März 2021

DIN SPEC 13288

ICS 35.240.80

Leitfaden für die Entwicklung von Deep-Learning-Bilderkennungssystemen in der Medizin; Text Deutsch und Englisch

Guideline for the development of deep learning image recognition systems in medicine; Text in German and English

Guide pour le développement de systèmes de reconnaissance d'images d'apprentissage profond en médecine; Texte en allemand et anglais

Gesamtumfang 35 Seiten

Dieses Dokument wurde durch die im Vorwort genannten Verfasser erarbeitet und verabschiedet.

www.din.de
www.beuth.de
3235648

Inhalt

Seite

Vorwort

Diese DIN SPEC wurde nach dem PAS-Verfahren erarbeitet. Die Erarbeitung von DIN SPEC nach dem PAS-Verfahren erfolgt in DIN SPEC (PAS)-Konsortien und nicht zwingend unter Einbeziehung aller interessierten Kreise.

Die Erarbeitung und Verabschiedung des Dokuments erfolgten durch die nachfolgend genannten Initiatoren und Verfasser:

— MindPeak GmbH: Felix Faber, Nora Bartels, Marc Päpper und Dr. Tobias Lang

— PSIORI GmbH: Dr. Sascha Lange, Dr. Christian Kaul und Lars Eickmeier

— FUSE-AI GmbH: Matthias Steffen und Dr. Sabrina Reimers-Kipping

— Hochschule Düsseldorf: Prof. Dr.-Ing. Thomas Zielke

— Quality Match GmbH: Dr. Daniel Kondermann

— IABG mbH: Bastian Bernhardt

Für dieses Thema bestehen derzeit keine Normen im Deutschen Normenwerk.

DIN SPEC (PAS) sind nicht Teil des Deutschen Normenwerks.

Für diese DIN SPEC wurde kein Entwurf veröffentlicht.

Trotz großer Anstrengungen zur Sicherstellung der Korrektheit, Verlässlichkeit und Präzision technischer und nicht-technischer Beschreibungen kann das DIN SPEC (PAS)-Konsortium weder eine explizite noch eine implizite Gewährleistung für die Korrektheit des Dokuments übernehmen. Die Anwendung dieses Dokuments geschieht in dem Bewusstsein, dass das DIN SPEC (PAS)-Konsortium für Schäden oder Verluste jeglicher Art nicht haftbar gemacht werden kann. Die Anwendung der vorliegenden DIN SPEC (PAS) entbindet den Nutzer nicht von der Verantwortung für eigenes Handeln und geschieht damit auf eigene Gefahr.

Es wird auf die Möglichkeit hingewiesen, dass einige Elemente dieses Dokuments Patentrechte berühren können. DIN ist nicht dafür verantwortlich, einige oder alle diesbezüglichen Patentrechte zu identifizieren.

Die kostenfreie Bereitstellung dieses Dokuments als PDF-Version über den Beuth WebShop wurde im Vorfeld finanziert.

Aktuelle Informationen zu diesem Dokument können über die Internetseiten von DIN (www.din.de) durch eine Suche nach der Dokumentennummer aufgerufen werden.

Einleitung

Bilderkennung spielt in vielen Feldern der Medizin eine zentrale Rolle, wie etwa in der Radiologie und der Histopathologie. Diese Bilderkennung wird heute von menschlichen Experten geleistet. Gerade repetitive Standard-Bilderkennungsaufgaben, die eine hohe Genauigkeit und komplexe Auswertungen erfordern, nehmen dabei einen erheblichen Teil des Arbeitstages von Ärzten ein.

Automatisierte Bilderkennung mit Deep Learning bietet enormes Potenzial für Effizienz- und Qualitätsgewinne in der Medizin [1]. Deep Learning bildet den neuen State-of-the-Art der Bilderkennung und wird bereits mit Erfolg in verschiedenen Feldern von automatisiertem Fahren bis zur Gesichtserkennung angewendet [2]. Deep Learning ist eine Methode der Künstlichen Intelligenz (KI) und basiert auf tiefen künstlichen neuronalen Netzen. Anstatt mit viel Aufwand fest codierte Regeln durch einen menschlichen Experten festzulegen, werden mit einem Deep-Learning-System (DL-System) statistische Muster aus Beispieldaten gelernt. Diese Technik erlaubt es, auch in der Medizin bisher automatisiert nicht-lösbare komplexe Aufgaben in der Bilderkennung künftig mit Computerunterstützung zu lösen.

Bilderkennungssysteme mit Deep Learning können potenziell in folgenden medizinischen Feldern angewendet werden:

— Radiologie;

— Histopathologie;

— Augenheilkunde;

— Hämatologie;

— Mikrobiologie;

— Dermatologie;

— Tiermedizin;

— Pharmaindustrie;

— Forschung (z. B. Alzheimer-Krankheit);

— Chirurgie (z. B. operierende Roboter müssen auch „sehen“ können).

Bilderkennungsaufgaben in der Medizin können verschiedene Zweckbestimmungen haben:

a) diagnostisch zur Feststellung von Krankheiten;

b) prognostisch zur Vorhersage von Krankheitsverläufen;

c) therapeutisch zur Empfehlung von Therapieentscheidungen.

Dabei nimmt der Einfluss des DL-Systems auf den Patienten von diagnostisch über prognostisch zu therapeutisch zu.

Bilderkennung kann in verschiedenen Automatisierungsgraden eingesetzt werden. Bei voll automatisiertem Einsatz wird das Ergebnis nicht mehr von einem menschlichen Experten überprüft. Ein voll automatisierter Einsatz liegt auch dann vor, wenn bei einer hohen Anzahl von Auswertungen nur die mit Unsicherheiten behafteten Ergebnisse von einem menschlichen Experten überprüft werden. Alternativ kann ein DL-System unterstützend eingesetzt werden: ein menschlicher Experte überprüft jedes Ergebnis und ist für die finale Entscheidungsfindung allein verantwortlich. Diese Vorgehensweise hat insbesondere bei Screening-Programmen eine große Bedeutung. Bei Mammographie-Screening konnte gezeigt werden, dass der Einsatz eines KI-Systems den Aufwand für eine Überprüfung durch menschliche Experten um über 80 % reduzieren kann, ohne dass die Erkennungsraten verringert wurden [3]. Außerdem ist es möglich, dass ein DL-System parallel und unabhängig vom Experten im Hintergrund eine automatisierte KI-Zweitmeinung erstellt.

Die Nutzer von Bilderkennungssystemen können sowohl Ärzte als auch Patienten sein. Hämatologen können zum Beispiel bei der visuellen Bestimmung des Blutbilds durch ein DL-System unterstützt werden. Patienten können Hautveränderungen durch eine Mobiltelefon-App screenen lassen oder sich für Röntgenbilder online automatisierte Zweitmeinungen einholen.

Für den praktischen Einsatz von DL-Systemen sind verschiedene Szenarien denkbar: zur Diagnose; zur Triagierung, bei der das DL-System Fälle entsprechend der Analyse-Ergebnisse für den Menschen priorisiert; zum Ausfiltern von eindeutig negativen Fällen; für eine automatisierte Zweitmeinung, die als Sicherheitsnetz die Erstmeinung des menschlichen Experten überprüft und gegebenenfalls einen Warnhinweis gibt.

Die konkreten Bilderkennungsaufgaben lassen sich technisch in folgende Problemarten unterteilen:

a) Klassifikation von Bildern (z. B. Ist auf dem Röntgenbild eine Knochenfraktur zu erkennen?);

b) Segmentierung (z. B. Wo befindet sich Tumorgewebe?);

c) Objekterkennung (z. B. Gibt es Malaria-Erreger im Blutausstrich?);

d) Objektlokalisierung (z. B. Wo befinden sich Malaria-Erreger im Blutausstrich?).

1 Anwendungsbereich

Dieses Dokument gibt die Anforderungen an, unter denen Bilderkennungsprobleme in der Medizin mit Hilfe eines Deep–Learning-Bilderkennungssystems bearbeitet werden können. Es erlaubt Entscheidungsträgern, Kenntnisse über die Anwendungsmöglichkeiten eines Deep-Learning-Bilderkennungssystems in der Medizin und seine Struktur zu gewinnen.

Mit Hilfe dieses Dokumentes kann die Schätzung des Aufwandes und des Nutzens eines Deep-Learning-Bilderkennungssystems unterstützt werden sowie eine genauere Erfolgsprognose erstellt werden.

Dieses Dokument gibt Leitlinien zur praktischen Entwicklung eines Deep-Learning-Bilderkennungssystems in der Medizin vom Vorgehen bei der Datensammlung über die Strukturierung der Daten zum Lernen der KI-Bilderkennung bis zur Ablaufstruktur von Lern-Experimenten, insbesondere mit Rücksicht auf die erhöhten Qualitätsmaßstäbe und regulativen Vorgaben in der Medizin.

