv:1703.10090.

17. Hassanpour S, Langlotz CP. Information extraction from multi-institutional radiology reports. Artif Intell Med. 2016;66:29-39.

18. Pons E, Braun LM, Hunink MG, Kors JA. Natural Language Processing in Radiology: A Systematic Review. Radiology. 2016;279(2):329-43.

19. Chen PH. Essential Elements of Natural Language Processing: What the Radiologist Should Know. Acad Radiol. 2020;27(1):6-12.

20. Kriegeskorte N, Golan T. Neural network models and deep learning. Curr Biol. 2019;29(7):R231-r6.

21. Dörn S. Neuronale Netze. Programmieren für Ingenieure und Naturwissenschaftler: Intelligente Algorithmen und digitale Technologien. Berlin, Heidelberg: Springer Berlin Heidelberg; 2018. p. 89-148.

22. Joshi P. How do Transformers Work in NLP? A Guide to the Latest State-of-the-Art Models. Analytics Vidhya 2019. Zuletzt abgerufen am 07.11.2020 unter https://www.analyticsvidhya.com/blog/2019/06/understanding-transformers-nlp-state-of-the-art-models/.

23. Thomas C. Recurrent Neural Networks and Natural Language Processing. Towards Data Science 2019. Zuletzt abgerufen am 01.11.2020 unter https://towardsdatascience.com/recurrent-neural-networks-and-natural-language-processing-73af640c2aa1.

24. Yu Y, Si X, Hu C, Zhang J. A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. Neural Comput. 2019;31(7):1235-70.

25. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin, I. Attention Is All You Need. arXiv preprint. 2017; arXiv:170603762.

26. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint. 2019; arXiv:181004805v2.

27. Huesch MD, Cherian R, Labib S, Mahraj R. Evaluating Report Text Variation and Informativeness: Natural Language Processing of CT Chest Imaging for Pulmonary Embolism. J Am Coll Radiol. 2018;15(3 Pt B):554-62.

28. Cai T, Giannopoulos AA, Yu S, Kelil T, Ripley B, Kumamaru KK, Rybicki FJ, Mitsouras D. Natural Language Processing Technologies in Radiology Research and Clinical Applications. Radiographics. 2016;36(1):176-91.

29. Bressem KK, Adams LC, Gaudin RA, Tröltzsch D, Hamm B, Makowski MR, Schüle CY, Vahldiek JL, Niehues SM. Highly accurate classification of chest radiographic reports using a deep learning natural language model pre-trained on 3.8 million text reports. Bioinformatics. 2021;36(21):5255-61.

30. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234-40.

31. Rivera Zavala R, Martinez P. The Impact of Pretrained Language Models on Negation and Speculation Detection in Cross-Lingual Medical Text: Comparative Study. JMIR Med Inform. 2020;8(12):e18953.

32. Liu F, Zhou P, Baccei SJ, Masciocchi MJ, Amornsiripanitch N, Kiefe CI, Rosen MP. Qualifying Certainty in Radiology Reports through Deep Learning-Based Natural Language Processing. AJNR Am J Neuroradiol. 2021;42(10):1755-1761.

33. Chen L, Shah R, Link T, Bucknor M, Majumdar S, Pedoia V. Bert model fine-tuning for text classification in knee OA radiology reports. Osteoarthritis and Cartilage. 2020;28:S315-S6.

34. Zhang X, Zhang Y, Zhang Q, Ren Y, Qiu T, Ma J, Sun Q. Extracting comprehensive clinical information for breast cancer using deep learning methods. Int J Med Inform. 2019;132:103985.

35. Colón-Ruiz C, Segura-Bedmar I. Comparing deep learning architectures for sentiment analysis on drug reviews. J Biomed Inform. 2020;110:103539.

36. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735-80.

37. Merity S, Shirish Keskar N, Socher R. Regularizing and Optimizing LSTM Language Models. arxiv preprint. 2017; arXiv:1708.02182.

38. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arxiv preprint. 2019; arXiv:1910.01108.

