ITUPublications

International Telecommunication Union
Standardization Sector

ITU-T Focus Group Deliverable

(03/2023)

Focus Group on Artificial Intelligence for Health (FG-AI4H)

FG-AI4H DEL5.3

Data annotation specification

ITU-T FG-AI4H Deliverable DEL5.3

Data annotation specification

Summary

Data annotation would be one of the most dependable factors on model performance, it serves as one important aspect of data quality control on artificial intelligence for health (AI4H). ITU-T FG-AI4H Deliverable DEL5.3 gives a general guideline of data annotation specification, including definition, background and goals, framework, standard operating procedure, scenario classifications and corresponding criteria, as well as recommended metadata and so on. A questionnaire is attached to seek input and collaboration with topic groups regarding data annotation.

Keywords

Artificial intelligence, dataset, data annotation, health, metadata, testing data.

Note

This is an informative ITU-T publication. Mandatory provisions, such as those found in ITU-T Recommendations, are outside the scope of this publication. This publication should only be referenced bibliographically in ITU-T Recommendations.

Change log

This document contains Version 1 of the Deliverable DEL5.3 on "*Data annotation specification*" approved on 16 March 2023 via the online approval process for the ITU-T Focus Group on AI for Health (FG-AI4H).

Editor:	Shan Xu CAICT, China	E-mail: xushan@caict.ac.cn
	Sebastian Bosse Fraunhofer HHI, Germany	E-mail: sebastian.bosse@hhi.fraunhofer.de
	Jianrong Wu Tencent Healthcare, China	E-mail: edwinjrwu@tencent.com

Contributors: (in alphabetical order)

Yanwu Xu
Intelligent Healthcare Unit,
Baidu, China
E-mail: xuyanwu@baidu.com

Nathan Guo
ShuKun Technology, China
E-mail: guoning@shukun.net

Yajun Zhang
Tencent Technology (Shenzhen),
China
E-mail: yajunzhang@tencent.com

Chunyang Niu
Zhejiang Lab, China
E-mail: niucy@zhejianglab.com

Guoqiang Li
Zhejiang Lab, China
E-mail: gli@zhejianglab.com

Huihui Fang
Artificial Intelligence Innovation
Business, Baidu, China
E-mail: fanghuihui@baidu.com

Harpreet Singh
ICMR, India
E-mail: hsingh@bmi.icmr.org.in

Table of Contents

ITU-T FG-AI4H Deliverable DEL5.3

Data annotation specification

1 Scope

Within the context of data quality for artificial intelligence applied in health, this deliverable gives a general guideline of data annotation specification, including inter alia definition, background and goals, framework, standard operating procedure (SOP), scenario classifications and corresponding criteria, as well as recommended metadata.

2 References

[ISO/IEC 2382] ISO/IEC 2382:2015, *Information technology – Vocabulary*.
https://www.iso.org/standard/63598.html

[GHTF/SG1/N071] GHTF/SG1/N071:2012, *Definition of the Terms 'Medical Device' and 'In Vitro Diagnostic (IVD) Medical Device'*.
https://www.imdrf.org/sites/default/files/docs/ghtf/final/sg1/technical-docs/ghtf-sg1-n071-2012-definition-of-terms-120516.pdf

[SAMD/N12] IMDRF/SaMD WG/N12FINAL:2014, *"Software as a Medical Device": Possible Framework for Risk Categorization and Corresponding Considerations*.
https://www.imdrf.org/sites/default/files/docs/imdrf/final/technical/imdrf-tech-140918-samd-framework-risk-categorization-141013.pdf

3 Definitions

3.1 Terms defined elsewhere

This Technical Report uses the following terms defined elsewhere:

3.1.1 artificial intelligence [ISO/IEC 2382]: Branch of computer science devoted to developing data processing systems that perform functions normally associated with human intelligence such as reasoning, learning, and self-improvement.

3.1.2 machine learning [ISO/IEC 2382]: Automatic learning process by which a functional unit improves its performance by acquiring new knowledge or skills, or by reorganizing existing knowledge or skills.

3.1.3 medical device [GHTF/SG1/N071]: Any instrument, apparatus, implement, machine, appliance, implant, reagent for in vitro use, software, material or other similar or related article, intended by the manufacturer to be used, alone or in combination, for human beings, for one or more of the specific medical purpose(s) of: a) diagnosis, prevention, monitoring, treatment or alleviation of disease; b) diagnosis, monitoring, treatment, alleviation of or compensation for an injury; c) investigation, replacement, modification, or support of the anatomy or of a physiological process; d) supporting or sustaining life; e) control of conception, f) disinfection of medical devices; g) providing information by means of in vitro examination of specimens derived from the human body; and does not achieve its primary intended action by pharmacological, immunological or metabolic means, in or on the human body, but which may be assisted in its intended function by such means.

3.1.4 software as a medical device [SaMD/N12]: Software intended to be used for one or more medical purposes that perform these purposes without being part of a hardware medical device.

3.2 Terms defined in this Technical Report

This Technical Report defines the following terms:

3.2.1 controlled vocabulary: An organized arrangement of words and phrases used to index content and to retrieve content through browsing or searching.

