Data preparation for artificial intelligence in medical imaging: A comprehensive guide to open-access platforms and tools

Oliver Diaz a,*, Kaisar Kushibar a, Richard Osuala a, Akis Linardos a, Lidia Garrucho a, Laura Igual b, Petia Radeva a, Fred Prior b, Polyxeni Gkontra a, Karim Lekadir a

a Faculty of Mathematics and Computer Science, University of Barcelona, Barcelona, Spain
b Department of Biomedical Informatics, University of Arkansas for Medical Sciences, Arkansas, USA

ARTICLE INFO

Keywords:
Artificial intelligence
Medical image datasets
Data curation
Annotation
Segmentation
Image storage
Open access
Image quality assessment

ABSTRACT

The vast amount of data produced by today’s medical imaging systems has led medical professionals to turn to novel technologies in order to efficiently handle their data and exploit the rich information present in them. In this context, artificial intelligence (AI) is emerging as one of the most prominent solutions, promising to revolutionise every day clinical practice and medical research. The pillar supporting the development of reliable and robust AI algorithms is the appropriate preparation of the medical images to be used by the AI-driven solutions. Here, we provide a comprehensive guide for the necessary steps to prepare medical images prior to developing or applying AI algorithms. The main steps involved in a typical medical image preparation pipeline include: (i) image acquisition at clinical sites, (ii) image de-identification to remove personal information and protect patient privacy, (iii) data curation to control for image and associated information quality, (iv) image storage, and (v) image annotation. There exists a plethora of open access tools to perform each of the aforementioned tasks and are hereby reviewed. Furthermore, we detail medical image repositories covering different organs and diseases. Such repositories are constantly increasing and enriched with the advent of big data. Lastly, we offer directions for future work in this rapidly evolving field.

1. Introduction

The term artificial intelligence (AI) was first coined in 1956 [1], but it was only recently that AI technologies showed their potential to reach or even surpass human performance in clinical practice. Roughly speaking, AI is a wide concept that refers to a set of computer algorithms that can perform human behaviour tasks, such as learning. On the other hand, machine learning is a sub-area of AI that covers machines with the ability to learn a given task by learning from past experience or past data, without the need for specific programming. Currently, AI has become a prominent topic in the healthcare sector [2], particularly in medical imaging [3,4]. By leveraging the latest innovations in computing power (e.g. Graphics Processing Units (GPU)), emerging AI technologies are expected to increase the quality and reduce the costs of medical imaging in future healthcare. This includes delivering enhanced image reconstruction [5–7], automated image segmentation [8–10], quality assurance approaches [11,12] and adequate image sequence selection [13], as well as by developing advanced image quantification for patient diagnosis and follow-up [14,15].

The development of AI solutions that are reproducible as well as transferable to clinical practice will require access to large scale data for model training and optimisation [16–18] (otherwise known as big data, and also referred to as the "oil" of the 21st century [19]). However, despite the acquisition of large volumes of imaging data routinely in clinical settings, access to big data in medical imaging poses significant challenges in practice. Hence, many researchers and developers have dedicated focused on the development of methods, tools, platforms and standards to facilitate the process of providing high-quality imaging data from clinical sites for technological developments, while complying with the relevant data regulations. To this end, data preparation pipelines [20] should cover a number of key steps as described in Fig. 1, including (i) image acquisition at clinical sites, (ii) image de-identification to remove personal information and protect patient privacy, (iii) data curation to control for image and non-image information quality, (iv) image storage, and (v) image annotation.

* Corresponding author.
E-mail address: oliver.diaz@ub.edu (O. Diaz).

https://doi.org/10.1016/j.ejmp.2021.02.007
Received 30 November 2020; Received in revised form 27 January 2021; Accepted 15 February 2021
Available online 5 March 2021
This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
In more detail, after getting approval from the corresponding ethical committees at the clinical sites, data de-identification is key to obtain anonymised images and to comply with local data protection regulations (e.g. GDPR, HIPAA). Subsequently, data curation is required to ensure that the associated data, e.g. metadata included in the image headers, are correct. For data storage and management, the FAIR guiding principles [21] recommend that data should be Findable, Accessible, Interoperable and Reusable (FAIR). Last but not least, medical annotations, including anatomical boundaries and lesion descriptions, are highly important not only for training, but also for testing the AI algorithms.

The purpose of this article is to provide a comprehensive guide to the main image data preparation stages, tools and platforms that can be leveraged before and during the design, implementation and validation of AI technologies. More precisely, we review existing solutions for image de-identification, data curation, centralised and decentralised image storage, as well as medical image annotation. We focus this survey on open-access tools that can benefit both clinical researchers and AI scientists at large scale, thus enabling community-wide and standardised data preparation and AI development in the medical imaging sector. Furthermore, we provide examples of medical imaging datasets and open-access data platforms that already cover different anatomical organs, diseases and applications in medical imaging.

2. Image de-identification tools

Tools in this category are also regarded as patient privacy preserving tools. Patient privacy is arguably of value by itself, but also ensures further values considered as fundamental such as dignity, respect, individuality and autonomy. More practically, privacy, and guarantees of such, enable patients to provide complete disclosures of their conditions, relationships between patients and their health providers [22].

Consensus exists to consider patient data as a highly sensitive resource with a need for privacy protection and secure communication. Legal regulation imposed by data protection authorities determine distribution, ownership and usage rights of such patient-specific data. Apart from legal considerations, organisations and individuals responsible for the collection and distribution of such data are encouraged to also apply ethical reasoning [23] to ensure a morally solid decision-making process guiding their actions of sharing or using these sensible data resources.

We consider sensitive patient data resources as any Protected Health Information (PHI) and Personally Identifiable Information (PII) linked to patient health information. The latter may include data from Electronic Health Records (EHR), medical images, clinical and biological data, and any other data collected by health providers that can add towards identifying a subject.

There are four major file formats in medical imaging [24]. DICOM (Digital Imaging and Communication On Medicine) format [25] is the international standard in this domain and covers all imaging modalities and organs, and it is supported by all vendors of clinical imaging systems. DICOM images contain a header, often referred to as metadata, with information regarding the image sequence, hospital, vendor, clinician or patient information among other information.

Despite the widely spread of DICOM, alternative (imaging or non-strictly-imaging) formats (developed specifically for neuro imaging) are also available, such as NIfTI, MINC and ANALYZE (first version of NIfTI). More recently, the format BIDS is rapidly replacing NIfTI.

Such patient data are an essential resource not only inside their natural environment, i.e. clinical site, but also outside. For example, patient data can be used for clinical trials and teaching [26], for research, and for training and validating AI solutions. Before using such data in these scenarios, they must comply with aforementioned data protection regulations such as HIPAA [27] or GDPR [28]. Techniques such as de-identification, anonymisation and pseudo-anonymisation can be applied to make the data available while simultaneously preserving patient privacy by removing all personal information which could identify an individual from medical images [29] before they leave the clinical centre.

The process of de-identification consists in removing or substituting all patient identifiers such as name, address, hospital identification number, from the patient data, i.e. the image metadata in the DICOM header, according to the local data protection regulation body. Anonymisation removes all patient identification data, thus they cannot be identified. The name, address, and full postcode must be removed, together with any other information which, in conjunction with other data held by or disclosed to the recipient, could identify the patient. Pseudonymisation is a procedure in which the PII is key coded using a unique identifier (i.e. a pseudonym). Such identifiers do not have relation with the individual, but could be traced back (if needed) through a well-secured and separately stored re-identification table. ISO 25237 [30] exemplifies one of several existing common standards for privacy preservation methods.