Dieses Dokument ist insbesondere für Hersteller von DL-Systemen und die Beteiligten an Forschungs- und Entwicklungsprojekten zum Einsatz von Deep-Learning-Bilderkennungssystemen in der Medizin.

Dieses Dokument legt keine spezifischen Angaben zu aktivem Lernen, mentalem Lernen, automatischem kontinuierlichem Lernen und dem bestimmungskonformen Einsatz des DL-Systems in der Praxis fest.

2 Normative Verweisungen

Die folgenden Dokumente werden im Text in solcher Weise in Bezug genommen, dass einige Teile davon oder ihr gesamter Inhalt Anforderungen des vorliegenden Dokuments darstellen. Bei datierten Verweisungen gilt nur die in Bezug genommene Ausgabe. Bei undatierten Verweisungen gilt die letzte Ausgabe des in Bezug genommenen Dokuments (einschließlich aller Änderungen).

DIN SPEC 13266:2020-04, *Leitfaden für die Entwicklung von Deep-Learning-Bilderkennungssystemen*

3 Begriffe

Für die Anwendung dieses Dokuments gelten die folgenden Begriffe.

DIN und DKE stellen terminologische Datenbanken für die Verwendung in der Normung unter den folgenden Adressen bereit:

— DIN-TERMinologieportal: verfügbar unter https://www.din.de/go/din-term

— DKE-IEV: verfügbar unter http://www.dke.de/DKE-IEV

ISO und IEC stellen terminologische Datenbanken für die Verwendung in der Normung unter den folgenden Adressen bereit:

— ISO Online Browsing Platform: verfügbar unter https://www.iso.org/obp

— IEC Electropedia: verfügbar unter http://www.electropedia.org/

3.1
Bilderkennung
Erfassung und Analyse eines Bildes, der Objekte, aus denen es besteht, sowie deren Eigenschaften und räumlichen Beziehungen durch eine Funktionseinheit

Anmerkung 1 zum Begriff: Bilderkennung schließt Szenenanalyse ein.

[QUELLE: DIN ISO/IEC 2382-28:1998-04, 28.01.14]

3.2
Bilderkennungsproblem
Problem ein bestimmtes Ziel der Bilderkennung zu lösen, z. B. ein Bild zu klassifizieren

[QUELLE: DIN SPEC 13266:2020-04, 3.2]

3.3
Deep Learning
DL
Lernen statistischer Muster durch tief verschachtelte neuronale Netze

[QUELLE: DIN SPEC 13266:2020-04, 3.4, modifiziert — Ein zusätzlicher Begriff wurde ergänzt.]

3.4
Deep-Learning-Architektur
DL-Architektur
Beschreibung der verschiedenen genutzten Recheneinheiten des neuronalen Netzes sowie deren Verbindung

Anmerkung 1 zum Begriff: Aus der Architektur ergibt sich die Anzahl der lernbaren Gewichte im Netzwerk. Je höher diese Anzahl, desto schwierigere Probleme sind lösbar; aber desto mehr Daten sind dazu notwendig.

[QUELLE: DIN SPEC 13266:2020-04, 3.5]

3.5
Deep-Learning-Bilderkennungssystem
DL-System
das zu entwickelnde Computermodell, welches das Bilderkennungsproblem löst

[QUELLE: DIN SPEC 13266:2020-04, 3.6]

3.6
Domäne
spezifischer Wissensbereich oder bereichsspezifisches Expertenwissen; genauer der Bereich, in dem das DL-Modell eingesetzt wird

[QUELLE: DIN SPEC 13266:2020-04, 3.7]

3.7
Grundwahrheiten
zu den Daten zugehörige ideale Ausgabewerte

[QUELLE: DIN SPEC 13266:2020-04, 3.8]

Anmerkung 1 zum Begriff Grundwahrheiten sind z. B. Klassifikationen (Bild 1 — Hund, Bild 2 — Katze).

3.8
Initialdaten
Gesamtmenge der Daten, die zum Start der Entwicklung des DL-Systems zur Verfügung stehen oder gesammelt werden, und die in Trainings-, Test- und Validierungsdaten aufgeteilt werden

Anmerkung 1 zum Begriff: In der Regel bestehen Initialdaten aus Bildern und den zugehörigen Grundwahrheiten/ Annotationen.

[QUELLE: DIN SPEC 13266:2020-04, 3.9]

3.9
künstliche Intelligenz
KI
Fähigkeit einer Funktionseinheit, solche Funktionen auszuführen, die im allgemeinen menschlicher Intelligenz zugeordnet werden, wie z. B. Schlussfolgern und Lernen

[QUELLE: DIN ISO/IEC 2382-28:1998-04, 28.01.02]

3.10
Modell
DL-Modell
neuronales Netz, das mit den Trainingsdaten trainiert wird, um das Bilderkennungsproblem zu lösen

[QUELLE: DIN SPEC 13266:2020-04, 3.11]

3.11
neuronales Netz
künstliches neuronales Netz
Computermodell unter Verwendung verteilter und paralleler Verarbeitung vor Ort, das aus einem Netzwerk einfacher als künstliche Neuronen bezeichneter Verarbeitungselemente, die ein komplexes umfassendes Verhalten darstellen können, besteht

[QUELLE: DIN ISO 18115-1:2017-07, 8.1, modifiziert — Begriff „ANN“ und alle Anmerkungen zum Begriff wurden nicht übernommen.]

3.12
Qualitätssicherungsdaten
Datensatz bestehend aus einer Teilmenge des Testdatensatzes sowie hinzukommender Daten während des Live Betriebs des Systems; der Datensatz kann sich mit der Zeit verändern

[QUELLE: DIN SPEC 13266:2020-04, 3.13]

3.13
Trainingsdaten
Teilmenge der Initialdaten, die dazu dient unterschiedliche Modelle zu trainieren

[QUELLE: DIN SPEC 13266:2020-04, 3.14]

3.14
Testdaten
Teilmenge der Initialdaten, mit deren Hilfe die Analysequalität des finalen Modells evaluiert werden kann

[QUELLE: DIN SPEC 13266:2020-04, 3.15, modifiziert — „das finale Modell ausgewählt“ wurde durch „die Analysequalität des finalen Modells evaluiert“ ersetzt.]

3.15
Validierungsdaten
Teilmenge der Initialdaten, die dazu dient unterschiedliche auf den Trainingsdaten trainierte Modellvarianten zu validieren

[QUELLE: DIN SPEC 13266:2020-04, 3.16]

4 Allgemeine Anforderungen

4.1 Allgemeines

DIN SPEC 13266:2020-04, Abschnitt 5 bis Abschnitt 7 müssen angewendet werden.

ANMERKUNG 1 Anwendungen in der Medizin zeichnen sich durch eine Vielzahl an Besonderheiten aus.

DL-Systeme in der Medizin können Medizinprodukte oder Teile davon sein.

ANMERKUNG 2 Aktuell gibt es weder in der EU noch in den USA allgemeingültige Gesetze und Normen, die DL-Systeme regulieren [4].

Für Medizinprodukte muss eine Zweckbestimmung festgelegt werden (z. B. diagnostisch oder prognostisch).

Die Produktentwicklung sollte auf etablierten Managementsystemstandards beruhen (z. B. DIN EN ISO 9001, DIN EN ISO/IEC 27001, DIN EN ISO/IEC 27701, DIN EN ISO 13485); allerdings decken solche Standards, soweit sie zur Zeit veröffentlicht sind, lediglich Teilbereiche der Entwicklung und Nutzung von KI ab (Qualität, Sicherheit, Datenschutz, usw.).

Das Risikomanagement sollte nach DIN EN ISO 14971 erfolgen und muss sowohl Risiken in der Produktentwicklung als auch Risiken hinsichtlich falscher Analyseergebnisse beachten. Die vorgesehenen Nutzer sowie der genaue Anwendungskontext müssen bestimmt werden. Es muss festgelegt werden, für die Bilder welcher Hardware das DL-System ausgelegt ist und bei welchen Eingabebildern keine Erkennungsgüte des DL-Systems garantiert werden kann.

4.2 Qualitätsanforderungen

Die Qualitätsanforderungen sind in der Medizin besonders hoch, da das Wohl von Patienten sichergestellt werden muss. Falsche Vorhersagen wie zum Beispiel Falsch-Negative bei der Diagnose von Krebsfällen haben außerordentlich hohe Kosten.

Um Nutzerakzeptanz zu finden, muss die Erkennungsgüte mindestens auf dem Niveau menschlicher Experten sein. Diese Erkennungsgüte muss in klinischen Studien validiert worden sein (in retrospekiven oder prospektiven Studien) [5].

Die Analyseergebnisse von DL-Systemen müssen reproduzierbar sein.