3.2.2 data annotation: Perform operations such as categorizing, sorting, editing, marking, and annotating on the data to be labelled such as images, and add tags to the data to generate machine-readable data codes that meet the requirements of machine learning training.

3.2.3 metadata: Data that provides information about other data.

3.2.4 supervised learning: The machine learning task of learning a function that maps an input to an output based on example input-output pairs.

3.2.5 unsupervised learning: A type of machine learning that looks for previously undetected patterns in a data set with no pre-existing labels and with a minimum of human supervision.

4 Abbreviations and acronyms

This Technical Report uses the following abbreviations and acronyms:

AI	Artificial Intelligence
AI4H	Artificial Intelligence for Health
FG-AI4H	Focus Group on Artificial Intelligence for Health
JSON	JavaScript Object Notation
ML	Machine Learning
SOP	Standard operating procedure
XML	Extensible Markup Language

5 Conventions

None.

6 Background and goals

The great potential of digital technologies especially machine learning (ML) and artificial intelligence (AI) are recognized to revolutionize the fields of medicine and public health in an unprecedented manner. While holding great promise, this rapidly developing field raises a number of uncertainties, for example, if the model is poorly designed or the underlying training data are biased or incorrect, errors or problematic results can occur. AI technology can only be used with complete confidence if it has been quality controlled through a rigorous evaluation in a standardized way. Among all the quality controls, data annotation would be one of the most dependable factors on model performance. In the case of mislabelled or inaccurate training instances, it is difficult for the supervised model to obtain the expected results. Many annotation tools exist, as exemplified in [b-Sangkuhl et al., 2020] to [b-Kipp 2001] but lack a consistent approach. The US Food and Drug Administration addressed some of these issues in [b-US FDA].

Quality control on data annotation is a factor that is easily overlooked but crucial to the model performance. It is especially critical to models based on large-scale datasets. Therefore, this deliverable addresses the following points:

– To assist the quality control of data annotation from standard operating procedure.

– To reduce model performance problems caused by inconsistent data annotations.

– To enable large-scale dataset projects on a high diversity of data formats and multi-annotators.
– To facilitate the training and education for non-professional annotators and improve common understandings.

7 Framework

Data annotation is one of the most dependable factors in the performance of supervised machine models. If the annotation for machine learning is incorrect, the decision rules built by the machine will be biased. As part of the entire artificial intelligence for health (AI4H) project, data annotation works as shown in Figure 1. During the testing and evaluation of the supervised machine models, unqualified annotations may be identified, which should be relabelled or deleted from the dataset.

With the help of annotators and annotation tools, a standard operating procedure of data annotation can convert input dataset into qualified annotations for supervised machine learning. This standard operating procedure is discussed in clause 8 in detail.

The information from the data annotation process and the raw dataset can be used for training dataset for supervised machine learning and optimization, as well as testing dataset for the evaluation process. Therefore, data annotation has a very close relationship to the above core process of the AI4H model, and as a result, is recognized as one of the most dependable factors on the model performance.

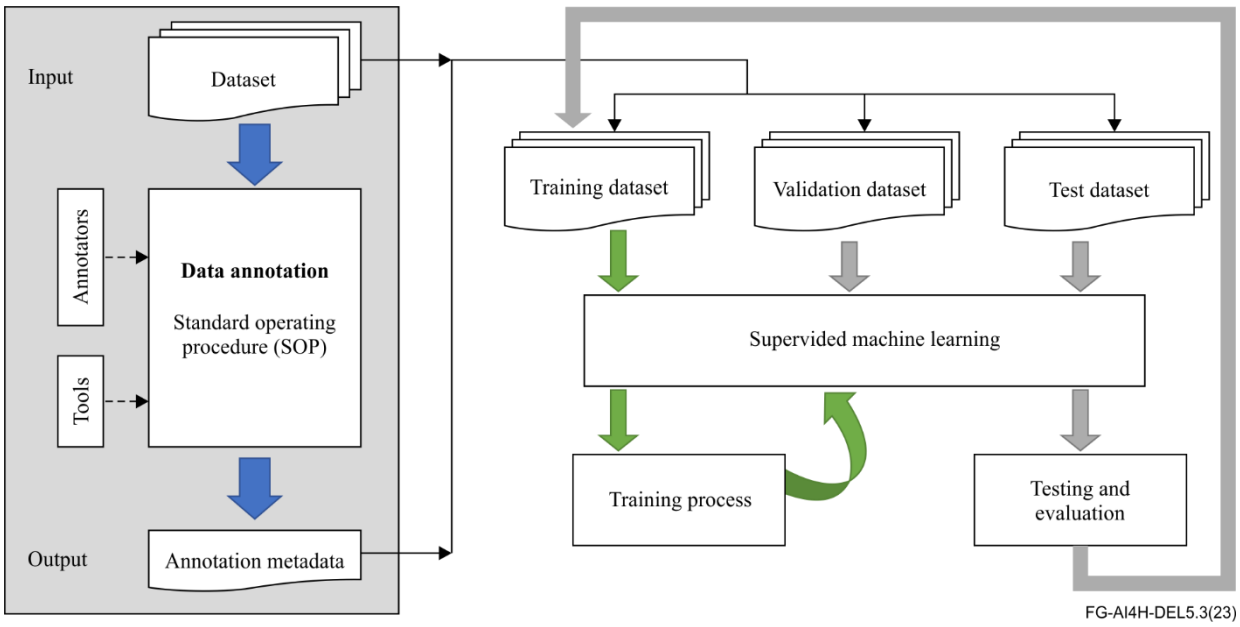


Figure 1 – Framework of data annotation and its external relations

8 Standard operating procedure

To establish a unified understanding and quality control mechanism, a standard operating procedure (SOP) is recommended. Figure 2 illustrates a formulated process of data annotation with much feasibility through variables and configurable thresholds.