Apart from sensitive image metadata, the pixel/voxel data of the images also needs to be de-identified if it contains burned in PHI/PII data annotations or if it illustrates body features that can aid to identifying a patient such as for example facial features as depicted in Fig. 2.

In the DICOM Information Object Definitions (IODs) such facial features are referred to as recognisable visual features if they allow the identification of a patient from an image or from a reconstruction of images [31]. Schwarz et al. [32] showed that computer vision algorithms could be used to identify individuals from their cranial MRIs. Their study, hence, demonstrated that patient identification via facial features poses a significant threat to patient privacy in clinical datasets. For 70 out of 84 participants (83%), they were able to match a photograph of the participant with the reconstructed facial image from the participant’s MRI using publicly available face-recognition software. Defacing tools address this issue by de-identifying a patient’s facial

![Fig. 1. Data preparation pipeline prior developing and/or evaluation of AI solution.](image-url)
features by removing the pixels/voxels in an image that correspond to the facial features of the patient. However, defacing tools contain trade-offs between the facial feature removal and the respective information loss and, thus, cannot guarantee perfect de-identification nor usability of the resulting images. It is recommended to visually inspect the results from defacing tools and to be aware of edge cases and limitations i.e. no defacing algorithm can handle head/neck cancers when the lesions are on the face. Examples of defacing tools include FreeSurfer’s deface [33] command line tool, pydeface [34] and mrdefacer [35], both Python libraries under MIT license for defacing of MRI, and Quickshear [36], a Python library under BSD-3-Clause license for defacing of neuroimages.

As illustrated in Table 1, different tools exist for applying privacy preserving methods to medical imaging datasets. Gonzalez et al. [38] provide a list of requirements for data de-identification tools that are to be considered when selecting and applying such a tool for personal or professional use. Sampling from this list, we encourage the prospective user to consider the following recommendations and to incorporate them in decision-making processes when evaluating de-identification tools:

- Refrain from de-identifying the data in a place different from its natural environment (clinical site).
- Apart from DICOM header anonymisation, also handle and validate the de-identification of burned in annotations and of identifying facial features (i.e. deface brain magnetic resonance images (MRI) as demonstrated in Fig. 2) when feasible. Note that, in selected cases (e.g. radiation therapy or radiosurgery treatment planning, head and neck cancers with lesions on the face), facial features information is necessary for the AI model.
- Actively evaluate your tool’s compliance with DICOM standards, i.e. by validating conformance with DICOM Application Level Confidentiality Profile Attributes [25].
- Define concrete privacy preservation requirements for the specific use-case and data at hand.
- Ensure traceability and audit compliance i.e. by keeping a record of software, version, affected data portion, results, etc, for every usage event.

For further analysis of a selection of tools highlighted in Table 1, we point to the analyses provided by Lien et al. [47], Aryanto et al. [48], and Gonzalez et al. [38] that may further assist the interested reader to match specific tool requirements with specific tool capabilities.

Moreover, with notable advances in AI research in recent years, novel approaches have appeared in the development of AI solutions that avoid and circumvent patient identification. These approaches compete with or complement the tools and techniques described in Table 1 and include federated learning [49] and synthetic data generation, i.e. using generative adversarial networks [50]. For example, Abadi et al. [51] demonstrated how to generate synthetic COVID CT images while Alyafi et al. [52] produced synthetic breast lesions from mammography images. Federated learning is a privacy-preserving approach, where AI models are trained at the clinical site, hence eliminating the requirement of transporting sensitive clinical data out of their natural environment. The resulting models from different sites are combined in a centralised location.

3. Curation tools

We define data curation as the entirety of procedures and actions after data gathering that refer to data management, creation, modification, verification, extraction, integration, standardisation, conversion, maintenance, quality assurance, integrity, validation, traceability and reproducibility. According to this broad definition, multiple tools and applications could arguably be serving the goal of data curation. To provide a concise yet comprehensive list of curation tools, we, hence, focus on tools with broad application that cover the most frequent use-cases of curation in medical imaging such as DICOM conversion, modification and validation. Table 2 illustrates the list of curation tools. Many of these tools are platform-independent and support various different technologies and integration options. It is to be noted that de-identification tools and processes could be seen as part of the data curation workflow, as they, for instance, modify and standardise the gathered data. It is in this regard that de-identification tools of Table 1 may also be found in Table 2, in particular if such tools offer additional curation capabilities alongside their de-identification or anonymisation features.

Data curation tools are important due to their function of investigating, detecting, preventing and solving issues in the datasets. Thus, in the absence of data curation processes, various issues are prone to arise in the later stages of AI development such as errors stemming from unreliable data or introduction of bias and uncertainty of the validity of prediction results. For instance, prediction results on the test dataset lack validity in case of non-curated image duplicates, where one image is added to the training dataset and its duplicate to the test datasets [40].

Bennett et al. [40] present further examples that highlight the
| Name                     | Core features                                           | Documentation/verification | Reference                             |
|--------------------------|---------------------------------------------------------|----------------------------|---------------------------------------|
| DicomCleaner™             | Java-based web UI, header cleaning                      | detailed website, open source code | https://www.pixelmed.com/cleaner.html |
| Posda Tools               | perl tool, de-identification in annotations            | readme, open source code, publication | https://posda.com/                     |
| DicomAnonymizer           | Python library, cleaning of headers DICOM standards, BSD 3-Clause license | readme, open source code | https://github.com/KitwareMedical/dicom-anonymizer/ |
| DVTK DICOM Anonymizer     | C# library with GUI, de-identification, DICOM standards, configurable header modification | open source code, publication | https://github.com/dicom-anonymizer | ||
| Niffler                   | Python3 library, DICOM header cleaning, real-time PACS receiver, BSD-3-Clause license | readme, open source code, publication | https://github.com/Emory-HITI/Niffler |
| DICOM Anonymizer          | GUI tool for Mac OS, configurable DICOM anonymisation, GPLv3 license | simple website | https://dicomanonymizer.com/         |
| MIRC CTP                  | Java tool, DICOM header and burned in annotation anonymisation | detailed wiki, publications | https://mirwiki.rsna.org             |
| DCMTK (DICOM Toolkit)     | ANSI C, C++ library, DICOM standards BSD-3-Clause license | website, wiki, code, publication | https://www.dcmtk.org/                |
| PrivacyGuard              | Java library, flexible and extensible de-identification, Apache V2.0 license | website, open source code, publication | https://sourceforge.net/projects/privacyguard/ |
| Dicomanon                 | Matlab library, DICOM header cleaning                   | detailed website | https://www.mathworks.com/help/images/ref/dicom.html |
| Horos                     | GUI tool for Mac OS, DICOM viewer with anonymisation option, LGPL-3.0 license | detailed website | https://horosproject.org             |
| GDCM/Gdcmanon             | C++ library, anonymisation burned in annotation removal, BSD license | detailed wiki | http://gdcm.sourceforge.net        |
| PET/CT viewer             | Plugin for FIU, Orthanc DICOM anonymisation and sharing | detailed website with user manual | http://petctviewer.org/index.php/dicom-operations |
| Conquest                  | Web server for Windows and                             | detailed website with forum | https://github.io/deid               |

Table 1 (continued)

| Name                     | Core features                                           | Documentation/verification | Reference                             |
|--------------------------|---------------------------------------------------------|----------------------------|---------------------------------------|
| De-identification Toolbox | Java tool, header anonymisation de-identification audit and validation | simple website, publication | https://enginenum.home.xs4all.nl/dicom.html |
| Pydicom Deid             | Python library, configurable header and pixel data anonymisation, MIT license | detailed website, open source code | https://pydicom.github.io/deid |

The importance of data curation in medical imaging. This includes issues such as spatial information loss due to separate frame of reference unique identifiers of associated image slices, and DICOM inconsistencies in shared attributes across a given entity such as a patient, study or series. They also report problems with data normalisation, missing change reproducibility that causes uncertainty, and the need for verifying DICOM metadata conformity to enable interoperable data exchange.

