Künstliche Intelligenz zur Detektion von fokaler kortikaler Dysplasie

Ein multizentrischer Vergleich existierender und neu trainierter Modelle

Dissertation

zur Erlangung des Doktorgrades (Dr. med.)

der Medizinischen Fakultät

der Rheinischen Friedrich-Wilhelms-Universität

Bonn

Lennart Nils Kersting

aus Bonn

2026

Angefertigt mit der Genehmigung
der Medizinischen Fakultät der Universität Bonn

1. Gutachter: PD Dr. Theodor Rüber
2. Gutachter: Prof. Dr. Julian Alexander Luetkens

Tag der mündlichen Prüfung: 08.06.2026

Aus der Klinik und Poliklinik für Epileptologie

Für meine Familie

Inhaltsverzeichnis

Abkürzungsverzeichnis

2D, 3D	zweidimensional, dreidimensional
95 %-KI	95 %-Konfidenzintervall
BIDS	Brain Imaging Data Structure (Datenstruktur für Neurobildgebung)
CNN	Convolutional Neural Network (faltendes neuronales Netz)
DICOM	Digital Imaging and Communications in Medicine (digitale Bildgebung und -kommunikation in der Medizin)
FCD	Focal Cortical Dysplasia (Fokale Kortikale Dysplasie)
FLAIR	Fluid Attenuated Inversion Recovery (Flüssigkeitsunterdrückte Inversionswiederherstellung)
ILAE	International League against Epilepsy (Internationale Liga gegen Epilepsie)
MAP18	Morphometric Analysis Program (Version 2018)
MELD	Multi-centre Epilepsy Lesion Detection (Multizentrische Epilepsie-Läsionsdetektion)
MRT	Magnetresonanztomographie
NIFTI	Neuroimaging Informatics Technology Initiative (Dateiformat für Neurobildgebung)
nnU-Net	no new U-Net (nicht neues U-Net)
T1	T1-gewichtete MRT-Aufnahme
T2	T2-gewichtete MRT-Aufnahme

1. Deutsche Zusammenfassung

Die vorliegende Dissertation beruht auf der Publikation „Detection of focal cortical dysplasia: Development and multicentric evaluation of artificial intelligence models“ (Kersting et al., 2025). Die Publikation wurde in einem international anerkannten Journal mit Begutachtungssystem veröffentlicht („Epilepsia“, Impact-Faktor: 6.6 (2024)).

1.1 Einleitung

Fokale kortikale Dysplasien (FCDs) sind epileptogene Läsionen, die durch entwicklungsbedingte Fehlbildungen der Großhirnrinde entstehen. Die epileptischen Anfälle beginnen dabei häufig schon im Kindes- oder Jugendalter (Blümcke et al., 2017). Seit der ersten Beschreibung einer FCD im Jahre 1971 (Taylor et al., 1971), kam es im Verlauf der Jahrzehnte immer wieder zu Weiterentwicklungen der auf der Histopathologie basierenden Klassifikation (Blümcke et al., 2011; Najm et al., 2022; Palmini et al., 2004). In der aktuellen Überarbeitung der Klassifikation der Internationalen Liga gegen Epilepsie (ILAE) (Najm et al., 2022), werden FCDs in drei histopathologische Haupttypen unterteilt (FCD Typ I-III). Zusätzlich wurden neue Kategorien („White Matter“, „No definitive FCD on histopathology“) eingeführt, um auch Läsionen der weißen Substanz sowie unklare oder grenzwertige Befunde adäquat einordnen zu können (Najm et al., 2022).

FCDs zählen zu den häufigsten Ursachen für pharmakoresistente fokale Epilepsien. Dabei stellt die chirurgische Resektion der Läsion oft eine effektive Behandlungsoption dar und kann bei geeigneten Patientinnen und Patienten in bis zu 70 % der Fälle zu Anfallsfreiheit führen (Lamberink et al., 2020). Ein entscheidender Bestandteil der präoperativen Diagnostik ist die Magnetresonanztomographie (MRT), da der Nachweis der Läsion mittels MRT-Bildgebung der stärkste prädiktive Faktor für ein erfolgreiches Operationsergebnis ist (Wagstyl et al., 2022). Charakteristische MRT-Merkmale einer FCD umfassen u.a. einen verbreiterten Kortex, eine verminderte Abgrenzbarkeit von grauer zu weißer Substanz, eine abnormale Gyrierung, und das sogenannte T2/FLAIR-hyperintense „transmantle sign“ (Urbach et al., 2022). Da die bildmorphologischen Merkmale oft nur schwach oder teilweise ausgeprägt und auch unterschiedlich in

Abhängigkeit vom histopathologischen Typ sein können, stellt das Erkennen einer FCD im klinischen Alltag oft eine Herausforderung dar. Dabei bleiben in bis zu 30 % der Fälle Läsionen bei der radiologischen Befundung unentdeckt (Urbach et al., 2022; Walger et al., 2024).

Um Radiologinnen und Radiologen bei der Detektion von FCDs zu unterstützen, wurden in den letzten Jahren bereits verschiedene KI-basierte Ansätze speziell für diese Aufgabe entwickelt, darunter MAP18 (David et al., 2021), MELD (Spitzer et al., 2022) und deepFCD (Gill et al., 2021). Darüber hinaus haben Zhang et al. (2024) die Anwendbarkeit von nnU-Net (Isensee et al., 2021), einem universell neu trainierbaren Deep-Learning-Framework zur Segmentierung verschiedener medizinischer Bilddaten, für den Zweck der FCD-Detektion untersucht. Dabei verzichteten sie jedoch auf eine umfassende Evaluation und stellten das Modell auch nicht öffentlich zur Verfügung (Zhang et al., 2024).

Die Architekturen der Modelle aus den oben genannten Arbeiten unterscheiden sich dabei grundlegend: MAP18 verwendet T1-gewichtete MRT-Aufnahmen und morphometrische Merkmalskarten, um pro Voxel die Wahrscheinlichkeit für das Vorliegen einer FCD vorherzusagen. MELD nutzt FreeSurfer (Fischl, 2012) zur Extraktion oberflächenbasierter Merkmale aus T1-gewichteten und FLAIR MRT-Aufnahmen, um dann mittels eines neuronalen Netzwerks die Wahrscheinlichkeit für das Vorliegen einer FCD für jeden Vertex der kortikalen Oberfläche vorherzusagen. deepFCD basiert auf einem Deep Convolutional Neural Network, das Vorhersagen auf 16 x 16 x 16 Voxel großen Teilstücken (3D-Patches) des Gesamtvolumens trifft, die im Nachhinein wieder zusammengesetzt werden müssen, während nnU-Net mit deutlich größeren Volumina trainiert werden kann.

Neben den Architekturen unterscheiden sich auch die verwendeten Kriterien, ab wann eine Läsion als „gefunden“ gilt. So wurde exemplarisch bei Gill et al. (2021) eine FCD bereits dann als erkannt gewertet, wenn die Vorhersage des Modells die manuelle Annotation (Ground-Truth) an mindestens einem Voxel überlappt. In einer Arbeit von Walger et al. (2024) wurden einheitliche Evaluationskriterien basierend auf dem Metrics Reloaded Framework (Maier-Hein et al., 2024) festgelegt, um Vorhersagen von Modellen oder auch Annotation von Menschen im Kontext der FCD-Detektion zu vergleichen. Dabei wurden die Vorhersagen von den drei oben genannten Modellen, die speziell für die FCD-

Detektion entwickelt wurden (MAP18, MELD, deepFCD) mit denen von Menschen auf einem monozentrischen Datensatz verglichen und es zeigte sich eine große Spannweite bzgl. der Detektionsraten (von 31 % bis 73 %).

Das Ziel der Arbeit von Kersting et al. (2025) war es, einen systematischen und fairen Vergleich zwischen Modellen zur FCD-Detektion durchzuführen. Zusätzlich zu den bestehenden Modellen wurden drei weitere Deep-Learning-Modelle ausgewählt („2d“-, „3d_fullres“-Variante von nnU-Net, FastSurferCNN) und neu trainiert. Alle sechs Modelle wurden anschließend erstmals auf einem multizentrischen Datensatz unter Verwendung einheitlicher Kriterien verglichen.

1.2 Material und Methoden

1.2.1 Datensatz

Die Datengrundlage der Arbeit von Kersting et al. (2025) bildete ein multizentrischer, retrospektiv erhobener Datensatz von Patientinnen und Patienten mit fokaler kortikaler Dysplasie (FCD) sowie 85 gesunde Kontrollen aus einem zuvor veröffentlichten Datensatz (Schuch et al., 2023). Die Bildgebung und die klinischen Daten der Patientinnen und Patienten wurden an vier spezialisierten Epilepsiezentren in Bonn, Berlin, Frankfurt am Main und Zürich erhoben. Die Daten des Universitätsklinikums Bonn stammen aus dem Zeitraum von 2006 bis 2021, die der Charité in Berlin aus den Jahren von 2017 bis 2020, die der Goethe-Universität Frankfurt am Main aus 2007 bis 2020 und die der Swiss Epilepsy Clinic in Zürich aus 2008 bis 2019. Für die Verwendung der Daten lag an allen beteiligten Standorten ein Ethikvotum vor (siehe Abschnitt 2.1 der Originalpublikation). Alle Untersuchungen wurden auf Grundlage der revidierten Deklaration von Helsinki und den geltenden gesetzlichen Bestimmungen durchgeführt.

Eingeschlossen wurden alle Patienten mit vollständigen MRT-Bilddaten, bestehend aus einer T1-gewichteten Sequenz sowie einer FLAIR-Sequenz, aufgenommen an MRT-Systemen mit einer Feldstärke von 3 Tesla. Ausgeschlossen wurden Fälle mit multiplen FCD-Läsionen, nicht eindeutig lokalisierbaren Läsionen, fehlenden oder unzureichenden Bilddaten (z. B. Artefakte, unvollständige Erfassung des Gehirns), sowie Fälle, bei denen

die Vorverarbeitung der Bilddaten aufgrund z. B. qualitativer Mängel nicht möglich war. Eine Übersicht über die Ausschlusskriterien kann Abbildung 1 entnommen werden.

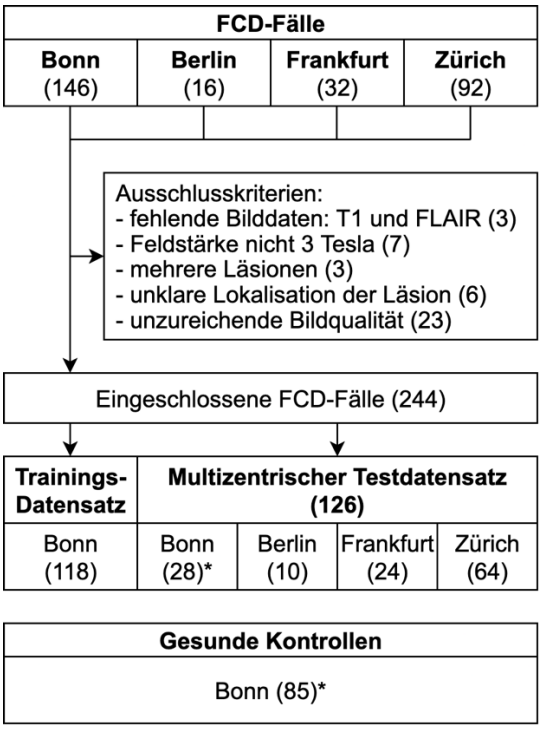

Abb. 1: Übersicht über die FCD-Fälle aus den vier beteiligten Zentren, die nach Anwendung der Ausschlusskriterien verbleibenden Fälle mit ihrer Aufteilung in Trainings- und Testdatensätze sowie die eingeschlossenen gesunden Kontrollen. Mit (*) gekennzeichnet sind Probandinnen und Probanden (FCD-Fälle und gesunde Kontrollen), die bereits Teil eines publizierten Datensatzes waren (Schuch et al., 2023). Abkürzungen: FCD, fokale kortikale Dysplasie; T1, T1-gewichtet; FLAIR, fluid-attenuated inversion recovery. Modifiziert nach: Kersting et al., 2025

Zusätzlich zu den Bilddaten wurden das Alter zum Zeitpunkt der MRT-Aufnahme, das biologische Geschlecht sowie, falls vorhanden, die histopathologische Klassifikation nach den Kriterien der ILAE erfasst. Es wurden keine Fälle aufgrund fehlender demographischer Daten ausgeschlossen. Die manuelle Erstellung der Läsionsmasken erfolgte standortweise durch erfahrene klinische Expertinnen und Experten, die wie im Rahmen der regulären prächirurgischen Diagnostik auf alle weiteren verfügbaren klinischen Informationen zugreifen konnten. Damit wurde sichergestellt, dass die Ground-Truth-Masken eine möglichst realitätsnahe Referenz für die automatisierte Läsionserkennung darstellen. Ein Beispiel für die Bilddaten mit Expertenmaske ist in Abbildung 2 A) dargestellt.

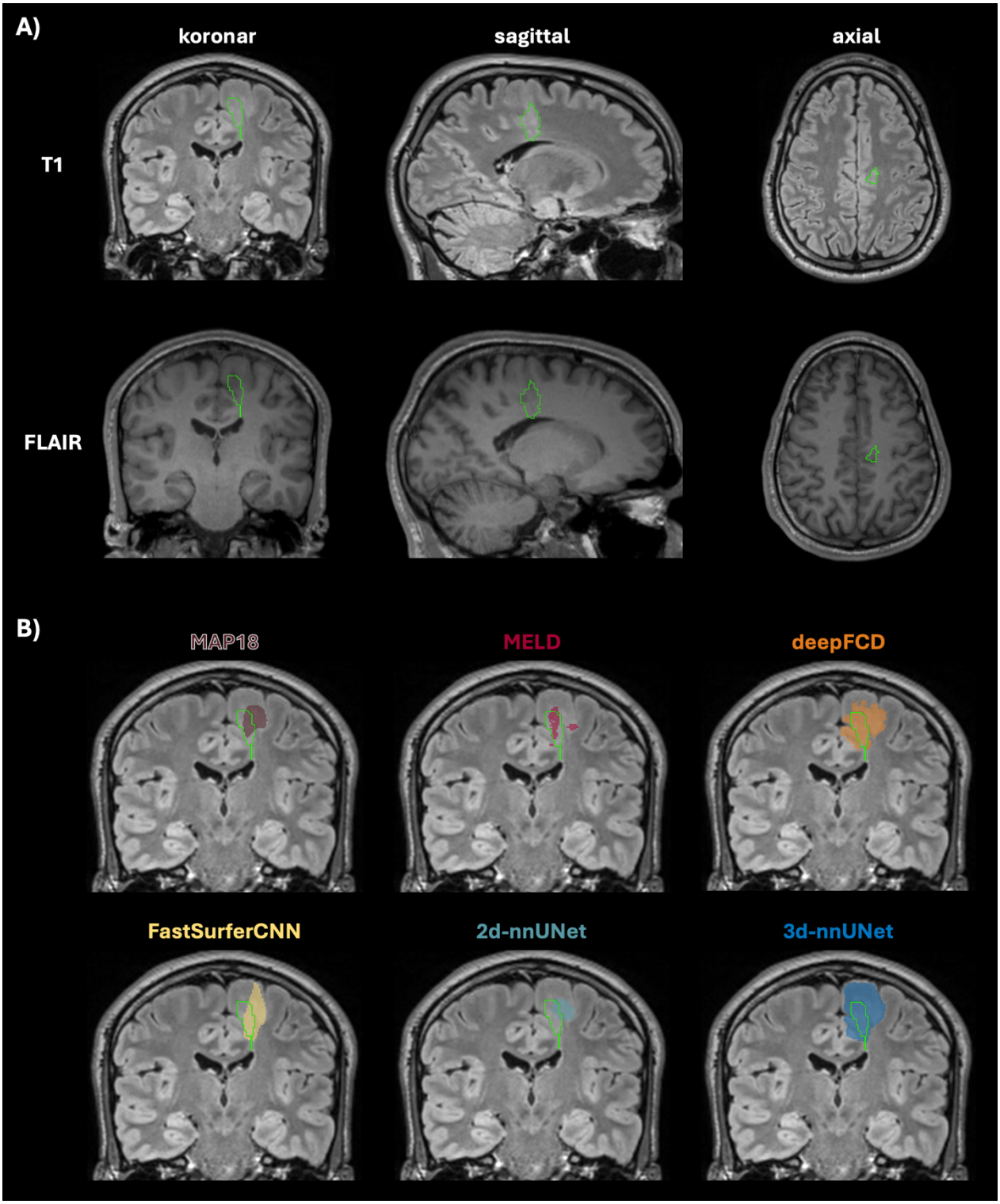

Abb. 2: Exemplarische Darstellung einer Fokalen kortikalen Dysplasie (FCD). In **A)** sind T1- und FLAIR-Aufnahmen in den drei Schnittebenen (koronar, sagittal, axial) mit der Expertenmaske als grüne Umrandung gezeigt. In **B)** sind die Vorhersagen der sechs Modelle für die in **A)** dargestellte Läsion in einem exemplarischen koronaren Schnitt dargestellt. Die Expertenmaske ist jeweils grün umrandet, die farbigen Flächen repräsentieren die Modellvorhersagen.