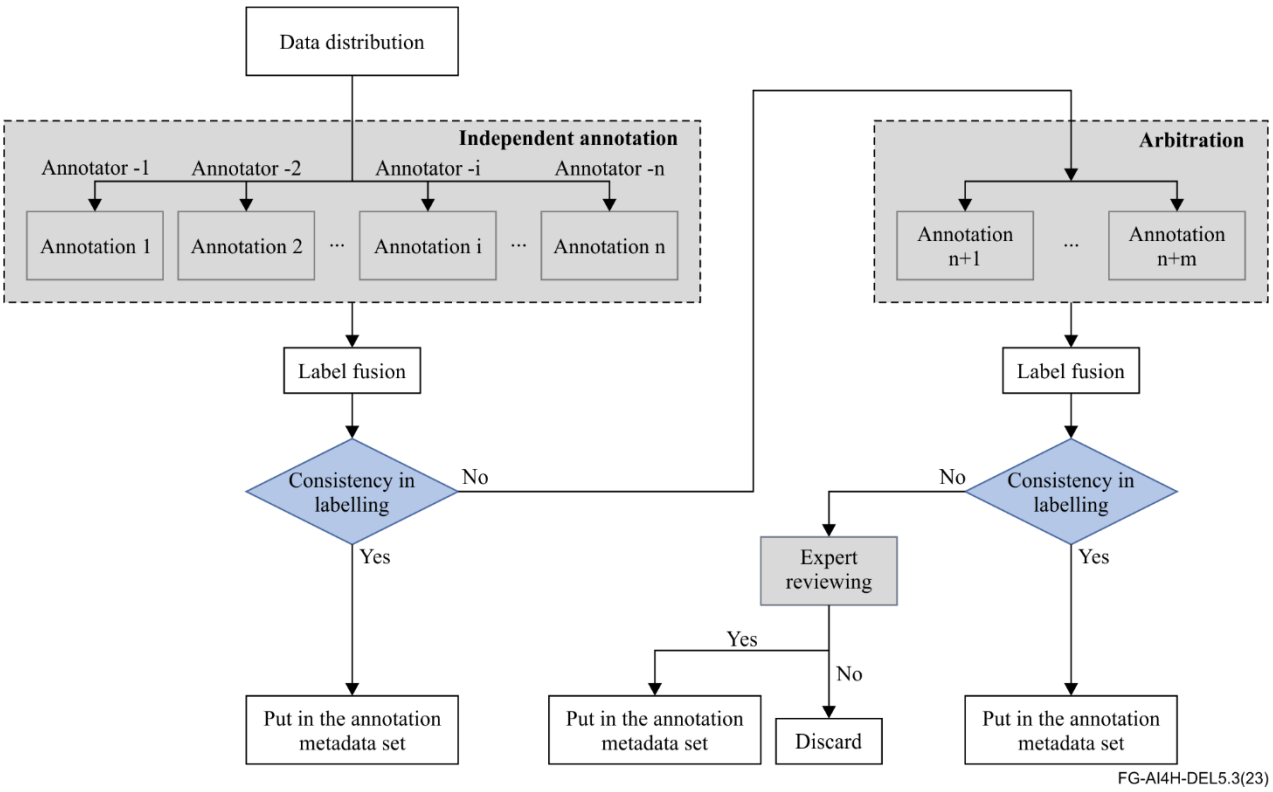


Figure 2 – Data annotation procedure

8.1 Independent annotation

The data annotation process starts with an independent annotation, represented by the left grey box in Figure 2, each instance in the dataset needs to be labelled by all or part of the annotators independently. To avoid bias in the data distribution, it is suggested that the process is carried out by grouping and crossing, and ensuring the effective resolution of the inconsistencies. Several annotators (represented as variable *n* in the figure) are invited to label the raw dataset. Certain qualifications are required on the annotators, for example, doctors and trained annotators in specific case domains.

However, for cost considerations, some projects will also set up one annotator (Set *n* to 1) in this parallel independent annotation part, and then goes to arbitration if it encounter difficulties.

8.2 Arbitration

In the above independent annotation part, if there is an inconsistency that is unacceptable, or there are difficulties and uncertainties in a single annotator setting, an additional annotator with more experience should be introduced for the arbitration, i.e., represented by the upper right grey box in Figure 2. Stricter requirements on the arbitration annotator qualifications include for example, a doctor with more than three years of experience in the case domain.

8.3 Expert reviewing

The expert reviewing is represented by the lower right grey box in Figure 2. This final review deals with some very tricky cases that cannot reach an agreement in the previous steps. Annotations confirmed by review experts will be marked as the final answer, and cases not approved could be considered to be sent back to the arbitration process and arbitrated by another arbitration expert. Stricter requirements on the expert qualifications include for example, having 5 years of experience or more.