The efforts of standardising the DICOM format is reflected in the development and progress of DICOM curation tools such as DCMTK [45]. For instance, DCMtk’s DCMCHECK [53] and dcm4che/dcmvaliddate test the adherence of DICOM files to DICOM Information Object Definitions [54, 55]. Alongside offering standardised implementation of medical imaging formats such as DICOM or BIDS [56], curation tools provide further standardisation aiding methods. These include conformity, inconsistency and referential integrity tests as implemented in Posda [40], as well as attribute, multiplicity, consistency and encoding validation available in dicom3tools, and in DVTK’s DICOM Validation Tool (DVT) [41].

A further distinctive feature of curation tools is accessibility, which is characterised by the capability of a tool to inclusively enable prospective users to access its functionality. Accessibility, hence, determines whether broad and heterogeneous user groups can readily adopt a tool into their workflows. Facilitating user interaction, a graphical user interface (GUI) can improve accessibility and simplifies adoption for users not familiar with a tool’s backbone technologies. Curation tools providing a GUI for user interaction include Posda, DVTk, dcm4che, BrainVoyager, LONI Debabeler, and dcm2niix via its GUI MRICroGL. Providing a user interface or in absence thereof, some curation tools require user to be familiar with underlying technologies such as command line tools (dicom3tools), docker (dcmqi) or programming languages such as Java (Java DICOM Toolkit) or ruby (ruby-dicom). These tools are less accessible to a broad user population, but arguably equally or more accessible to a specialised sub-population of prospective users such as software engineers. Also, access to the backbone technologies allows such specialised user groups to reproduce, configure, or extend the functionality of the respective tool. Optimal, a curation tool provides accessibility to both, specialised user groups and broad heterogeneous user populations, through suitable user interaction channels. For example, Posda, DVTk and dcm4che provide users with the option to choose between using web-based GUIs or programming language interfaces such as perl [40], C# [41] and Java [55], respectively.

Given a rapid technological progress, software upgrades are needed to provide users with the benefits of the latest technological advancements. The absence of such upgrades reinforces the risk of technological obsolescence of curation tools. Obsolescence is either caused by outdated methodology within a software tool (technological obsolescence) or by an outdated modality targeted by a software tool (functional obsolescence) [61]. Awareness of technological changes and paradigm shifts enables responding in time to a shift in demands to avoid
DCMTK

Additionally, data curation is a suitable part of AI development to introduce fairness evaluation according to the FAIR Guiding Principles. As previously stated, a fair dataset has been assembled once the datasets fulfill the FAIR requirements comprising findability, accessibility, interoperability, and reusability. Among others, following the FAIR principles includes providing well-described, searchable, uniquely identifiable and standardised image metadata, a data usage license, alongside creation, version and attribute details of the dataset [21].

4. Image storage

To address the limitations of hard-copy based medical image management, the first basic form of Picture Archiving and Communication Systems (PACS) was developed [62]. The first large scale installation of PACS took place in 1982 at the University of Kansas, Kansas City, USA [62,63]. At its core, PACS is an environment that facilitates the storage of multi-modal medical images (MRI, computerised tomography CT, positron emission tomography PET, etc) in central databases and their communication through browser-based formats (i.e. DICOM) that are easily accessible within entire hospitals as well as across multiple devices and locations. Furthermore, PACSs interface with other medical automation systems, including Hospital Information System (HIS), Electronic Medical Record (EMR) and Radiology Information System (RIS).

The PACS workflow (Fig. 3) begins with the packaging of multi-modal images to DICOM format which is then sent to a series of devices for relevant processing. The whole pipeline is summarised as follows:

- Quality Assurance (QA) workstation: verifies patient demographics and any other important attributes to the study. Note that this step might not be a common practise in some countries such as the USA.
- Archive (central storage device): stores the verified images along with any reports measurements and other relevant information relating to them.
- Reading Workstations: the place where a radiologist reviews the data and formulates their diagnosis.

An important extra step that one has to take with PACS is backup, i.e. ensuring facilities and means to recover images in the event of an error or disaster. As any critical data storage and management system, PACS data should be protected by maintaining them in several locations while adhering to privacy regulations. Traditionally, this has been done by transferring data physically through removable media (e.g. hard drives) off-site. However, with the advent of cloud computing, an increasing number of centres is migrating to a Cloud-based PACS paradigm [64]. That means that rather than a central storage device, image data are safely stored within the cloud, whose physical storage spans multiple servers, often in multiple locations.

The PACS revolution has sparked a boom in the production of commercial tools for image storage overcoming the limitations of the clinically-focused tool. Apart from commercial solutions, there also exist several open source solutions that require minimal investment from researchers and clinicians to follow. All of these open source solutions are frameworks from which to build your own server, but do not maintain a free storage service which would be costly to maintain. However, they are expandable and offer a variety of plugins, allowing

Table 2

| Name          | Core features                                                                 | Documentation/ verification | Reference                                                                 |
|---------------|-------------------------------------------------------------------------------|----------------------------|-------------------------------------------------------------------------|
| Ponda Tools   | perl tool, revision tracking, roll-back, bulk edit duplicate, inconsistency and conformity tests | readme, open source code, publication [40] | https://ponda.com/                                                      |
| dicom3tools   | Command line tool, DICOM conversions validation, attribute modification, BSD License | website, open source code | http://www. dcmtk.org/dicom3tools.html                                  |
| DCMTK         | ANSI C C++ library, DICOM standards converting, network, BSD-3-clause license   | website, wiki, code, publication | https://www. dcmtk.org                                                  |
| DVTk          | C++ library with GUI, DICOM standards, edit, validate, compare, network analysis | open source code, publication [41] | https://github. com/151706061/ DVTk                                   |
| dcm4che       | Java tool with web UI, production-ready DICOM standards and conversions | website, detailed wiki, publication | https://www. cm4che.org                                                 |
| dcmqi         | C++ library, DICOM conversions, querying, format for sharing, 3-clause BSD license | website, open source code, publication | https://github.io/dcmqi-guide                                            |
| dcm2nii       | C++ library, extension of dcm2nii, DICOM to NIfTI conversion | website, open source code, publication | https://www. nitr.org/plugins/ mwiki/index. php/dcm2nii/MainPage       |
| LONI Debabeler| Java tool with GUI, configurable conversions metadata extraction | website, publication [59] | https://resource. loni.usc.edu/resources/downloads/loni-debabeler        |
| BDS Tools     | Tools for neuroimaging dataset standardisation DICOM to BIDS conversion | website, open source code, publication | https://github. com/bids-standard                                        |
| BrainVoyager  | C++ neuroimaging tool with GUI, analysis normalisation, coregistration | detailed website, publication [60] | http://www. brainvoyager.com                                            |
| ruby-dicom    | Ruby library, editing, writing, network support for DICOM, GPL-3.0 license | readme, open source code | https://github. com/dicom/ruby-dicom                                    |
| Java DICOM ToolKit | Java stand-alone tool, write, validate, query DICOM network support, BSD license | website, javadoc, user guide | http://www. pixelmed.com/dicomtoolkit.html                               |
one to store medical images in the cloud through a separate provider which can comply with the data protection regulations.

