Akustische Covid-Diagnose durch Husten

Praktikum Künstliche Intelligenz

Wintersemester 2024/2025

Paula Schwalm
paula.schwalm@mni.thm.de
Technische Hochschule Mittelhessen

Abstract

Die vorliegende Arbeit untersucht den Einsatz von Künstlicher Intelligenz (KI) zur Diagnose von COVID-19 anhand von Hustengeräuschen. Der Fokus liegt dabei auf der Entwicklung und Evaluation verschiedener neuronaler Netzwerke mit besonderem Schwerpunkt auf dem Vergleich kausaler und nicht-kausaler Modelle, um die Relevanz der zeitlichen Abfolge bei der Diagnose zu untersuchen. Dafür werden Convolutional Neural Networks (CNN) und Long Short-Term Memory (LSTM) Modelle betrachtet und auf einem Datensatz von 15377 Hustengeräuschen trainiert und getestet.Die Ergebnisse zeigen, dass beide Architekturen für die Problemstellung geeignet sind. Beim Vergleich zwischen kausalen und nicht-kausalen Modellen konnte festgelegt werden, dass die zeitliche Abfolge für die Diagnose hier als nicht entscheidend betrachtet werden kann.

Keywords Künstliche Intelligenz · Audioklassifizierung · COVID-19

1 Einleitung

> „Die Art des Geräusches ändert sich, wenn Sie Covid haben – selbst wenn Sie noch asymptomatisch sind.“
>
> — Brian Subirana (Übersetzung: [1] Original: [2])

Die COVID-19-Pandemie, ausgelöst durch das SARS-CoV-2-Virus im Dezember 2019 [3], wurde im März 2020 von der Weltgesundheitsorganisation (englisch: Word Health Organization, WHO) offiziell zur Pandemie erklärt [4] und konnte erst 2022 durch Impfstoffentwicklung eingedämmt werden [5].

Angesichts der großen Symptomvielfalt – von leichten Erkältungserscheinungen bis zu lebensbedrohlichen Lungenentzündungen [6] – haben Forscher:innen einen neuen Diagnoseansatz entwickelt. Da 67,7% der untersuchten Patienten laut einer Studie der WHO unter trockenem Husten litten [7], untersuchen wissenschaftliche Studien die Möglichkeit, COVID-19 anhand von Hustengeräuschen mittels Künstlicher Intelligenz (KI) zu diagnostizieren. Das Ergebnis wird in vielversprechenden Studien ([8], [9], [10]) dargestellt, die zeigen, dass eine solche Audioklassifizierung möglich ist. Dieser Ansatz bietet somit eine kostengünstige, nicht-invasive und schnelle Alternative zu herkömmlichen Testverfahren und demonstriert das Potenzial von Deep Learning in der medizinischen Diagnostik.
Diese Arbeit zielt auf die Entwicklung und Untersuchung verschiedener neuronaler Netzwerke für diese Problemstellung ab. Der Fokus liegt nicht auf der Ermittlung des optimalen Ansatzes, sondern auf der Erforschung unterschiedlicher Lösungswege. Der Schwerpunkt liegt auf dem Vergleich kausaler und nicht-kausaler Modelle, die sich in der zeitlichen Verarbeitung der für die Diagnose verwendeten Daten unterscheiden. Bei einer kausalen Implementierung werden die Daten streng in ihrer zeitlichen Reihenfolge verarbeitet, d.h. es werden nur Informationen aus der Vergangenheit und dem aktuellen Moment betrachtet. Im Gegensatz dazu haben nicht-kausale Modelle Zugriff auf die gesamte Sequenz und können beim Analysieren eines bestimmten Zeitpunkts auch „in die Zukunft schauen“. Diese Untersuchung soll zeigen, ob die zeitliche Struktur der Hustengeräusche für die Diagnose relevant ist und ob sich die Genauigkeit der Modelle durch die Berücksichtigung der zeitlichen Abfolge verbessert.

Um realitätsnahe Ergebnisse zu gewährleisten, werden die Audiodaten ohne Vorverarbeitung (keine Filterung von Hintergrundgeräuschen o.ä.) und ohne Metadaten verwendet, was die künftige Erweiterung des Datensatzes erleichtert.