8.4 Decision making box

Represented by the blue boxes in Figure 2, the judgment and decision making on labelling consistency cannot be avoided anywhere in the independent annotation, arbitration and expert reviewing. The simple mechanism should be, if the consistency satisfies the specific requirements, such as reaching a certain threshold or a combination of conditions, the annotation shall be saved with confidence; if the consistency does not satisfy the specific requirements, i.e., does not reach a certain threshold or a combination of conditions, the annotation will be discarded. Therefore, the criteria of consistency and corresponding requirements should be identified, and they are usually designed according to different scenarios, with more details given in clause 9.

8.5 Annotators training and assessment

With the continuous popularity of the AI4H model, we may expect a future with more mature and extensive mechanisms for annotators' engagement. In addition to the experienced doctors mentioned above, candidates with no professional qualifications but are well-trained and are quantitatively assessed are also possible to be invited in the process of data annotation.

The training and assessment of annotators may include the following ways:

– Gold standard materials: Data annotation made by review experts or arbitration groups can be seen as gold standards, a unified document with examples can be developed as a reference to teach candidates on how to achieve the tasks.

– Training courses: In addition to paper documents, training courses are also an effective way to educate candidates and reach a common understanding of data annotation tasks, especially in large-scale dataset and numerous annotators.

– Quantitative assessment: To evaluate the performance of different annotators, examinations and certifications with specific evaluation metrics can be conducted. Only after the corresponding evaluation, metrics are calculated with the gold standard and the annotator's results reaches a certain requirement. For example, beyond a certain threshold, a candidate can be assigned to the annotation task being certificated.

8.6 Variable description

Variables and configurable threshold in this procedure are listed here for your convenience.

– Number of independent annotators

– Number of arbitration experts

– Number of review experts

– Different options on consistency criteria (usually the same in clauses 8.1-8.3)

– Configurable requirement or threshold on consistency criteria in the independent annotation

– Configurable requirement or threshold on consistency criteria in the arbitration

– Configurable requirement or threshold on consistency criteria in the expert reviewing

9 Consistency judgement

For the decision box in Figure 2, different criteria for consistency are selected according to different application scenarios. The main considerations are from two perspectives, first is the input data type, elaborated in clause 9.1; the second is the output requirement for AI4H models, elaborated in clause 9.2. Under these two different classification dimensions, the options on consistency criteria will be different, as elaborated in clause 9.3.

9.1 Input data type classification

Biomedical information evolved with the medicine practice and engineering technologies at an unprecedented speed through the medical images obtained by human body imaging, high-resolution viewing of cells, and pathological specimens. Modalities covered in common measurement include X-ray, ultrasound, magnetic resonance (MR), X-ray computed tomography (CT), nuclear medicine, and high-resolution microscopy, etc. Table 1 refers to their specific information.

Table 1 – Summary of common medical measurement modalities

	Dimensionality	Description	Anatomies
X-ray	2D, 2D+t	Produces images by measuring the attenuation of X-rays through the body, via a detector array [b-Bushberg et al., 2012]	Most organs
CT	2D, 3D, 3D+t	Creates 2D cross-sectional images of the body by using a rotating X-ray source and detector [b-Hsieh 2009]	Most organs
Ultrasound	2D, 2D+t, 3D, 3D+t	A transducer array emits acoustic pulses and measures the echoes from the tissue scatters [b-Bushberg et al., 2012]	Most organs
MRI	3D, 3D+t	Use a magnetic field to align protons; RF and gradient pulses are used to selectively excite protons in the tissues and blood in order to measure their spatially encoded unclear magnetic resonance signals [b-Panayides et al., 2020]	Most organs
Nuclear	2D, 3D, 3D+t	Measures the emission of gamma rays through the decay of radioisotopes introduced into the body via external detectors / gamma cameras. [b-Bushberg et al., 2012]	All organs with radioactive tracer uptake
Microscopy	2D, 3D, 3D+t	Typically uses an illumination source and lenses to magnify specimens before capturing an image [b-Bushberg et al., 2012]	Primarily biopsies and surgical specimens

Based on the above common medical measurement modalities, a classification of input data modalities for AI4H tasks are given in Table 2, with text and numbers added in specific cases of case history descriptions and blood pressures or respiratory rates, etc.

Table 2 – Summary of input data modalities for AI4H tasks

Data	Dimensionality	Description	Examples
Image	2D	Two-dimensional medical imaging	– Fundus photos
3D images	3D	Three-dimensional spatial imaging	– Sets of CT slices
4D	4D (3D+t)	3D space imaging changes over time	– Heart film imaging
Video	2D+t	Camera or monitor recording	– Falls among the elderly
Audio/ signal	1D+t	Sound or transmitted in signal form	– Heart sound / ECG
Text	1D, 2D	Structured / unstructured description in words	– Case history, diagnosis extraction
Single number	1D	Single measurement data	– Blood pressure or respiratory rate

9.2 Output requirement classification

When the final output requirements of models are different, even if it is the same input data format, data annotations will be different. Different output requirements include classification, detection, segmentation, localization, etc. Corresponding descriptions and examples are given in Table 3.