Dcm4che (https://www.dcm4che.org) [55] is the most popular tool for clinical data management with Kheops (https://docs.kheops.online/) gaining ground rapidly. On the other hand, the Extensible Neuro-imaging Archive Toolkit (XNAT) https://www.xnat.org/ [65] is arguably the leading open source solution for management and storage of large heterogeneous data in research. It is a highly extensible Java web application that runs on an Apache TomCat server. It is able to manage many kinds of data, but engineered with a focus on imaging and clinical research data. Project owners have complete control to grant data access and user rights. Data can be stored indexed, searchable in a Postgres database or as resource files on a file system. With an extensible XML-based data model, it can support any kind of tabular data. The core is the RESTful API which can be used for handling data (requesting, displaying, uploading, downloading, removing) and is tied to the same user permission model providing the same level of access as the front end assuring the security of the stored data. XNAT contains rich documentation and a video series ‘XNAT Academy’. A wide range of institutes are using XNAT and most prominently the Human Connectome Project [66].

Dicoogle is an extensible, open source PACS archive. It has an modular architecture that allows for pluggable indexing and retrieval mechanisms to be developed separately and integrated in deployment. Storage, indexing and querying capabilities are entrusted to plugins. Therefore, Dicoogle does not offer a database of its own, but rather abstractions over these resources. Documentation for its functionalities includes a learning pack for getting started and developing with comprehensive examples and snippets of code.

Orthanc provides a lightweight standalone server and, similarly to XNAT and Dicoogle, it is expandable through plugins. It has an extremely detailed guide, the ‘Orthanc Book’, including guides for both users and developers as well as useful information on plugins, working with DICOMs, integrating other software and more. Orthanc also includes a professional version with extended capabilities.

The reader may refer to Table 3 for more open source solutions. For commercial solutions some examples include ORCA, iQ-WEB, PostDICOM and Dicom Director.

| Programming language/API | Documentation | Application | Reference |
|--------------------------|--------------|-------------|-----------|
| XNAT                     | Java, RESTful API | XNAT Academy: video series | Research | https://www.xnat.org/about/ |
| Kheops                   | RESTful API | API resources documented, includes user guide | Clinical | https://kheops.online/ |
| Dicoogle                 | Java, RESTful API | Learning pack, snippets and examples | Research | https://www.dicoogle.com/ |
| Orthanc-Server           | C++, RESTful API | Orthanc Book, rich documentation | Both | https://www.orthanc-server.com/ |
| OHIF                     | JavaScript (Meteor) | Detailed guide including deployment instruction | Clinical | https://ohif.org/ |
| JVS DICOM                | C++         | Lacking | Research | https://wiser.server.jilab.fi/old-jysmicopt-software/ |
| EasyPACS                 | Java, dcm4che API | One github page | Research | https://mehmetsen80.github.io/EasyPACS/ |
| PacsOne                  | MySQL, RESTFUL | PDF manual may be downloaded | Research | https://www.pacszone.net/solutions.htm? |
| Dcm4Che                  | Java         | One github page | Clinical | https://www.dcm4che.org/ |
| NeurDICOM               | Python, (plugins also in C, C++), RESTful | One github page | Research | https://github.com/reactmed/neurdicom |

Table 3
List of open source image storage (PACS) solutions. They are all cross-platform except for JVS DICOM which runs only on Windows. Some include premium options.

5. Image annotations

In order to train and validate AI algorithms, image annotations are pivotal. Image annotation corresponds to the process of labelling the images with essential information (e.g. spatial location, classification), and are often refereed as ground truth. This data is often contained inside the same DICOM file or in a separate text report and should be converted to a more readable and standard format, such as JSON or CSV, for later processing and AI development.

Medical image annotations depend on the type of image (2D, 3D, 4D) and the image technology used (e.g. MRI, US, CT). Therefore, there is a need for tools that are able to handle and annotate different image modalities and formats. More precisely, the type of image annotation will vary depending on the task to be performed by the AI algorithm.

For example, in cases where solely localisation information is needed, bounding boxes or contours (e.g. circles, ellipses, polygons) are typically used to depict the spatial location of an object of interest. If the AI task requires pixel-wise labels, then more detailed contours are created to segment the image into the regions or volumes of interest. However, obtaining such detailed segmentation masks is more challenging and time-consuming (Fig. 4).

Manual medical image annotation is a tedious task, especially on 3D images. Semi-automated or automated image annotation tools can alleviate the workload of human observers, i.e. clinicians, and increase the number of available annotations necessary for the development of AI solutions in the field of medical imaging.

In addition of image localisation information, annotations can also include lesion descriptions, cancer diagnoses and stages, patient outcomes and other clinical data of interest (Fig. 5).

Although previous works have also provided a detailed list of tools for data annotation [69,70], this section provides a selection of open-source, user-friendly and medical image oriented tools and platforms, as described in Table 4. A description of the most common tools is provided below.3D Slicer [37] is a well known open source software platform widely used for medical image informatics, image processing, and three-dimensional visualisation. 3D Slicer capabilities include handling a large variety of imaging formats (e.g. DICOM images),

![Fig. 3. PACS workflow.](image-url)
Fig. 4. ITK-SNAP workspace. Brain MRI with corresponding brain structure segmentation masks from MICCAI 2012 Grand-Challenge is shown as an example [67].

Fig. 5. Sample GUI (ImageJ) where observers are requested to evaluate the realism of a (real or simulated) breast lesion [68].

| Name            | 3D visualisation | Medical imaging oriented | Open-source | Web-based | Annotations                | Reference                  |
|-----------------|------------------|--------------------------|-------------|-----------|----------------------------|----------------------------|
| 3D Slicer       | ✓                | ✓                        | ✓           | ✓         | manual, auto               | www.slicer.org             |
| ePAD            |                  |                          |             |           | manual                     | epad.stanford.edu          |
| Horos Viewer    | ✓                |                          |             |           | manual, semi-auto          | www.horosproject.org       |
| ImageJ/ FIJI    | ✓                |                          |             |           | manual                     | fiji.sc                    |
| InVesalius      | ✓                |                          |             |           | manual, auto               | invesalius.github.io       |
| ITK-SNAP        | ✓                |                          |             |           | manual, semi-auto          | www.itksnap.org            |
| MedSeg          |                  |                          |             |           | manual                     | www.medseg.ai              |
| MeVisLab        |                  |                          |             |           | manual, auto               | www.mevislab.de            |
| MITK            |                  |                          |             |           | manual                     | www.mitk.org               |
| ParaView        |                  |                          |             |           | manual                     | www.paraview.org           |
| Seg3D           |                  |                          |             |           | manual, semi-auto          | www.sci.utah.edu/cibc-software/seg3d |

Crowdsourcing Platforms

| Name            | 3D visualisation | Medical imaging oriented | Open-source | Web-based | Annotations | Reference                  |
|-----------------|------------------|--------------------------|-------------|-----------|-------------|----------------------------|
| CMRAD           | ✓                | ✓                        |             |           | manual      | www.cmrad.com              |
| Crowds Cure     |                  |                          |             |           | manual      | www.crowds-cure.org        |
interactive visualisation of 3D images, and manual and automated image segmentation features, among others. It is a powerful tool with a large user community with a complete documentation including a wiki page, a discussion forum, and user and developer manuals. Slicer supports different types of modular development and has extensions for improved segmentation, registration, time series, quantification and radiomic feature extraction available on their web page. As an example, the EMSegment Easy module performs intensity-based image segmentation automatically.