DL-Systeme müssen verlässlich sein. Die Systeme dürfen im Klinikalltag nicht ausfallen, da sonst der darauf angepasste klinische Arbeitsablauf (en: workflow) bedroht sein kann.

Die Nutzerakzeptanz ist in der Medizin besonders entscheidend; dazu muss die KI-Bilderkennung sich nahtlos in medizinische Arbeitsabläufe einbinden. Außerdem müssen die Vorhersagen nachvollziehbar und erklärbar sein, damit sich menschliche Experten auf die Ergebnisse verlassen können.

Das Qualitätsmanagement sowie Prozessprüfungen sollten auf etablierten Managementsystemstandards beruhen (z. B. DIN EN ISO 9001, DIN EN ISO/IEC 27001, DIN EN ISO/IEC 27701, DIN EN ISO 13485).

4.3 Repräsentativer Datensatz zur Abbildung des Problems

Eine große Herausforderung in der Bilderkennung in der Medizin ist die Heterogenität der Bilddaten. Bilddaten werden mit Geräten und Softwares verschiedener Hersteller erstellt.

ANMERKUNG 1 Beispielsweise gibt es in der Pathologie verschiedene Scanning-Systeme für pathologische Proben von verschiedenen Herstellern. Die verschiedenen Scanning-Systeme und Softwares führen zu verschiedenartigen Bildcharakteristiken, etwa in Kontrast und Schärfe. Selbst bei Systemen desselben Modelltyps gibt es häufig Varianzen. Weiterhin gibt je nach Krankenhaus oder Labor unterschiedliche Bedingungen, unter denen die Bilder entstehen:

beispielsweise können die Lichtverhältnisse variieren oder die chemischen Färbemittel bei pathologischen Proben haben unterschiedliche Zusammensetzungen.

Krankenhäuser und Labore haben oft unterschiedliche Zusammensetzungen von Patientengruppen, die sie versorgen. Labore können auf bestimmte Krankheiten spezialisiert sein. Dies führt zu unterschiedlichen Datenverteilungen an Krankheitsbildern. Bei der Datensatzerstellung muss berücksichtigt werden, dass durch die Laborauswahl kein Bias in die Datenerhebung kommt.

ANMERKUNG 2 Allgemein spielt in der Medizin die Heterogenität der Patienten eine Rolle, so dass die resultierenden Bilder für die Bilderkennung häufig starke individuelle Varianzen aufweisen.

Bei der DL-Entwicklung muss analysiert werden, welchen Einfluss Hardware- und Software-Beschaffenheiten auf die Datengüte und die Analysegüte haben.

Der Hersteller des DL-Produkts zur Bildverarbeitung muss festlegen, für welche Typen von Eingabebildern und Bildquellen Vorhersagen gemacht werden können. Für eine große Anwendungsbreite ist in der Regel ein entsprechend breiter und repräsentativer Trainingsdatensatz notwendig.

4.4 Datensammlung

Die Datengewinnung für die KI-Systementwicklung ist in der Medizin eine besonders große Herausforderung, da hierfür medizinisches Fachwissen und hochqualitative Daten notwendig sind und sich medizinische Daten durch eine große Heterogenität auszeichnen. Die Qualität der gewonnenen Daten ist von herausragender Bedeutung, da nur mit solchen Daten eine hohe Erkennungsleistung erreicht werden kann.

Zur Erzeugung eines geeigneten Datensatzes muss sichergestellt werden, dass die Auswahl der Daten repräsentativ für die Anwendung sind, dass die Qualität sowohl der Daten als auch der Annotationen hinreichend sind, und dass die gewählten Annotationsregeln möglichst eindeutig sind. Eine detaillierte Diskussion des Themas Datensatzqualität befindet sich in DIN SPEC 13266.

Ein DL-System wird typischerweise mit überwachtem Lernen trainiert. Auch andere DL-Trainingsmethoden sind anwendbar.

ANMERKUNG 1 Es ist möglich, durch den zusätzlichen Einsatz von unüberwachtem Lernen geeignete Feature-Repräsentationen für das überwachte Lernen zu lernen.

Das Training mit überwachtem Lernen erfordert Annotationen (Labels) der Daten. Im Gegensatz zu anderen Domänen müssen Annotationen in der Medizin mit Expertenwissen erstellt werden.

ANMERKUNG 2 Somit ist es schwierig und teuer, Daten zu erhalten, da hierfür eine Vielzahl an Spezialisten erforderlich ist.

Die Kompetenz von Annotatoren muss durch Eignungstests sichergestellt werden. Dies kann mittels eines vorgehaltenen Testdatensatzes geschehen, für den von Experten die Grundwahrheit eindeutig bestimmt wurde.

Der Annotationsprozess muss genau definiert und überwacht werden, um eine konsistente Güte der Daten zu erreichen. Dazu müssen Annotatoren entsprechend geschult werden. Die Arbeit der Annotatoren muss konstant überwacht werden, um die Qualität zu sichern. Dies kann beispielsweise dadurch geschehen, dass in den Annotationsprozess bereits bekannte Bilder eingestreut werden, für die man die Analyse des Annotators mit der bereits bekannten Grundwahrheit vergleicht. Für eine hohe Qualität ist es wichtig, dass die Incentivierung der Annotatoren nicht falsche Anreize setzt.

BEISPIEL Ein falscher Anreiz kann eine Vergütung pro abgeschlossenem Fall sein.

Eine weitere Schwierigkeit in der Medizin ist, dass häufig eine eindeutige Grundwahrheit fehlt. Die Datensammlung muss solche Ungenauigkeiten abbilden.

ANMERKUNG 3 Beispielsweise können in der Histopathologie oftmals Zellen nicht mit eindeutiger Gewissheit klassifiziert werden, da diese Information in den Bilddaten fehlt.

Um eine hohe Datenqualität sicherzustellen, ist es häufig notwendig, mehrere Annotatoren dieselben Daten annotieren zu lassen. Es sollten dann nur solche Daten zum Trainieren und Testen verwendet werden, bei denen die Annotatoren übereinstimmen.

Für die Definition des Annotationsprozesses müssen Abschätzungen zur Inter- und Intraobserver-Variabilität gewonnen werden. Diese sagen aus, wie sehr die Meinungen verschiedener Annotatoren (inter) bzw. wie sehr die Meinung desselben Annotators zu verschiedenen Zeitpunkten (intra) abweicht. Damit erhält man eine Abschätzung der Schwierigkeit des Annotationsproblems und damit der Analyseaufgabe. Außerdem kann man abschätzen, wie viele Experten man zur Annotation braucht.

Der Annotationsprozess inklusive des Qualitätssicherungsprozesses muss detailliert dokumentiert werden. Eine Dokumentation muss insbesondere enthalten:

a) Informationen zu Ausbildung, relevanten Qualifikationen und Hintergrundwissen des Annotators;

b) Informationen zu Datenschutzmaßnahmen und IT-Sicherheit;

c) angewendete Incentivierungsmaßnahmen;

d) Instruktionen für die Annotatoren in der Form von mindestens Tool- und Annotations-Instruktionen);

e) Dokumentation des Annotationstools (z. B. Screenshots, Bildschirmaufnahmen);

f) Methodik zur Überprüfung der Korrektheit der Annotationen (z. B. Clustering, Wiederholungen);

g) Methoden zur Qualitätssicherung, z. B. Methode der Aggregation mehrfach wiederholter Annotationen durch unterschiedliche Annotatoren und Überwachung der Annotatoren.

Datenschutz und Anonymisierung bzw. Pseudonymisierung spielen in der Medizin eine herausragende Rolle. Hierfür muss definiert und dokumentiert werden, ob für das DL-System Meta-Daten in Form persönlicher Patientendaten vonnöten sind.

Bei der Datensammlung muss definiert werden, für welche Eingabedaten (Formate, Bildgrößen, Farbkodierungen usw.) das DL-System ausgelegt sein muss.

Bei der Datensammlung muss darauf geachtet werden, einen systematischen Bias zu vermeiden sowie so genannte Datenleckagen auszuschließen, bei denen korrelierte, aber nicht kausale Daten vom DL-System fälschlicherweise als Grundlage der Bildanalyse gelernt werden (z. B. Bilder mit der Klasse “Krebs”, bei denen zusätzlich immer ein Lineal zu sehen ist).

4.5 Einbettung in den medizinischen Arbeitsablauf

Veränderungen in medizinischen Arbeitsprozessen sind besonders aufwendig. Vor der Entwicklung eines DL-Systems muss geklärt werden, ob die Veränderungen zum Einsatz eines DL-Systems und die Eingliederung in klinische Prozesse realistisch sind.

Die Resultate der KI-Bilderkennung müssen für die Anwender leicht verständlich sein. Es dürfen keine Unklarheiten entstehen, wie mit den Ergebnissen umgegangen wird.

ANMERKUNG Beispielsweise führen Wahrscheinlichkeitsangaben etwa in Form von Wahrscheinlichkeits-Heatmaps häufig zu Unklarheiten für Anwender.