39. Honnibal M, Montani I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. 2017.

40. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, Davison J, Shleifer S, von Platen P, Ma C, Jernite Y, Plu J, Xu C, Scao TL, Gugger S, Drame M, Lhoest Q, Rush AM. HuggingFace's Transformers: State-of-the-art Natural Language Processing. arxiv preprint. 2019; arXiv:1910.03771.

41. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Köpf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala, S. PyTorch: An Imperative Style, High-Performance Deep Learning Library. Advances in Neural Information Processing Systems 32 Curran Associates, Inc. 2019: 8024–35.

42. Howard J, Gugger S. fastai: A Layered API for Deep Learning. arxiv preprint. 2020; arXiv:2002.04688.

43. Liu H, Perl Y, Geller J. Transfer Learning from BERT to Support Insertion of New Concepts into SNOMED CT. AMIA Annu Symp Proc. 2019;2019:1129-38.

44. Wei Q, Dunbrack RL, Jr. The role of balanced training and testing data sets for binary classifiers in bioinformatics. PLoS One. 2013;8(7):e67863.

45. Kim YM, Lee TH. Korean clinical entity recognition from diagnosis text using BERT. BMC Med Inform Decis Mak. 2020;20(Suppl 7):242.

46. Nakamura Y, Hanaoka S, Nomura Y, Nakao T, Miki S, Watadani T, Yoshikawa T, Hayashi N, Abe O. Automatic detection of actionable radiology reports using bidirectional encoder representations from transformers. BMC Med Inform Decis Mak. 2021;21(1):262.

47. Martins PH, Marinho Z, Martins AFT. ∞-former: Infinite Memory Transformer. arxiv preprint. 2021; arXiv:2109.00301.

48. Ettinger A. What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. Transactions of the Association for Computational Linguistics. 2020;8:34-48.

49. Cornegruta S, Bakewell R, Withey S, Montana G. Modelling Radiological Language with Bidirectional Long Short-Term Memory Networks. arxiv preprint. 2016; arXiv:1609.08409.

50. Moon JH, Lee H, Shin W, Choi E. Multi-modal Understanding and Generation for Medical Images and Text via Vision-Language Pre-Training. arxiv preprint. 2021; arXiv:2105.11333.

51. Smit A, Jain S, Rajpurkar P, Pareek A, Ng AY, Lungren MP. CheXbert: Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT. arxiv preprint. 2020; arXiv:2004.09167.

52. Datta S, Si Y, Rodriguez L, Shooshan SE, Demner-Fushman D, Roberts K. Understanding spatial language in radiology: Representation framework, annotation, and spatial relation extraction from chest X-ray reports using deep learning. J Biomed Inform. 2020;108:103473.

53. Chen MC, Ball RL, Yang L, Moradzadeh N, Chapman BE, Larson DB, Langlotz CP, Amrhein TJ, Lungren MP. Deep Learning to Classify Radiology Free-Text Reports. Radiology. 2018;286(3):845-52.

54. Yuan J, Zhu H, Tahmasebi A. Classification of Pulmonary Nodular Findings based on Characterization of Change using Radiology Reports. AMIA Jt Summits Transl Sci Proc. 2019;2019:285-94.

55. Olthof AW, van Ooijen PMA, Cornelissen LJ. Deep Learning-Based Natural Language Processing in Radiology: The Impact of Report Complexity, Disease Prevalence, Dataset Size, and Algorithm Type on Model Performance. J Med Syst. 2021;45(10):91.

56. Bustos A, Pertusa A, Salinas J-M, de la Iglesia-Vayá M. PadChest: A large chest x-ray image dataset with multi-label annotated reports. arxiv preprint. 2019; arXiv:1901.07441.

57. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, Marklund H, Haghgoo B, Ball R, Shpanskaya K, Seekins J, Mong DA, Halabi SS, Sandberg JK, Jones R, Larson DB, Langlotz CP, Patel BN, Lungren MP, Ng AY. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. arxiv preprint. 2019; arXiv:1901.07031.

58. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. JAMA Intern Med. 2018;178(11):1544-7.

59. Titano JJ, Badgeley M, Schefflein J, Pain M, Su A, Cai M, Swinburne N, Zech J, Kim J, Bederson J, Mocco J, Drayer B, Lehar J, Cho S, Costa A, Oermann EK. Automated deep-neural-network surveillance of cranial images for acute neurologic events. Nat Med. 2018;24(9):1337-1341.