ITK-SNAP [71] is the product of a collaboration between the universities of Pennsylvania and Utah, and focuses specifically on the problem of image segmentation offering a user friendly interface. ITK-SNAP provides tools for manual delineation of anatomical structures. Labelling can take place in all three orthogonal cut planes (axial, coronal and sagittal) and visualised as a 3D rendering. ITK-SNAP also provides automated segmentation using the level-set method [72] which allows segmentation of structures that appear homogeneous in medical images using little human interaction. This tool has been widely applied in many studies focusing on facial pathologies and anatomical studies [73], carotid artery segmentation [74], diffusion MRI analysis [75], and virtual reality in Medicine [76], among others. A screenshot of ITK-SNAP is shown in Fig. 4.

The Medical Imaging Interaction Toolkit (MITK) [78], developed by the German Cancer Research Center, is an open-source software focused on the development of interactive medical image processing. MITK combines the Insight Toolkit (ITK) and the Visualization Toolkit (VTK) within an application framework. The leading GPU manufacturer NVIDIA has created the NvidiaAnnotation plugin for MITK Workbench which can be used to leverage the power of an NVIDIA AI-Assisted Annotation Server and perform automated segmentation.

Horos Viewer is a free, open source medical image viewer. The goal of the Horos Project is to develop a fully functional, 64-bit medical image viewer for Mac OS X. Horos is based upon OsiriX™ and other open source medical image libraries.

Another popular tool in the medical imaging field is ImageJ [79], a Java-based image processing and analysis tool developed at the National Institute of Health and the Laboratory for Optical and Computational Instrumentation (LOCI, University of Wisconsin). ImageJ has built-in support for reading DICOM files. However, Fiji (Fiji Is Just ImageJ) [80], which is an image processing distribution of ImageJ, bundles a lot of plugins to facilitate scientific image analysis, including Dicom support. Fiji plugins useful for medical image annotation include, among others:

- Bio-Formats: an open source solution for converting proprietary microscopy image data and metadata into standardised, open formats.
- SciView: 3D visualisation and virtual reality capabilities for images and meshes.
- MaMuT: combines TrackMate and BigDataViewer for browsing, annotating, and curating annotations for large image data.
- Trainable Weka Segmentation: a tool that combines a collection of machine learning algorithms with a set of selected image features to produce pixel-based segmentation.

All the tools described above are typically used locally to annotate images. However, there are crowd-sourcing portals, such as Crowds Cure and CMRAID platform, which allow experts to collaboratively annotate medical images on the cloud across sites. One of the main challenges of AI in medical imaging is the lack of consensus on image annotations across clinicians and clinical sites (i.e. intra- and inter-observer variability). Thus semi-automated and automated tools with little supervision from clinicians can be a potential tool to reduce the time dedicated to this task.

6. Medical image repositories

The previous sections have shown tools to prepare data before they are used for developing or evaluating AI solutions, or to create an image repository. However, there exists many medical imaging repositories, often open-access or controlled-access, which can be used to enrich datasets (e.g. multicentre, multi-vendor, multi-disease) or directly develop own solutions.

Due to the data-driven nature of ML algorithms – in particular, DL approaches – there have been some initiatives that significantly advanced the data collection and availability for the research community. Table 5 summarises various sources of open-access medical imaging databases categorised by target organ and disease. Note that while some sources are open-access, with online registration being required in some cases, the others require permission to access the data. The latter is often attainable via an online request.

One of the well-known resources is the TCIA data repository [81], which offers a variety of curated imaging collections for multiple organs specifically oriented towards cancer imaging. Recently, X-ray and Computed Tomography (CT) images for COVID-19 patients were added to the resource. The images from this repository can be downloaded as collections grouped by a common disease or imaging modality. The primary data format used in TCIA is DICOM. However, as detailed in Table 2, there are different open-access tools that could be used to convert into other different data formats such as NIfTI, which store an image in a single compressed file. Moreover, TCIA also provides clinical data for some of the cases such as patient outcomes, treatment details, genomics and expert analyses. TCIA uses the National Biomedical Imaging Archive (NBAI) software (https://imaging.nci.nih.gov) as a backbone and extends its utilisation by providing more curated datasets, user support, and wiki-guides.

UK Biobank is another resource that achieved an outstanding impact in medical data collection and research. Apart from a wide variety of clinical data such as EHR, it hosts imaging collections of more than 100,000 participants including scans of brain, heart, abdomen, bones and carotid artery.

Over the past few years, several challenges have been organised on different medical imaging topics as part of the Medical Image Computing and Computer Assisted Intervention (MICCAI) conference. These challenges are organised by different institutions, which provide a dataset to solve a particular medical imaging problem. These datasets often become benchmarks for evaluating novel AI approaches, which is crucial in terms of reproducibility and comparison with the state-of-the-art approaches. The Grand-Challenges repository provides a compilation of references to the original resources where the datasets can be downloaded or requested from the challenge organisers. Moreover, starting from 2018, MICCAI organisation developed an online platform for challenge proposal submissions with structured descriptions of challenge protocols (https://www.biomedical-challenges.org/). This allows more transparent evaluation, reproduction, and adequate interpretation of the challenge results [82].

While the MICCAI challenges are yearly organised events, Kaggle offers ongoing challenges on different ML topics including Medical Imaging. Recently, Kaggle introduced a dataset usability rating for each available challenge, which indicates how easy-to-use is its data. Datasets with high usability rating are often processed and curated for the participants to download and immediately proceed with their experiments. This is particularly important for the newcomers to AI in the Medical Imaging field.

The aforementioned imaging repositories host datasets for multiple organs and various medical conditions. However, there are other data collection initiatives that focus on a specific organ. For example, large sets of neuroimaging datasets could be accessed from IDA, OASIS, NITRC, and CQ500 [83] repositories (see Table 5) that include imaging data for healthy young, adult, ageing, and patient data for various neurological disorders. Most of the datasets from these collections are
benchmarks to evaluate AI-based approaches for image segmentation or disease classification. Similarly, STARE, DRIVE (part of Grand-Challenges), and HRF [84] datasets are commonly used to evaluate automatic methods for retinal fundus image segmentation as they provide expert annotations for human eye disease studies, just to mention a few. The International Skin Imaging Collaboration (ISIC) provides a collection of digital images of skin lesions for teaching and to promote development of automated diagnostic tools by organising public challenges.

OPTIMAM (OMI-DB) is an on-going project collected over 2.5 million breast cancer screening mammography images [85]. EchoNet-Dynamic is another resource with over 10K echocardiography videos with corresponding labelled clinical measurements and human expert annotations [86]. In addition, euCanSHare – a joint EU-Canada project funded by European Horizon 2020 programme – is establishing a cross-border data sharing and multi-cohort cardiovascular research platform.