Der Anbieter des DL-Systems muss sicherstellen, dass das System konstant verfügbar im medizinischen Arbeitsablauf ist.

4.6 Methodik

Auch auf der methodischen Seite gibt es Besonderheiten für DL-Systeme in der Medizin. Datenaugmentierungen werden häufig zu einer Verbreiterung des Lerndatensatzes eingesetzt, müssen in der Medizin jedoch mit Bedacht eingesetzt werden, damit keine unrealistischen Daten entstehen.

ANMERKUNG 1 Die Bilddaten sind häufig nicht standardisiert. Bildgrößen, Schärfe/n und Farbkontraste variieren stärker als in den typischen Datensätzen der allgemeinen DL-Forschung.

ANMERKUNG 2 Die Bilddaten sind häufig sehr groß, z. B. sind pathologische Bilder häufig 10 000-mal größer als Bilder etwa in der Gesichtserkennung und in der Produktsuche.

Ein DL-System muss mit Varianzen (siehe ANMERKUNG 1 und ANMERKUNG 2) sowohl im Training als auch in der Inferenz im praktischen Einsatz umgehen.

4.7 Erklärbarkeit und Nutzervertrauen

Ärzte und Patienten müssen der DL-Bildanalyse vertrauen können, um sie anzuwenden. Neuronale Netze sind im Gegensatz zu anderen Verfahren nicht transparent. Es gibt jedoch zunehmend mehr Ansätze zur Erklärbarkeit von DL-Modellen, die dabei helfen können, Anwendervertrauen zu gewinnen [6].

ANMERKUNG Die Datenschutz-Grundverordnung ermöglicht es Anwendern, Informationen über Entscheidungen unterliegender KI-Logiken zu erhalten.

Es gilt zwei Arten von Erklärbarkeiten zu unterscheiden:

1) Interpretation von Modellen; sowie

2) die Erklärung konkreter Einzelvorhersagen.

Typischerweise wird heute 2) gemeint, wenn von Erklärbarkeit gesprochen wird.

ANMERKUNG Die Forschung zur Interpretation von Modellen in DL-Systemen steckt noch in den Anfängen, auch wenn es interessante erste Ansätze gibt. Bei der Erklärung konkreter Einzelvorhersagen von DL-Systemen gibt es vielversprechende Ansätze in der aktuellen Forschungsliteratur, wie die Vorhersagen von DL-Modellen plausibilisiert werden können. Zum Beispiel kann man bei der Klassifikation von Bildern nachvollziehen, welche Input-Pixel für die Klassifikation entscheidend sind. Es haben sich jedoch noch keine Standard-Methoden etabliert, da die derzeitigen Verfahren unterschiedliche Stärken und Schwächen haben und sich der aktuelle Zustand in einer heuristischen Phase befindet. Es ist jedoch davon auszugehen, dass die Forschung in diesem Bereich in den nächsten Jahren weitere Fortschritte Richtung Erklärbarkeit machen wird.

Es gibt verschiedene Abstufungen der Erklärbarkeit:

a) Wenn wichtige Pixel zur Vorhersage hervorgehoben werden, bleibt mehr geistige Leistung beim Anwender zu bestimmen, ob dies Sinn macht.

b) Alternativ kann das DL-System relevante Zwischenstrukturen bestimmen und visualisieren, beispielsweise Zellen im Kontext der Bestimmung von Tumor, die zeigen, dass das DL-System entsprechend relevante Objekte gefunden hat.

Die Ergebnisse sollten visualisiert werden, um Anwendern die Ergebnisse anschaulicher zu vermitteln und ihnen die Möglichkeit zu geben, die Güte der Ergebnisse einschätzen zu können.

Für das Nutzervertrauen ist die Durchführung von Validierungs- und Verifikationsprozessen in der Software- und Produktentwicklung notwendig. Typischerweise können große Teile der Anwendersoftware nach konventionellen Standards getestet werden, z. B. ISO/IEC/IEEE 29119-1. Für KI-Softwarekomponenten existieren entsprechende Standards noch nicht. Die Durchführung von klinischen Studien ist deshalb die

wichtigste Voraussetzung für Nutzervertrauen. Die Ergebnisse von klinischen Studien sowie initiale Erkennungsgütemetriken und -spezifikationen müssen angegeben werden.

DL-Systeme sind prinzipiell anfällig für Fehlfunktionen, die von sogenannten „adversarial examples“ ausgelöst werden [7]. „Adversarial examples“ sind Bildbeispiele, die eine hohe Ähnlichkeit mit bedeutsamen Bildmustern aufweisen, dennoch gänzlich falsch interpretiert werden. Solche Bildbeispiele verringern nicht nur die Erkennungsgüte, sie können auch ein Sicherheitsproblem darstellen. Das Problem kann die Robustheit eines DL-Systems generell in Frage stellen [8], und damit auch das Nutzervertrauen beeinträchtigen. Bei der Entwicklung von DL-Systemen müssen Maßnahmen ergriffen werden um die Wahrscheinlichkeit zu minimieren, dass „adversarial examples“ ein grobes Fehlverhaltens des Systems hervorrufen können.

5 Technische Umsetzung

5.1 Allgemeines

Die technische Umsetzung ist zusätzlich zu den Anforderungen in diesem Dokument nach DIN SPEC 13266:2020-04, 8.2 bis 8.9 und Anhang A durchzuführen. Alle Arbeitsschritte müssen dokumentiert werden. Dies gilt insbesondere, falls das DL-System in einem Produkt für die praktische Anwendung vorgesehen ist und eine entsprechende Zertifizierung angestrebt wird. Dazu muss die Datengewinnung dokumentiert werden (Quelle, Annotator/en). Der Trainings- und Validierungsprozess muss beschrieben werden. In einer analytischen Studie muss gezeigt werden, dass das entwickelte DL-System ausreichende Erkennungsgüte besitzt, um im nächsten Schritt eine klinische Studie durchzuführen. Dazu muss das Bilderkennungsmodell auf Testdaten evaluiert werden. Die Auswahl der Testdaten, die Metriken und die Auswertung müssen dokumentiert werden.

5.2 Formalisierung und Problemdefinition

Es gilt DIN SPEC 13266:2020-04, 8.2.

Vor dem Beginn der technischen Umsetzung muss formalisiert werden, was genau die DL-Bilderkennung leisten muss. Es muss ein eindeutiger Satz an Fehlermaßen definiert werden und es muss definiert werden, was die Mindestanforderungen in der Analysegüte sind.

5.3 Datenakquisitionsprozess und Erstellung des Datensatzes

Es gilt DIN SPEC 13266:2020-04, 8.3.

Die Daten-Annotation sollte von einer Expertengruppe statt einzelner Experten gemacht werden. Eine Herausforderung hierbei ist, festzulegen, was ein Experte ist. Dies können sowohl alle Ärzte sein als auch die Top-Experten auf dem Gebiet. Die Annotatoren arbeiten mit bereitgestellten Tools, auf deren Anwendung sie zunächst geschult werden müssen.

Manche Annotationsaufgaben sind zeitaufwändig und repetitiv. Solche Annotationen können nach kurzer Lernzeit oft auch von Laien angefertigt werden. So können Kosten und Zeit gespart werden, da Laien in der Regel leichter für diese Aufgabe zu finden sind. Beim Einsatz von Laien muss immer geprüft werden, ob die Qualität der Annotationen vergleichbar mit jenen der Expertengruppe sind. Darüber hinaus müssen Laien-annotationen immer von Experten validiert werden, bevor sie zum Training des DL-Systems genutzt werden.

Entsprechend der Definition des Anwendungskontexts muss ein repräsentativer Datensatz für Training und Validierung gesammelt werden.

Bei der Sammlung der Daten sollten Inter- und Intraobservervarianz gemessen werden.

Bei der Datensammlung müssen seltene Klassen gesondert berücksichtigt werden. Diese sind in der Medizin typisch. Häufig ist ein Großteil der Patienten unauffällig und es gibt seltene Krankheiten.

Für die Datensammlung ist es nötig, dass alle relevanten Tools validiert sind.

Der Umfang der Daten hängt ab von der konkreten Bilderkennungsaufgabe.

Ein wichtiger Schritt ist die Definition eines Testdatensatzes. Dieser darf nicht mit den Beispielen in Trainings- und Validierungsdaten überlappen. Der Testdatensatz dient dazu, am Ende des gesamten Entwicklungsprozesses die Vorhersagequalität des finalen Modells abzuschätzen.

5.4 Software-Entwicklung

In der Software-Entwicklung müssen die besten Praktiken in der Code-Entwicklung in der Medizin berücksichtigt werden (Unit Tests, Integration Tests, usw.), beispielsweise DIN EN 62304 für medizinische Geräte-Software.