Zum Training der Modelle wurden bei Kersting et al. (2025) nur die FCD-Fälle aus Bonn verwendet. Zur Vermeidung potenzieller Verzerrungen durch Überschneidung der Daten erfolgte zunächst eine Aufteilung des Datensatzes in einen Trainings- und Testdatensatz. Den Testdatensatz aus Bonn bildeten 28 repräsentative FCD-Fälle aus einem veröffentlichten Datensatz (Schuch et al., 2023), die bereits in einer vorherigen Arbeit festgelegt wurden (Walger et al., 2025). Der multizentrische Testdatensatz setze sich zusätzlich zu den 28 FCD-Fällen aus Bonn, aus den eingeschlossenen FCD-Fällen der drei anderen Zentren sowie aus 85 gesunden Kontrollen zusammen.

1.2.2 Datenvorverarbeitung

Zunächst wurden von Kersting et al. (2025) die vorliegenden Bilddaten im DICOM-Format mit dem Tool „dcm2niix“ (Li et al., 2016) in das NIFTI-Dateiformat konvertiert und für die weitere Verarbeitung gemäß der Brain Imaging Data Structure (BIDS) (Gorgolewski et al., 2016) abgespeichert. Die T1-gewichteten und FLAIR MRT-Bilddaten wurden anschließend mittels „synthseg“ (Billot et al., 2023a, 2023b) und „mri_easyreg“ (Hoffmann et al., 2022; Iglesias, 2023) koregistriert.

1.2.3 Modellauswahl und Modelltraining

In der Arbeit von Kersting et al. (2025) wurde zunächst eine Auswahl bestehender Modelle zur automatisierten Detektion fokaler kortikaler Dysplasien (FCD) getroffen. Der Quellcode musste öffentlich zugänglich sein, um die Vorhersage auf neuen Datensätzen zu ermöglichen. Außerdem sollten die Modelle bereits auf externen Datensätzen validiert worden sein. Ausgewählt wurden die drei Modelle MAP18 (David et al., 2021), MELD (Spitzer et al., 2022) und deepFCD (Gill et al., 2021).

Um ein breites Spektrum methodischer Ansätze abzubilden, wurden ergänzend zu diesen etablierten Verfahren drei neue Modelle im Rahmen dieser Studie selbst trainiert. Besonderes Augenmerk lag darauf, verschiedene Modellarchitekturen einzubeziehen, die sich in der Art der Datenrepräsentation innerhalb des Modells unterscheiden. Ausgewählt wurden daher die „2d“ und „3d_fullres“-Varianten von nnU-Net (Isensee et al., 2021) sowie FastSurferCNN (Henschel et al., 2020). Letzteres wird am DZNE in Bonn entwickelt und

stellt eine Deep Learning basierte Alternative zu FreeSurfer dar, um eine anatomische Segmentierung des gesamten Gehirns zu erzeugen. Es basiert auf einem 2,5D-Ansatz, bei dem sogenannte „thick slices“ (sieben Voxel dicke Scheiben) zur Vorhersage genutzt werden. Dieses Modell wurde zum Zeitpunkt der Arbeit von Kersting et al. (2025) noch nicht zur Läsionsdetektion getestet.

Für das Training von nnU-Net und FastSurferCNN wurden die zuvor koregistrierten T1- und FLAIR-Aufnahmen verwendet. Als Zielgebiet dienten die von Expertinnen und Experten erstellten binären Läsionsmasken. Das Training erfolgte auf einem Linux-System mit einer NVIDIA RTX 3090 Ti Grafikkarte.

Für das Training von FastSurferCNN wurden ausschließlich Scheiben verwendet, die gemäß Läsionsmaske tatsächlich eine FCD enthielten. Die jeweils zusammengehörigen T1- und FLAIR-Scheiben wurden zusammengefügt und dienten dem Modell als Eingabevolumen. Das Training wurde mittels 5-facher Kreuzvalidierung (“5-fold crossvalidation”) für jeweils 150 Epochen durchgeführt. Als Verlustfunktion wurde ein “Binary-Crossentropy-Loss” verwendet. Es wurden für jede anatomische Ebene (koronar, axial, sagittal) separate Modelle trainiert.

Für nnU-Net wurden die „2d“ und „3d_fullres“-Varianten ohne weitere Modifikationen trainiert. Das Training wurde auf 400 Epochen begrenzt. Auf die automatische Nachprozessierung wurde verzichtet, um sämtliche vorhergesagten Cluster zu behalten und nicht nur z.B. das größte zusammenhängende Cluster.

1.2.4 Generierung der Vorhersagen und Nachverarbeitung

Die Vorhersagen der etablierten Modelle wurden bei Kersting et al. (2025) gemäß den Vorgaben der jeweiligen Veröffentlichungen erstellt. Abbildung 2 B) zeigt beispielhaft die Vorhersagen jedes Modells für eine ausgewählte FCD.

Um das MELD-Modell auf Daten eines neuen Zentrums bzw. Scanners anwenden zu können, musste jedoch zunächst eine “Harmonisierung” vorgenommen werden. Diese soll Unterschiede durch den Einsatz verschiedener Scanner und Sequenzen ausgleichen. Um die Harmonisierung durchzuführen, wird empfohlen mindestens 20 Patientinnen oder

Patienten oder gesunde Kontrollen zu verwenden. Um diese Anforderung zu erfüllen, wurden auch nicht-FCD-Fälle der anderen Zentren eingesetzt.

Für die finalen Vorhersagen von FastSurferCNN wurden die Vorhersagen der fünf Trainingsdurchläufe je Ebene gemittelt und anschließend mit einem Schwellenwert von 0,5 binarisiert. Die 3 binären Vorhersagen wurden daraufhin so kombiniert, dass alle Voxel beibehalten wurden, bei denen in mindestens einer der drei Ebenen eine FCD klassifiziert wurde. Um zufällige oder sehr kleine Cluster zu vermeiden, wurden alle Cluster mit einer Ausdehnung von weniger als 100 Voxel verworfen.

Für die Auswertung und den Vergleich aller Modelle mussten die Vorhersagen sowohl als binäre Masken, als auch als durchnummerierte Cluster vorliegen. Dafür wurden benachbarte Voxel einer vorhergesagten Läsion zu zusammenhängenden Regionen gruppiert. MELD und deepFCD produzierten bereits eine solche Vorhersage. Für MAP18, nnU-Net und FastSurferCNN wurde die Clustering-Prozedur von deepFCD angewandt. Bei MAP18 war es außerdem notwendig, zunächst die nicht binäre Vorhersage entsprechend der Publikation mit einem Schwellenwert von 0,5 in eine binäre Maske umzuwandeln (David et al., 2021). Ein schematischer Überblick über den Ablauf und die anschließende Auswertung ist in Abbildung 3 dargestellt.

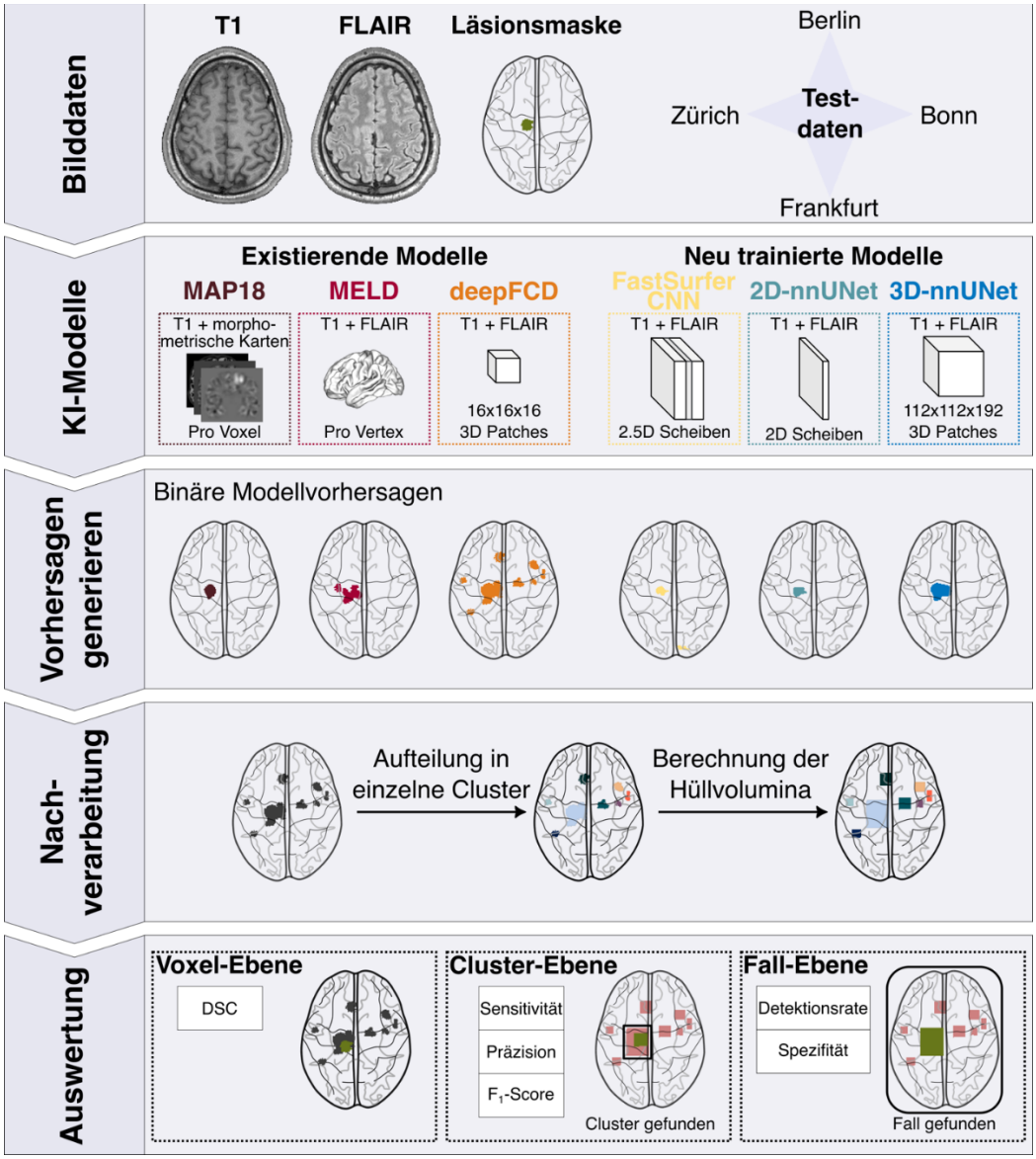

Abb. 3: Dargestellt ist der Ablauf der Studie von Kersting et al. (2025). Zur Generierung der Vorhersagen lagen Bilddaten aus vier Zentren (Berlin, Bonn, Frankfurt, Zürich) vor, bestehend aus T1-gewichteten Sequenzen, FLAIR-Bildern und den Läsionsmasken (in grün dargestellt). Verglichen wurden insgesamt sechs Modelle zur Detektion fokaler kortikaler Dysplasien (FCD): drei bereits publizierte Ansätze (MAP18, MELD, deepFCD) sowie drei in der Studie neu trainierte Modelle (FastSurferCNN, 2D-nnUNet, 3D-nnUNet). Für jedes Modell ist dargestellt, welche Bildinformationen als Eingabe benötigt wurden und auf welcher Datenstruktur die Vorhersagen basierten (Voxel, Vertex, 2D-Scheiben, 2.5D-Scheiben, 3D-Patches). Die Modellvorhersagen wurden anschließend in Cluster unterteilt und für jedes Cluster Hüllvolumina berechnet (dargestellt am Beispiel von deepFCD, jede Farbe entspricht einem einzelnen Cluster). Die Auswertung erfolgte auf drei Ebenen (Voxel-, Cluster- und Fall-Ebene). Die binären, nicht geclusterten Vorhersagen dienten der Berechnung des Dice-Koeffizienten (DSC) auf Voxel-Ebene, während auf Cluster- und Fall-Ebene die Hüllvolumina herangezogen wurden. Abkürzungen: T1, T1-gewichtet; FLAIR, fluid-attenuated inversion recovery; KI, Künstliche Intelligenz; DSC, Dice-Koeffizient. Modifiziert nach: Kersting et al., 2025

1.2.5 Auswertung und Statistik

Die Auswertung und der Vergleich der Modellvorhersagen erfolgten in der Arbeit von Kersting et al. (2025) in einem mehrstufigen Prozess. Dazu wurden die Vorhersagen mit den durch Expertinnen und Experten erstellten Läsionsmasken auf Voxel-, Cluster- und Fall-Ebene verglichen. Außerdem wurden die Auswertungen für die gesamte Test-Kohorte sowie für jedes der vier Zentren vorgenommen. Zum Vergleich wurden sämtliche Masken auf eine isotrope räumliche Auflösung von 1 x 1 x 1 mm vereinheitlicht.

Auf Voxel-Ebene wurde die binäre Vorhersagemasken jedes Modells mit der Expertenmaske verglichen. Als zentrales Maß für die Übereinstimmung wurde der Dice-Koeffizient verwendet, definiert als

$$\mathrm{DSC} = \frac{2 \cdot |\mathrm{P} \cap \mathrm{G}|}{|\mathrm{P}| + |\mathrm{G}|} \tag{1}$$

wobei P die Menge der von einem Modell als Läsion vorhergesagten Voxel und G die Menge der in der Expertenmaske als FCD markierten Voxel bezeichnet. Ein Dice-Wert von 1 weist auf eine vollständige Übereinstimmung hin, ein Wert von 0 auf keinerlei Überlappung. Zusätzlich wurde die Anzahl der Fälle erfasst, in denen ein Modell eine vollständig leere Maske erzeugt hat und somit keine Voxel als FCD klassifiziert wurden.

Für die Auswertung auf Cluster-Ebene wurden für die vorhergesagten Cluster der Modelle und die der Expertenmasken minimale quaderförmige Hüllvolumina („Bounding Box“ bzw. „Bounding Volume“) berechnet. Anschließend wurde für jedes Modell die durchschnittliche Anzahl der vorhergesagten Cluster pro Fall und die durchschnittliche Größe der Cluster bestimmt. Zur Klassifizierung, ob ein Cluster als richtig-positiv oder falsch-positiv galt, wurden zwei verschiedene Kriterien herangezogen. Zum einen wurde ein vorhergesagtes Cluster als richtig-positiv gewertet, wenn der Dice-Score, berechnet für die Hüllvolumina von vorhergesagtem Cluster und Expertenmaske, größer als 0,22 war. Dieser Schwellenwert wurde in einer Vorarbeit empirisch ermittelt (Walger et al., 2024). Dieses Kriterium wurde bei Kersting et al. (2025) als Detektions-Kriterium bezeichnet. Zum anderen wurde ermittelt, ob sich der Massenmittelpunkt des vorhergesagten Clusters innerhalb der Expertenmaske befand. Dieses zweite Kriterium wurde bei Kersting et al. (2025) als Pinpointing-Kriterium bezeichnet. Obwohl in den zugrunde liegenden Daten

lediglich eine Läsion pro Fall vorhanden war, konnten unter Umständen mehrere vorhergesagte Cluster eines Modells die festgelegten Kriterien erfüllen. Für die Auswertung auf Cluster-Ebene wurden alle Cluster in die Berechnung einbezogen. Auf dieser Grundlage wurden Präzision_C, Sensitivität_C und der F_1-Score auf Cluster-Ebene berechnet:

$$\text{Präzision}_C = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$\text{Sensitivität}_C = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$\text{F}_1 = \frac{2 \cdot \text{Präzision}_C \cdot \text{Sensitivität}_C}{\text{Präzision}_C + \text{Sensitivität}_C} \tag{4}$$

wobei TP (richtig-positiv) die Anzahl korrekt erkannter Cluster, FP (falsch-positiv) die Anzahl fälschlich als Läsion klassifizierter Cluster und FN (falsch-negativ) die Anzahl nicht erkannter Läsionen bezeichnet. Die Präzision_C beschreibt somit den Anteil, der durch das Modell gefundenen Cluster an allen durch das Modell vorhergesagten Clustern. Durch viele falsch-positive und wenig gefundene Cluster sinkt daher die Präzision_C. Die Sensitivität_C hingegen beschreibt den Anteil der durch das Modell gefunden Clustern an allen zu findenden Clustern. Hier gilt, je mehr Läsionen gefunden wurden, desto höher ist die Sensitivität_C. Der F_1-Score, als harmonisches Mittel, kombiniert diese beiden Metriken.