2 Forschungsstand

Eines der ersten Forschungsergebnisse zur Erkennung von COVID-19 durch die Analyse von Hustengeräuschen wurde von Wissenschaftler:innen des Massachusetts Institute of Technology veröffentlicht [2]. Zur Erkennung entwickelten sie ein Sprachverarbeitungsframework, das akustische Biomarker-Funktionsextraktoren nutzt und auf einem Convolutional Neural Network (CNN) basiert. Dadurch konnten sie eine Erkennungsquote von 98,5% bei nachweislich infizierten Personen erreichen und 100% bei symptomfreien Probanden:innen [11], wobei das Modell an 4256 Probanden:innen getestet worden ist. Aufgrund der vielversprechenden Ergebnisse wurden während der COVID-19-Pandemie diverse Ansätze zur Problemlösung entwickelt. Neben Einzelbeobachtungen wurden auch Wettbewerbsanalysen durchgeführt, um die besten Modelle zu identifizieren, wie in der Studie „Cough Audio Analysis for COVID-19 Diagnosis“ [8].

Die Literaturrecherche zeigt, dass sich die Ansätze in wesentlichen Faktoren unterscheiden: in der Datenvorverarbeitung, der Modellwahl und -komplexität sowie der Auswahl des Datensatzes. Aufgrund des Endes der Pandemie und der damit verbundenen sinkenden Relevanz von COVID-19 ist die Forschung in diesem speziellen Anwendungsfall rückläufig oder wird nicht mehr weiterverfolgt. Forschende und Unternehmen verlagern ihren Fokus auf andere medizinische Bereiche, etwa die KI-gestützte Erkennung von Tuberkulose und anderen Krankheiten [12].

3 Datensatz

Es stehen verschiedene Datensätze mit Aufnahmen von Husten zur Verfügung. Die größten Datensätze, wie etwa die der Cambridge University, können aufgrund ihrer eingeschränkten Lizensierbarkeit [13] hier aber nicht weiter betrachtet werden. Aus diesem Grund wurden die Open-Source-Datensätze von Virufy [14], Coswara [15] und Coughvid [16] verwendet.

Nach der Bereinigung der Daten und dem Entfernen ungültiger Elemente konnte ein Datensatz von 15377

Hustengeräuschen erstellt werden. Die Datensätze von Coswara enthielten teilweise detaillierte Metadaten sowie eine erweiterte Klassifizierung des Gesundheitszustandes. Um eine binäre Klassifizierung zu ermöglichen, wurden diese Daten einer der beiden Klassen (COVID-19 positiv oder negativ) zugeordnet. Der Datensatz von Coswara bietet zudem bei einigen Datensätzen noch Daten zu starkem Husten (die Audiodateien mit dem Namen „cough-heavy.wav“ in den Coswara) an, was ebenfalls einbezogen wurde, um die Datenbasis zu erweitern. Eine detaillierte Übersicht über die Datenverteilung ist in der Tabelle 1 dargestellt.

Datensatz	**Positiver Husten**	**Negativer Husten**
Virufy	7	9
Coswara	1356	3098
Coughvid	924	9983
Gesamt	2287	13090

Tabelle 1: Datenverteilung der verwendeten Datensätze nach der Bereinigung

Da viele Audioaufnahmen mehrere wiederholte Huster enthalten, während andere nur einen Huster enthalten, könnte man den Datensatz auch aufteilen, um einzelne Huster als unabhängige Datenpunkte zu betrachten. Auf diese Weise könnte der Datensatz drastisch erweitert werden. Diese Möglichkeit wurde aber hier nicht weiter verfolgt, da zu Beginn festgelegt wurde, dass die Daten so wenig wie möglich vorverarbeitet werden sollen. Der potenzielle Nutzen einer solchen Erweiterung wird aber hier nicht verkannt.

Der Datensatz ist in Bezug auf die Größe immer noch begrenzt und weist zudem ein sehr unausgewogenes Verhältnis zwischen positiven und negativen Klassen auf. Angesichts des kleinen Datensatzes werden die Daten im Verhältnis 80-20 aufgeteilt, wobei 80% für das Training und 20% für den Test verwendet werden.

4 Methoden

4.1 Feature-Extraktion mit MFCC (Mel-Frequency Cepstral Coefficients)

Um die Audiodaten in neuronalen Netzwerken zu verarbeiten, müssen diese zunächst in ein geeignetes Format umgewandelt werden. Ein bewährtes Verfahren stellt dabei die Umwandlung in Mel-Frequency-Cepstral-Koeffizienten (MFCC) dar. Diese Methode extrahiert die wichtigsten akustischen Merkmale des Audiosignals [17] und ermöglicht dadurch eine kompakte Darstellung des Frequenzspektrums. Die MFCCs liegen hier in Form von 128x128-Pixel-Bildern vor, die anschließend in eindimensionale Feature-Vektoren der Länge 128 umgewandelt werden, indem der Mittelwert entlang der vertikalen Achse berechnet wird. Dadurch entsteht ein (1D) Feature-Vektor, der als Eingabe für die neuronalen Netzwerke dient.