5.5 Trainieren des DL-Systems

Das Trainieren des DL-Systems in der Medizin muss wie folgt ablaufen:

1) Dokumentierte Strukturierung der Daten zum Trainieren und Testen des DL-Systems nach DIN SPEC 13266:2020-04, 8.4;

2) Identifikation von potenziellen DL-Architekturen nach DIN SPEC 13266:2020-04, 8.5;

3) Festlegen der Ablaufstruktur von Lern-Experimenten nach DIN SPEC 13266:2020-04, 8.6;

4) Modelltraining, -validierung und -selektion ausschließlich mit Trainings- und Validierungsdaten, ggf. mit Kreuzvalidierung;

5) Modellselektion nach DIN SPEC 13266:2020-04, 8.7 und nach Hyperparameteroptimierung;

6) Festsetzen einfacher Baselines-Modelle;

7) Evaluierung des finalen DL-Modells und der Baselines auf dem Testdatensatz;

8) Begründung warum einfache und leichter interpretierbare Modelle nicht verwendet werden können;

9) Überprüfung der Reproduzierbarkeit der Testergebnisse;

10) Überprüfung der Reproduzierbarkeit des DL-Modells;

11) Versionsverwaltung von Daten.

5.6 Modell-Deployment

Es gilt DIN SPEC 13266:2020-04, 8.8.

Für das Deployment des DL-Systems ist die nahtlose Eingliederung in das klinische Umfeld notwendig. Dazu muss sichergestellt werden, dass die notwendige Performance und Effizienz auf der Zielhardware erreicht wird. Im Live-Betrieb muss die Qualität des DL-Systems durch kontinuierliche umfangreiche Tests im Rahmen einer Post-Market-Überwachung und Qualitätssicherung sichergestellt werden.

In der Praxis muss validiert und verifiziert werden, ob Nutzer die DL-Bilderkennung im Sinne der medizinischen Zweckbestimmung nutzen. Es muss somit validiert werden, ob Nutzer die Ergebnisse verstehen und ob Nutzer dem System blind vertrauen und ob dies der Zweckbestimmung gerecht wird.

5.7 Qualitätssicherung des Deep-Learning-Bilderkennungssystems

Es gilt DIN SPEC 13266:2020-04, 8.9.

5.8 Kontinuierliche Verbesserung und Nachlernen

Eine Stärke eines DL-Systems zur Bilderkennung ist das kontinuierliche Lernen. Das DL-System kann anhand der Bilder, die es verarbeitet, weiterlernen. Dies ermöglicht einerseits die kontinuierliche Verbesserung sowie andererseits eine Anpassung, falls sich die Verteilung über die Zeit ändert. Es kann jedoch auch dazu führen, dass sich das DL-System verschlechtert mit entsprechenden Konsequenzen für Patienten. Kontinuierliches Lernen, welches automatisch Veränderungen an einem medizinischen DL-System durchführt, schließt sich deshalb nach dem heutigen Stand der Technik aus. Aber auch der Einsatz von kontinuierlichem Lernen im Sinne von regelmäßigem kontrollierten Nachlernen durch den Hersteller oder den Anwender bringt erhebliche Komplexitäten mit sich, in Technik [9] als auch Regulierung [10]. Medizinische Qualitätssicherung und -management müssen dabei vom DL-Hersteller sichergestellt werden.

Es muss jederzeit sichergestellt werden, dass sich das DL-System nicht verschlechtert. Dafür ist die kontinuierliche Validierung auf einem Referenzdatensatz, den sogenannten Qualitätssicherungsdaten, notwendig. Vor dem Deployment des DL-Modells sollten Grenzen definiert werden, innerhalb derer sich Modelle verändern dürfen. Aktualisierte Modelle dürfen nur in den praktischen Einsatz gebracht werden, wenn kontinuierliche Tests auf den Qualitätssicherungsdaten durchgeführt werden. Es muss die Möglichkeit bestehen, zu früheren Versionen zurück zu wechseln. Bei Updates der KI müssen alle Nutzer informiert werden.

ANMERKUNG Für kontinuierliches Lernen gibt es bisher weder in der EU noch in den USA eindeutige Regulierungen. Auf regulatorischer Seite hat sich die FDA (en: Food and Drug Administration) in den USA jedoch mit einem Diskussionspapier positioniert [12]. Danach legen DL-Hersteller in Form eines „Algorithm Change Protocol“ (ACP) vorab fest, in welchem Umfang sich DL-Modelle ändern können; des Weiteren legen sie eindeutige Änderungen und Weiterlernprozesse fest.

Literaturhinweise

[1] ISO/IEC/IEEE 29119-1, *Software and systems engineering — Software testing — Part 1: Concepts and definitions*

[2] Litjens, G., Kooi, T., Bejnordi, B. et al. (2017), A survey on deep learning in medical image analysis. Medical Image Analysis, 42

[3] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature 521, p. 436–444 doi:10.1038/nature14539

[4] McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. C., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., Ledsam, J. R., Melnick, D., Mostofi, H., Peng, L., Reicher, J. J., Romera-Paredes, B., Sidebottom, R., Suleyman, M., Tse, D., Young, K. C., Fauw, J. D., Shetty, S. (2020): International evaluation of an AI system for breast cancer screening. Nature. 577, 7788, p. 89–94. [1]

[5] Johner et al.: Leitfaden zur KI bei Medizinprodukten.[2]

[6] Santos, I.C., Gazelle, G.S., Rocha, L.A., Tavares, J.M.R.: Medical device specificities: opportunities for a dedicated product development methodology. Expert, Review of Medical Devices 9(3), 299–311 (May 2012)

[7] Christoph Molnar: Interpretable Machine Learning.[3]

[8] Xu, H., Ma, Y., Liu, H.-C., Deb, D., Liu, H., Tang, J.-L., Jain, A. K. (2020): Adversarial Attacks and Defenses in Images, Graphs and Text: A Review. International Journal of Automation and Computing. 17, 2, p. 151–178.

[9] Dietterich, T. G. (2017): Steps Toward Robust Artificial Intelligence. AI Magazine. 38, 3, p. 3–24

[10] Parisi, G. I., Kemker, R., Part, J. L. (2019), Kanan, C., Wermter, S.: Continual lifelong learning with neural networks: A review. Neural Networks. 113, p. 54–71. [4]

[11] Perspectives and Good Practices for AI and Continuously Learning Systems in Healthcare. [5]

[12] Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD), FDA, April 2019

---

[1] Verfügbar unter: https://www.nature.com/articles/s41586-019-1799-6

[2] Verfügbar unter: https://github.com/johner-institut/ai-guideline/blob/master/Guideline-AI-Medical-Devices_DE.md

[3] Verfügbar unter: https://christophm.github.io/interpretable-ml-book/cnn-features.html

[4] Verfügbar unter: https://doi.org/10.1016/j.neunet.2019.01.012

[5] Verfügbar unter: https://static1.squarespace.com/static/58d0113a3e00bef537b02b70/t/5cf13dec8f88ee0001f7c5c7/1559313901740/AI_WhitePaper_GoodPractices.pdf

[13] DIN EN ISO 9001, *Qualitätsmanagementsysteme — Anforderungen*

[14] DIN EN ISO 13485, *Medizinprodukte — Qualitätsmanagementsysteme — Anforderungen für regulatorische Zwecke*

[15] DIN EN ISO 14971, *Medizinprodukte — Anwendung des Risikomanagements auf Medizinprodukte*

[16] DIN EN ISO/IEC 27001, *Informationstechnik — Sicherheitsverfahren — Informationssicherheitsmanagementsysteme — Anforderungen*

[17] DIN ISO 18115-1:2017-07, *Chemische Oberflächenanalyse — Vokabular — Teil 1: Allgemeine Begriffe und Begriffe für die Spektroskopie (ISO 18115-1:2013)*

[18] DIN ISO/IEC 2382-28:1998-04, *Informationstechnik — Begriffe — Teil 28: Künstliche Intelligenz — Grundbegriffe und Expertensysteme (ISO/IEC 2382-28:1995)*

[19] DIN EN 62304 (VDE 0750-101), *Medizingeräte-Software — Software-Lebenszyklus-Prozesse*

[20] DIN EN ISO/IEC 27701, *Sicherheitstechniken - Erweiterung zu ISO/IEC 27001 und ISO/IEC 27002 für das Management von Informationen zum Datenschutz - Anforderungen und Richtlinien*

— Leerseite —

March 2021

DIN SPEC 13288

Titel en:	Guideline for the development of deep learning image recognition systems in medicine; Text in German and English
Titel de:	Leitfaden für die Entwicklung von Deep-Learning-Bilderkennungssystemen in der Medizin; Text Deutsch und Englisch
Titel fr:	Guide pour le développement de systèmes de reconnaissance d'images d'apprentissage profond en médecine; Texte en allemand et anglais

Contents

Foreword

This DIN SPEC has been developed according to the PAS procedure. The development of a DIN SPEC according to the PAS procedure is carried out in DIN SPEC (PAS)-consortiums and does not require the participation of all stakeholders.