Auf Fall-Ebene galt ein Fall als korrekt erkannt, sobald mindestens ein vom Modell vorhergesagtes Cluster als gefunden gewertet wurde. Auch hier wurde wieder zwischen den beiden oben genannten Kriterien differenziert, um die Sensitivität auf Fall-Ebene zu berechnen. Die Spezifität wurde anhand der gesunden Kontrollgruppe bestimmt. Ein Kontrollfall galt als richtig-negativ, wenn die vorhergesagte Maske des Modells leer war. Wurde hingegen ein Cluster vorhergesagt, wurde der Fall als falsch-positiv gewertet.

Die Auswertung erfolgte mit Python (Version 3.8, Python Software Foundation, Delaware, Vereinigte Staaten) im Wesentlichen mit den Paketen NumPy (Version 1.23.5), pandas (Version 1.4.3), Matplotlib (Version 3.5.1), SciPy (Version 1.8.0), PyTorch (Version 2.4.1), PyBIDS (Version 0.15.5), NiBabel (Version 3.2.2), ANTsPy (Version 0.5.4). Angegeben wurden, wenn nicht anders beschrieben, immer der Mittelwert mit 95 %-

Konfidenzintervall. Die Abschätzung der 95 %-Konfidenzintervalle für die Auswertung pro Zentrum wurde mittels eines Bootstrapping-Verfahren vorgenommen.

1.3 Ergebnisse

1.3.1 Datensatz

Nach Anwendung der Ausschlusskriterien, verblieben bei Kersting et al. (2025) insgesamt 244 FCD-Fälle und 85 gesunde Kontrollen (Abb. 1). Der multizentrische Testdatensatz setzte sich aus 28 FCD-Fällen aus Bonn, 10 aus Berlin, 24 aus Frankfurt, 64 aus Zürich (insgesamt 126 FCD-Fälle) sowie den 85 gesunden Kontrollen zusammen. Die restlichen 118 FCD-Fälle aus Bonn wurden im Vorfeld zum Training von nnU-Net und FastSurferCNN eingesetzt und waren nicht Teil des multizentrischen Test-Datensatzes. Für 49% aller FCD-Fälle lag eine histopathologische Klassifizierung vor, wobei FCD-Typ II mit 124 Fällen am häufigsten vertreten war. Die Kohorte aus Berlin umfasste mit einem mittleren Alter von 7,5 ± 2,9 Jahren nur pädiatrische Patientinnen und Patienten. Eine zusammenfassende Übersicht zur Demographie sowie der histopathologischen Klassifikation ist in Tabelle 1 dargestellt.

Die Bilddaten wurden an unterschiedlichen Scannern der Hersteller Siemens (Siemens Healthineers, Erlangen, Deutschland), Philips (Philips Medical Systems, Hamburg, Deutschland) und GE (GE HealthCare Technologies, Chicago, Illinois, USA) an den 4 Zentren akquiriert. Alle T1-gewichteten Sequenzen und auch alle FLAIR-Aufnahmen (bis auf 27 aus Bonn) hatten eine isotrope Auflösung zwischen 0,5 mm und 1,0 mm. Genaue Details zu den verwendeten Scannern und Sequenzen sind „Supplementary Table 1” der Veröffentlichung von Kersting et al. (2025) zu entnehmen.

Tab. 1: Demographische Charakteristika und histopathologische Klassifikation des Trainings- und Testdatensatzes. Das Alter ist angegeben als Mittelwert mit Standardabweichung. Die histopathologische Einteilung erfolgte nach ILAE-Klassifikation. Die Daten für die Fälle aus Zürich lagen nur für die Gesamtkohorte (92 Fälle) vor und nicht auf Einzelfall-Ebene, entsprechend beziehen sich die Daten in der Tabelle auf die Gesamtkohorte. Abkürzungen: FCD, fokale kortikale Dysplasie; m, männlich; w, weiblich; n. a., nicht angegeben; histo., histopathology. Modifiziert nach: Kersting et al. 2025

	Trainings-Datensatz	**Multizentrischer Test-Datensatz**				
		Bonn	Berlin	Frankfurt	Zürich	Kontrollen
Demographische Daten						
Anzahl (n)	118	28	10	24	92	85
Alter bei MRT [Jahre]	29,5 ± 14,2	28,0 ± 10,7	7,5 ± 2,9	28,3 ± 15,5	27,2 ± 14,3	33,3 ± 11,9
Geschlecht m/w (n)	67/51	16/12	5/5	20/4	39/53	42/43
Histopathologie						
FCD-Typ I	3	0	0	0	2	-
FCD-Typ II (a/b)	69 (19/50)	20 (6/14)	1 (0/1)	9 (1/8)	25 (n. a.)	-
FCD-Typ IIIb	0	0	0	1	0	-
No definitive FCD on histo.	1	0	0	0	0	-
Nicht operiert	45	8	9	13	59	-
Nicht klassifiziert	0	0	0	0	4	-
Keine Information vorliegend	0	0	0	1	2	-

1.3.2 Modell-Vorhersagen

In der Auswertung auf Voxel-Ebene zeigten sich deutliche Unterschiede zwischen den getesteten Modellen. Mit einem Dice-Koeffizienten von 0,36 (95 %-KI: 0,30-0,41) erreichte 3D-nnUNet den höchsten Wert, FastSurferCNN hingegen mit 0,06 (95 %-KI: 0,03-0,08) den niedrigsten. Auffällig war außerdem, dass deepFCD in allen 126 FCD-Fällen Voxel als FCD klassifizierte, wohingegen die anderen Modelle auch leere Masken produzierten. Mit 65 % produzierte 2D-nnUNet die meisten leeren Masken. Die restlichen Modelle lagen zwischen 12 % und 26 %. Die vollständigen Metriken auf Voxel-Ebene können Tabelle 2 entnommen werden.

Tab. 2: Ergebnisse für die gesamte multizentrische Test-Kohorte. Für die Auswertung der Vorhersagen von MAP18 wurden die Fälle aus Zürich exkludiert, da diese im Training von MAP18 verwendet wurden. Die Präzision, Sensitivität und der F_1-Score auf Cluster-Ebene wurden anhand des Detektions-Kriteriums berechnet, so wurden die Cluster als richtig-positiv gewertet, die einen Dice-Wert von 0,22 überschritten. Die 85 gesunden Kontrollen wurden nur zur Berechnung der Spezifität auf Fall-Ebene herangezogen. Die Werte in eckigen Klammern sind 95%-Konfidenzintervalle, die in runden Klammern stellen Zähler und Nenner dar. Modifiziert nach: Kersting et al., 2025

	MAP18	**MELD**	**deepFCD**	**FastSurfer CNN**	**2D-nnUNet**	**3D-nnUNet**
Voxel-Ebene						
Leere Vorhersage	21 % (13/62)	12 % (15/126)	0 % (0/126)	26 % (33/126)	65 % (82/126)	21 % (27/126)
Dice-Koeffizient	0,15 [0,10; 0,19]	0,21 [0,18; 0,24]	0,15 [0,12; 0,17]	0,06 [0,03; 0,08]	0,07 [0,04; 0,10]	0,36 [0,30; 0,41]
Cluster-Ebene						
Anzahl Cluster pro Fall	3,5 [2,7; 5,2]	2,1 [1,8; 2,5]	24,7 [22,4; 27,2]	2,3 [1,9; 2,6]	0,5 [0,3; 0,6]	0,9 [0,8; 1,0]
Größe [ml]	0,44 [0,31; 0,79]	0,97 [0,80; 1,44]	0,92 [0,87; 0,97]	0,86 [0,71; 1,02]	0,75 [0,48; 1,28]	2,44 [1,94; 3,18]
$Präzision_C$	0,08 (18/219)	0,26 (70/266)	0,03 (107/3109)	0,06 (16/284)	0,26 (16/62)	0,63 (69/110)
$Sensitivität_C$	0,29 (18/62)	0,52 (70/134)	0,82 (107/130)	0,13 (16/126)	0,13 (16/126)	0,55 (69/126)
F_1-Score	0.13	0.35	0.07	0.08	0.17	0.58
Fall-Ebene						
Pinpointing-Kriterium	66 % (41/62)	48 % (61/126)	80 % (101/126)	25 % (31/126)	29 % (36/126)	64 % (81/126)
Detektions-Kriterium	29 % (18/62)	49 % (62/126)	82 % (103/126)	13 % (16/126)	13 % (16/126)	55 % (69/126)
Spezifität	51 % (43/85)	55 % (47/85)	0 % (0/85)	22 % (19/85)	95 % (81/85)	86 % (73/85)

Auf Cluster-Ebene prognostizierten die Modelle unterschiedlich viele Cluster pro Fall. Die meisten sagte deepFCD mit durchschnittlich 24,7 (95 %-KI: 22,4-27,2) Clustern vorher, die wenigsten 2D-nnUNet mit 0,5 (95 %-KI: 0,3-0,6). Die kleinsten Cluster produzierte MAP18 mit einer mittleren Größe von 0,44 ml (95 %-KI: 0,31-0,79), die größten 3D-nnUNet mit 2,44 ml (95 %-KI: 1,94-3,18). Unter Anwendung des Detektions-Kriteriums, erzielte deepFCD mit 82 % die höchste Sensitivität auf Cluster-Ebene, allerdings bei einer gleichzeitig sehr geringen Präzision von 0,03 und damit einhergehend dem niedrigsten F_1-Score von 0,07. Dies erklärt sich vor allem durch die große Anzahl an falsch-positiven

Clustern. 3D-nnUNet erzielte den höchsten F_1-Score mit 0,58 bei einer Sensitivität von 55 % und einer Präzision von 0,63. Den zweithöchsten F_1-Score erreichte MELD mit 0,35 bei einer ähnlichen Sensitivität von 52 %, aber einer geringeren Präzision von 0,26. Unter der Bedingung, dass nur richtig-positive Cluster betrachtet werden, erreichte 3D-nnUNet mit 0,62 (95 %-KI = 0,58-0,66) den höchsten Dice-Koeffizienten. Die Metriken auf Cluster-Ebene nach dem Detektions-Kriterium können Tabelle 2 entnommen werden. Abbildung 4 A) zeigt die Präzision und Sensitivität auf Cluster-Ebene für beide Kriterien.

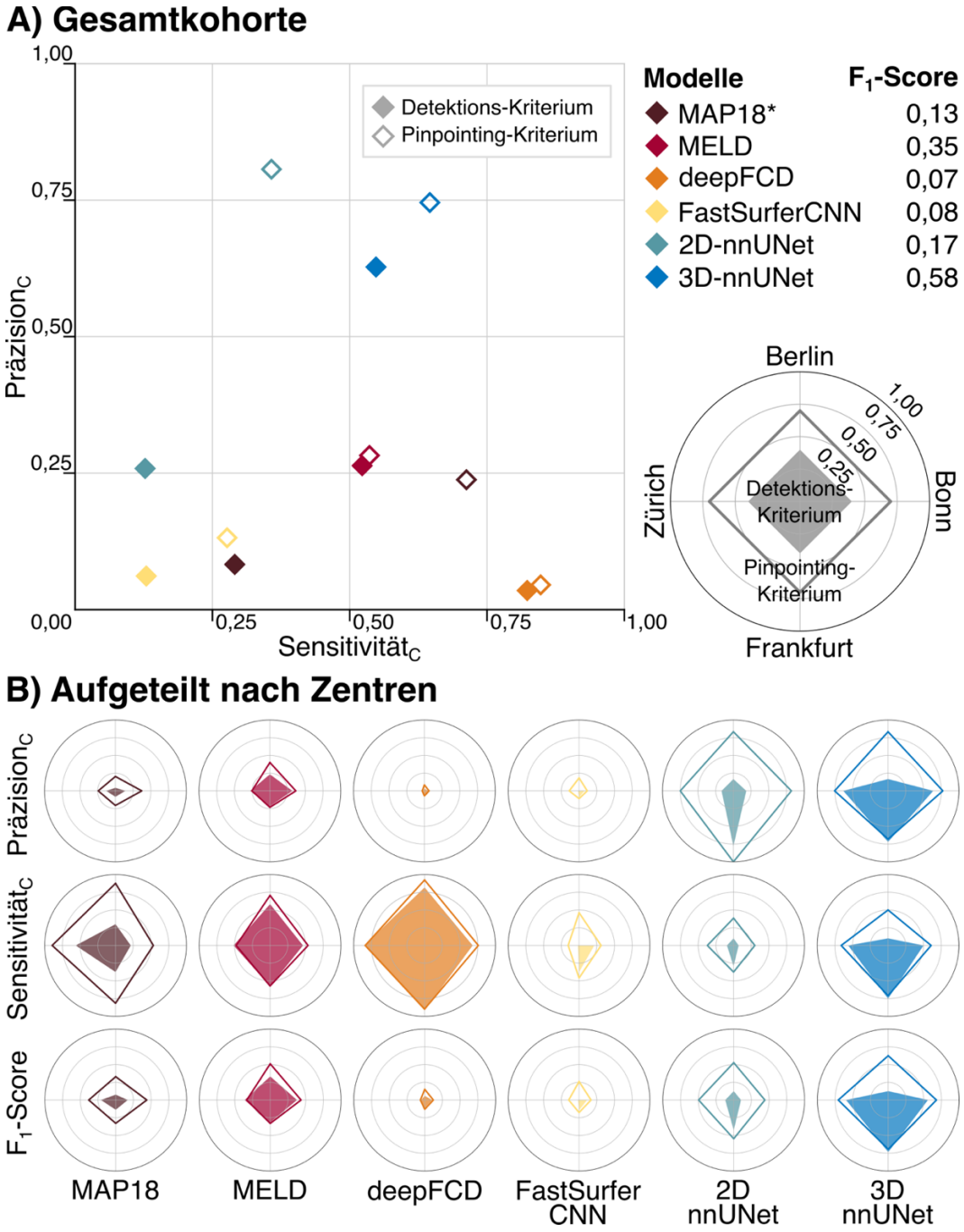

Abb. 4: Dargestellt sind die Metriken auf Clusterebene (Präzision$_C$, Sensitivität$_C$, F_1-Score) für die Gesamtkohorte in (A) sowie nach Zentren in (B) für alle sechs Modelle (MAP18, MELD, deepFCD, FastSurferCNN, 2D-nnUNet, 3D-nnUNet). Gefüllte Marker und Flächen beziehen sich auf die Auswertung nach dem Detektions-Kriterium, nicht gefüllte auf die Auswertung nach dem Pinpointing-Kriterium. Beim Detektions-Kriterium gilt ein Cluster als korrekt erkannt, wenn die Überlappung mit der Expertenmaske einen Schwellenwert von 0,22 des Dice-Koeffizienten überschreitet, während beim Pinpointing-Kriterium der Massenmittelpunkt des Clusters innerhalb der Läsionsmaske liegen muss. In (A) ist der F_1-Score für das Detektions-Kriterium in der rechts angefügten Tabelle aufgeführt. In den Netzdiagrammen in (B) repräsentiert jede Achse eines der vier Zentren, wie in der rechtsseitigen Legende angegeben. Modifiziert nach: Kersting et al., 2025.

Auf Fall-Ebene klassifizierte deepFCD nach dem Detektions-Kriterium mit 82 % die meisten FCD-Fälle korrekt, FastSurferCNN und 2D-nnUNet mit jeweils 13 % die wenigsten. 3D-nnUNet hatte die zweithöchste Detektionsrate mit 55 %, MELD und MAP18 erreichte 49 % bzw. 29 %. Nach dem Pinpointing-Kriterium wurden bei allen Modellen, außer deepFCD und MELD, mehr FCD-Fälle korrekt klassifiziert. deepFCD und MELD waren um 2 % bzw. 1 % schlechter. Nach beiden Kriterien erkannte deepFCD jedoch die meisten FCD-Fälle. Besonders groß war der Einfluss durch die Wahl des Kriteriums bei MAP18, statt 29 % (Detektions-Kriterium) erreichte das Modell 66 % (Pinpointing-Kriterium) und schnitt damit nach dem Pinpointing-Kriterium am zweitbesten ab. Bei der Betrachtung der Modellvorhersagen für die gesunden Kontrollen, erreichte 2D-nnUNet mit 95 % die höchste Spezifität, gefolgt von 3D-nnUNet mit 86 %. Die geringste Spezifität von 0 % hatte deepFCD, welches in 100 % der gesunden Kontrollen Läsionen vorhersagte. Die Detektions- bzw. Pinpointing-Raten und die Spezifität aller Modelle für den multizentrischen Testdatensatz können Tabelle 2 entnommen werden.