Table 3 – Output requirements

Task	Description	Examples
Classification	The problem of classifying instances into two or more classes	– Identify abnormal tissue – Diabetic retinopathy grade
Detection	Identify an object, usually marked with a rectangle for further processing	– Detect the position of a coronary plaque for further processing
Segmentation	Separate certain lesions and draw the specific outline of the lesion	– Tumour segmentation
Localization	Calculate the central coordinate of the anatomical structure	– Localize the optic disc or macular fovea for further analysis of ocular fundus diseases

9.3 Criteria option matrix

With the above two dimensions, a matrix can be developed according to different data input formats and model output requirements. This matrix can act as a reference for selecting criteria options. Details are shown in Table 4, and other scenarios are to be added to cover all possible use cases in the FG and the AI4H industry.

Table 4 – Criteria options in different scenarios

<table>
<tr><th>Task
Data type</th><th>Classification</th><th colspan="2">Detection</th><th>Segmentation</th><th>Localization</th></tr>
<tr><td>Image</td><td rowspan="8">Type 1: Classification</td><td colspan="2">Type 2: Detection and segmentation of images</td><td></td><td>Type 3: Localisation</td></tr>
<tr><td>3D images</td><td colspan="2">(a) slicing 3D data into different 2D views before fusing to obtain a final detection or segmentation regions
(b) exploit the 3D data by using architectures that perform 3D convolutions and then train the network from scratch on 3D medical images
[b-Dou et al., 2016][b-Ding et al., 2017][b-Korolev et al., 2017][b-Nie et al., 2016].</td><td></td><td>(a) slicing 3D data into 2D views to obtain the regions of the target object before calculation of the final position coordinate
(b) exploit the 3D data by using architectures that perform 3D convolutions and then train the network from scratch on 3D medical images</td></tr>
<tr><td>4D</td><td colspan="2">Condensing the 4D data into three dimensions</td><td></td><td>–</td></tr>
<tr><td>Video</td><td colspan="2">Condensing the 2D+t data into three dimensions</td><td></td><td>–</td></tr>
<tr><td>Audio / signal</td><td>–</td><td>–</td><td>–</td><td>–</td></tr>
<tr><td>Text</td><td>–</td><td>–</td><td>–</td><td>–</td></tr>
<tr><td>Single number</td><td>–</td><td>–</td><td>–</td><td>–</td></tr>
</table>

Type 1: Classification

For this type criteria like Cohen's kappa, weighted kappa, Fleiss' kappa, and Krippendorff's alpha are recommended for classification tasks. The detailed calculation methods are shown in Table 5.

– Cohen's kappa: Cohen's kappa coefficient (κ) is a statistic that is used to measure inter-rater reliability for qualitative items. It is generally thought to be a more robust measure than simple percent agreement calculation, as κ considers the possibility of the agreement occurring by chance.

– Weighted kappa: Weighted kappa allows disagreements to be weighted differently and is especially useful when codes are ordered. Three matrices are involved, the matrix of observed scores, the matrix of expected scores based on chance agreement, and the weight matrix.

– Fleiss' kappa: Fleiss' kappa is a statistical measure for assessing the reliability of agreement between a fixed number of raters when assigning categorical ratings to a number of items or classifying items. This contrasts with other kappas such as Cohen's kappa, which only work when assessing the agreement between not more than two raters or the interrater reliability for one appraiser versus themselves.

– Krippendorf's alpha: Krippendorf's alpha is an assessment of inter-rate reliability dealing with missing data, various sample sizes, categories and numbers of raters, and any type of measurement level. It can be seen as a generalization of Fleiss' kappa (and others).

Type 2: Detection and segmentation of images

For this type, criteria like the Jaccard index and Dice's coefficient are recommended to use for the detection and segmentation of images. Detailed calculation methods are shown in Table 6.

– The Jaccard index: Jaccard index is also known as the intersection over union (IoU) and the Jaccard similarity coefficient, which is a statistic used for gauging the similarity and diversity of sample sets.

– Dice's coefficient: Dice's coefficient is the quotient of similarity and ranges between 0 and 1. This coefficient is not very different in form from the Jaccard index, and they have a connection as J=D/(2-D), D=2J/(1+J).

Type 3: Localization

For this type, criteria like euclidean distance (ED) are recommended to use for localisation. Euclidean distance is a commonly used definition of distance, which refers to the true distance between two points in the m-dimensional space.

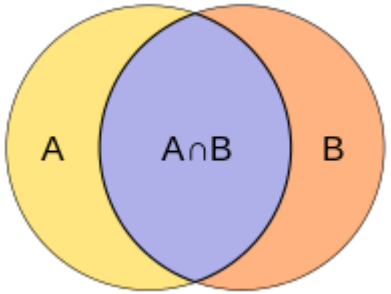
– In 2D space: $ED\big((x_1, y_1), (x_2, y_2)\big) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$.