This document has been developed and adopted by the initiators and authors named below:

— MindPeak GmbH: Felix Faber, Nora Bartels, Marc Päpper and Dr. Tobias Lang

— PSIORI GmbH: Dr. Sascha Lange, Dr. Christian Kaul and Lars Eickmeier

— FUSE-AI GmbH: Matthias Steffen and Dr. Sabrina Reimers-Kipping

— Hochschule Düsseldorf: Prof. Dr.-Ing. Thomas Zielke

— Quality Match GmbH: Dr. Daniel Kondermann

— IABG mbH: Bastian Bernhardt

At present, there are no standards covering this topic in the body of German Standards.

DIN SPEC (PAS)s are not part of the body of German Standards.

A draft of this DIN SPEC (PAS) has not been published.

Despite great efforts to ensure the accuracy, reliability and precision of technical and non-technical information, the DIN SPEC (PAS)-consortium cannot give any explicit or implicit assurance or warranty in respect of the accuracy of the document. Users of this document are hereby made aware that the consortium cannot be held liable for any damage or loss. The application of this DIN SPEC (PAS) does not release users from the responsibility for their own actions and is applied at their own risk.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. DIN shall not be held responsible for identifying any or all such patent rights.

Provision of this document free of charge as a PDF via the Beuth WebShop has been financed in advance.

For current information on this document, please go to DIN's website (www.din.de) and search for the document number in question.

Introduction

Image recognition plays a central role in many fields of medicine, such as radiology and histopathology. Today, this image recognition is performed by human experts. Especially repetitive standard image recognition tasks, which require high accuracy and complex evaluations, take up a considerable part of the working day of physicians.

Automated image recognition with Deep Learning offers enormous potential for efficiency and quality gains in medicine [1]. Deep Learning is the new state-of-the-art in image recognition and is already being used successfully in various fields from autonomous driving to face recognition [2]. Deep Learning is an artificial intelligence (AI) method based on deep artificial neural networks. Instead of a human expert setting hard coded rules, a Deep Learning (DL) system is used to learn statistical patterns from sample data. This technique allows one to solve complex tasks in image recognition with computer support in the future, which did not have an automatic solution in medicine so far.

Image recognition systems with Deep Learning can potentially be applied in the following medical fields:

— radiology;

— histopathology;

— ophthalmology;

— hematology;

— microbiology;

— dermatology;

— veterinary medicine;

— pharma industry;

— research (e.g. Alzheimer's disease);

— surgery (e.g. operating robots need to be able to “see” as well).

Image recognition tasks in medicine can have different purposes:

a) diagnostic to detect diseases;

b) prognostic for the prediction of disease progression;

c) therapeutic for the recommendation of therapy decisions.

The impact of the DL system on the patient increases from diagnostic to prognostic to therapeutic.

Image recognition can be used in different degrees of automatization. When fully automated, the result is no longer checked by a human expert. A fully automated application also exists if, in the case of a large number of evaluations, only the results with uncertainties are checked by a human expert. Alternatively, a DL system can be used as a support: a human expert checks each result and is solely responsible for the final decision. This approach is particularly important for screening programs. In mammography screening, it has been shown that the use of an AI system can reduce the effort required for human expert review by more than 80 % without reducing detection rates [3]. In addition, it is possible for a DL system to generate an automated AI second opinion in parallel and independently of the expert in the background.

The users of image recognition systems can be both physicians and patients. Hematologists, for example, can be supported in the visual determination of blood counts by a DL system. Patients can have skin lesions screened by a cell phone app or obtain automated second opinions online for X-rays.

For the practical use of DL systems, various scenarios are conceivable: for diagnosis; for triage, in which the DL system prioritizes cases for the human according to the analysis results; for filtering out clearly negative cases; for an automated second opinion, which as a safety net checks the first opinion of the human expert and, if necessary, issues a warning.

The concrete image recognition tasks can be divided technically into the following types of problems:

a) Classification of images (e.g. is there a bone fracture on the X-ray image?);

b) Segmentation (e.g. Where is tumor tissue located?);

c) Object recognition (e.g. Are there malaria pathogen in the blood smear?);

d) Object localization (e.g. Where are malaria pathogen in the blood smear?).

1 Scope

This document gives the requirements, under which image recognition problems in medicine can be addressed using a Deep Learning image recognition system. It allows decision makers to gain knowledge about the application possibilities of a Deep Learning image recognition system in medicine and its structure.

With the help of this document, the estimation of the effort and benefit of a Deep Learning image recognition system can be supported and a more accurate forecast of success can be made.

This document gives guidelines for the practical development of a Deep Learning image recognition system in medicine, from the procedure of data collection to the structuring of data for learning AI image recognition and the procedural structure of learning experiments, especially with regard to the increased quality standards and regulatory requirements in medicine.

This document is particularly for producers of DL-systems and those participating in research and development projects for the application of Deep Learning image recognition systems in medicine.

This document does not specify specific information about active learning, mental learning, automatic continuous learning and the intended use of the DL-System in practice.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

DIN SPEC 13266:2020-04, *Guidelines for the development of deep learning image recognition systems*

3 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

DIN and DKE maintain terminological databases for use in standardization at the following addresses:

— DIN-TERMinologieportal: available at https://www.din.de/go/din-term

— DKE-IEV: available at http://www.dke.de/DKE-IEV

ISO and IEC maintain terminological databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at http://www.electropedia.org/

3.1
image recognition
perception and analysis, by a functional unit, of an image, its constituent objects, their properties, and their spatial relationships

Note 1 to entry: Image recognition includes scene analysis.

[SOURCE: DIN ISO/IEC 2382-28:1998-04, 28.01.14]

3.2
image recognition problem
problem to solve a specific goal of image recognition, e.g. to classify an image

[SOURCE: DIN SPEC 13266:2020-04, 3.2]

3.3
Deep Learning
DL
learning statistical patterns through deeply nested neural networks

[SOURCE: DIN SPEC 13266:2020-04, 3.4, modified — An additional term was added.]

3.4
Deep-Learning Architecture
DL Architecture
description of the different computing units used in the neural network and their connection

Note 1 to entry: The architecture determines the number of learnable weights in the network. The higher this number, the more difficult problems can be solved; but the more data is needed to do so.

[SOURCE: DIN SPEC 13266:2020-04, 3.5]

3.5
Deep Learning image recognition system
DL System
computer model to be developed to solve the image recognition problem

[SOURCE: DIN SPEC 13266:2020-04, 3.6]

3.6
domain
specific field of knowledge or expertise; more precisely, the area in which the DL model is used

[SOURCE: DIN SPEC 13266:2020-04, 3.7]

3.7
ground truths
ideal output values associated with the data

[SOURCE: DIN SPEC 13266:2020-04, 3.8]

Note 1 to entry: Ground truths are, e.g., classifications (image 1 — dog, image 2 — cat).

3.8
initial data
total amount of data available or collected at the start of DL system development, divided into training, test and validation data

Note 1 to entry: Usually initial data consist of images and the associated ground truths/annotations.

[SOURCE: DIN SPEC 13266:2020-04, 3.9]

3.9
artificial intelligence
AI
capability of a functional unit to perform functions that are generally associated with human intelligence such as reasoning and learning

[SOURCE: DIN ISO/IEC 2382-28:1998-04, 28.01.02]

3.10
model
DL model
neural network, which is trained with the training data to solve the image recognition problem

[SOURCE: DIN SPEC 13266:2020-04, 3.11]

3.11
neural network
artificial neural network
computational model utilizing distributed, parallel local processing, and consisting of a network of simple processing elements called artificial neurons, which can exhibit complex global behaviour

[SOURCE: DIN ISO 18115-1:2017-07, 8.1, modified — Term “ANN” and all notes were not taken over.]

3.12
quality assurance data
data set consisting of a subset of the test data set and data that are added during live operation of the system; the data set can change over time

[SOURCE: DIN SPEC 13266:2020-04, 3.13]

3.13
training data
subset of the initial data, which serves to train different models

[SOURCE: DIN SPEC 13266:2020-04, 3.14]

3.14
test data
subset of the initial data, which can be used to evaluate the analysis quality of the final model

[SOURCE: DIN SPEC 13266:2020-04, 3.15, modified — “the final model selected” was replaced by “the analysis quality of the final model evaluated”].

3.15
validation data
subset of the initial data used to validate different model variants trained on the training data

[SOURCE: DIN SPEC 13266:2020-04, 3.16]

4 General requirements

4.1 General

DIN SPEC 13266:2020-04, Clause 5 to Clause 7 shall be followed.

NOTE 1 Applications in medicine are characterized by a multitude of special features.