Die Auswertung der Modellvorhersagen unter Berücksichtigung der einzelnen Zentren zeigte, dass sich die Detektionsraten je nach Zentrum zum Teil deutlich unterschieden. deepFCD zeigte in allen Zentren auf Fall-Ebene die höchsten Detektionsraten mit Werten zwischen 68 % (Bonn) und 92 % (Frankfurt). Bis auf MAP18 hatten auch alle anderen Modelle die höchste Detektionsrate für die Daten aus Frankfurt. MAP18 erreichte die höchste Detektionsrate für die Daten aus Zürich, wobei zu beachten ist, dass diese in dessen Training verwendet wurden. Außerdem fiel auf, dass die Detektionsraten von MELD zwischen den Zentren am wenigsten variierten mit Werten zwischen 46 % (Bonn) und 54 % (Frankfurt). In der Berliner pädiatrischen Kohorte erkannten 2D-nnUNet und 3D-nnUNet nach dem Detektions-Kriterium jeweils nur einen Fall, lokalisierten jedoch nach dem Pinpointing-Kriterium zwei bzw. fünf von zehn FCDs. Beim Vergleich der Detektionsraten der neu trainierten Modelle für die Validierungsdaten (118 Fälle aus Bonn) und Testdaten aus Bonn (28 Fälle), fiel auf, dass FastSurferCNN (21 % auf 53 %, $p < 0,01$), 3D-nnUNet (50 % auf 77 %, $p < 0,01$) und 2D-nnUNet (7 % auf 14 %, $p = 0,53$) für die Validierungsdaten höhere Detektionsraten erreichten. Die anderen Modelle, die nicht mit diesen Daten trainiert wurden, zeigten alle keine signifikanten Unterschiede zwischen Bonner Validierungs- und Testdatensatz (MAP18: $p = 0,36$; MELD: $p = 0,83$; deepFCD: $p = 0,63$). Alle Detektions- und Pinpointing-Raten für die einzelnen Modelle und Zentren

sind in Tabelle 3 aufgeführt. Abbildung 4 B) zeigt die Metriken auf Cluster-Ebene gruppiert nach Zentren. Zentrumsbezogene Metriken auf Voxel- und Cluster-Ebene finden sich in „Table 3“ der Veröffentlichung von Kersting et al. (2025).

Tab. 3: Detektions- und Pinpointing-Raten auf Fall-Ebene aller Modelle, aufgeteilt nach Zentren. Die Werte in runden Klammern stellen Zähler und Nenner dar. Modifiziert nach: Kersting et al., 2025

	MAP18	**MELD**	**deepFCD**	**FastSurfer CNN**	**2D-nnUNet**	**3D-nnUNet**
Bonn						
Pinpointingrate	50 % (14/28)	54 % (15/28)	75 % (21/28)	29 % (8/28)	25 % (7/28)	61 % (17/28)
Detektionsrate	21 % (6/28)	46 % (13/28)	68 % (19/28)	21 % (6/28)	7 % (2/28)	50 % (14/28)
Berlin						
Pinpointingrate	80 % (8/10)	50 % (5/10)	80 % (8/10)	30 % (3/10)	20 % (2/10)	50 % (5/10)
Detektionsrate	30 % (3/10)	50 % (5/10)	80 % (8/10)	0 % (0/10)	10 % (1/10)	10 % (1/10)
Frankfurt						
Pinpointingrate	79 % (19/24)	50 % (12/24)	88 % (21/24)	46 % (11/24)	38 % (9/24)	71 % (17/24)
Detektionsrate	38 % (9/24)	54 % (13/24)	92 % (22/24)	38 % (9/24)	29 % (7/24)	75 % (18/24)
Zürich						
Pinpointingrate	88 % (56/64)	45 % (29/64)	80 % (51/64)	14 % (9/64)	28 % (18/64)	66 % (42/64)
Detektionsrate	56 % (36/64)	48 % (31/64)	84 % (54/64)	2 % (1/64)	9 % (6/64)	56 % (36/64)
Validierungsdaten						
Pinpointingrate	64 % (76/118)	57 % (67/118)	82 % (97/118)	64 % (75/118)	42 % (50/118)	81 % (95/118)
Detektionsrate	32 % (38/118)	51 % (60/118)	74 % (87/118)	53 % (63/118)	14 % (17/118)	77 % (91/118)

1.4 Diskussion

In der Arbeit von Kersting et al. (2025) wurden sechs Modelle zur automatisierten FCD-Detektion auf einem multizentrischen Datensatz verglichen. Die höchsten Detektionsraten erreichten deepFCD (82 %) und 3D-nnUNet (55 %). deepFCD erreichte diesen Wert jedoch nur auf Kosten einer sehr niedrigen Präzision (0,03) aufgrund zahlreicher falsch-positiver Cluster. Im Schnitt produzierte deepFCD 24,7 (95 %-KI: 22,4-27,2) Cluster pro Fall. Dies führte auch zum niedrigsten F_1-Score (0,07) unter allen Modellen. 3D-nnUNet mit der zweithöchsten Detektionsrate erreichte bei einer Präzision von 0,63 den höchsten F_1-Score (0,58). Bei Betrachtung der gesunden Kontrollen sagte deepFCD in allen Fällen fälschlicherweise Läsionen vorher und erreichte damit eine Spezifität von 0 %, während 3D-nnUNet eine Spezifität von 86 % erreichte.

Auch die benötigte Rechenzeit spielt für den klinischen Einsatz eine bedeutende Rolle. Die Modelle wiesen diesbezüglich erhebliche Unterschiede auf. MELD war mit Abstand am zeitintensivsten und benötigte etwa 7 bis 9 Stunden pro Fall. Davon entfielen allein 6 bis 8 Stunden auf die FreeSurfer-Rekonstruktion, hinzu kamen 45 bis 60 Minuten für Merkmalsextraktion und Harmonisierung sowie weitere 5 bis 10 Minuten für die eigentliche Vorhersage (Spitzer et al., 2022). Für MAP18 ergab sich in exemplarischen Durchläufen eine Gesamtdauer von ungefähr 20 Minuten pro Fall, während die Verarbeitung mit deepFCD mit etwa zwei Stunden deutlich länger dauerte. Schneller waren dagegen FastSurferCNN mit etwa 8 Minuten sowie die beiden nnU-Net-Varianten, die jeweils ungefähr 6 Minuten pro Fall benötigten.

In der Arbeit von Kersting et al. (2025) wurden zwei unterschiedliche Kriterien angewendet, um eine Läsion als gefunden zu werten. An den unterschiedlich hohen Detektionsraten von MAP18 (29% Detektions-Kriterium vs. 66% Pinpointing-Kriterium) zeigte sich, dass die Bewertung eines Modells stark vom angewendeten Kriterium abhängen kann. Ein weiteres Beispiel war 3D-nnUNet für die pädiatrischen Fälle aus Berlin. Nach dem Pinpointing-Kriterium wurden 50% gefunden, während nach dem Detektions-Kriterium nur 10 % der Fälle erkannt wurden. Deshalb war es für einen umfassenden Vergleich notwendig, die Leistung der Modelle anhand verschiedener Kriterien auszuwerten. In den Originalarbeiten von MAP18 und deepFCD wurde eine FCD bereits dann als gefunden gewertet, wenn die Vorhersage und Läsionsmaske mit

mindestens einem Voxel überlappten (David et al., 2021; Gill et al., 2021). In der Arbeit, in der MELD vorgestellt wurde, genügte ebenfalls eine beliebige Überlappung der Vorhersage mit der Expertenmaske, wobei die Auswertung dort auf Oberflächenkarten erfolgte und damit eine Überlappung von einem Vertex ausreichte (Spitzer et al., 2022). Da ein Kriterium, bei dem eine beliebige Überlappung ausreicht, auch durch eine unspezifische Vorhersage erfüllt werden kann, beispielsweise bei einer Vorhersage des gesamten Gehirns als Läsion, erscheint es sinnvoll, robustere Kriterien heranzuziehen. In der Arbeit von Kersting et al. (2025) wurden daher das Detektions-Kriterium und Pinpointing-Kriterium herangezogen, welche im Kontext der FCD-Detektion zuvor bei Walger et al. (2024) etabliert wurden. Hinzu kommt, dass bei Betrachtung der reinen Detektionsraten nicht berücksichtigt wird, wie viele falsch-positive Cluster durch ein Modell vorhergesagt werden. Deshalb wurde in der Arbeit von Kersting et al. (2025) der F_1-Score auf Cluster-Ebene eingesetzt, da dieser Sensitivität und Präzision in einer Metrik kombiniert und damit sowohl die Erkennungsrate wahrer Läsionen als auch die Vermeidung falsch-positiver Vorhersagen berücksichtigt. Es bleibt allerdings zu untersuchen, wie viele falsch-positive Vorhersagen im klinischen Alltag tolerierbar sind und welchen Einfluss sie auf die befundende Person haben. Zusätzlich sind auch Metriken auf Voxel-Ebene, wie z.B. der Dice-Koeffizient in der Modellbewertung verbreitet. Auch hier bleibt offen, welchen Stellenwert diese Metriken in der FCD-Detektion tatsächlich haben. So kann die präzise Abgrenzung einer Läsion beispielsweise für die operative Planung entscheidend sein. Eine mehrstufige Analyse auf Voxel-, Cluster- und Fall-Ebene, wie sie in der Studie von Kersting et al. (2025) durchgeführt wurde, ist daher erforderlich, um die Modelle fair und umfangreich zu vergleichen.

Die Ergebnisse zeigten, dass die Modelle nicht in allen Zentren gleich gute Ergebnisse lieferten. Dabei zeigte MELD über alle Zentren hinweg die stabilste Leistung. Dies könnte darauf zurückzuführen sein, dass MELD eine Harmonisierung der Daten mittels ComBat (Fortin et al., 2018) durchführt, um den Effekt unterschiedlicher Scanner und Protokolle zu minimieren. Außerdem zeigte sich, dass die neu trainierten Modelle für die Validierungsdaten aus Bonn besser abschnitten als für die Bonner Testkohorte. Dieser Trend war auch bei MAP18 für die Daten aus Zürich zu beobachten, da diese in dessen Training verwendet wurden. Dies verdeutlicht erneut die Bedeutung einer strikten Trennung zwischen Trainings- und Testdatensatz. Während die Frankfurter Fälle von allen

Modellen am zuverlässigsten erkannt wurden, stellte insbesondere die Berliner pädiatrische Kohorte für beispielsweise nnUNet eine Herausforderung dar. Dies könnte neben z.B. Scanner bedingten Unterschieden auch an den Charakteristika pädiatrischer Bilddaten liegen. Sicherlich ist es in zukünftigen Studien sinnvoll, Modelle gezielt auf Subgruppen wie pädiatrische Fälle auszurichten oder diese in den Trainingsdaten stärker zu repräsentieren. Die beobachteten Unterschiede verdeutlichten insgesamt, dass eine multizentrische Evaluation unverzichtbar ist, um die Generalisierbarkeit und Zuverlässigkeit eines Modells beurteilen zu können. Für die externe Validierung ist es zudem wichtig, dass trainierte Modelle öffentlich verfügbar gemacht werden. Um studienübergreifend vergleichbar zu bleiben, ist es sinnvoll, die Vorhersagegenauigkeit eines Modells auch auf einem öffentlich zugänglichen Datensatz anzugeben. In der Studie von Kersting et al. (2025) wurden daher die Ergebnisse aller Modelle sowohl für die 85 veröffentlichten Bonner FCD-Fälle (Schuch et al., 2023) als auch für den von Walger et al. (2025) vorgeschlagenen Test-Datensatz angegeben.

Die untersuchten Modelle unterschieden sich deutlich hinsichtlich ihrer Architektur. Besonders die Modelle, die auf dreidimensionalen Daten basierten, zeigten dabei gute Ergebnisse in der FCD-Erkennung. deepFCD, welches mit 16 x 16 x 16 großen 3D-Patches arbeitet, wies die höchste Sensitivität auf, erreichte jedoch nur eine sehr geringe Präzision. 3D-nnU-Net hingegen, welches mit 112 x 112 x 192 großen 3D-Patches arbeitet, erzielte den höchsten F_1-Score und die zweithöchste Sensitivität. Auch Avesta et al. (2023) berichteten von einem Vorteil dreidimensionaler Ansätze gegenüber 2D- und 2,5D-Modellen bei der Segmentierung verschiedener anatomischer Hirnstrukturen. Ein möglicher Grund für diese Überlegenheit könnte die bessere Nutzung räumlicher Kontextinformation sein, die sich durch volumetrische Eingaben ergibt. Eine Besonderheit von MAP18 und MELD besteht darin, dass zunächst durch eine umfangreiche Vorverarbeitung relevante Merkmale extrahiert werden, die dem Modell anschließend zur Verfügung gestellt werden. Ein weiterer zu untersuchender Ansatz wäre, kontextsensitive 3D-Architekturen mit solchen vorab erstellten Merkmalskarten zu kombinieren, um diese als zusätzliche Informationen zu nutzen. Außerdem sind alle untersuchten Modelle Segmentierungsmodelle, d.h. sie geben für jeden Voxel die Wahrscheinlichkeit an, eine FCD darzustellen. Sollte es klinisch ausreichend sein, lediglich auf verdächtige Regionen hinzuweisen, wäre es zudem interessant, Ansätze der Objekterkennung zu untersuchen.

Darüber hinaus haben neuere Arbeiten gezeigt, dass auch quantitative Bildgebungsverfahren die Detektionsleistung verbessern können. Ding et al. (2024) nutzten Magnetic Resonance Fingerprinting (MRF) in Kombination mit MAP18 und konnten dadurch die Zahl falsch-positiver Cluster deutlich senken. In einer weiteren Studie nutzten sie MRF-basierte Bildgebungsdaten von 40 FCD-Fällen zum Training von nnU-Net und erreichten dabei eine Sensitivität von 80 % bei durchschnittlich 1,7 falsch-positiven Clustern (Ding et al., 2025). Bei der klinischen Bewertung einer FCD stehen den Ärztinnen und Ärzten neben den Bilddaten weitere Informationen zur Verfügung, wie z.B. EEG-Daten oder Anfallssemiologie. Es bleibt daher auch zu untersuchen, inwieweit multimodale Ansätze, die diese weiteren Informationen mit einbeziehen die Detektionsleistung weiter verbessern können.

In der Studie von Kersting et al. (2025) sind mehrere Limitationen zu beachten. Es wurden ausschließlich 3-Tesla-MRTs in die Analyse einbezogen. Ob sich die Modellleistung bei anderen Feldstärken verändert, bleibt damit offen und sollte in zukünftigen Studien untersucht werden. Zudem lag nur bei etwa der Hälfte der Patientinnen und Patienten eine histologische Diagnosesicherung vor. Frühere Analysen für die Bonner Kohorte konnten keine wesentlichen Unterschiede für die Erkennungsraten zwischen histologisch bestätigten und nicht bestätigten Fällen nachweisen (Walger et al., 2024). Da die meisten histologisch bestätigten Fälle FCD Typ II waren, lassen sich nur eingeschränkt Rückschlüsse auf die Modellleistung bei anderen Subtypen, wie z.B. FCD Typ I ziehen. Darüber hinaus ist die exakte Detektion von FCDs selbst für Expertinnen und Experten anspruchsvoll (Walger et al., 2024), sodass eine gewisse Unsicherheit in den Masken unvermeidbar bleibt. Da die Annotationen in dieser Arbeit aus verschiedenen Zentren stammten, könnten zusätzlich Unterschiede im Annotierungsstil eingeflossen sein. Die Verwendung der zwei unterschiedlichen Kriterien (Pinpointing-Kriterium und Detektions-Kriterium) sollte dazu beitragen, der damit verbundenen Variabilität Rechnung zu tragen.

Zusammenfassend stellt die Arbeit von Kersting et al. (2025) den ersten multizentrischen Vergleich frei verfügbarer KI-Modelle zur Detektion von FCDs dar. Unter den untersuchten Verfahren zeigte das neu trainierte 3D-nnU-Net gemessen am F_1-Score die günstigste Balance zwischen Sensitivität und Präzision auf Cluster-Ebene. Ein zusätzlicher Vorteil dieses Ansatzes lag in der kurzen Vorhersagedauer von nur wenigen Minuten pro Fall,

was die Übertragbarkeit in den klinischen Alltag erleichtern könnte. Des Weiteren wiesen Deep-Learning-Modelle mit einer 3D-Architektur die höchste Sensitivität auf und sollten daher zukünftig weiter untersucht werden.