– In 3D space: $ED\big((x_1, y_1, z_1), (x_2, y_2, z_2)\big) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}$

Table 5 – Criteria calculation for classification

Criteria	Situation	Calculation method	Parameter explanation
Cohen's kappa	Assessing the agreement between not more than two raters or the interrater reliability for one appraiser versus themselves.	$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e}$ If the raters are in complete agreement then kappa =1; If there is no agreement among the raters other than what would be expected by chance kappa =0. It is possible for the statistic to be negative which implies that there is no effective agreement between the two raters or the agreement is worse than random.	where p_o is the relative observed agreement among raters (identical to accuracy), and p_e is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly seeing each category. $p_e = \frac{1}{N^2} \sum_k n_{k1} n_{k2}$
Weighted kappa	Allows disagreements to be weighted differently and is especially useful when codes are ordered.	$\kappa = 1 - \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} x_{ij}}{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} m_{ij}}$	where k is the number of codes and w_{ij}, x_{ij}, and m_{ij} are elements in the weight, observed, and expected matrices, respectively. The weights in the diagonal cells are all 1 (i.e., $w_{ij} = 1$, for all i), and the weights in the off-diagonal cells range from 0 to <1 (i.e., $0 \leq w_{ij} < 1$, for all $i \neq j$).
Fleiss' kappa	Assessing the reliability of agreement between a fixed number of raters when assigning categorical ratings to a number of items or classifying items.	$\kappa = \frac{\overline{P} - \overline{P}_e}{1 - \overline{P}_e}$ If the raters are in complete agreement, then Fleiss' kappa =1. If there is no agreement among the raters (other than what would be expected by chance) then Fleiss' kappa <0.	The factor $1 - \overline{P}_e$ gives the degree of agreement that is attainable above chance, and $\overline{P} - \overline{P}_e$ gives the degree of agreement actually achieved above chance.
Krippendorf's alpha	Assessment of inter-rate reliability dealing with missing data, various sample sizes, categories and numbers of raters, and any type of measurement level. Generalization of Fleiss' kappa (and others)	$\alpha = 1 - \frac{D_o}{D_e} D_0 = \frac{1}{n} \sum_{c \in R} \sum_{k \in R} \delta(c,k) \sum_{u \in U} m_u \frac{n_{cku}}{P(m_u, 2)} D_e$ $= \frac{1}{P(n,2)} \sum_{c \in R} \sum_{k \in R} \delta(c,k) P_{ck} P_{ck}$ $= \begin{matrix} n_c n_k, if c \neq k \\ n_c(n_c - 1), if c = k \end{matrix}$	D_o: Disagreement observed D_e: Disagreement expected by chance δ: Metric function n: Number of pairable elements m_u: Number of items per unit/sample $n_{c,k,u}$: Number of pairs in unit u P: Permutation function $P_{c,k}$: Number of permutations in pair (c,k)

Table 6 – Criteria calculation for image detection and segmentation

Criteria	Calculation method	Graphical representation
Jaccard index	Numerator represents the area of overlap between two annotations; Denominator represents the area encompassed by two annotations. Dividing the area of overlap by the area of union yields our final score. $$J(A,B) = \frac{\lvert A \cap B \rvert}{\lvert A \cup B \rvert} = \frac{\lvert A \cap B \rvert}{\lvert A \rvert + \lvert B \rvert - \lvert A \cap B \rvert}$$	 J=D/(2-D), D=2J/(1+J)
Dice's coefficient	Numerator represents the double area of overlap between two annotations; Denominator represents the sum of two annotation areas. Dividing the area of overlap by the sum area yields our final score. $$D = \frac{2\lvert A \cap B \rvert}{\lvert A \rvert + \lvert B \rvert}$$	

9.4 Post-processing of the annotations

After the criteria calculation and consistency judgment, different post-processing methods on annotations that are acceptable as consistent will also cause different results. For example, calculate the average value of the marked results (x, y, w, h) or a maximum area with a consistency above the threshold in an image.

10 Recommended metadata

Metadata is considered to be the output of the data annotation process, all the necessary information for the annotation process should be included in the metadata. A metadata format is given in Table 7, further details are for further study.

11 Output file

The output files include the annotation documents and the origin images. The formats of the annotation documents include but are not limited to extensible markup language (XML), JavaScript object notation (JSON), text, etc. The annotation documents should include at least three items, i.e., image path, object name, and object coordinates. Supporting documents may be given if it's necessary to interpret the annotation information. Annex B gives an example of the annotation document for endoscopic images.

12 File saving

Both the images and the documents should be named according to the same rules for easy querying. For example, they can be named with the number of the classification of the object, and the document's name is the same as the corresponding origin images.