60. Galbusera F, Cina A, Bassani T, Panico M, Sconfienza LM. Automatic Diagnosis of Spinal Disorders on Radiographic Images: Leveraging Existing Unstructured Datasets With Natural Language Processing. Global Spine J. 2021:21925682211026910.

61. Short RG, Bralich J, Bogaty D, Befera NT. Comprehensive Word-Level Classification of Screening Mammography Reports Using a Neural Network Sequence Labeling Approach. J Digit Imaging. 2019;32(5):685-92.

62. Fu S, Leung LY, Wang Y, Raulli AO, Kallmes DF, Kinsman KA, Nelson KB, Clark MS, Luetmer PH, Kingsbury PR, Kent DM, Liu H. Natural Language Processing for the Identification of Silent Brain Infarcts From Neuroimaging Reports. JMIR Med Inform. 2019;7(2):e12109.

63. Groot OQ, Bongers MER, Karhade AV, Kapoor ND, Fenn BP, Kim J, Verlaan JJ, Schwab JH. Natural language processing for automated quantification of bone metastases reported in free-text bone scintigraphy reports. Acta Oncol. 2020;59(12):1455-60.

64. Wu X, Zhao Y, Radev D, Malhotra A. Identification of patients with carotid stenosis using natural language processing. Eur Radiol. 2020;30(7):4125-33.

65. Wang Y, Sohn S, Liu S, Shen F, Wang L, Atkinson EJ, Amin S, Liu H. A clinical text classification paradigm using weak supervision and deep representation. BMC Med Inform Decis Mak. 2019;19(1):1.

66. Li AY, Elliot N. Natural language processing to identify ureteric stones in radiology reports. J Med Imaging Radiat Oncol. 2019;63(3):307-10.

67. Senders JT, Cho LD, Calvachi P, McNulty JJ, Ashby JL, Schulte IS, Almekkawi AK, Mehrtash A, Gormley WB, Smith TR, Broekman MLD, Arnaout O. Automating Clinical Chart Review: An Open-Source Natural Language Processing Pipeline Developed on Free-Text Radiology Reports From Patients With Glioblastoma. JCO Clin Cancer Inform. 2020;4:25-34.

68. Huhdanpaa HT, Tan WK, Rundell SD, Suri P, Chokshi FH, Comstock BA, Heagerty PJ, James KT, Avins AL, Nedeljkovic SS, Nerenz DR, Kallmes DF, Luetmer PH, Sherman KJ, Organ NL, Griffith B, Langlotz CP, Carrell D, Hassanpour S, Jarvik JG. Using Natural Language Processing of Free-Text Radiology Reports to Identify Type 1 Modic Endplate Changes. J Digit Imaging. 2018;31(1):84-90.

69. Chen L, Song L, Shao Y, Li D, Ding K. Using natural language processing to extract clinically useful information from Chinese electronic medical records. Int J Med Inform. 2019;124:6-12.

70. Dahl FA, Rama T, Hurlen P, Brekke PH, Husby H, Gundersen T, Nytrø Ø, Øvrelid L. Neural classification of Norwegian radiology reports: using NLP to detect findings in CT-scans of children. BMC Med Inform Decis Mak. 2021;21(1):84.

71. Liu H, Zhang Z, Xu Y, Wang N, Huang Y, Yang Z, Jiang R, Chen H. Use of BERT (Bidirectional Encoder Representations from Transformers)-Based Deep Learning Method for Extracting Evidences in Chinese Radiology Reports: Development of a Computer-Aided Liver Cancer Diagnosis Framework. J Med Internet Res. 2021;23(1):e19689.

72. Casey A, Davidson E, Poon M, Dong H, Duma D, Grivas A, Grover C, Suárez-Paniagua V, Tobin R, Whiteley W, Wu H, Alex B. A systematic review of natural language processing applied to radiology reports. BMC Med Inform Decis Mak. 2021;21(1):179.

73. Davidson EM, Poon MTC, Casey A, Grivas A, Duma D, Dong H, Suárez-Paniagua V, Grover C, Tobin R, Whalley H, Wu H, Alex B, Whiteley W. The reporting quality of natural language processing studies: systematic review of studies of radiology reports. BMC Med Imaging. 2021;21(1):142.