### Table 5
List of medical image repositories for different anatomical organs and diseases. Open Access (OA) and Review Committee (RC) datasets are shown. Letters ‘s’, ‘f’, ‘d’ before MRI indicate structural, functional, and diffusion, respectively. TBI stands for Traumatic Brain Injury.

| Organ                  | Disease       | Name                                      | Access | Images          | Reference                              |
|------------------------|---------------|-------------------------------------------|--------|-----------------|----------------------------------------|
| Multiple               | Cancer        | The Cancer Imaging Archive (TCIA)         | OA/RC  | Radiology, histopathology | https://www.cancerimagingarchive.net |
| Multiple               | Multi         | National Biomedical Imaging Archive (NBA) | OA/RC  | Radiology       | https://imaging.nci.nih.gov            |
| Multiple               | Multi         | UKBiobank                                  | RC     | MRI, DIXA       | https://www.ukbiobank.ac.uk            |
| Multiple               | Multi         | Grand-Challenges                           | OA     | Multi-domain    | https://grand-challenge.org            |
| Multiple               | Multi         | Kaggle                                     | RC     | Multi-domain    | http://www.kaggle.com                  |
| Multiple               | Multi         | VISCERAL: Visual Concept Extraction Challenge in Radiology | RC | Multi-domain | http://www.visceral.eu/benchmarks     |
| Multiple               | Multi         | Medical Segmentation Decathlon             | OA/RC  | CT, MRI         | http://medicaldecalthion.com           |
| Brain                  | Multi         | OpenNeuro                                  | OA/RC  | Multi-domain    | https://openneuro.org                  |
| Brain                  | Multi         | Image and Data Archive (IDA)               | OA/RC  | s/f/dMRI, CT/PET/SPEC | https://ida.loni.usc.edu |
| Brain                  | Normal, dementia, Alzheimer’s          | OASIS Brains Dataset                       | OA     | MRI             | https://www.oasis-brains.org           |
| Brain                  | Multi         | NITRC: Neuroimaging Tools and Resources Collaboratory | OA | s/fMRI | https://nitr.org                      |
| Brain                  | TBI           | The Federal Interagency TBI Research (FITBIR) | RC | MRI, PET, Contrast | http://fitbir.nih.gov                  |
| Brain                  | TBI, Stroke   | CQ500                                      | OA/RC  | CT              | http://headctstudy.qure.ai/dataset     |
| Brain                  | Multi         | NDA                                        | RC     | MRI             | https://nda.nih.gov                    |
| Brain                  | Multi         | Connectome                                 | RC     | sMRI, fMRI      | https://www.humanconnectome.org        |
| Breast                 | Cancer screening | MIAO mini-database                      | OA     | MG, US          | http://peipa.essex.ac.uk/info/mias.html |
| Breast                 | Cancer screening | BCDR                                      | RC     | MG, US          | https://bcdr.eu                        |
| Breast                 | Cancer        | DDSM                                       | OA     | MG              | http://www.eng.usf.edu/cvprg/ Mammothography/Database.html |
| Breast                 | Cancer        | OMI-DB                                     | OA/RC  | MG              | https://medphys.royalsurrey.nhs.uk/omiDB |
| Breast                 | Cancer        | INbreast                                   | OA/RC  | MG              | http://medicalresearch.inespporto.pt/breastresearch/index.php/Get_INbreast_Database |
| Cardiac                | Clinical routine care | EchoNet-Dynamic                      | OA/RC  | Echocardiogram videos | https://echonet.github.io/dynamic/camers |
| Cardiac                | Multi-abnormal | CAMUS project                             | OA/RC  | Echocardiogram  | https://www.creatis.insa-lyon.fr/Challenge/camus |
| Cardiac                | Multi         | EuCanShare                                 | RC     | MRI             | http://www.eucareshare.eu              |
| Cardiac                | Multi         | Cardiac Atlas Project                      | OA/RC  | MRI             | http://www.cardiacatlas.org            |
| Full body              | Healthy, unknown | Thorax                                    | OA     | CT, MRI         | https://www.nlm.nih.gov/research/visible/NIHOC |
| Lung                   | Multi         | Cornell Engineering: Vision and Image Analysis lab | OA | CT            | http://www.via.cornell.edu/databases   |
| Lung                   | COVID19       | MostMedData                                | OA     | CT              | https://rostemed.ai/en                  |
| Lung                   | COVID19       | COVID-19 CT segmentation                   | OA     | CT              | http://medicalsegmentation.com/covid19  |
| Lung                   | COVID19       | BIMCV COVID-19                             | OA     | CT, CXR         | https://github.com/BIMCV-CSUSP/BIMCV-COVID19 |
| Lung                   | COVID19       | COVID-19 Image Data Collection             | OA     | CT, CXR         | https://github.com/isee8023/covid-chestxray-dataset |

**Note:** The link to the COVID-19 Image Data Collection dataset is not directly accessible from the text provided. The dataset is described as COVID19 Image Data Collection OA CT, CXR placed in the table, but the specific link is not mentioned.
NHS Chest X-ray [87] provides annotated X-ray images of 14 common thorax diseases. Also, collections of diagnosed lung CT images of pulmonary diseases are available from the Cornell-Engineering: Vision and Image Analysis lab repository. Moreover, the worldwide impact of the recent COVID19 pandemic on healthcare systems required an immediate reaction to develop automated diagnosis methods. Thus, open-source initiatives accumulated a vast amount of crowd-sourced Chest X-ray and CT images of COVID19 along with healthy and other pulmonary cases. Examples of such datasets include BIMCV COVID19 [88], COVID-19 Image Data Collection [89]. As of 2020, these datasets are continuously expanding (see Table 5) and the images are aggregated from different sources, hence, some overlapping cases may occur. Nonetheless, it is an enormous effort that the research community has put together in battling against the pandemic. Additionally, a more curated dataset with CT images could be requested from a MICCAI endorsed COVID19 lesion segmentation challenge in https://covid-segmentation.grand-challenge.org as well as in TCIA.

Although listed in the Grand-Challenges, it is worth mentioning the medical image segmentation decathlon challenge that provides expert annotated images for ten different tasks: brain tumour (BraTS) [90], heart (LASC) [91], and liver (LITS) [92] compiled from previous MICCAI challenges and TCIA databases; and brain hippocampus structure, prostate [93], lung [94,95], pancreas [96], hepatic vessel [97], spleen [98], and colon datasets were acquired from various healthcare institutions [99]. In addition, detailed CT and MRI full body post-mortem scans of a male and a female subjects are available in the Visible Human Project to study human anatomy as well as to evaluate medical imaging algorithms.

7. Discussion

As confirmed by many healthcare professionals [100-104], AI is revolutionising the field of medicine in general and medical imaging in particular. For instance, AI is suggested to be included in the Medical Physics Curricular and Professional Programme [105]. Although there are still several open issues to be addressed, AI has already demonstrated significant potential to overcome human performance in selected tasks, such as image segmentation. Moreover, AI provides key information in the clinical decision-making process. Without AI, this information would have been extremely difficult, if not infeasible, to extract and combine in an optimal way.

The success of AI is in part thanks to the increasing number of (open-access) medical image datasets becoming available. The images are used by AI networks to extract the most informative features for identifying the boundaries of anatomical structures or for predicting the presence of a disease. However, prior to this step, medical images need to be adequately prepared in order to be safely used and to maximise their potential in the AI development or assessment.