4.2 Modellauswahl

Es wurden zwei verschiedene Architekturen für die Entwicklung von neuronalen Netzwerken verwendet: CNN und Long Short-Term Memory (LSTM). Die Auswahl erfolgte aufgrund ihrer Eignung für die Problemstellung, aber auch wegen ihrer häufigen Verwendung in der Praxis. Ursprünglich war auch die Verwendung von Transformer-Modellen geplant, jedoch konnte dies aufgrund von Zeitmangel nicht weiter verfolgt werden. Je Architektur werden zwei Modelle K, N entwickelt, wobei die Hyperparameter gleich bleiben. Bei K handelt es sich um eine kausale Implementierung und bei N um eine nicht-kausale Implementierung der Architektur. Damit stehen die folgende Mengen $K_{\{\text{Architektur}\}}$ zur Verfügung:

$$K_{\{\text{Architektur}\}} = \left\{K_{\{\text{CNN}\}}, N_{\{\text{CNN}\}}, K_{\{\text{LSTM}\}}, N_{\{\text{LSTM}\}}\right\}$$

5 Experimente und Ergebnisse

Für die Experimente wurden die Modelle auf den Trainingsdaten trainiert und auf den Validierungsdaten getestet. Bei jeder Erstellung eines Modells wurden verschiedene Hyperparameter und Architekturen getestet, um möglichst gute Ergebnisse zu erzielen. Am Ende jeder Testphase wurde das Modell mit den besten Ergebnissen ausgewählt und auf den Testdaten getestet.

5.1 CNN (Convolutional Neural Network)

Die Architektur des CNN ist auf Abbildung 1 dargestellt und zeigt, dass sich beide Elemente $K_{\{\text{CNN}\}}$ und $N_{\{\text{CNN}\}}$ in ihrer Padding-Strategie unterscheiden.

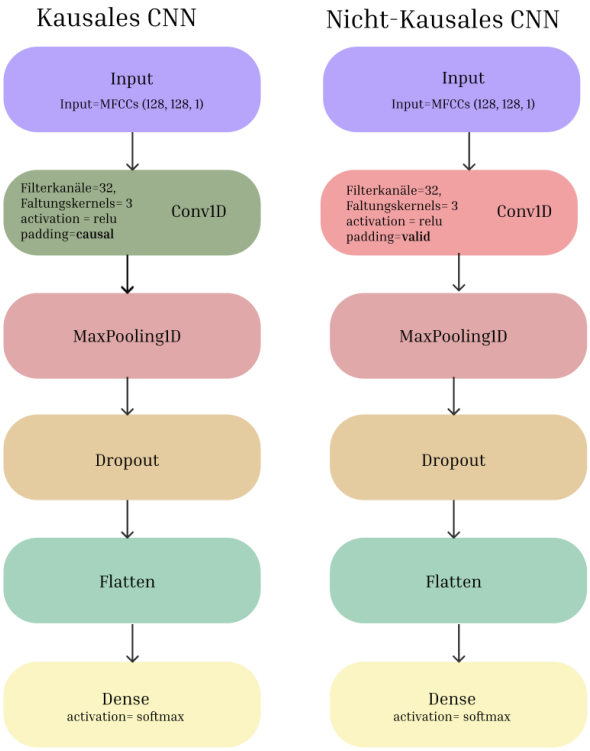

Abbildung 1: Architektur der LSTM-Modelle

Es handelt sich dabei um unterschiedliche Ansätze zur Behandlung von Sequenzgrenzen in den Eingabedaten. Modell $K_{\{\text{CNN}\}}$ verwendet kausales Padding („causal“), während Modell $N_{\{\text{CNN}\}}$ nicht-kausales Padding („valid“) einsetzt. Das kausale Padding von Modell $K_{\{\text{CNN}\}}$ gewährleistet die temporale Kausalität der Modellausgabe, indem sichergestellt wird, dass die Voraussage zu einem spezifischen Zeitpunkt t ausschließlich von Informationen bis einschließlich dieses Zeitpunkts beeinflusst wird, niemals jedoch von nachfolgenden Datenpunkten. Im Gegensatz dazu ermöglicht die nicht-kausale Konfiguration von $N_{\{\text{CNN}\}}$ den bidirektionalen Zugriff auf die vollständige Sequenz. Die zwei Modelle wurden anhand der Genauigkeit auf den Testdaten verglichen, wobei 1 für eine perfekte Vorhersage und 0 für eine vollständig falsche Vorhersage steht. Das Ergebnis zeigt, dass $N_{\{\text{CNN}\}}$ eine höhere Genauigkeit (0.7412) imm Vergleich zu $K_{\{\text{CNN}\}}$ (0.7350) aufweist, wobei der Unterschied als vernachlässigbar betrachtet wird (0.0062).