74. Kim C, Zhu V, Obeid J, Lenert L. Natural language processing and machine learning algorithm to identify brain MRI reports with acute ischemic stroke. PLoS One. 2019;14(2):e0212778.

75. Deshmukh N, Gumustop S, Gauriau R, Buch V, Wright B, Bridge C, Naidu R, Andriole K, Bizzo B. Semi-Supervised Natural Language Approach for -Grained Classification of Medical Reports. arxiv preprint. 2019; arXiv:1910.13573.

76. Miao S, Xu T, Wu Y, Xie H, Wang J, Jing S, Zhang Y, Zhang X, Yang Y, Zhang X, Shan T, Wang L, Xu H, Wang S, Liu Y. Extraction of BI-RADS findings from breast ultrasound reports in Chinese using deep learning approaches. Int J Med Inform. 2018;119:17-21.

77. Hassanpour S, Bay G, Langlotz CP. Characterization of Change and Significance for Clinical Findings in Radiology Reports Through Natural Language Processing. J Digit Imaging. 2017;30(3):314-22.

78. Chen PH, Zafar H, Galperin-Aizenberg M, Cook T. Integrating Natural Language Processing and Machine Learning Algorithms to Categorize Oncologic Response in Radiology Reports. J Digit Imaging. 2018;31(2):178-84.

79. Hong Y, Zhang J. Investigation of Terminology Coverage in Radiology Reporting Templates and Free-text Reports. International Journal of Knowledge Content Development & Technology. 2015;5:5-14.

80. Brown AD, Marotta TR. A Natural Language Processing-based Model to Automate MRI Brain Protocol Selection and Prioritization. Acad Radiol. 2017;24(2):160-6.

81. Dalal S, Hombal V, Weng WH, Mankovich G, Mabotuwana T, Hall CS, Fuller J 3rd, Lehnert BE, Gunn ML. Determining Follow-Up Imaging Study Using Radiology Reports. J Digit Imaging. 2020;33(1):121-30.

82. Carrodeguas E, Lacson R, Swanson W, Khorasani R. Use of Machine Learning to Identify Follow-Up Recommendations in Radiology Reports. J Am Coll Radiol. 2019;16(3):336-43.

83. Callen AL, Dupont SM, Price A, Laguna B, McCoy D, Do B, Talbott J, Kohli M, Narvid J. Between Always and Never: Evaluating Uncertainty in Radiology Reports Using Natural Language Processing. J Digit Imaging. 2020;33(5):1194-201.

Eidesstattliche Versicherung

„Ich, Lina Xu, versichere an Eides statt durch meine eigenhändige Unterschrift, dass ich die vorgelegte Dissertation mit dem Thema: Klassifikation von computertomographischen Befundtexten des Thorax anhand von Deep Learning / Classification of Computed Tomography Findings of the Chest Based on Deep Learning selbstständig und ohne nicht offengelegte Hilfe Dritter verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel genutzt habe.
Alle Stellen, die wörtlich oder dem Sinne nach auf Publikationen oder Vorträgen anderer Autoren/innen beruhen, sind als solche in korrekter Zitierung kenntlich gemacht. Die Abschnitte zu Methodik (insbesondere praktische Arbeiten, Laborbestimmungen, statistische Aufarbeitung) und Resultaten (insbesondere Abbildungen, Graphiken und Tabellen) werden von mir verantwortet.

Ich versichere ferner, dass ich die in Zusammenarbeit mit anderen Personen generierten Daten, Datenauswertungen und Schlussfolgerungen korrekt gekennzeichnet und meinen eigenen Beitrag sowie die Beiträge anderer Personen korrekt kenntlich gemacht habe (siehe Anteilserklärung). Texte oder Textteile, die gemeinsam mit anderen erstellt oder verwendet wurden, habe ich korrekt kenntlich gemacht.

