Entwicklung und Evaluation von Deep-Learning-Modellen für die dermatopathologische Routinediagnostik am Beispiel aktinischer Keratosen

Julius Balkenhol [1,2]; Maximilian Schmidt [3,4]; Jean Le'Clerc Arrastia [3,4]; Daniel Otero Baguer [3,4]; Johannes Leuscher [3,4]; T. Schnauder [3], J. Langhorst [3], Georgia Gilbert [5], Fabian Hörst [4]; Thomas Dirschka [1,2]; Lutz Schmitz [1,6]

[1] CentroDerm GmbH, Praxis für Dermatologie, Institut für Dermatopathologie
[2] Fakultät für Gesundheit, Universität Witten-Herdecke
[3] Zentrum für Industrielle Mathematik, Universität Bremen
[4] Aisencia GmbH, Lise-Meitner-Str. 4, 28359 Bremen
[5] NHS Lothian, Edinburgh, United Kingdom
[6] Klinik für Dermatologie, Venereologie und Allergologie, Ruhr Universität Bochum

Einleitung

Deep Learning Modelle (DLM) erreichen in der Dermatopathologie hohe diagnostische Genauigkeiten zwischen 95 und 98% [1–4]. Bisher fehlen jedoch prozessorientierte Ansätze für den Einsatz in der Routinediagnostik. Ziel dieser Arbeit war die Entwicklung und Evaluation von DLM entlang zentraler diagnostischer Teilschritte. Dazu wurde die Aktinische Keratose (AK) als häufige dermatopathologische Diagnose mit einem breiten morphologischen Spektrum gewählt.

Methode

Es wurden >1.000 Einzelfälle mit AK aus dem CentroDermPath-Institut für Dermatopathologie Wuppertal mit einem Hamamatsu S210 Scanner digitalisiert. Pixelgenaue Annotationen erfolgten durch zwei Dermatopathologen mit Konsensus-Validierung. Auf dieser Grundlage wurden mittels supervised learning modifizierte U-Net-Architekturen trainiert [5]. Entsprechend des diagnostischen Workflows wurden mehrere Klassifikationsaufgaben definiert: Detektion von AK (*Abb. 1*), Grading nach AKI–III und PROI–III und die Erfassung histomorphologischer Varianten (Atrophe, Akantholytische, Bowenoide, Hyperkeratotische, Lichenoide, Pigmentierte).

Literatur

1. Le'Clerc Arrastia J, Heilenkötter N, Otero Baguer D, Hauberg-Lotte L, Boskamp T, Hetzer S, et al. Deeply Supervised UNet for Semantic Segmentation to Assist Dermatopathological Assessment of Basal Cell Carcinoma. 2021
2. Jansen P, Baguer DO, Duschner N, Le'Clerc Arrastia J, Schmidt M, Wiepjes B, et al. Evaluation of a Deep Learning Approach to Differentiate Bowen's Disease and Seborrheic Keratosis. Cancers. 2022
3. Duschner N, Baguer DO, Schmidt M, Griewank KG, Hadaschik E, Hetzer S, et al. Applying an artificial intelligence deep learning approach to routine dermatopathological diagnosis of basal cell carcinoma. 2023
4. Jansen P, Arrastia JL, Baguer DO, Schmidt M, Landsberg J, Wenzel J, et al. Deep learning based histological classification of adnex tumors. European Journal of Cancer. 2024
5. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF. 2015

Ergebnisse & Diskussion

Die grundsätzliche Anwendbarkeit auf AKs wurde in einer Fallkontrollstudie mit 977 Fällen durchgeführt (Tab.1). Das auf AK trainierte Modell erreichte eine Gesamtgenauigkeit von 98,9% mit pixelgenauer Übereinstimmung von 78,8% auf dem Patch Level (1024x1024 Px). Für das Grading der AK wurde ein zweites Modell trainiert, das eine Spezifizität von 91,5% und Sensitivität von 86,7% über alle sieben Klassen erreichte (Abb. 2). Die Klasse AK-I zeigte im Vergleich eine deutlich schlechtere Sensitivität von 46,0%, was die Gesamtleistung über alle Klassen negativ beeinflusste. Dabei war AK-I diejenige Klasse, die als einzige Klasse eine deutlich geringere Menge an Trainingsdaten aufwies. Für die histomorphologischen Varianten wurden sechs individuelle Modelle trainiert. Dort zeigte sich ein ähnliches Verhalten der Metriken im Verhältnis zur Datenmenge. Während die histologischen Elemente der hyperkeratotischen und der lichenoiden AK Sensitivität von 96,9% respektive 87,1% und Spezifizität von 97,0% und 94,0% aufwiesen, zeigten die restlichen Klassen Abweichungen bis nur 17,4%. Die Ergebnisse waren dabei direkt abhängig von der im Training verwendeten Pixelanzahl.