5.2 LSTM (Long Short-Term Memory)

$K_{\{\text{LSTM}\}}$ und $N_{\{\text{LSTM}\}}$ unterscheiden sich durch eine Schicht in ihrer Architektur voneinander, wie auf Abbildung 2 dargestellt und dadurch auch in ihrer Informationsverarbeitung.

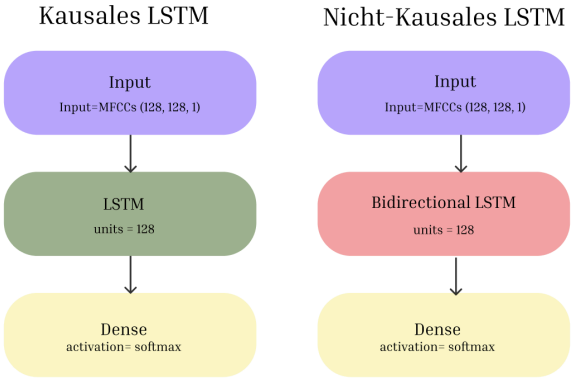

Abbildung 2: Architektur der LSTM-Modelle

Das Modell $K_{\{\text{LSTM}\}}$ verwendet eine unidirektionale LSTM-Schicht, die Sequenzdaten nur in vorwärts gerichteter Richtung verarbeitet. Dadurch stellt es sicher, dass jede Vorhersage ausschließlich auf vergangenen und aktuellen Informationen basiert, ohne zukünftige Datenpunkte zu berücksichtigen – es handelt sich also um eine kausale Implementierung. $K_{\{\text{LSTM}\}}$ wird durch eine Dense-Ausgangsschicht mit Softmax-Aktivierung für die Klassifikation abgeschlossen. Im Gegensatz dazu nutzt $N_{\{\text{LSTM}\}}$ eine bidirektionale LSTM-Schicht, die Sequenzdaten sowohl vorwärts als auch rückwärts verarbeitet. Dies ermöglicht dem Modell, Informationen aus der gesamten Sequenz – einschließlich zukünftiger Zeitpunkte – zu verwenden, wodurch es nicht-kausal ist.

Die experimentellen Ergebnisse zeigen, dass $N_{\{\text{LSTM}\}}$ eine minimale höhere Genauigkeit (0.7796) im Vergleich zu $K_{\{\text{LSTM}\}}$ (0.7783) aufweist, wobei der Unterschied als vernachlässigbar betrachtet wird (0.0013).

5.3 Ergebnis

Die Abbildung 3 veranschaulicht die Genauigkeit der verschiedenen Modelle.

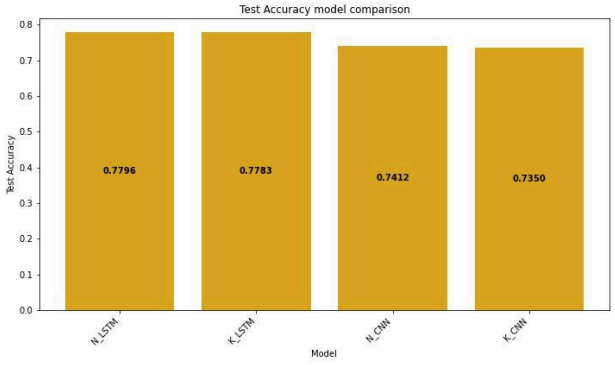

Abbildung 3: Genauigkeit der verschiedenen Modelle

Insgesamt deuten diese Ergebnisse darauf hin, dass beide Modellvarianten bei beiden Architekturen A vergleichbare Leistungen erbringen, da die

Unterschiede in der Genauigkeit minimal sind. Die LSTM-basierten Modelle erreichten insgesamt eine höhere Genauigkeit (0.7796) als die CNN-basierten Modelle (0.7412). Da die Literaturrecherche zeigte, dass CNN-Modellvarianten häufig in der Audioklassifizierung eingesetzt werden und dabei hohe Genauigkeiten erzielen, ist davon auszugehen, dass die aktuelle Implementierung noch starkes Optimierungspotenzial aufweist. Weitere Experimente sind daher notwendig, um die CNN-Modellarchitektur zu verfeinern und ihre Leistungsfähigkeit für diese spezifische Anwendung zu steigern.