Meine Anteile an etwaigen Publikationen zu dieser Dissertation entsprechen denen, die in der untenstehenden gemeinsamen Erklärung mit dem/der Erstbetreuer/in, angegeben sind. Für sämtliche im Rahmen der Dissertation entstandenen Publikationen wurden die Richtlinien des ICMJE (International Committee of Medical Journal Editors; www.icmje.og) zur Autorenschaft eingehalten. Ich erkläre ferner, dass ich mich zur Einhaltung der Satzung der Charité – Universitätsmedizin Berlin zur Sicherung Guter Wissenschaftlicher Praxis verpflichte.

Weiterhin versichere ich, dass ich diese Dissertation weder in gleicher noch in ähnlicher Form bereits an einer anderen Fakultät eingereicht habe.

Die Bedeutung dieser eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unwahren eidesstattlichen Versicherung (§§156, 161 des Strafgesetzbuches) sind mir bekannt und bewusst.

Datum Unterschrift

Lebenslauf

Mein Lebenslauf wird aus datenschutzrechtlichen Gründen in der elektronischen Version meiner Arbeit nicht veröffentlicht.

Danksagung

An dieser Stelle möchte ich den Menschen meinen großen Dank aussprechen, die mich bei der Anfertigung der Doktorarbeit unterstützt haben.

Zunächst möchte ich Frau PD Dr. med. Lisa Adams für die Ermöglichung der Promotion und Vermittlung des Themas danken. Ihre sehr freundliche und zugewandte Art war mir eine große Hilfe bei dem Verfassen der Dissertation.

Mein besonderer Dank gilt Herrn PD Dr. med. Keno-Kyrill Bressem für die enorme Unterstützung bei der Umsetzung der gesamten Arbeit. Ich bin äußerst dankbar für den häufigen Austausch und die Motivation zu jedem Zeitpunkt der Promotion.

Abschließend möchte ich meinen Eltern und meiner Schwester für ihre Begleitung meines Ausbildungswegs und ihren stetigen Zuspruch danken.

Bescheinigung Statistik

CHARITÉ

CharitéCentrum für Human- und Gesundheitswissenschaften

Charité | Campus Charité Mitte | 10117 Berlin

Institut für Biometrie und Klinische Epidemiologie (iBikE)

Direktorin: Prof. Dr. Geraldine Rauch

Postantschrift:
Charitéplatz 1 | 10117 Berlin
Besucheranschrift:
Reinhardtstr. 58 | 10117 Berlin

Tel. +49 (0)30 450 562171
geraldine.rauch@charite.de
https://biometrie.charite.de/

Name, Vorname: Xu, Lina

Emailadresse: lina.xu@charite.de

Matrikelnummer: 224376

Promotionsbetreuer*in: Bressem, Kyrill

Promotionsinstitution/Klinik: Institut für Radiologie

Bescheinigung

Hiermit bescheinige ich, dass Frau *Lina Xu* innerhalb der Service Unit Biometrie des Instituts für Biometrie und Klinische Epidemiologie (iBikE) bei mir eine statistische Beratung zu einem Promotionsvorhaben wahrgenommen hat. Folgende Beratungstermine wurden wahrgenommen:

- *Termin 1: 19.04.2021*

Folgende wesentliche Ratschläge hinsichtlich einer sinnvollen Auswertung und Interpretation der Daten wurden während der Beratung erteilt:

- *F1-Score zur Bewertung der Güte des Klassifikationsmodells heranziehen*

Diese Bescheinigung garantiert weder die richtige Umsetzung der in der Beratung gemachten Vorschläge, die korrekte Durchführung der empfohlenen statistischen Verfahren noch die richtige Darstellung und Interpretation der Ergebnisse. Die Verantwortung hierfür obliegt allein dem Promovierenden. Das Institut für Biometrie und Klinische Epidemiologie übernimmt hierfür keine Haftung.

Datum: 19.04.2021 Name des Beraters: Claus Nowak

[redacted]

Unterschrift Berater, Institutsstempel

CHARITÉ
UNIVERSITÄTSMEDIZIN BERLIN
Institut für Biometrie und
Klinische Epidemiologie
Campus Charité Mitte
Charitéplatz 1 | D-10117 Berlin
Sitz: Reinhardtstr. 58

Claus Peter Nowak
Digital unterschrieben von Claus Peter Nowak
Datum: 2021.04.19 08:41:22 +02'00'