1.5 Zusammenfassung

In der Arbeit von Kersting et al. (2025) wurde die KI-gestützte Detektion von fokaler kortikaler Dysplasie (FCD) in MRT-Aufnahmen untersucht. FCDs sind angeborene Fehlbildungen der Großhirnrinde und eine der häufigsten Ursachen für medikamentös schwer behandelbare fokale Epilepsien (Blümcke et al., 2017). Die Magnetresonanztomographie (MRT) spielt in der präoperativen Diagnostik eine zentrale Rolle. Die korrekte Lokalisation stellt im klinischen Alltag jedoch eine erhebliche Herausforderung dar, da bis zu 30 % übersehen werden (Urbach et al., 2022; Walger et al., 2024). In den letzten Jahren wurden verschiedene Algorithmen veröffentlicht, die bei der FCD-Detektion unterstützen sollen. Es wurden sechs verschiedene KI-Modelle verglichen, darunter drei bereits veröffentlichte (MAP18, MELD, deepFCD) sowie drei neu zur FCD-Detektion trainierte Modelle (FastSurferCNN, 2D-nnUNet, 3D-nnUNet).

Insgesamt wurden MRT-Datensätze von 329 Personen ausgewertet, darunter 244 mit FCD und 85 gesunde Kontrollpersonen. 118 FCD-Fälle aus Bonn dienten dem Training der Modelle, die übrigen 126 FCD-Fälle aus vier Zentren (Bonn, Berlin, Frankfurt, Zürich) bildeten den unabhängigen Testdatensatz.

Um die Modelle fair und umfassend zu vergleichen, wurden ihre Vorhersagen mit den von Expertinnen und Experten markierten Läsionen anhand einheitlicher Kriterien auf Voxel-, Cluster- und Fall-Ebene bewertet. Dazu wurden hauptsächlich der Dice-Koeffizient, der F_1-Score auf Cluster-Ebene und die Detektionsrate sowie Spezifität auf Fall-Ebene betrachtet.

Das neu trainierte 3D-nnUNet erreichte den höchsten F_1-Score mit 0,58 (Sensitivität 0,55, Präzision 0,63) sowie den höchsten Dice-Score mit 0,36. Es wies die zweithöchste Detektionsrate von 55 % und eine Spezifität von 86 % auf. Die höchste Detektionsrate erreichte deepFCD mit 82 %, jedoch mit der geringsten Spezifität von 0 %. Zudem produzierte es die meisten falsch-positiven Cluster und wies somit die geringste Präzision

von 0,03 und den geringsten F_1-Score von 0,07 auf. MELD erreichte eine Detektionsrate von 49 %, MAP18 von 29 %, FastSurferCNN und 2D-nnUNet je von 13 %. Weiterhin zeigte sich, dass sich die Detektionsraten der Modelle für die Fälle aus den vier Zentren unterschieden. MELD erreichte zwar nicht die höchste Detektionsrate, lieferte jedoch über alle vier Zentren hinweg die stabilste Leistung.

Die Studie von Kersting et al. (2025) stellte den ersten multizentrischen Vergleich öffentlich verfügbarer KI-basierter FCD-Detektionsmodelle dar. Die Performance variierte stark zwischen den Zentren und Modellen. Das neu trainierte 3D-nnUNet erzielte dabei das beste Gleichgewicht aus Präzision und Sensitivität bei gleichzeitig schneller Laufzeit. Zudem zeigten Modelle mit dreidimensionaler Datenverarbeitung insgesamt bessere Detektionsraten, was auf die Bedeutung des räumlichen Kontexts bei der FCD-Erkennung hindeutet.

1.6 Literaturverzeichnis der deutschen Zusammenfassung

Avesta A, Hossain S, Lin M, Aboian M, Krumholz HM, Aneja S. Comparing 3D, 2.5D, and 2D Approaches to Brain Image Auto-Segmentation. Bioengineering 2023; 10. https://doi.org/10.3390/bioengineering10020181.

Billot B, Greve DN, Puonti O, Thielscher A, Van Leemput K, Fischl B, Dalca AV, Iglesias JE. SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining. Med Image Anal 2023a; 86: 102789. https://doi.org/10.1016/j.media.2023.102789.

Billot B, Magdamo C, Cheng Y, Arnold SE, Das S, Iglesias JE. Robust machine learning segmentation for large-scale analysis of heterogeneous clinical brain MRI datasets. Proc Natl Acad Sci 2023b; 120: e2216399120. https://doi.org/10.1073/pnas.2216399120.

Blümcke I, Kobow K, Holthausen H. Die ILAE-Klassifikation fokaler kortikaler Dysplasien im klinischen Gebrauch. Z Für Epileptol 2017; 30: 200–7. https://doi.org/10.1007/s10309-017-0119-0.

Blümcke I, Thom M, Aronica E, Armstrong DD, Vinters HV, Palmini A, et al. The clinicopathologic spectrum of focal cortical dysplasias: a consensus classification

proposed by an ad hoc Task Force of the ILAE Diagnostic Methods Commission. Epilepsia 2011; 52: 158–74. https://doi.org/10.1111/j.1528-1167.2010.02777.x.

David B, Kröll-Seger J, Schuch F, Wagner J, Wellmer J, Woermann F, Oehl B, Van Paesschen W, Breyer T, Becker A, Vatter H, Hattingen E, Urbach H, Weber B, Surges R, Elger CE, Huppertz H-J, Rüber T. External validation of automated focal cortical dysplasia detection using morphometric analysis. Epilepsia 2021; 62: 1005–21. https://doi.org/10.1111/epi.16853.

Ding Z, Hu S, Su T-Y, Choi JY, Morris S, Wang X, Sakaie K, Murakami H, Huppertz H-J, Blümcke I, Jones S, Najm I, Ma D, Wang ZI. Combining magnetic resonance fingerprinting with voxel-based morphometric analysis to reduce false positives for focal cortical dysplasia detection. Epilepsia 2024; 65: 1631–43. https://doi.org/10.1111/epi.17951.

Ding Z, Morris S, Hu S, Su T-Y, Choi JY, Blümcke I, Wang X, Sakaie K, Murakami H, Alexopoulos AV, Jones SE, Najm IM, Ma D, Wang ZI. Automated Whole-Brain Focal Cortical Dysplasia Detection Using MR Fingerprinting With Deep Learning. Neurology 2025; 104: e213691. https://doi.org/10.1212/WNL.0000000000213691.

Fischl B. FreeSurfer. NeuroImage 2012; 62: 774–81. https://doi.org/10.1016/j.neuroimage.2012.01.021.

Fortin J-P, Cullen N, Sheline YI, Taylor WD, Aselcioglu I, Cook PA, Adams P, Cooper C, Fava M, McGrath PJ, McInnis M, Phillips ML, Trivedi MH, Weissman MM, Shinohara RT. Harmonization of cortical thickness measurements across scanners and sites. NeuroImage 2018; 167: 104–20. https://doi.org/10.1016/j.neuroimage.2017.11.024.

Gill RS, Lee H-M, Caldairou B, Hong S-J, Barba C, Deleo F, et al. Multicenter Validation of a Deep Learning Detection Algorithm for Focal Cortical Dysplasia. Neurology 2021; 97: e1571–82. https://doi.org/10.1212/WNL.0000000000012698.

Gorgolewski KJ, Auer T, Calhoun VD, Craddock RC, Das S, Duff EP, et al. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. Sci Data 2016; 3: 160044. https://doi.org/10.1038/sdata.2016.44.

Henschel L, Conjeti S, Estrada S, Diers K, Fischl B, Reuter M. FastSurfer - A fast and accurate deep learning based neuroimaging pipeline. NeuroImage 2020; 219: 117012. https://doi.org/10.1016/j.neuroimage.2020.117012.

Hoffmann M, Billot B, Greve DN, Iglesias JE, Fischl B, Dalca AV. SynthMorph: Learning Contrast-Invariant Registration Without Acquired Images. IEEE Trans Med Imaging 2022; 41: 543–58. https://doi.org/10.1109/TMI.2021.3116879.

Iglesias JE. A ready-to-use machine learning tool for symmetric multi-modality registration of brain MRI. Sci Rep 2023; 13: 6657. https://doi.org/10.1038/s41598-023-33781-0.

Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 2021; 18: 203–11. https://doi.org/10.1038/s41592-020-01008-z.

Kersting LN, Walger L, Bauer T, Gnatkovsky V, Schuch F, David B, Neuhaus E, Keil F, Tietze A, Rosenow F, Kaindl AM, Hattingen E, Huppertz H-J, Radbruch A, Surges R, Rüber T. Detection of focal cortical dysplasia: Development and multicentric evaluation of artificial intelligence models. Epilepsia 2025; 66: 1165–76. https://doi.org/10.1111/epi.18240.

Lamberink HJ, Otte WM, Blümcke I, Braun KPJ, Aichholzer M, Amorim I, et al. Seizure outcome and use of antiepileptic drugs after epilepsy surgery according to histopathological diagnosis: a retrospective multicentre cohort study. Lancet Neurol 2020; 19: 748–57. https://doi.org/10.1016/S1474-4422(20)30220-9.

Li X, Morgan PS, Ashburner J, Smith J, Rorden C. The first step for neuroimaging data analysis: DICOM to NIfTI conversion. J Neurosci Methods 2016; 264: 47–56. https://doi.org/10.1016/j.jneumeth.2016.03.001.

Maier-Hein L, Reinke A, Godau P, Tizabi MD, Buettner F, Christodoulou E, et al. Metrics reloaded: recommendations for image analysis validation. Nat Methods 2024; 21: 195–212. https://doi.org/10.1038/s41592-023-02151-z.

Najm I, Lal D, Alonso Vanegas M, Cendes F, Lopes-Cendes I, Palmini A, et al. The ILAE consensus classification of focal cortical dysplasia: An update proposed by an ad hoc task

force of the ILAE diagnostic methods commission. Epilepsia 2022; 63: 1899–919. https://doi.org/10.1111/epi.17301.

Palmini A, Najm I, Avanzini G, Babb T, Guerrini R, Foldvary-Schaefer N, Jackson G, Lüders HO, Prayson R, Spreafico R, Vinters HV. Terminology and classification of the cortical dysplasias. Neurology 2004; 62: S2-8. https://doi.org/10.1212/01.wnl.0000114507.30388.7e.

Schuch F, Walger L, Schmitz M, David B, Bauer T, Harms A, et al. An open presurgery MRI dataset of people with epilepsy and focal cortical dysplasia type II. Sci Data 2023; 10: 475. https://doi.org/10.1038/s41597-023-02386-7.

Spitzer H, Ripart M, Whitaker K, D'Arco F, Mankad K, Chen AA, et al. Interpretable surface-based detection of focal cortical dysplasias: a Multi-centre Epilepsy Lesion Detection study. Brain 2022; 145: 3859–71. https://doi.org/10.1093/brain/awac224.

Taylor DC, Falconer MA, Bruton CJ, Corsellis JA. Focal dysplasia of the cerebral cortex in epilepsy. J Neurol Neurosurg Psychiatry 1971; 34: 369–87. https://doi.org/10.1136/jnnp.34.4.369.

Urbach H, Kellner E, Kremers N, Blümcke I, Demerath T. MRI of focal cortical dysplasia. Neuroradiology 2022; 64: 443–52. https://doi.org/10.1007/s00234-021-02865-x.

Wagstyl K, Whitaker K, Raznahan A, Seidlitz J, Vértes PE, Foldes S, et al. Atlas of lesion locations and postsurgical seizure freedom in focal cortical dysplasia: A MELD study. Epilepsia 2022; 63: 61–74. https://doi.org/10.1111/epi.17130.

Walger L, Bauer T, Kügler D, Schmitz MH, Schuch F, Arendt C, et al. A Quantitative Comparison Between Human and Artificial Intelligence in the Detection of Focal Cortical Dysplasia. Invest Radiol 2024: 10.1097/RLI.0000000000001125. https://doi.org/10.1097/RLI.0000000000001125.

Walger L, Schmitz MH, Bauer T, Kügler D, Schuch F, Arendt C, et al. A public benchmark for human performance in the detection of focal cortical dysplasia. Epilepsia Open 2025; 10: 778–86. https://doi.org/10.1002/epi4.70028.

Zhang S, Zhuang Y, Luo Y, Zhu F, Zhao W, Zeng H. Deep learning-based automated lesion segmentation on pediatric focal cortical dysplasia II preoperative MRI: a reliable approach. Insights Imaging 2024; 15: 71. https://doi.org/10.1186/s13244-024-01635-6.

2. Veröffentlichung

Dieser Publikationsdissertation liegt die folgende, unabhängig begutachtete Veröffentlichung zugrunde:

Kersting LN, Walger L, Bauer T, Gnatkovsky V, Schuch F, David B, Neuhaus E, Keil F, Tietze A, Rosenow F, Kaindl AM, Hattingen E, Huppertz H-J, Radbruch A, Surges R, Rüber T. Detection of focal cortical dysplasia: Development and multicentric evaluation of artificial intelligence models. Epilepsia 2025; 66: 1165-1176

https://doi.org/10.1111/epi.18240

3. Erklärung zum Eigenanteil

Die Arbeit wurde in der Klinik für Epileptologie und Neuroradiologie unter Betreuung von PD Dr. med. Theodor Rüber durchgeführt.

Die Konzeption der Arbeit erfolgte durch mich, Lennart Kersting, in Zusammenarbeit mit meinem Doktorvater PD Dr. med. Theodor Rüber sowie Lennart Walger, geteilter Erstautor der Originalpublikation.

Die Bilddaten einschließlich der Expertenmasken lagen für das Zentrum Bonn bereits vor bzw. wurden von den kooperierenden Zentren zur Verfügung gestellt.

Die Anpassung und das Training von FastSurferCNN erfolgten eigenständig durch mich. Das Training von nnU-Net wurde in Zusammenarbeit mit Lennart Walger durchgeführt.

Das Generieren der Vorhersagen mit allen Modellen (mit Ausnahme von MAP18) erfolgte für die Daten aus Bonn in Zusammenarbeit mit Lennart Walger, für die Daten aus den anderen Zentren überwiegend durch mich. Die Vorhersagen mit MAP18 wurden durch Dr. med. Vadym Gnatkovsky, PhD erstellt.

Die statistische Auswertung und Visualisierung erfolgte eigenständig durch mich mit Unterstützung durch Lennart Walger.

Das Verfassen des englischen Originalentwurfs der Publikation erfolgte eigenständig durch mich.

Ich versichere, die Dissertationsschrift selbständig verfasst zu haben und keine weiteren als die von mir angegebenen Quellen und Hilfsmittel verwendet zu haben.

4. Danksagung

Zuerst möchte ich meinem Doktorvater, PD Dr. Theodor Rüber für die ausgezeichnete Betreuung und die Möglichkeit danken, diese Arbeit in den Kliniken für Neuroradiologie und Epileptologie durchführen zu können. Die fachkundige Beratung und Expertise, sowie die Unterstützung und die Geduld während meiner Forschungsarbeit waren immer sehr motivierend und hilfreich.

Außerdem möchte ich mich bei Lennart Walger für die wertvollen Diskussionen, die Beratung und die Unterstützung bedanken.

Dem gesamten Forschungsteam danke ich für die gute Zusammenarbeit und die nette Atmosphäre, in der ich mich immer sehr gut aufgehoben gefühlt habe.

Abschließend möchte ich meiner Familie für ihre Geduld und Unterstützung während der gesamten Forschungsarbeit danken.

