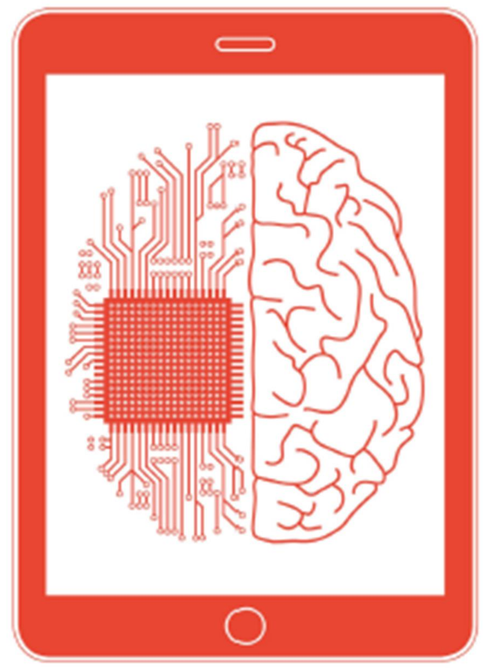
Fachhochschule Nordwestschweiz
Hochschule für Life Sciences

# Automatische Datenextraktion aus Anamnesebögen

**Lukas Kamber**
Bachelor-Thesis, Studienrichtung Medizininformatik

Auftraggeber: Prof. Dr. Bernd Stadlinger, UZH Zentrum für Zahnmedizin
Expert/in: Prof. Dr. Björn Menze
Verantwortliche/r: Prof. Dr. Abdullah Kahraman



**Einleitung**:

Ein Großteil der Patentendaten wird auch heute noch manuell und auf Papier erfasst und verarbeitet. Dies ist mit einem beträchtlichen Zeitaufwand sowie einer hohen Fehleranfälligkeit verbunden. Die manuelle Erfassung von Daten birgt das Risiko von Übertragungsfehlern, Lesefehlern und einer verzögerten Verarbeitung. Des Weiteren ist die manuelle Verarbeitung großer Datenmengen ineffizient und kann zu Engpässen führen, wodurch die Qualität der Patientenversorgung beeinträchtigt wird [1], [2].

Das Ziel der vorliegenden Untersuchung bestand in der Extraktion relevanter Daten aus Patientenfragebögen mithilfe von Bildverarbeitungsalgorithmen und Optical Character Recognition (OCR). Die erfolgreiche Automatisierung der Datenextraktion würde nicht nur die Effizienz der Datenverarbeitung erheblich steigern, sondern auch die Qualität der Datenanalyse verbessern und letztlich zu einer besseren Patientenversorgung beitragen. In der Arbeit soll die Frage beantwortet werden, wie gut sich aktuelle Open-Source OCR-Technologien zur Extraktion von Patientendaten in Anamnesefragebögen eignen.

**Umsetzung**:

Für die Bildvorverarbeitung wurde die Python-Bibliothek “OpenCV” verwendet. Dies beinhaltet die Binarisierung (schwarz/weiß) des Bildes und das Geraderücken, falls das Bild schräg eingescannt wurde. Zusätzlich wurde OpenCV zur Erkennung und Zuordnung der Checkboxmarkierungen verwendet.

Als OCR-Engine wurde Tesseract eingesetzt. Mit Tesseract konnte der Text in den Dokumenten erfasst und extrahiert werden. Das Skript wurde dann auf über 31’000 Patienten angewendet. Dafür war der Einsatz eines High Performance Computers (HPC) sinnvoll.

Die erkannten Checkboxen wurden anhand ihrer X- und Y-Koordinaten der entsprechenden Frage zugewiesen und durch die Koordinaten der “Ja”- und “Nein”-Worte in einem Formular mit dem Wert gekennzeichnet. Eine Markierung oder Nichtmarkierung wurde durch die Anzahl Pixel in den gefundenen Rechtecken definiert.

**Ergebnisse**:

Mit OpenCV wurden die Checkboxen mit dem Wert “Ja” grün, die Checkboxen mit dem Wert “Nein” rot und die nicht angekreuzten Checkboxen gelb markiert.