7.1. Remaining challenges of data preparation

As shown in this paper, a wide range of open access image repositories and open source tools have been developed over the last decade to promote standardised best practices for data preparation in medical imaging. However, there are still many challenges that will require further attention and research, as discussed below.

7.1.1. Anonymisation

Image anonymisation (or de-identification) is key to preserving patient privacy. In the last years, several international regulatory environments (e.g. GDPR, HIPPA) have been updated and such changes must be regularly reflected within the image de-identification tools. Therefore, mechanisms should be put in place to automate this process (currently performed in a semi-automated way), and more importantly, to demonstrate that these anonymisation tools actually meet regulatory requirements. Special emphasis should be put into synthetic data generation and validation procedures, which can overcome many of the current limitations.

As previously discussed, a 3D reconstruction of the head should not identify a person, thus certain spatial information should be also removed (image defacing). However, the challenge is to remove identifiable facial features while preserving essential scientific data without distortion. This is difficult for several diseases such as head/neck cancers, or radiation therapy planning where currently key data is destroyed.

Although all anonymisation efforts are focused on the image, anonymisation strategies should also consider the associated clinical data, annotations and other forms of labelled data.

7.1.2. Curation

Data curation is another important step to ensure the data is properly organised and structured. Once the data collection is defined in a fair way, our efforts should focus on improving the quality of the data. This can be achieved by setting standards and guidelines in the entire process of medical image preparation from the de-identification step to the data annotation step and specially in the data curation step. Initiatives in the data collection and availability are also crucial for AI research in medical imaging because they allow to create benchmarks for multi-centre and multi-scanner AI evaluation.

Automated tools and standards for measuring image quality, particularly for quantitative analyses, are greatly needed. Furthermore, approaches to automatically detect and correct image artefacts will be vital to guarantee certain quality in the images used to train AI algorithms. Similar efforts (quality standards and tools) should be also promoted to assure the quality of annotations, labels and image derived features.

7.1.3. Annotation

Image annotations are also pivotal to ensure a correct training of AI algorithms and they should be performed with care. However, having accurate delineations or annotations is challenging as well as very time consuming, especially in 3D imaging modalities. Annotators (typically clinicians) do not have the time to annotate the hundred or thousand images needed to train AI algorithms. At the same time, despite the advances made with the advent of deep learning, automated segmentation tools that can be robustly applied across varying imaging protocols and clinical sites are still lacking, in particular as open-source. As an illustration, our experience in the EuCanImage Horizon 2020 project, focused on the annotation of liver, colorectal and breast cancer images, suggests that additional research is needed to develop the next generation of open-access image segmentation tools for the wide benefit of the community. Current and future crowdsourcing and collaborative annotation platforms will be of high value to capture the heterogeneity of annotation based on several clinicians.

7.1.4. Storage

Since most of the image datasets are scattered in the web, large scale integration of distributed data repositories are needed to centralise its access and to reuse the image derived features. This opens new questions as to whether researchers should process their data on the cloud (typically a paying option) or if they should download the data to work on their own computing environment. The availability of such platforms should also allows the cross-linkage and semantic integration of radiology/pathology/clinical/-omics data with the images.

Furthermore, data accessibility is very important to promote a good standardisation of AI development. This is already covered by the FAIR principles [21] and it is a requirement for open science, open access to de-identify data. A key component to enhance data accessibility (currently adopted by the TCIA, the largest cancer imaging archive), is the use of a digital object identifier (DOI) to each data collection so that collection can be cited and directly accessed from the DOI.

From the image standardisation point of view, the DICOM format
already provides an international standard for image data, although other image formats have appeared in the last years (mostly for neuro imaging). However, in order to facilitate the distribution of AI models, standardised containers could be of great benefits for the field. Once developed, AI tools should be made open source to facilitate their distribution. In addition, the development of an open source community that supports that software is pivotal to guarantee its maintenance and upgrade.

7.2. Future research in AI

In addition to data preparation, which is the focus of this paper, it is worth discussing future directions for AI in medical imaging: (i) Data augmentation and synthesis, (ii) Federated learning, (iii) ethical issues of AI, and (iv) Uncertainty estimation.

Data augmentation has shown promise in AI and to enrich the data preparation stage. State-of-the-art data augmentation techniques range from basic strategies using feasible geometric transformations, flipping, colour modification, cropping, rotation, noise injection and random erasing [106] to other more advanced techniques that involve the creation of new synthetic images, such as generative adversarial networks [50].

Recently, new privacy-preserving approaches, known as Federated learning [49,107], have been proposed in clinical research for privacy-preserving AI and, somehow, to enrich the datasets used by the AI technology. These techniques train an algorithm across multiple decentralised clinical sites holding local data samples, without exchanging them. Therefore, locally-trained AI results are later combined in a centralised location. Such novel paradigms should enable larger and more representative samples while also assisting in protection of patient privacy [108,109].

Ethical use of AI tools in medicine is a major concern. The statement presented in [110] highlights the consensus of several international societies that ethical use of AI in radiology should promote well-being, minimise harm, and ensure that the benefits and harms are distributed among stakeholders. Important issues of ethics, fairness and inclusion can arise from the pitfalls and biases during data preparation. Routine clinical data collected by clinical sites can be deficient, biased (e.g. gender imbalance [111]), or prone to noise (e.g. presence of image artifacts [112]). Several advancements can be achieved by defining algorithms that can track these issues efficiently [113]. However, in order to generalise the AI-generated results to the human population, large-scale, multi-centre training and test datasets of sufficient quality are often required [18].

As shown in this paper, there are many parameters that affect the data preparation and the quality of the data compiled for AI training. As a result, in addition to accuracy, AI models for medical image analysis should be assessed on the level of confidence in the predictions. Uncertainty estimation is of particular importance since data preparation is imperfect. If the uncertainty is too high, the clinicians should be notified in order to take this information into account in the final decision-making process.

Kendall and Gal [114] have demonstrated two types of uncertainties in computer vision that are also applicable in medical image analysis: the epistemic uncertainty caused by lack of knowledge in the AI model (reduced training data) and the aleatoric uncertainty (inherent to the data such as acquisition artefacts, patient movement, radiofrequency spikes, etc.).

Uncertainty estimation in deep learning scenarios is often obtained by approximating Bayesian posterior using neural networks [115–117], although non-Bayesian approaches have also been addressed [118]. The uncertainties are usually represented as a heat map defined by certain uncertainty measures such as entropy, variance, and mutual information.

Overall, the predictive model uncertainty should be always considered in making critical decisions which is always part of common clinical routines. We believe that this emerging research topic will improve the applicability of AI towards practical scenarios by increasing the trustworthiness of such methods that are currently considered as black-box.

8. Conclusions

In this article, we reviewed open access tools and platforms for the different steps of medical image preparation. This process is essential before starting the design or application of any AI model. More precisely, the steps comprising a typical medical image pipeline are: de-identification, data curation, centralised and decentralised medical image storage, and data annotations tools. The presented structured summary provides users and developers a comprehensive guide to choose among the plethora of currently available tools and platforms to prepare medical images prior to developing or applying AI algorithms. Furthermore, we provided a detailed list of medical imaging datasets covering different anatomical organs and diseases.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 952103, from the Spanish Ministry (grant RTI2018-095232-B-C21) and Catalan government (grant SGR1742).