5. Publikation (PDF-Version)

Received: 27 September 2024 | Revised: 12 December 2024 | Accepted: 12 December 2024

DOI: 10.1111/epi.18240

Epilepsia®

RESEARCH ARTICLE

Detection of focal cortical dysplasia: Development and multicentric evaluation of artificial intelligence models

Lennart N. Kersting[1,2] | **Lennart Walger**[1,2] | **Tobias Bauer**[1,2,3] | **Vadym Gnatkovsky**[2] | **Fabiane Schuch**[2] | **Bastian David**[2] | **Elisabeth Neuhaus**[4,5] | **Fee Keil**[4] | **Anna Tietze**[6] | **Felix Rosenow**[5,7] | **Angela M. Kaindl**[8,9,10,11] | **Elke Hattingen**[4,5] | **Hans-Jürgen Huppertz**[12] | **Alexander Radbruch**[1,3,13] | **Rainer Surges**[2] | **Theodor Rüber**[1,2,3,13]

[1]Department of Neuroradiology, University Hospital Bonn, Bonn, Germany
[2]Department of Epileptology, University Hospital Bonn, Bonn, Germany
[3]German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany
[4]Department of Neuroradiology, Goethe University Frankfurt, Frankfurt am Main, Germany
[5]LOEWE Center for Personalized Translational Epilepsy Research (CePTER), Goethe-University, Frankfurt am Main, Germany
[6]Charité-Universitätsmedizin Berlin, Institute of Neuroradiology, Berlin, Germany
[7]Epilepsy Center Frankfurt Rhine-Main and Department of Neurology, Goethe-University, Frankfurt am Main, Germany
[8]Charité-Universitätsmedizin Berlin, Department of Pediatric Neurology, Berlin, Germany
[9]Charité-Universitätsmedizin Berlin, Center for Chronically Sick Children, Berlin, Germany
[10]Charité-Universitätsmedizin Berlin, German Epilepsy Center for Children and Adolescents, Berlin, Germany
[11]Charité-Universitätsmedizin Berlin, Institute of Cell and Neurobiology, Berlin, Germany
[12]Swiss Epilepsy Clinic, Klinik Lengg AG, Zurich, Switzerland
[13]Center for Medical Data Usability and Translation, University of Bonn, Bonn, Germany

Correspondence

Theodor Rüber, Department of Neuroradiology and Epileptology, University Hospital Bonn, Campus Venusberg 1, 53127 Bonn, Germany.
Email: theodor.rueber@ukbonn.de

Abstract

Objective: Focal cortical dysplasia (FCD) is a common cause of drug-resistant focal epilepsy but can be challenging to detect visually on magnetic resonance imaging. Three artificial intelligence models for automated FCD detection are publicly available (MAP18, deepFCD, MELD) but have only been compared on single-center data. Our first objective is to compare them on independent multi-center test data. Additionally, we train and compare three new models and make them publicly available.

Methods: We retrospectively collected FCD cases from four epilepsy centers. We chose three novel models that take two-dimensional (2D) slices (2D-nnUNet), 2.5D slices (FastSurferCNN), and large 3D patches (3D-nnUNet) as inputs and trained them on a subset of Bonn data. As core evaluation metrics, we used

Lennart N. Kersting and Lennart Walger share first authorship.

voxel-level Dice similarity coefficient (DSC), cluster-level F_1 score, subject-level detection rate, and specificity.

Results: We collected 329 subjects, 244 diagnosed with FCD (27.7 ± 14.4 years old, 54% male) and 85 healthy controls (7.1 ± 2.4 years old, 51% female). We used 118 subjects for model training and kept the remaining subjects as an independent test set. 3D-nnUNet achieved the highest F_1 score of .58, the highest DSC of .36 (95% confidence interval [CI] = .30–.41), a detection rate of 55%, and a specificity of 86%. deepFCD showed the highest detection rate (82%) but had the lowest specificity (0%) and cluster-level precision (.03, 95% CI = .03–.04, F_1 score = .07). MELD showed the least performance variation across centers, with detection rates between 46% and 54%.

Significance: This study shows the variance in performance for FCD detection models in a multicenter dataset. The two models with 3D input data showed the highest sensitivity. The 2D models performed worse than all other models, suggesting that FCD detection requires 3D data. The greatly improved precision of 3D-nnUNet may make it a sensible choice to aid FCD detection.

KEYWORDS

computer-aided detection, epilepsy, lesion detection, model comparison, MRI

1 | INTRODUCTION

Focal cortical dysplasia (FCD) is the third most common cause of drug-resistant focal epilepsy.[1] Surgical intervention yields seizure freedom in up to 70% of eligible candidates.[2] As part of the preoperative diagnostic workup, magnetic resonance imaging (MRI) is an important modality.[3] Accurate detection of the lesion on MRI is the best clinical predictor for postoperative seizure freedom.[4] FCDs typically exhibit specific features on MRI, including abnormal gyration, transmantle sign, cortical thickening, and gray–white matter blurring.[5] However, FCDs are still difficult to detect, with a high interrater variability and up to 30% of cases missed by conventional visual assessment.[5,6]

To aid the detection of FCDs, various studies have introduced artificial intelligence (AI)-based approaches.[7–9] A prior study introduced specific evaluation criteria for FCD detection based on the Metrics Reloaded Framework[10] and compared three state-of-the-art models (MAP18: Morphometric Analysis Program, version of 2018,[7] MELD: Multi-centre Epilepsy Lesion Detection,[8] deepFCD: deep learning-based model for FCD detection[9]) to human readers with different levels of expertise on single-center data.[6] Detection rates of these models varied from 31% to 73%, with the best model matching the sensitivity of experts, albeit being much less precise. More recently, Zhang and colleagues trained nnUNet, a widely used medical image segmentation framework,[11] for FCD detection, while omitting a detailed evaluation, and not making the model publicly available.[12] All of these models employ substantially different approaches. MAP18 predicts FCDs based on single voxels from T1-weighted (T1w) images and morphometric feature maps, whereas MELD uses surface-based features from both T1w and fluid-attenuated inversion recovery (FLAIR) images. deepFCD yields predictions on small three-dimensional (3D) patches (16×16×16 voxels), whereas 3D-nnUNet uses larger 3D patches (112×112×192 voxels). In summary, the advantages of the widely used nnUNet have not been leveraged, and a detailed, multicenter comparison of the various state-of-the-art FCD detection approaches has not been conducted.

Here, we trained three deep learning architectures for FCD detection on an in-house dataset of 118 individuals

Key points

1. A multicenter cohort of 329 subjects was used to evaluate six AI models, three state-of-the-art and three new models, for the detection of focal cortical dysplasia.
2. The two models with 3D input data performed best, with a detection rate of up to 82%.
3. The newly trained 3D-nnUNet demonstrated superior balance between precision and sensitivity.

with epilepsy and FCD. For this, we selected the 2D and 3D full-resolution versions of nnUNet,[11] as well as FastSurferCNN, a 2.5D approach.[13] We then compared our newly trained approaches and the three state-of-the-art models (MAP18, MELD, and deepFCD) on 126 FCD cases from four different centers, including 28 publicly available subjects,[14] and an additional 85 published healthy controls (HCs). We hypothesized that there would be significant differences in the overall performance of these models and aimed to determine which approaches are best suited for the automated detection of FCDs.

2 | MATERIALS AND METHODS

2.1 | Participants

We retrospectively ascertained subjects with epilepsy and FCD who underwent presurgical evaluation at four different centers: Bonn (2006–2021), Berlin (2017–2020), Frankfurt (2007–2020), and Zurich (2008–2019). Exclusion criteria were missing whole brain T1w or FLAIR scans acquired at 3 T, multiple FCDs, inconclusive location, and failed preprocessing due to poor image quality. Lesion masks were created by clinicians experienced in the diagnosis of FCD at each individual center, with access to all other clinical information, such as electroencephalographic recordings, if available. Demographic and clinical information included age at scan, sex, and histopathological diagnosis according to International League Against Epilepsy classification.[15,16] No subjects were excluded due to incomplete demographic or clinical information. In addition, we included 85 HCs from a published dataset from Bonn.[14] To meet the subject requirement for MELD harmonization, non-FCD cases from Berlin and Frankfurt were used. Figure 1 illustrates the inclusion/exclusion process. The study was approved by the internal review boards in Bonn (no. 136/19), Berlin (no. EA2/084/18), and Frankfurt (no. 20–649) and in compliance with the internal review guidelines in Zurich, and written informed consent was obtained from all participants.

FCD cases from Bonn overlap with previous studies.[6,7,14,17–20] Utilizing all cases, Walger and colleagues compared human and AI model performance.[6] In this study, we introduce three new models and compare performances on multicenter data. Eighty-five FCD cases from Bonn have previously been published[14] of which 28 were defined as a representative test set, with an average expert detection rate of 49%.[20] Cases from Frankfurt overlap with previous studies.[21,22] Cases from Zurich were used for the training of MAP18,[7] and had to be excluded from the evaluation of MAP18 in this study due to data leakage, which could lead to overly optimistic performance estimates.

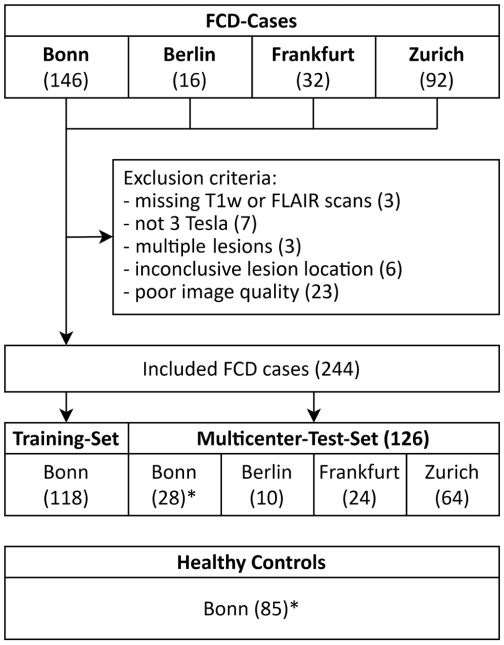

FIGURE 1 Flowchart illustrating the inclusion and exclusion process. *These subjects are part of the published focal cortical dysplasia (FCD) dataset from Bonn.[14] FLAIR, fluid-attenuated inversion recovery; T1w, T1-weighted.

2.2 | AI models

Inclusion criteria for existing AI models for FCD detection were publicly available code and external validation. They were met by MAP18,[7] MELD,[8] and deepFCD.[9] We chose three models representing a 2.5D, a 2D, and a 3D approach with open-source implementations, namely FastSurferCNN and nnUNet to train from scratch. FastSurferCNN has been developed for whole brain segmentation,[13] and nnUNet is widely used for segmentation tasks in medical imaging.[11] Figure 2 provides a schematic overview of the study processing pipeline. All models were trained and/or run on a Linux machine with an Nvidia 3090 Ti GPU and an Intel Core i9 CPU, except for MAP18, which was run on a Windows computer with an Intel Core i5 CPU.

2.2.1 | Model training

Training of nnUNet and FastSurferCNN was performed with coregistered T1w and FLAIR images as inputs. We used synthseg[23,24] and mri_easyreg[24,25] for coregistration.

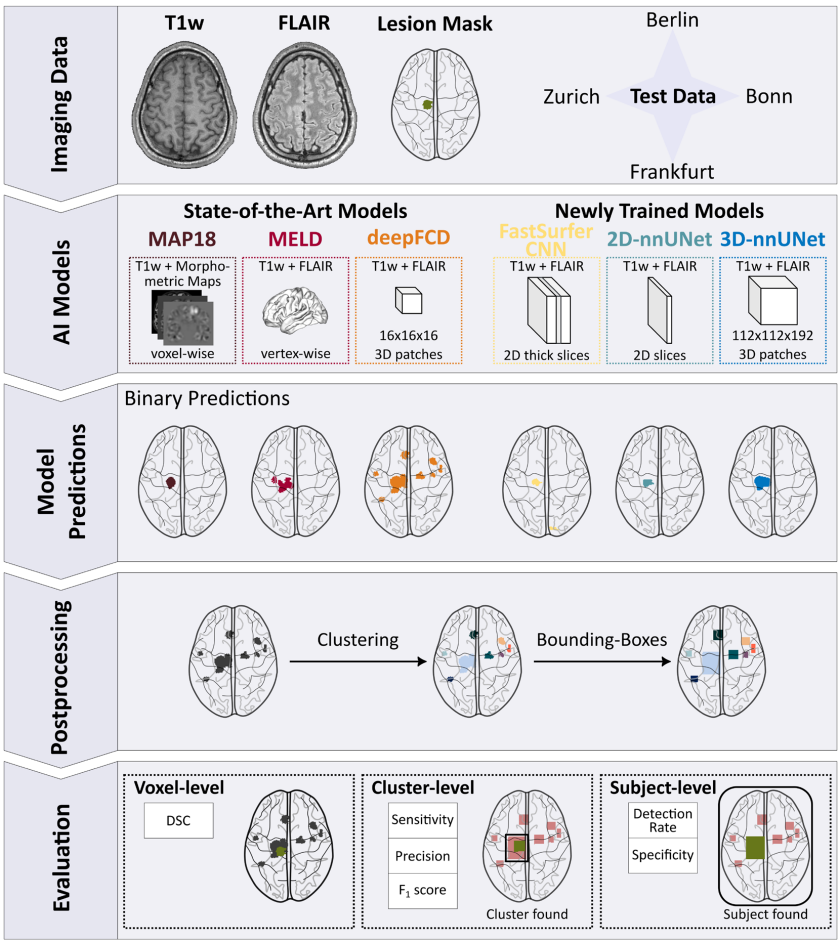

FIGURE 2 Processing pipeline illustrating the comparison between three newly trained models (FastSurferCNN, 2D-nnUNet, 3D-nnUNet) and three state-of-the-art models (MAP18, MELD, deepFCD). Each model's internal data structure is visualized (one-voxel, one-vertex, two-dimensional [2D] slices, thick slices, 3D patches). The ground truth lesion mask is shown in green, and each model is represented by a unique color. Binary predictions are used for voxel-level evaluation, measured by the Dice similarity coefficient (DSC), whereas bounding boxes of clustered predictions are used for cluster- and subject-level evaluation. The F_1 score is a prediction measure describing the harmonic mean of precision and sensitivity. AI, artificial intelligence; FLAIR, fluid-attenuated inversion recovery; T1w, T1-weighted.

For FastSurferCNN, we concatenated T1w and FLAIR slices as combined inputs and sampled only slices containing lesional voxels. We performed fivefold cross-validation, training each fold for 150 epochs using binary cross-entropy loss. The number of epochs was empirically chosen so the model converged for each fold. Separate models were trained for all three anatomical views. We trained nnUNet without any modifications except for reducing the maximum number of epochs to 400 (from 1000), for which the model converged across all folds. We omitted the postprocessing step included in its pipeline (which potentially removes all but the largest predicted cluster). We did not change any other hyperparameters.

2.2.2 | Model postprocessing

To arrive at a binary prediction for each model, some postprocessing had to be conducted. For FastSurferCNN, we averaged the output for each view across all folds,

thresholded it with .5, and combined the three views keeping the maximum value for each voxel. The output of MAP18 was binarized as described in the original publication.[7] The outputs of FastSurferCNN, 2D- and 3D-nnUNet, and MAP18 were clustered using deepFCD's clustering procedure, which groups together voxels that are immediately connected. MELD and deepFCD already produced a clustered output. The clustered output was not further processed, except for FastSurferCNN, where clusters smaller than 100 voxels were removed (the size was determined empirically).

2.3 | Metrics

We evaluated all models at voxel, cluster, and subject levels based on the comparison of model predictions and ground truth lesion masks resampled to an isotropic resolution of 1.0 mm. At voxel level, we report the Dice similarity coefficient (DSC). At cluster level, we first calculated a bounding box around each predicted cluster and the ground truth lesion, that is, the smallest rectangular region enclosing the entire cluster. We then applied two different criteria to determine true positive clusters, as defined in prior work.[6] The first criterion is based on single point localization, which we refer to as "pinpointing." It is met if the clusterwise center of mass falls within the lesion mask. The second criterion is called "detecting," which is met if the DSC score between two bounding boxes exceeds a threshold of .22. To not confuse this score with the voxel-level DSC, we will refer to it as "boxDSC." The specific threshold was determined by Walger and colleagues to specifically reflect the detection performance of experts for FCD detection.[6] Note that these criteria could be met by multiple clusters per subject, even though each had only a single lesion and all clusters were considered in the calculation of the cluster-level metrics. We report cluster-level sensitivity, precision, and F_1 score (a prediction measure combining precision and sensitivity into a single performance metric). At subject level, we report detection rates and pinpointing rates. An FCD case was considered "found" if at least one cluster met the criterion. Subject-level specificity was determined by evaluating model predictions for HCs. An HC was considered "true negative" if the prediction was empty. As primary metrics for comparing model performance, we use voxel-level DSC, cluster-level F_1 score, and subject-level detection rate.

2.4 | Statistical analysis

Values were reported as mean and 95% confidence interval (CI). We applied bootstrapping to estimate CIs for the performance variance across centers. All statistical analyses were performed using Stata.[27]

2.5 | Code availability

Instructions for setting up the trained nnUNet models to generate predictions for new individuals are available on GitLab (https://gitlab.com/lab_tni/projects/nnunet_fcd). Adapted FastSurferCNN code can be made available upon reasonable request to the corresponding author.