Table 7 – Recommended metadata

<table>
<tr><th>Type</th><th colspan="5">Content</th></tr>
<tr><td>General information</td><td colspan="5">a. Institution and responsible or corresponding PI
b. Construction dates of annotation dataset
c. Regulatory aspects (e.g., data privacy)
d. Description of use case</td></tr>
<tr><td>Annotation procedure information</td><td colspan="5">e. Details on data annotation process (annotator number, experts group setting, tools, etc.)
f. Achieved consistency and criteria employed
g. Post-processing method employed on annotations
h. Ontology employed
i. Label list or description</td></tr>
<tr><td>Data acquisition information</td><td colspan="5">j. Collection device model
k. Collection frame rate / sampling rate</td></tr>
<tr><td>Instance information</td><td colspan="5">l. Instance identification code
m. Patient information (age, gender)
n. Diagnosis information (symptoms)</td></tr>
<tr><td rowspan="8">Annotation information</td><td>Task
Data type</td><td>Classification</td><td>Detection</td><td>Segmentation</td><td>Localization</td></tr>
<tr><td>– Image</td><td rowspan="6">class labels</td><td rowspan="5">– signal instance</td><td rowspan="5">– label per instance</td><td rowspan="5">– coordinate label of signal instance</td></tr>
<tr><td>– 3D images</td></tr>
<tr><td>– 4D</td></tr>
<tr><td>– Video</td></tr>
<tr><td>– Audio / signal</td></tr>
<tr><td>– Text</td><td>–</td><td>– label per word, intent, or sentence</td><td>–</td></tr>
<tr><td>– Single number</td><td>–</td><td>–</td><td>–</td><td>–</td></tr>
</table>

Annex A

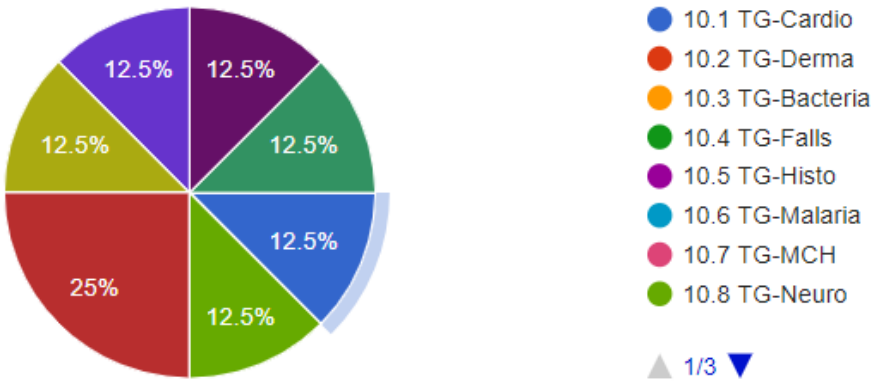
Questionnaire on data annotation

By Google Form: https://forms.gle/3fYrm3SZSrNQu3eeA

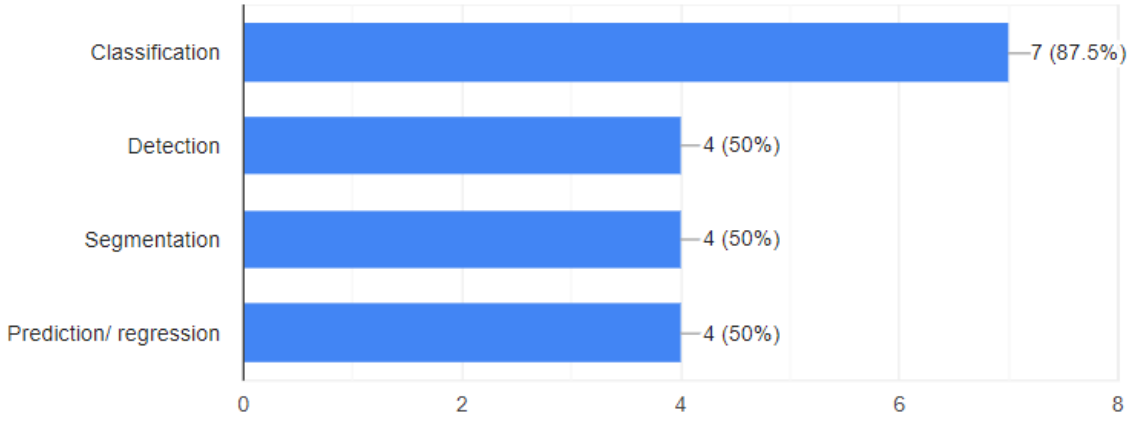
This questionnaire aims to gather insights into the current practices, and the specific requirements of data annotation in the Focus Group on Artificial Intelligence for Health (FG-AI4H) topic groups and AI4H products.

Your input and suggestions will be of great value to us in forming a data annotation specification together, as one of the deliverables with the FG-AI4H. We would appreciate it if you could take the time to complete the questionnaire, or if you have further ideas, please feel free to contact us. (xushan@caict.ac.cn; sebastian.bosse@hhi.fraunhofer.de, edwinjrwu@tencent.com)

1. To which topic group are you contributing?



2. Which annotation task category is relevant for your project within the topic group?



3. Is there any gold standard (or state-of-the-art task intervention method) relevant to your project within the topic group?