DL systems in medicine can be medical devices or parts thereof.

NOTE 2 Currently, there are no generally applicable laws and standards in the EU or the USA that regulate DL systems [4].

For medical devices a purpose shall be defined (e.g. diagnostic or prognostic).

Product development should be based on established management system standards (e.g. DIN EN ISO 9001, DIN EN ISO/IEC 27001, DIN EN ISO/IEC 27701, DIN EN ISO 13485); however, such standards, as far as they are currently published, only cover parts of the development and use of AI (quality, security, privacy, etc.).

Risk management should be carried out in accordance with DIN EN ISO 14971 and shall take into account both risks in product development and risks regarding incorrect analysis results. The intended users as well as the exact application context shall be determined. It shall be determined for which hardware the DL system is designed and for which input images no sufficient recognition quality of the DL system can be guaranteed.

4.2 Quality requirements

The quality requirements are particularly high in medicine, as the well-being of patients shall be ensured. False predictions, such as false negatives in the diagnosis of cancer cases, have extraordinarily high costs.

To find user acceptance, the recognition quality shall be at least on the level of human experts. This recognition quality shall have been validated in clinical studies (in retrospective or prospective studies) [5].

The analysis results of DL systems shall be reproducible.

DL systems shall be reliable. The systems shall not fail in the daily routine of a hospital, as otherwise the clinical workflow adapted to it can be threatened.

User acceptance is particularly crucial in medicine; to achieve this, AI image recognition shall be seamlessly integrated into medical workflows. In addition, the predictions shall be comprehensible and explainable so that human experts can rely on the results.

Quality management and process audits should be based on established management system standards (e.g. DIN EN ISO 9001, DIN EN ISO/IEC 27001, DIN EN ISO/IEC 27701, DIN EN ISO 13485).

4.3 Representative dataset to illustrate the problem

A major challenge in image recognition in medicine is the heterogeneity of image data. Image data are created with devices and software from different manufacturers.

NOTE 1 In pathology, for example, there are different scanning systems for pathological samples from different manufacturers. The different scanning systems and software lead to different image characteristics, for example in contrast and sharpness. Even with systems of the same model type there are often variances. Furthermore, the conditions under which the images are generated vary from hospital to hospital or laboratory to laboratory: for example, the lighting conditions can vary or the chemical dyes used in pathological samples have different compositions.

Hospitals and laboratories often have different compositions of patient groups they care for. Laboratories can be specialized in certain diseases. This leads to different data distributions of disease patterns. When creating data sets, it shall be taken into account that no bias is introduced into the data collection due to laboratory selection.

NOTE 2 In medicine in general, the heterogeneity of patients plays a role, so that the resulting images for image recognition often show strong individual variances.

During DL development, the influence of hardware and software properties on data and analysis quality shall be analyzed.

The producer of the image processing DL product shall specify for which types of input images and image sources predictions can be made. To get a wide range of applications, a correspondingly broad and representative training data set is usually required.

4.4 Data collection

Data acquisition for AI system development is a particularly challenging task in medicine, since it requires medical expertise and high-quality data, and medical data is characterized by a high degree of heterogeneity. The quality of the acquired data is of outstanding importance, since only with such data a high recognition performance can be achieved.

In order to generate a suitable data set, it shall be ensured that the selection of data is representative for the application, that the quality of both the data and the annotations is sufficient, and that the selected annotation rules are as unambiguous as possible. A detailed discussion of the topic data set quality can be found in DIN SPEC 13266.

A DL system is typically trained with supervised learning. Other DL training methods are also applicable.

NOTE 1 It is possible to learn suitable feature representations for supervised learning by the additional use of unsupervised learning.

Training with supervised learning requires annotations (labels) of the data. In contrast to other domains, annotations in medicine shall be created with expert knowledge.

NOTE 2 This makes it difficult and expensive to obtain data, as it requires a large number of specialists.

The competence of annotators shall be ensured by aptitude tests. This can be done by means of a test data set, for which the ground truth has been clearly determined by experts.

The annotation process shall be precisely defined and monitored in order to achieve consistent data quality. For this purpose, annotators shall be trained accordingly. The work of the annotators shall be constantly monitored to ensure quality. This can be done, for example, by interspersing already known images into the annotation process, for which the analysis of the annotator is compared to the already known ground truth. For high quality, it is important that the incentilation of the annotators does not create false incentives.

EXAMPLE A false incentive can be compensation per completed case.

A further difficulty in medicine is that a clear ground truth is often missing. The data collection shall reflect such inaccuracies.

NOTE 3 In histopathology, for example, cells often cannot be classified with a clear certainty because this information is missing in the image data.

To ensure high data quality, it is often necessary to have several annotators annotate the same data. In this case, only those data should be used for training and testing where the annotators match.

For the definition of the annotation process, estimates of inter- and intra-observer variability shall be obtained. These estimates show how much the opinions of different annotators (inter) or how much the opinion of the same annotator differs at different points in time (intra). This gives an estimate of the difficulty of the annotation problem and thus of the analysis task. Furthermore, it can be estimated how many experts are needed for the annotation.

The annotation process including the quality assurance process shall be documented in detail. A documentation shall contain in particular:

a) information on education, relevant qualifications and background knowledge of the annotator;

b) information on data protection measures and IT security;

c) applied incentive measures;

d) instructions for the annotators in the form of at least tool and annotation instructions;

e) documentation of the annotation tool (e.g. screenshots, screen recordings);

f) methodology for checking the correctness of annotations (e.g. clustering, repetitions);

g) methods for quality assurance, e.g. method of aggregating multiple repeated annotations by different annotators and monitoring the annotators.

Data protection and anonymization or pseudonymization play a prominent role in medicine. For this purpose, it shall be defined and documented whether metadata in the form of personal patient data are required for the DL system.

When collecting data, it shall be defined for which input data (formats, image sizes, color codes, etc.) the DL system shall be designed.

When collecting data, care shall be taken to avoid systematic bias and to exclude so-called data leaks, where correlated but not causal data are erroneously learned by the DL system as the basis for image analysis (e.g. images with the class "cancer", where a ruler is also always visible).

4.5 Embedding in the medical workflow

Changes in medical work processes are particularly complex. Before developing a DL system, it shall be clarified whether the changes for the use of a DL system and the integration into clinical processes are realistic.

The results of AI image recognition shall be easy for the user to understand. There shall be no ambiguity about how the results are handled.

NOTE For example, probability information in the form of probability heat maps often leads to uncertainty for users.

The provider of the DL system shall ensure that the system is constantly available in the medical workflow.

4.6 Methodology

Also on the methodical side there are special features for DL systems in medicine. Data augmentations are often used to broaden the learning data set, but shall be used with care in medicine to avoid unrealistic data.

NOTE 1 The image data is often not standardized. Image sizes, sharpness/es and color contrasts vary more than in the typical data sets of general DL research.

NOTE 2 The image data is often very large, e.g. pathological images are often 10 000 times larger than images used for example in face recognition and product searches.

A DL system shall deal with variances (see 5.6, NOTES 1 and 2) both in training and in inference in practical use.

4.7 Explainability and user confidence

Physicians and patients shall be able to trust the DL image analysis to use it. Neural networks, unlike other methods, are not transparent. However, there are more and more approaches to explain DL models that can help to gain user confidence [6].

NOTE The General Data Protection Regulation enables users to obtain information about decisions of underlying AI logics.

There are two types of explanations:

1) interpretation of models; and

2) the explanation of concrete individual predictions.

Typically today 2) is meant, if one speaks of explainability.

NOTE The research on the interpretation of models in DL systems is still in its infancy, even though there are interesting initial approaches. When explaining concrete individual predictions of DL systems, there are promising approaches in the current research literature on how to make the predictions of DL models plausible. For example, when classifying images, one can understand which input pixels are decisive for the classification. However, standard methods have not yet been established, since the current methods have different strengths and weaknesses and the current status quo is in a heuristic phase. However, it can be assumed that research in this area will make further progress towards explainability in the coming years.

There are different gradations of explainability:

a) If important pixels are highlighted for prediction, more mental effort remains for the user to determine if this makes sense.

b) Alternatively, the DL-System can determine and visualize relevant intermediate structures, e.g. cells in the context of tumor determination, which show that the DL-System has found relevant objects.

The results should be visualized to give users a clearer understanding of the results and to enable them to assess the quality of the results.

User confidence requires the implementation of validation and verification processes in software and product development. Typically, large parts of the user software can be tested according to conventional standards, e.g. ISO/IEC/IEEE 29119-1. For AI software components, corresponding standards do not yet exist. The performance of clinical studies is therefore the most important prerequisite for user confidence. The results of clinical studies and initial recognition quality metrics and specifications shall be provided.