3 | RESULTS

3.1 | Participants

A total of 244 FCD cases and 85 HCs were included in the study, of which 211 (126 FCD cases, 85 HCs) formed the multicenter test cohort, as shown in Figure 1. Of the included FCD cases, 49% were histologically confirmed, including five cases with FCD type I, 124 cases with FCD type II, and only one case with FCD type III. For Zurich, the pre-exclusion cohort is presented in Table 1, because demographic and clinical information is not available at the individual level for data protection reasons. The Berlin cohort differs from other centers in that it only contains pediatric subjects (mean age at scan = 7.5 ± 2.9 years). The youngest age at scan in the entire dataset was 3 years. Demographic and clinical information is summarized in Table 1. All T1w images and all but 27 FLAIR images from Bonn, as well as all images from Berlin, Frankfurt, and Zurich, had isotropic resolutions between .5 mm and 1.0 mm. Scanner-specific information is shown in Supplementary Table 1.

3.2 | Test set performance

All metrics except for subject-level specificity were calculated using the group of 126 FCD cases. The 85 HCs were only used to calculate subject-level specificity. The average number of clusters per subject varied from .5 (95% CI = .3–.6) for 2D-nnUNet to 24.7 (95% CI = 22.4–27.2) for deepFCD. 2D-nnUNet also generated the highest number of zero predictions with 65% and deepFCD the least with 0%. MAP18 predicted the smallest clusters on average with a volume of .44 mL (95% CI = .31–.79) and 3D-nnUNet the largest with 2.44 mL (95% CI = 1.94–3.18). Whereas Table 2 gives an overview over all metrics for all models, in the following we highlight key differences. Including all preprocessing steps, generating predictions

TABLE 1 Demographic and clinical Information.

Characteristic	FCD cases					HCs, Bonn
	Bonn, training	Bonn, test	Berlin, test	Frankfurt, test	Zurich, test[a]	
Number, *n*	118	28	10	24	92	85
Age at scan, years, mean ± SD	29.5 ± 14.2	28.0 ± 10.7	7.5 ± 2.9	28.3 ± 15.5	27.2 ± 14.3	7.1 ± 2.4
Sex, *n*						
Female	51	12	5	4	53	43
Male	67	16	5	20	39	42
Histopathology, *n*						
No surgery	45	8	9	13	59	-
I	3	0	0	0	2	-
II (a/b)	69 (19/50)	20 (6/14)	1 (0/1)	9 (1/8)	25 (NA)	-
IIIb	0	0	0	1	0	-
No definite FCD on histopathology	1	0	0	0	0	-
No classification	0	0	0	0	4	-
No information	0	0	0	1	2	-

Abbreviations: FCD, focal cortical dysplasia; HC, healthy control; NA, not available.

[a]Zurich information was only available for the pre-exclusion cohort (64 of these were included in this study).

TABLE 2 Performance metrics for the multicenter test cohort.

Level	Metric	MAP18[a]	MELD	deepFCD	FastSurferCNN	2D-nnUNet	3D-nnUNet
Voxel	Empty	21% (13/62)	12% (15/126)	0% (0/126)	26% (33/126)	65% (82/126)	21% (27/126)
	DSC	.15 [.10–.19]	.21 [.18–.24]	.15 [.12–.17]	.06 [.03, .08]	.07 [.04–.10]	.36 [.30–.41]
Cluster	Number/subject	3.5 [2.7–5.2]	2.1 [1.8–2.5]	24.7 [22.4–27.2]	2.3 [1.9–2.6]	.5 [.3–.6]	.9 [.8–1.0]
	Volume, mL	.44 [.31–.79]	.97 [.80–1.44]	.92 [.87–.97]	.86 [.71–1.02]	.75 [.48–1.28]	2.44 [1.94–3.18]
	Precision	.08 (18/219)	.26 (70/266)	.03 (107/3109)	.06 (16/284)	.26 (16/62)	.63 (69/110)
	Sensitivity	.29 (18/62)	.52 (70/134)	.82 (107/130)	.13 (16/126)	.13 (16/126)	.55 (69/126)
	F_1 score	.13	.35	.07	.08	.17	.58
Subject	Pinpointing rate	66% (41/62)	48% (61/126)	80% (101/126)	25% (31/126)	29% (36/126)	64% (81/126)
	Detection rate	29% (18/62)	49% (62/126)	82% (103/126)	13% (16/126)	13% (16/126)	55% (69/126)
	Specificity	51% (43/85)	55% (47/85)	0% (0/85)	22% (19/85)	95% (81/85)	86% (73/85)

Note: Values in square brackets are 95% confidence intervals; values in parentheses are numerators/denominators. The F_1 score is a prediction measure describing the harmonic mean of precision and recall. The 85 healthy controls were only used to calculate subject-level specificity.

Abbreviation: DSC, Dice similarity coefficient.

[a]For MAP18, Zurich cases were excluded from evaluation, as they were used in training.

for a single subject took approximately 20 min for MAP18, 7–9 h for MELD (6–8 h for FreeSurfer reconstruction, 45–60 min for feature extraction and harmonization, and 5–10 min for inference), and approximately 2 h for deepFCD (<5 min for preprocessing and 120 min for inference). FastSurferCNN required approximately 8 min and both nnUNets approximately 6 min (for both, preprocessing took approximately 5 min and inference 3 min and <1 min, respectively).

3.2.1 | Voxel level

3D-nnUNet achieved the highest DSC score with .36 (95% CI = .30–.41), followed by MELD with .21 (95% CI = .18–.24). MAP18 and deepFCD achieved similar DSC scores of .15 (95% CI = .10–.19) and .15 (95% CI = .12–.17), respectively, for different reasons. Whereas MAP18 showed a low sensitivity of .11 (95% CI = .07–.15) and high precision of .51 (95% CI = .41–.60), the opposite was true

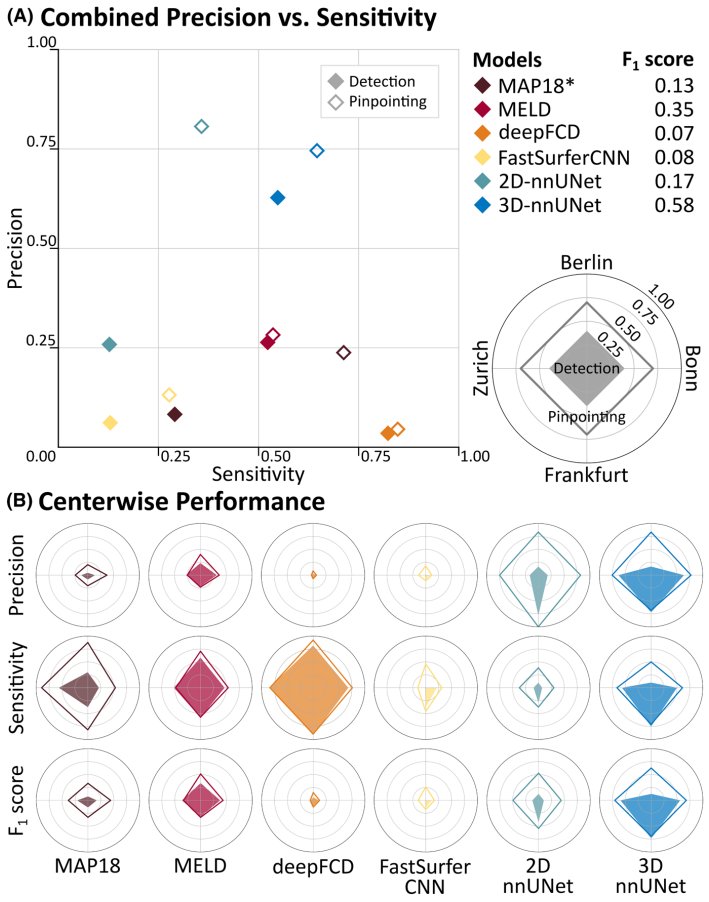

FIGURE 3 Cluster-level performance of all models for the entire test cohort (A) and centerwise (B). (A) Cluster-level precision and sensitivity for the entire test cohort. *MAP18 excludes Zurich data. (B) Radar plots showing cluster-level precision, sensitivity, and F_1 score for individual centers for detection criterion (shaded areas) and pinpointing (unfilled lines). The F_1 score is a prediction measure describing the harmonic mean of precision and sensitivity.

for deepFCD (sensitivity = .43, 95% CI = .38–.48; precision = .11, 95% CI = .09–.14). FastSurferCNN and 2D-nnUNet were lowest, with a DSC score of .06 (95% CI = .03–.08) and .07 (95% CI = .04–.10), respectively.

3.2.2 | Cluster level

deepFCD achieved the highest average DSC score, choosing the best-overlapping cluster per subject with .38 (95% CI = .34–.42), followed by 3D-nnUNet with .36 (95% CI = .30–.41). However, given that a cluster was detected, 3D-nnUNet showed the highest DSC of .62 (95% CI = .58–.66). Whereas 3D-nnUNet achieved the highest precision of .63, deepFCD showed the lowest with .03. deepFCD was most sensitive (.82), followed by 3D-nnUNet (.55) and MELD (.52). Combining precision and sensitivity, 3D-nnUNet showed the highest F_1 score with .58, followed by MELD with .35. deepFCD had the lowest F_1 score with .07. Figure 3A shows cluster-level precision and sensitivity.

3.2.3 | Subject level

deepFCD achieved the highest pinpointing rate (80%) and detection rate (82%). MAP18 achieved the second highest pinpointing rate (66%), followed by 3D-nnUNet (64%), which had the second highest detection rate (55%). 2D-nnUNet and FastSurferCNN showed both the lowest pinpointing (29% and 25%, respectively) and detection rates (both 13%). 2D-nnUNet produced the fewest nonzero predictions for HCs with 5% (4/85), followed by 3D-nnUNet with 14% (12/85). MELD, MAP18, and FastSurferCNN produced predictions in 45% (38/85), 49% (42/85), and 78% (66/85), respectively. deepFCD had the lowest specificity (0%), with nonzero predictions for 100% (85/85) of the HCs.

3.3 | Centerwise performance

The detection rate of the state-of-the-art models did not differ substantially between the test and validation set from Bonn (MAP18: 11%, MELD: 4%, deepFCD: 6%). Detection rates on the validation set were 32% higher for FastSurferCNN (test: 21%, validation: 53%) and 27% higher for 3D-nnUNet (test: 50%, validation: 77%); however, 2D-nnUNet differed by only 7% (test: 7%, validation: 14%). All performance metrics for the validation data are shown in Supplementary Table 2.

The Bonn test set was selected to have an expert performance of 49% by Walger, Schmitz, and colleagues.[20] The highest detection rate was achieved by deepFCD with 68%, followed by 3D-nnUNet with 50% and MELD with 46%. MAP18, FastSurferCNN, and 2D-nnUNet detected 21%, 21%, and 7% of the subjects, respectively. These test cases are also part of the previously published FCD dataset from Bonn. Supplementary Table 3 shows the model performance metrics for all published FCD cases.

Comparing performance per center, MELD showed the least variance in performance across all centers, with detection rates between 46% for Bonn and 54% for Frankfurt. deepFCD was the most sensitive for each center, with a minimum detection rate of 68% for Bonn and a maximum of 92% in Frankfurt. Overall, the Frankfurt cohort represented the "easiest" dataset, with the highest overall detection rate of 57% (Zurich: 42%, Bonn: 35%, Berlin: 33%) and all models achieving their highest detection rate. For Berlin, both 2D-nnUNet and 3D-nnUNet detected only a single subject, but 3D-nnUNet still pinpointed five of 10 (2D-nnUNet pinpointing 2/10). FastSurferCNN detected zero subjects from the Berlin cohort but pinpointed three of 10. The detection rates for all other models did not substantially differ for the Berlin cohort. Similar to the increased detection rates of FastSurferCNN and 3D-nnUNet on our validation data, we observed a 27% higher detection rate for MAP18 on the Zurich cohort. Values for all centerwise metrics are shown in Table 3. Figure 3B shows the centerwise cluster-level metrics.

4 | DISCUSSION

We trained three new models for the detection of FCDs using FastSurferCNN, 2D-nnNUnet, and 3D-nnUNet and compared them to the three state-of-the-art FCD detection AI models MAP18, MELD, and deepFCD on an independent multicenter dataset. The new models were trained on 118 FCD cases from Bonn. All models were tested on 85 HCs and 126 FCD cases from four centers, including a pediatric cohort. deepFCD showed the highest detection rate of 82%, followed by 3D-nnUNet with 55%. However, deepFCD also produced nonzero predictions in all HCs, that is, showed a specificity of 0%, whereas 3D-nnUNet still had a specificity of 86%. 2D-nnUNet and 2.5D FastSurferCNN showed the lowest detection rates, both at 13%. Cluster-level precision was highest for 3D-nnUNet with .63 and lowest for deepFCD with .03. In combination, 3D-nnUNet showed the highest cluster-level F_1 score with .58, followed by MELD with .35. deepFCD had the lowest F_1 score with .07.

Evaluation of AI models for FCD detection is highly dependent on the choice of the criterion for "finding" a lesion.[6] MAP18 was particularly poorly represented by the chosen criterion with 29% detected versus 66% pinpointed. In Berlin, 3D-nnUNet showed its lowest detection rate of 10%, but still pinpointed 50%. Criteria such as one-voxel overlap, as used in the original works of MAP18,[7] MELD,[8] and deepFCD,[9] but also pinpointing, can be exploited, for example, by predicting many clusters. Thus, comparing models based solely on any sort of detection rate does not provide a comprehensive evaluation.[28] We chose the cluster-level F_1 score to highlight the tradeoff between the number of lesions detected and the number of clusters predicted. However, the F_1 score may be less informative compared to the detection rate, especially from a clinical perspective. The acceptable number of false-positive clusters in clinical practice remains an open research question. Although voxel-level metrics are most commonly used to evaluate model performance, it remains unclear how important such metrics are for the care of people with epilepsy; for example, to capture the precise extent of a lesion could be critical for surgical planning. Performing evaluations at each of these levels is necessary to provide a comprehensive analysis and comparison.

Model performance varied between centers. The newly trained models showed better performance on validation data, compared to test data, and also MAP18 performed substantially better on the Zurich cohort, which was used in its training. MELD's performance showed the least variability across all centers, which may be due to the MRI data harmonization scheme used in MELD's pipeline. Regardless of a model's ability to generalize, a given cohort may be "easier" overall, as we observed for Frankfurt data, or a model may be specifically affected by some characteristics within a dataset, as we observed with nnUNet on the Berlin pediatric cohort. Future studies may want to specifically focus on such characteristics, for example, training a model on pediatric cases. Such variation across model performance and cohorts highlights the need for multicenter evaluation to get a better understanding of all the factors that may influence any specific model.

The models evaluated in our study use different methods to process MRI data and generate predictions. MAP18 operates at the single-voxel level,[7] MELD at the single-vertex level,[8] and deepFCD uses small 16×16×16 voxel

TABLE 3 Centerwise focal cortical dysplasia detection performance.