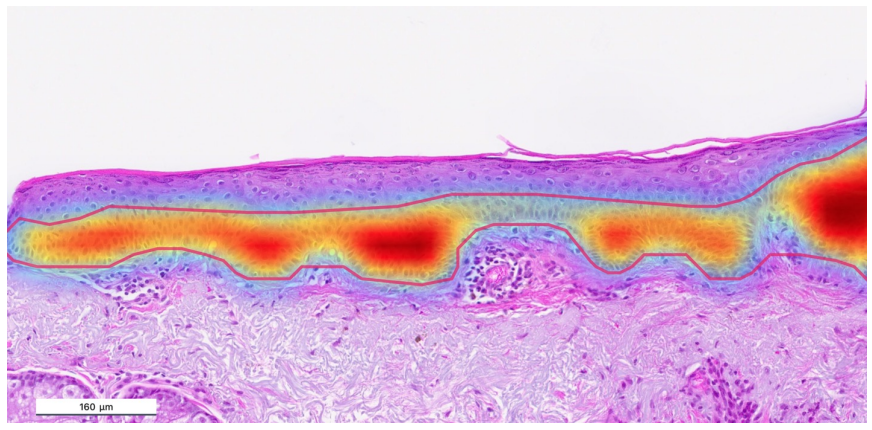

Figure 1

	Vorhersage nach Label		
	Testkohorte Negative Kontrollen	**Testkohorte (nur AK)**	**Übrige Kohorte (nur AK)**
Richtig Positive (TP)	-	104	613
Falsch Negative (FN)	-	1	13
Richtig Negative (TN)	79	-	-
Falsch Positive (FP)	5	-	-
Gesamt	84	105	626
	Vorhersagen nach Kohorte		
		Testkohorte	**Gesamt**
Richtig Positive (TP)	-	104	717
Falsch Negative (FN)	-	1	14
Richtig Negative (TN)	-	79	79
Falsch Positive (FP)	-	5	5
Gesamt	-	189	815
	Auswertungsergebnisse der Kohorte gesamt n = 815 (95% Konfidenzintervall)		
PPV	-	-	99,3% (98,4 - 99,7)
NPV	-	-	84,9% (76,3 - 90,8)
Sensitivität	-	-	98,1% (96,8 - 98,9)
Spezifität	-	-	94,0% (86,8 - 97,4)
Genauigkeit	-	-	97,7% (96,4 - 98,5)
F1-Score	-	-	98,7% (98,1 - 99,3)

Table 1

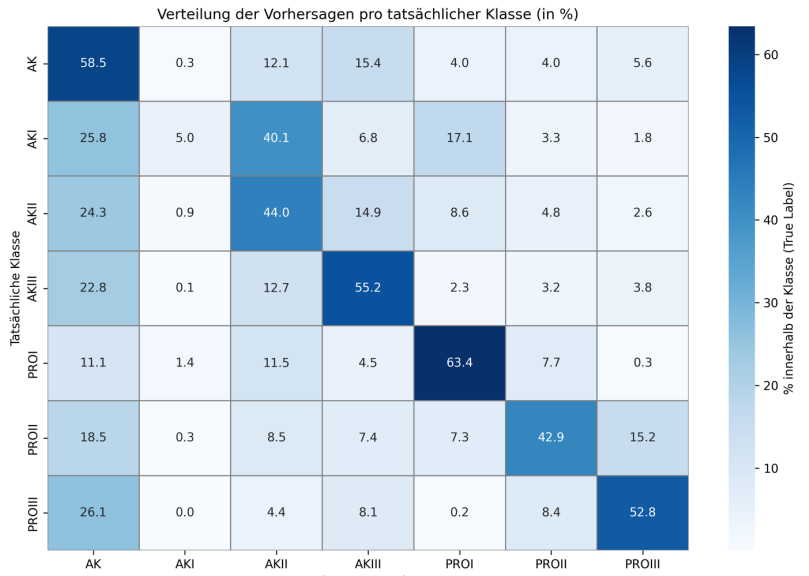

Figure 2

Fazit

Bei ausreichender Trainingsdatenmenge können für Dermatosen und Ihre einzelnen Aspekte, auf die ein DLM trainiert werden soll, hohe diagnostische Genauigkeiten erzielt werden. Für die Entwicklung von DLM in der Dermatopathologie stellt sich einerseits die Herausforderung der horizontalen Abdeckung eines breiten Diagnosespektrums, gleichzeitig auch die Sammlung und Aufbereitung von Daten in der vertikalen Abdeckung eines dermatopathologischen Krankheitsbildes.