DL systems are in principle susceptible to malfunctions triggered by so-called “adversarial examples” [7]. Adversarial examples are image examples that show a high similarity to significant image patterns, but are nevertheless completely misinterpreted. Such image examples not only reduce the recognition quality, they can also pose a security problem. The problem can question the robustness of a DL system in general [8], and thus also affect user confidence. In the development of DL systems, measures shall be taken to minimize the probability that adversarial examples can cause gross system malfunction.

5 Technical implementation

5.1 General

In addition to the requirements in this document, the technical implementation shall be carried out according to DIN SPEC 13266:2020-04, 8.2 to 8.9 and Annex A. All work steps shall be documented. This applies in particular, if the DL system is to be used in a product in practice and a corresponding certification is being sought. For this purpose, the data acquisition shall be documented (source, annotator/s). The training and validation process shall be described. In an analytical study it shall be shown that the developed DL-system has sufficient recognition quality to perform a clinical study in the next step. Therefore, the image recognition model shall be evaluated on test data. The selection of test data, the metrics and the evaluation shall be documented.

5.2 Formalization and problem definition

DIN SPEC 13266:2020-04, 8.2 applies.

Before starting the technical implementation, it shall be formalized what exactly the DL image recognition shall be able to do. A clear set of error measures shall be defined and it shall be defined what the minimum requirements in analysis quality are.

5.3 Data acquisition process and creation of the data set

DIN SPEC 13266:2020-04, 8.3 applies.

Data annotation should be performed by an expert group instead of individual experts. One challenge here is to define what an expert is. This can be all physicians as well as the top experts in the field. The annotators work with tools that are provided and shall first be trained to use them.

Some annotation tasks are time consuming and repetitive. Such annotations can often be done by laypersons after a short learning period. This can save costs and time, since laypersons are usually easier to find for this task. When using laypersons, it shall always be checked whether the quality of the annotations is comparable to that of the expert group. Furthermore, layman annotations shall always be validated by experts before they are used for training the DL system.

According to the definition of the application context, a representative data set shall be collected for training and validation.

Inter- and intra-observer variance should be measured during data collection.

When collecting data, rare classes shall be considered separately. These are typical in medicine. Frequently, the majority of patients are inconspicuous and there are rare diseases.

For data collection it is necessary that all relevant tools are validated.

The amount of data depends on the specific image recognition task.

An important step is the definition of a test data set. This shall not overlap with the examples in training and validation data. The test data set is used to estimate the prediction quality of the final model at the end of the entire development process.

5.4 Software Development

In software development, best practices in code development in medicine shall be considered (unit tests, integration tests, etc.), such as DIN EN 62304 for medical device software.

5.5 Training of the DL-System

The training of the DL-System in medicine shall be carried out as follows:

1) documented structuring of the data for training and testing the DL system according to DIN SPEC 13266:2020-04, 8.4

2) identification of potential DL architectures according to DIN SPEC 13266:2020-04, 8.5;

3) determination of the sequence structure of learning experiments according to DIN SPEC 13266:2020-04, 8.6;

4) model training, validation and selection exclusively with training and validation data, if necessary with cross-validation;

5) model selection according to DIN SPEC 13266:2020-04, 8.7 and according to hyperparameter optimization;

6) setting of simple Baselines models;

7) evaluation of the final DL model and baselines on the test data set;

8) reason why simple and easier to interpret models cannot be used;

9) checking of the reproducibility of the test results;

10) checking of the reproducibility of the DL model;

11) version management of data.

5.6 Model Deployment

DIN SPEC 13266:2020-04, 8.8 applies.

For the deployment of the DL system, seamless integration into the clinical environment is necessary. Therefore, it shall be ensured that the necessary performance and efficiency is achieved on the target hardware. In live operation, the quality of the DL system shall be ensured by continuous extensive testing as part of post-market surveillance and quality assurance.

In practice, it shall be validated and verified whether users use DL image recognition in the sense of the intended medical purposes. Thus, it shall be validated whether users understand the results and whether users blindly trust the system and whether this is in line with the intended purpose.

5.7 Quality assurance of the Deep Learning image recognition system

DIN SPEC 13266:2020-04, 8.9 applies.

5.8 Continuous improvement and relearning

A strength of a DL system for image recognition is continuous learning. The DL system can continue learning based on the images it processes. This allows for continuous improvement on one hand and adaptation, if the distribution changes over time, on the other hand. However, it can also cause the DL system to deteriorate with corresponding consequences for patients. Continuous learning, which automatically makes changes to a medical DL system, is therefore out of the question with the current state of technology. However, the use of continuous learning in the sense of regular controlled re-learning by the manufacturer or the user also involves considerable complexity, both in technology [9] and regulation [10]. Medical quality assurance and management shall be guaranteed by the DL manufacturer.

It shall always be ensured that the DL system does not deteriorate. For this purpose the continuous validation on a reference data set, the so-called quality assurance data, is necessary. Before deploying the DL model, limits should be defined within which models may change. Updated models may only be put into practical use, if continuous tests are performed on the quality assurance data. It shall be possible to switch back to earlier versions. All users shall be informed when the AI is updated.

NOTE So far, there are no clear regulations for continuous learning either in the EU or in the USA. On the regulatory side, however, the FDA (Food and Drug Administration) has positioned itself in the USA with a discussion paper [12]. According to this paper, DL manufacturers specifies in advance, in the form of an “Algorithm Change Protocol” (ACP), the extent to which DL models can change; they also specify clear change and further learning processes.

Bibliography

[1] ISO/IEC/IEEE 29119-1, *Software and systems engineering — Software testing — Part 1: Concepts and definitions*

[2] Litjens, G., Kooi, T., Bejnordi, B. et al. (2017), A survey on deep learning in medical image analysis. Medical Image Analysis, 42

[3] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature 521, p. 436–444 doi:10.1038/nature14539

[4] McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. C., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., Ledsam, J. R., Melnick, D., Mostofi, H., Peng, L., Reicher, J. J., Romera-Paredes, B., Sidebottom, R., Suleyman, M., Tse, D., Young, K. C., Fauw, J. D., Shetty, S. (2020): International evaluation of an AI system for breast cancer screening. Nature. 577, 7788, p. 89–94. [1]

[5] Johner et al.: Leitfaden zur KI bei Medizinprodukten. [2]

[6] Santos, I.C., Gazelle, G.S., Rocha, L.A., Tavares, J.M.R.: Medical device specificities: opportunities for a dedicated product development methodology. Expert, Review of Medical Devices 9(3), 299–311 (May 2012)

[7] Christoph Molnar: Interpretable Machine Learning.[3]

[8] Xu, H., Ma, Y., Liu, H.-C., Deb, D., Liu, H., Tang, J.-L., Jain, A. K. (2020): Adversarial Attacks and Defenses in Images, Graphs and Text: A Review. International Journal of Automation and Computing. 17, 2, p. 151–178.

[9] Dietterich, T. G. (2017): Steps Toward Robust Artificial Intelligence. AI Magazine. 38, 3, p. 3–24

[10] Parisi, G. I., Kemker, R., Part, J. L. (2019), Kanan, C., Wermter, S.: Continual lifelong learning with neural networks: A review. Neural Networks. 113, p. 54–71. [4]

[11] Perspectives and Good Practices for AI and Continuously Learning Systems in Healthcare. [5]

[12] Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD), FDA, April 2019

---

[1] Available at: https://www.nature.com/articles/s41586-019-1799-6

[2] Available at: https://github.com/johner-institut/ai-guideline/blob/master/Guideline-AI-Medical-Devices_DE.md

[3] Available at: https://christophm.github.io/interpretable-ml-book/cnn-features.html

[4] Available at: https://doi.org/10.1016/j.neunet.2019.01.012

[5] Available at: https://static1.squarespace.com/static/58d0113a3e00bef537b02b70/t/5cf13dec8f88ee0001f7c5c7/1559313901740/AI_WhitePaper_GoodPractices.pdf

[13] DIN EN ISO 9001, *Quality management systems — Requirements*

[14] DIN EN ISO 13485, *Medical devices — Quality management systems — Requirements for regulatory purposes*

[15] DIN EN ISO 14971, *Medical devices — Application of risk management to medical devices*

[16] DIN EN ISO/IEC 27001, *Information technology — Security procedures — Information security management systems — Requirements*

[17] DIN ISO 18115-1:2017-07, *Surface chemical analysis — Vocabulary — Part 1: General terms and terms used in spectroscopy (ISO 18115-1:2013)*

[18] DIN ISO/IEC 2382-28:1998-04, *Information technology — Vocabulary — Part 28: Artificial intelligence — Basic concepts and expert systems (ISO/IEC 2382-28:1995)*

[19] DIN EN 62304 (VDE 0750-101), *Medical device software — Software life cycle processes*

[20] DIN EN ISO/IEC 27701, *Security techniques — Extension to ISO/IEC 27001 and ISO/IEC 27002 for privacy information management — Requirements and guidelines*