Level	Metric	MAP18	MELD	deepFCD	FastSurferCNN	2D-nnUNet	3D-nnUNet
Bonn, $n=28$							
Voxel	Empty	29% (8/28)	29% (8/28)	0% (0/28)	11% (3/28)	68% (19/28)	29% (8/28)
	DSC	.12 [.08–.18]	.17 [.11–.24]	.15 [.10–.20]	.07 [.03, .13]	.03 [.01–.06]	.30 [.20–.40]
Cluster	Number/subject	1.5 [.9–2.6]	1.5 [1.0–1.9]	11.9 [10.6–13.2]	2.8 [2.1–3.5]	.4 [.2–.6]	.8 [.6–1.0]
	Volume, mL	.45 [.29–.67]	.70 [.52–.91]	.85 [.73–1.01]	.61 [.47–.87]	.35 [.14–.66]	1.75 [1.12–2.75]
	Precision	.14 (6/43)	.32 (13/41)	.06 (19/333)	.08 (6/78)	.18 (2/11)	.64 (14/22)
	Sensitivity	.21 (6/28)	.46 (13/28)	.68 (19/28)	.21 (6/28)	.07 (2/28)	.50 (14/28)
	F_1 score	.17	.38	.11	.11	.10	.56
Subject	Pinpointing rate	50% (14/28)	54% (15/28)	75% (21/28)	29% (8/28)	25% (7/28)	61% (17/28)
	Detection rate	21% (6/28)	46% (13/28)	68% (19/28)	21% (6/28)	7% (2/28)	50% (14/28)
Berlin, $n=10$							
Voxel	Empty	10% (1/10)	0% (0/10)	0% (0/10)	10% (1/10)	70% (7/10)	40% (4/10)
	DSC	.13 [.05–.26]	.16 [.07–.28]	.16 [.09–.25]	.02 [.00–.08]	.04 [.00–.16]	.09 [.01–.34]
Cluster	Number/subject	6.9 [4.3–8.8]	3.0 [2.1–3.7]	29.4 [25.8–33.0]	3.3 [2.1–4.1]	.6 [.1–1.8]	.6 [.2–.8]
	Volume, mL	.53 [.21–1.72]	1.90 [.70–6.36]	.97 [.80–1.34]	.63 [.43–1.23]	.50 [.14–1.10]	1.52 [.41–3.63]
	Precision	.04 (3/69)	.23 (7/30)	.03 (9/294)	.00 (0/33)	.17 (1/6)	.17 (1/6)
	Sensitivity	.30 (3/10)	.58 (7/12)	.82 (9/11)	.00 (0/10)	.10 (1/10)	.10 (1/10)
	F_1 score	.08	.33	.06	.00	.12	.12
Subject	Pinpointing rate	80% (8/10)	50% (5/10)	80% (8/10)	30% (3/10)	20% (2/10)	50% (5/10)
	Detection rate	30% (3/10)	50% (5/10)	80% (8/10)	0% (0/10)	10% (1/10)	10% (1/10)
Frankfurt, $n=24$							
Voxel	Empty	17% (4/24)	4% (1/24)	0% (0/24)	0% (0/24)	62% (15/24)	8% (2/24)
	DSC	.18 [.11–.27]	.25 [.18–.32]	.28 [.21–.36]	.13 [.07–.21]	.15 [.07–.27]	.51 [.37–.63]
Cluster	Number/subject	4.5 [3.0–9.0]	2.9 [2.0–4.1]	15.4 [12.2–18.9]	4.2 [3.3–5.1]	.4 [.2–.5]	1.0 [.8–1.2]
	Volume, mL	.37 [.24–.60]	1.13 [.83–1.56]	.85 [.73–1.00]	1.05 [.81–1.45]	1.51 [.74–2.93]	2.95 [2.01–4.23]
	Precision	.08 (9/107)	.22 (15/69)	.06 (22/370)	.09 (9/101)	.78 (7/9)	.72 (18/25)
	Sensitivity	.38 (9/24)	.58 (15/26)	.92 (22/24)	.38 (9/24)	.29 (7/24)	.75 (18/24)
	F_1 score	.14	.32	.11	.14	.42	.73
Subject	Pinpointing rate	79% (19/24)	50% (12/24)	88% (21/24)	46% (11/24)	38% (9/24)	71% (17/24)
	Detection rate	38% (9/24)	54% (13/24)	92% (22/24)	38% (9/24)	29% (7/24)	75% (18/24)
Zurich, $n=64$							
Voxel	Empty	8% (5/64)	9% (6/64)	0% (0/64)	45% (29/64)	64% (41/64)	20% (13/64)
	DSC	.31 [.26–.36]	.22 [.18–.27]	.09 [.07–.11]	.03 [.01–.05]	.07 [.04–.12]	.36 [.29–.44]
Cluster	Number/subject	4.4 [3.1–7.7]	2.0 [1.6–2.4]	33.0 [30.2–36.2]	1.1 [.8–1.6]	.6 [.4–.9]	.9 [.7–1.0]
	Volume, mL	.39 [293–541]	.76 [.62–.99]	.94 [.88–1.00]	.99 [.71–1.43]	.73 [.33–1.58]	2.58 [1.82–3.94]
	Precision	.13 (36/281)	.28 (35/126)	.03 (57/2112)	.01 (1/72)	.17 (6/36)	.63 (36/57)
	Sensitivity	.56 (36/64)	.51 (35/68)	.85 (57/67)	.02 (1/64)	.09 (6/64)	.56 (36/64)
	F_1 score	.21	.36	.05	.01	.12	.60
Subject	Pinpointing rate	88% (56/64)	45% (29/64)	80% (51/64)	14% (9/64)	28% (18/64)	66% (42/64)
	Detection rate	56% (36/64)	48% (31/64)	84% (54/64)	2% (1/64)	9% (6/64)	56% (36/64)

Note: Values in square brackets are 95% confidence intervals; values in parentheses are numerators/denominators. The F_1 score is a prediction measure describing the harmonic mean of precision and recall.

Abbreviation: DSC, Dice similarity coefficient.

3D patches.[9] FastSurferCNN uses a 2.5D approach,[13] whereas nnUNet uses either a 2D or 3D approach,[11] using a large 3D patch size (112×112×192 voxel) compared to deepFCD. Our results suggest that models using a 3D framework, especially large 3D patches, are the most promising approach for FCD segmentation. deepFCD was the most sensitive, but least precise model, whereas 3D-nnUNet had the highest F_1 score and second highest sensitivity. A study by Avesta and colleagues compared the performance of 3D models and their 2D or 2.5D counterparts in segmenting three anatomical regions in brain MRI and showed that 3D approaches were superior, even with limited training data.[29] However, differences in input data dimensions or shape may be only one of many factors that affect model performance. The heavy preprocessing employed in MAP18 and MELD may further aid detection ability. Combining preprocessing, for example, generating feature maps of MAP18, with 3D patch-based models may be worth exploring in future approaches.

Several limitations should be acknowledged. First, the localization of FCDs remains challenging even for experts,[6] leading to uncertainty in the ground truth annotations. In this study, lesion masks were drawn by clinicians from different centers, which may introduce additional variations in the annotation style. The employed "pinpointing" and "detecting" criteria aim to counteract such voxel-level variance. Second, only approximately half of all FCD cases were histologically confirmed. Although in a previous study we found no significant differences in detection performance between confirmed and unconfirmed cases in the Bonn cohort,[6] this remains a limitation of our study. Third, because only five of the histologically confirmed cases were FCD type I, our study provides limited insight into the performance of automated lesion detection algorithms for FCD type I. Lastly, we have only included 3-T MRI data, and it remains unclear whether and how model performance is affected by different scanner field strengths.

In conclusion, we presented the first multicenter comparison of publicly available AI-based approaches for FCD detection. Our newly trained 3D-nnUNet offered the best tradeoff between cluster-level sensitivity and precision. Its comparatively fast runtime of only a few minutes per case may aid its integration into clinical practice. Future model development may focus on 3D models for high sensitivity while trying to maintain high precision. Furthermore, it has to be determined what the tradeoff between sensitivity and precision means with respect to how helpful a model is in the diagnostic workup.

AUTHOR CONTRIBUTIONS

Lennart Kersting and Lennart Walger contributed equally to all aspects of this work. Tobias Bauer, Alexander Radbruch, Rainer Surges, and Theodor Rüber contributed to the conception and design. Tobias Bauer, Vadym Gnatkovsky, Fabiane Schuch, Bastian David, Elisabeth Neuhaus, Fee Keil, Anna Tietze, Felix Rosenow, Angela M. Kaindl, Elke Hattingen, Hans-Jürgen Huppertz, and Theodor Rüber acquired the data and contributed to its analysis and interpretation. Theodor Rüber helped draft the manuscript, and Elisabeth Neuhaus, Anna Tietze, Felix Rosenow, Angela M. Kaindl, Elke Hattingen, Hans-Jürgen Huppertz, Alexander Radbruch, Rainer Surges, and Theodor Rüber contributed to its revision.

ACKNOWLEDGMENTS

This work was partially supported by a grant of the federal state of Hesse for the LOEWE Center for Personalized Translational Epilepsy Research, as well as by the Einstein Stiftung Fellowship through the Günter Endres Fond and the Sonnenfeld-Stiftung. Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

F.R. has received honoraria for lecturing and consultation from Angelini Pharma, Eisai, Jazz Pharma, Roche Pharma, Takeda, and UCB Pharma, and has received financial research support from Dr. Schär Deutschland, Vitaflo Deutschland, Nutricia Milupa, Desitin Pharma, Hamburg, Federal State of Hesse (via the LOEWE program), Chaja Foundation Frankfurt, Reiss Foundation Frankfurt, Dr. Senckenbergische Foundation Frankfurt, Ernst Max von Grunelius Foundation Frankfurt, and Detlev-Wrobel-Fonds for Epilepsy Research Frankfurt outside the submitted work. A.M.K. has served on the advisory boards of Angelini, Desitin, Jazz Pharmaceuticals, Novartis, and UCB. H.-J.H. is the author of the Morphometric Analysis Program v2018 (MAP18). A.R. has served on scientific advisory boards for GE Healthcare, Bracco, Bayer, Guerbet, and AbbVie; has received speaker honoraria from Bayer, Guerbet, Siemens, and Medscape; and has been a consultant for, and has received institutional study support from, Guerbet and Bayer. R.S. has received fees as speaker or for serving on advisory boards from Angelini, Arvelle, Bial, Desitin, Eisai, Janssen-Cilag, LivaNova, Novartis, Precisis, UCB Pharma, UNEEG, and Zogenix. These activities were not related to the content of this article. T.R. has received fees as a speaker from Eisai. The remaining authors have no conflicts of interest. We confirm that we have read the Journal's position on issues involved in ethical publication and affirm that this report is consistent with those guidelines.

DATA AVAILABILITY STATEMENT

A subset of the 146 FCD cases from Bonn, including the 28 test cases used in this study and the 85 HCs, is available to the public (doi: 10.18112/openneuro.ds004199.v1.0.5).

Instructions for installing the newly trained nnUNet models are available on GitLab (https://gitlab.com/lab_tni/projects/nnunet_fcd). Additional data can be made available upon reasonable request to the corresponding author.

ORCID

Lennart N. Kersting https://orcid.org/0009-0002-1983-4892
Lennart Walger https://orcid.org/0000-0002-3300-6877
Tobias Bauer https://orcid.org/0000-0002-0555-6214
Bastian David https://orcid.org/0000-0002-0146-0629
Angela M. Kaindl https://orcid.org/0000-0001-9454-206X
Theodor Rüber https://orcid.org/0000-0002-6180-7671

REFERENCES

1. Blümcke I, Kobow K, Holthausen H. Die ILAE-Klassifikation fokaler kortikaler Dysplasien im klinischen Gebrauch. Z Für Epileptol. 2017;30(3):200–7.
2. Lamberink HJ, Otte WM, Blümcke I, Braun KPJ, Aichholzer M, Amorim I, et al. Seizure outcome and use of antiepileptic drugs after epilepsy surgery according to histopathological diagnosis: a retrospective multicentre cohort study. Lancet Neurol. 2020;19(9):748–57.
3. Guerrini R, Duchowny M, Jayakar P, Krsek P, Kahane P, Tassi L, et al. Diagnostic methods and treatment options for focal cortical dysplasia. Epilepsia. 2015;56(11):1669–86.
4. Wagstyl K, Whitaker K, Raznahan A, Seidlitz J, Vértes PE, Foldes S, et al. Atlas of lesion locations and postsurgical seizure freedom in focal cortical dysplasia: a MELD study. Epilepsia. 2022;63(1):61–74.
5. Urbach H, Kellner E, Kremers N, Blümcke I, Demerath T. MRI of focal cortical dysplasia. Neuroradiology. 2022;64(3):443–52.
6. Walger L, Bauer T, Kügler D, Schmitz MH, Schuch F, Arendt C, et al. A quantitative comparison between human and artificial intelligence in the detection of focal cortical dysplasia. Investig Radiol. 2024;1125. https://doi.org/10.1097/RLI.0000000000001125
7. David B, Kröll-Seger J, Schuch F, Wagner J, Wellmer J, Woermann F, et al. External validation of automated focal cortical dysplasia detection using morphometric analysis. Epilepsia. 2021;62(4):1005–21.
8. Spitzer H, Ripart M, Whitaker K, D'Arco F, Mankad K, Chen AA, et al. Interpretable surface-based detection of focal cortical dysplasias: a multi-centre epilepsy lesion detection study. Brain. 2022;145(11):3859–71.
9. Gill RS, Lee H-M, Caldairou B, Hong SJ, Barba C, Deleo F, et al. Multicenter validation of a deep learning detection algorithm for focal cortical dysplasia. Neurology. 2021;97(16):e1571–e1582.
10. Maier-Hein L, Reinke A, Godau P, Tizabi MD, Buettner F, Christodoulou E, et al. Metrics reloaded: recommendations for image analysis validation. Nat Methods. 2024;21(2):195–212.
11. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods. 2021;18(2):203–11.
12. Zhang S, Zhuang Y, Luo Y, Zhu F, Zhao W, Zeng H. Deep learning-based automated lesion segmentation on pediatric focal cortical dysplasia II preoperative MRI: a reliable approach. Insights Imaging. 2024;15(1):71.
13. Henschel L, Conjeti S, Estrada S, Diers K, Fischl B, Reuter M. FastSurfer - a fast and accurate deep learning based neuroimaging pipeline. NeuroImage. 2020;219:117012.
14. Schuch F, Walger L, Schmitz M, David B, Bauer T, Harms A, et al. An open presurgery MRI dataset of people with epilepsy and focal cortical dysplasia type II. Sci Data. 2023;10(1):475.
15. Blümcke I, Thom M, Aronica E, Armstrong DD, Vinters HV, Palmini A, et al. The clinicopathologic spectrum of focal cortical dysplasias: a consensus classification proposed by an ad hoc task force of the ILAE diagnostic methods commission. Epilepsia. 2011;52(1):158–74.
16. Najm I, Lal D, Alonso Vanegas M, Cendes F, Lopes-Cendes I, Palmini A, et al. The ILAE consensus classification of focal cortical dysplasia: an update proposed by an ad hoc task force of the ILAE diagnostic methods commission. Epilepsia. 2022;63(8):1899–919.
17. Rácz A, Becker AJ, Quesada CM, Borger V, Vatter H, Surges R, et al. Post-surgical outcome and its determining factors in patients operated on with focal cortical dysplasia type II—A retrospective Monocenter study. Front Neurol. 2021;12:666056. https://doi.org/10.3389/fneur.2021.666056/full
18. Salemdawod A, Wach J, Banat M, Borger V, Hamed M, Haberl H, et al. Predictors of postoperative long-term seizure outcome in pediatric patients with focal cortical dysplasia type II at a German tertiary epilepsy center. J Neurosurg Pediatr. 2022;29(1):83–91.
19. Wagner J, Weber B, Urbach H, Elger CE, Huppertz HJ. Morphometric MRI analysis improves detection of focal cortical dysplasia type II. Brain. 2011;134(10):2844–54.
20. Walger, L, Schmitz, MH, Bauer, T, Kügler, D, Schuch, F, Arendt, C, et al. A Public Benchmark for Human Performance in FCD Detection. 2024. https://doi.org/10.21203/rs.3.rs-4528693/v1.
21. Ahmad R, Maiworm M, Nöth U, Seiler A, Hattingen E, Steinmetz H, et al. Cortical changes in epilepsy patients with focal cortical dysplasia: new insights with T2 mapping. J Magn Reson Imaging JMRI. 2020;52(6):1783–9.
22. Maiworm M, Nöth U, Hattingen E, Steinmetz H, Knake S, Rosenow F, et al. Improved visualization of focal cortical dysplasia with surface-based multiparametric quantitative MRI. Front Neurosci. 2020;14:622.
23. Billot B, Greve DN, Puonti O, Thielscher A, van Leemput K, Fischl B, et al. SynthSeg: segmentation of brain MRI scans of any contrast and resolution without retraining. Med Image Anal. 2023;86:102789.
24. Billot B, Magdamo C, Cheng Y, Arnold SE, das S, Iglesias JE. Robust machine learning segmentation for large-scale analysis of heterogeneous clinical brain MRI datasets. Proc Natl Acad Sci. 2023;120(9):e2216399120.
25. Iglesias JE. A ready-to-use machine learning tool for symmetric multi-modality registration of brain MRI. Sci Rep. 2023;13(1):6657.
26. Hoffmann M, Billot B, Greve DN, Iglesias JE, Fischl B, Dalca AV. SynthMorph: learning contrast-invariant registration without acquired images. IEEE Trans Med Imaging. 2022;41(3):543–58.
27. Anon. Stata Statistical Software. 2023.
28. Walger L, Adler S, Wagstyl K, Henschel L, David B, Borger V, et al. Artificial intelligence for the detection of focal cortical

dysplasia: challenges in translating algorithms into clinical practice. Epilepsia. 2023;64(5):1093–112.
29. Avesta A, Hossain S, Lin M, Aboian M, Krumholz HM, Aneja S. Comparing 3D, 2.5D, and 2D approaches to brain image auto-segmentation. Bioengineering. 2023;10(2):181.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Kersting LN, Walger L, Bauer T, Gnatkovsky V, Schuch F, David B, et al. Detection of focal cortical dysplasia: Development and multicentric evaluation of artificial intelligence models. Epilepsia. 2025;66:1165–1176. https://doi.org/10.1111/epi.18240