5.4 Fazit und Ausblick

Das Ziel der Arbeit, die Entwicklung und Untersuchung verschiedener neuronaler Netzwerke für die Diagnose von COVID-19 durch Hustengeräusche, konnte erreicht werden. Die Ergebnisse zeigen, dass beide Architekturen, CNN und LSTM, für die Problemstellung geeignet sind und in der Lage sind, eine gute Genauigkeit zu erreichen. Die Untersuchung von kausalen und nicht-kausalen Modellen zeigt, dass die zeitliche Abfolge für die Diagnose hier als nicht entscheidend betrachtet werden kann. Um ein aussagekräftigeres Ergebnis zu erzielen, sollten weitere Experimente durchgeführt werden, um die Modelle weiter zu optimieren und zu validieren. Die weitere Betrachtung anderer Architekturen, wie etwa Transformer-Modelle, könnte ebenfalls weitere Erkenntnisse liefern. Außerdem sollte die Erweiterung des Datensatzes in Betracht gezogen werden, um die Datenbasis zu vergrößern und die Generalisierungsfähigkeit der Modelle zu verbessern. Die Ergebnisse dieser Arbeit zeigen jedoch, wie auch bereits vorangestellte Studien, dass die Diagnose von COVID-19 durch Hustengeräusche mittels KI möglich ist und ein großes Potenzial für die medizinische Diagnostik bietet.

6 Literaturverzeichnis

[1] Oliver Bünte, Zugegriffen: 25. November 2024. [Online]. Verfügbar unter: https://www.heise.de/news/KI-soll-COVID-19-Infizierte-am-Husten-erkennen-4945746.html

[2] Zoe Kleinman, Zugegriffen: 25. November 2024. [Online]. Verfügbar unter: https://www.bbc.com/news/technology-54780460

[3] Zugegriffen: 25. November 2024. [Online]. Verfügbar unter: https://www.dzif.de/de/glossar/sars-cov-2

[4] Zugegriffen: 25. November 2024. [Online]. Verfügbar unter: https://www.tagesschau.de/ausland/europa/coronavirus-317.html

[5] Zugegriffen: 25. November 2024. [Online]. Verfügbar unter: https://www.mdr.de/wissen/ende-corona-pandemie-weltweit-zweitausenddreiundzwanzig-deutschland-endemischer-zustand-100.html

[6] Zugegriffen: 25. November 2024. [Online]. Verfügbar unter: https://www.infektionsschutz.de/coronavirus/basisinformationen/symptome-und-krankheitsverlauf/

[7] Zugegriffen: 25. November 2024. [Online]. Verfügbar unter: https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-final-report.pdf

[8] T. P. &. B. G. Teghdeep Kapoor, Zugegriffen: 25. November 2024. [Online]. Verfügbar unter: https://doi.org/10.1007/s42979-022-01522-1

[9] [Online]. Verfügbar unter: http://dx.doi.org/10.3233/IDT-210206

[10] S. H. Rob Dunne Tim Morris, Zugegriffen: 3. Januar 2025. [Online]. Verfügbar unter: https://doi.org/10.21203/rs.3.rs-63796/v1

[11] B. S. Jordi Laguarta Ferran Hueto, Zugegriffen: 30. November 2024. [Online]. Verfügbar unter: https://www.embs.org/ojemb/articles/covid-19-artificial-intelligence-diagnosis-using-only-cough-recordings/

[12] Shravya Shetty, Zugegriffen: 28. Februar 2025. [Online]. Verfügbar unter: https://blog.google/technology/health/ai-model-cough-disease-detection/

[13] Zugegriffen: 15. Dezember 2024. [Online]. Verfügbar unter: https://www.covid-19-sounds.org/en/blog/neurips_dataset.html

[14] [Online]. Verfügbar unter: https://github.com/virufy/virufy_data

[15] [Online]. Verfügbar unter: https://github.com/iiscleap/Coswara-Data

[16] [Online]. Verfügbar unter: https://www.kaggle.com/datasets/andrewmvd/covid19-cough-audio-classification

[17] Neeraj Sharma, Zugegriffen: 28. Februar 2025. [Online]. Verfügbar unter: https://iiscleap.github.io/coswara-blog/coswara/tutorial/2020/08/20/mfcc.html