References

[1] Andresen SL, John McCarthy: father of AI. IEEE Intell Syst 2002;17(5):84-5.
[2] Patel VL, Shortliffe EH, Stellman M, Slezovits P, Berthold MR, Bellassa R, et al. The coming of age of artificial intelligence in medicine. Artif Intell Med 2009;46(1-3):5-17.
[3] Giger ML. Machine learning in medical imaging. J Am College Radiol 2018;15(3): S12-S20.
[4] Savad physique P, Joth J, Dohan A, Vakalopoulou M, Reinhold C, Paragiou N, et al. Demystification of ai-driven medical image interpretation: past, present and future. Eur Radiol 2019;29(1):1616-24.
[5] Teuwen J, Morastov N, Fedon C, Caballo M, Reine J, Bakic P, et al. Deep learning reconstruction of digital breast tomosynthesis images for accurate breast density and patient-specific radiation dose estimation. arXiv preprint arXiv:2006.050822020.
[6] Wang G, Ye JC, De Man B. Deep learning for tomographic image reconstruction. Nat Mach Intell 2020;2(12):737-48.
[7] Singh R, Wu W, Wang G, Kalra MK. Artificial intelligence in image reconstruction: The change is here. Phys Medica 2020;79:113-25.
[8] Agarwal R, Diaz O, Llado X, Gubner-Mérida A, Vilanova JC, Martí R. Lesion segmentation in automated 3d breast ultrasound: volumetric analysis. Ultrasom Imag 2018;40(2):97-112.
[9] Kubikar B, Valverde S, González-Villa S, Bernal J, Cabezas M, Oliver A, et al. Automated sub-cortical brain structure segmentation combining spatial and deep convolutional features. Med Image Anal 2018;48:177-86.
[10] Apte AP, Iyer A, Thor M, Pandya R, Haq R, Jiang J, et al. Library of deep-learning image segmentation and outcomes model-implementations. Phys Med 2020;73:190-6.
[11] Kimura Y, Kadoya N, Tomori S, Ouy K, Jin G. Error detection using a convolutional neural network with dose difference maps in patient-specific quality assurance for volumetric modulated arc therapy. Phys Med 2020;73:57-64.
[12] Olaciregui-Ruiz I, Torres-Xirau I, Teuwen J, van der Heide UA, Mans A. A Deep Learning-based correction to EPID dosimetry for attenuation and scatter in the unity MR Linac system. Phys Med 2020;71:124-31.
[13] Wang Y, Wang M. Selecting proper combination of mpMRI sequences for prostate cancer classification using multi-input convolutional neuronal network. Phys Med 2020;80:92-100.
[14] Leknadar K, Galimzianova A, Betin A, del Mar Vila M, Igual L, Rubin DL, et al. A convolutional neural network for automatic characterization of plaque composition in carotid ultrasound. IEEE J Biomed Health Inf 2016;21(1):48-55.
[15] Getin I, Raïs-Estahragh Z, Petersen SE, Napel S, Piekchn SK, Neubauer S, et al. Radiomics signatures of cardiovascular risk factors in cardiac MRI: results from the UK Biobank. Front Cardiovasc Med 2020;7;
O. Diaz et al.

Physica Medica 83 (2021) 25–37

[81] Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppell P, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digital Imag 2013;26(6):1045–57.

[82] Maier-Hein L, Eisemann M, Reineke A, Onorini S, Stankovic M, Scholz P, et al. Why rankings of biomedical image analysis competitions should be interpreted with care. Nat Commun 2018;9(1):1–13.

[83] Chilamkurthy S, Ghosh R, Tananala S, Biviji M, Campeau NG, Venugopal VK, et al. Deep learning algorithms for detection of critical findings in head ct scans: a retrospective study. The Lancet 2018;392(10162):2388–96.

[84] Budai A, Bock R, Maier A, Horngger J, Michelson G. Robust vessel segmentation in fundus images. Int J Biomed Image 2013:2013.

[85] Halling-Brown MD, Warren LM, Ward D, Lewis E, Mackenzie A, Wallis MG, et al. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digital Imag 2013;26(6):1045–57.

[86] Ouyang D, He B, Ghorbani A, Yuan N, Ebinger J, Langlotz CP, et al. Video-based AI for beat-to-beat assessment of cardiac function. Nature 2020;580(7802):252–6.

[87] Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8: hospital-scale chest x-ray dataset and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 1007–10.

[88] Vosyli Dill, Saborir JM, Montelli JA, Pertusa A, Buxts A, Cazurra M, et al. BMVC COVID-19+: a large annotated dataset of RX and CT images from COVID-19 patients. arXiv preprint arXiv:2006.01174 2020.

[89] Cohen JP, Morrison P, Dao L, Roth K, Duong TQ, Ghassemi M. COVID-19 Image Data Collection: Prospective Predictions Are the Future. arXiv preprint arXiv:2006.11988 2020.

[90] Bakas S, Reyes M, Jakab A, Bauer S, Rempfler M, Crimi A, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. arXiv preprint arXiv:1811.02629 2018.

[91] Tobon-Gomez C, Geis AJ, Peters J, Weese J, Pinto K, Karim R, et al. Benchmark for algorithms segmenting the left atrium from 3D CT and MRI datasets. IEEE Trans Med Imag 2015;34(7):1460–73.

[92] Bilic P, Christ PF, Vorontsov E, Chlebus G, Chen H, Dou Q, et al. The liver tumor segmentation benchmark (lits). arXiv preprint arXiv:1901.04056 2019.

[93] Litjens G, Wehmert P, Garcia AJ, van de Ven W, Karssemeijer N, Huisman H. A pattern localization of common thorax diseases. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 1007–10.

[94] Simpson AL, Leal JN, Pugalenthi A, Allen PJ, DeMatteo RP, Fong Y, et al. Survival prediction in pancreatic ductal adenocarcinoma by quantitative dataset of non-small cell lung cancer. Sci Data 2018;5(1):1–9.

[95] Antun V, Renna F, Poon C, Adcock B, Hansen AC. On instabilities of deep learning. In: International conference on machine learning. PMLR; 2016, p. 1050–6.

[96] Swinder N, De-Arteaga M, auz NTHI, Leisner MD, Kalai AT. What are the biases in my word embedding? arXiv preprint arXiv:1703.04977 2017.

[97] Antun V, Renna F, Poon C, Adcock B, Hansen AC. On instabilities of deep learning in image reconstruction and the potential costs of a. Proc Nat Acad Sci 2020.

[98] Swinder N, De-Arteaga M, auz NTHI, Leisner MD, Kalai AT. What are the biases in my word embedding? arXiv preprint arXiv:1812.08769 2019.

[99] Kendall A, Gal Y. What uncertainties do we need in bayesian deep learning for computer vision? arXiv preprint arXiv:1703.04977 2017.

[100] MacKay DJ. A practical bayesian framework for backpropagation networks. Neural Comput 1992;4(3):448–72.

[101] Gal Y, Ghahramani Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: International conference on machine learning. PMLR; 2016, p. 1050–1059.

[102] Blundell C, Cornebise J, Kavukcuoglu K, Wierstra D. Weight uncertainty in neural network. In: International conference on machine learning. PMLR; 2015, p. 1613–1622.

[103] Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. arXiv preprint arXiv:1612.01474 2016.