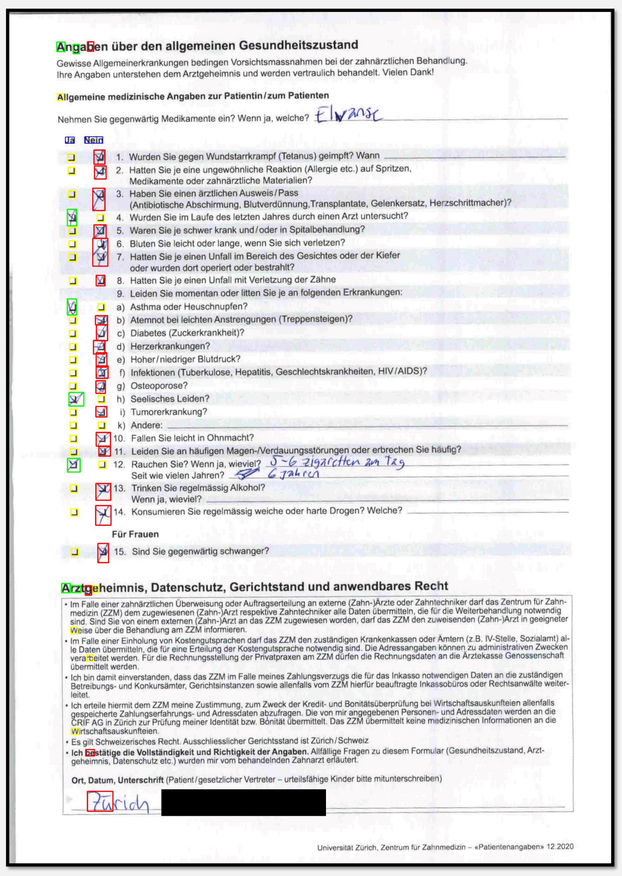


Abbildung 1: Markierte Checkboxen durch OpenCV

Der Accuracy Score über alle Daten für die Erkennung von “Ja”, “Nein”, “Not marked” und “Both marked” betrug 85.93%. Eine detailliertere Darstellung der Fehlerquellen findet sich in der Confusion Matrix.

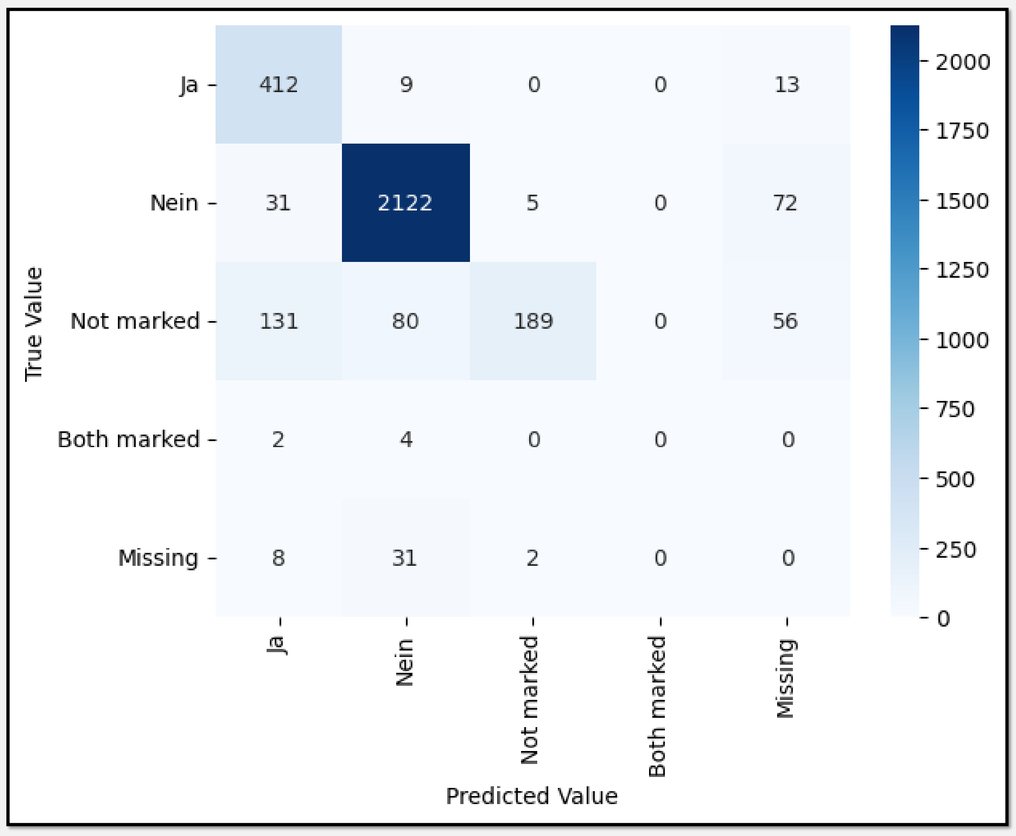



Abbildung 2: Confusion Matrix zur Auswertung der korrekten und falschen Ergebnissen

**Schlussfolgerung**:

Die Erkennung der Checkboxen erfolgt mit einer hohen Genauigkeit von etwa 86 Prozent. Die Erkennung von nicht markierten oder beidseitig markierten Feldern stellt sich jedoch als besondere Herausforderung dar. Eine weitere signifikante Anzahl von Fehlern resultierte aus der Fehlinterpretation von Textzeilen als Fragen.

Die vorliegende Untersuchung belegt, dass aktuelle Open-Source-OCR-Technologien zur Extraktion von Patientendaten aus Anamnesefragebögen grundsätzlich geeignet sind, jedoch noch Optimierungspotenzial aufweisen.

**Quellenverzeichnis:**

[1] N. Tavabi, M. Singh, J. Pruneski, und A. M. Kiapour, „Systematic evaluation of common natural language processing techniques to codify clinical notes“, PLOS ONE, Bd. 19, Nr. 3, S. e298892, März 2024, doi: 10.1371/journal.pone.0298892.

[2] K. Arnold, G. A. Mugisha, F.-M. Uzoka, S. Imanirakiza, C. Muhumuza, und J. N. Bukenya, „Development of an e-Health System for Improving Health-Care Access in Developing Countries“, Proceedings of the Future Technologies Conference (FTC) 2021, Volume 2, Bd. 359. Springer International Publishing, Cham, S. 607–616, 2022. doi: 10.1007/978-3-030-89880-9_45.