Histology, Cross-image validation, human annotations

Pathological report, Cross annotation by doctors

Snake expert (herpetologist) identification

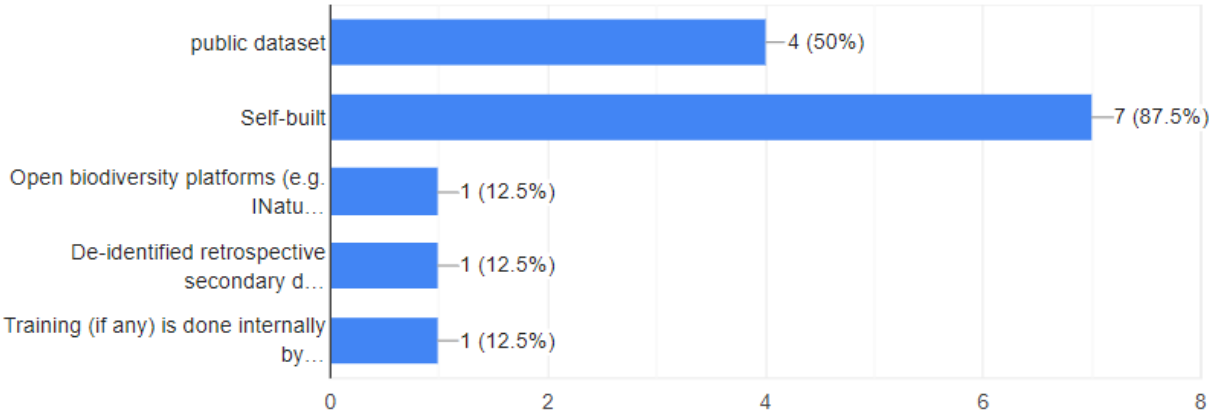
The gold standard photography method for the detection of Diabetic Retinopathy is stereoscopic color fundus photography in 7 standard fields (30°) as defined by the Early Treatment Diabetic Retinopathy Study (ETDRS) group.

Post-mortem pathology evaluation is the gold standard. unfortunately few data are available with post-mortem evaluation. In the absence of such data, biological markers provide a more reliable alternative to the more uncertain clinical diagnostic based on symptoms only.

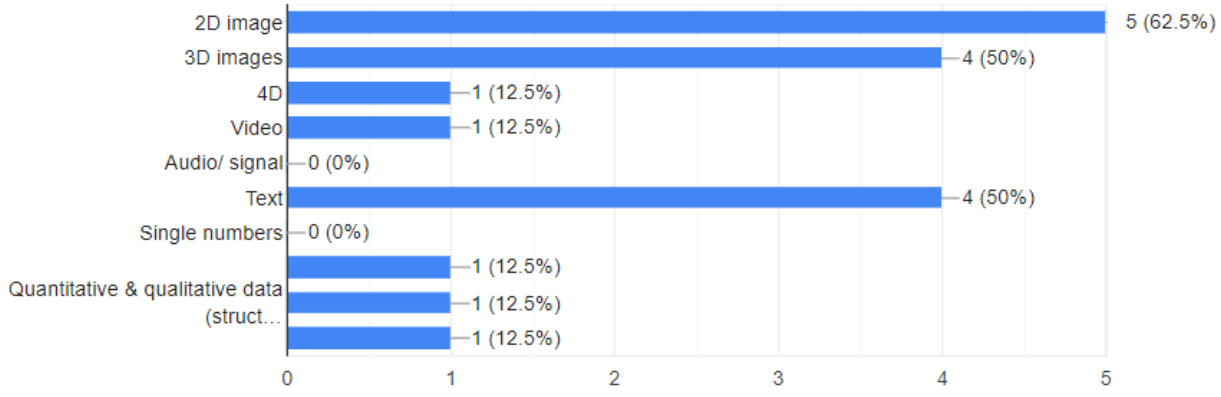
Yes

The average doctor opinion is the current gold-standard (even cases confirmed later with more evidence cannot be used to judge the "correct" answer for less evidence).

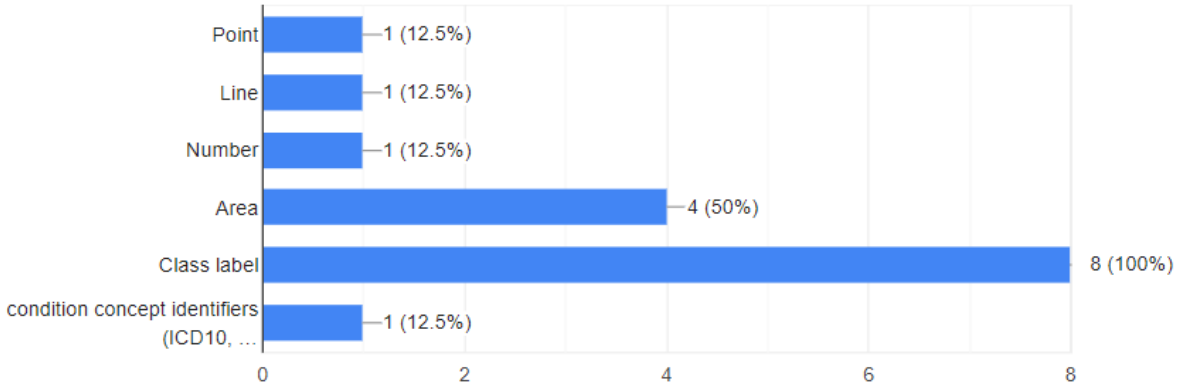
4. What is your data source for the training and testing dataset? (Appreciate it if you can provide more info.)



5. Which signal or data modalities are relevant for your project with the topic group?



6. What is the nature of the annotation?



7. What kind of annotation procedure are you using? any annotation tool that you use?

all of the above, custom made tool

Cross annotation, Self-built annotation tool

Expert identification, crowdsourcing

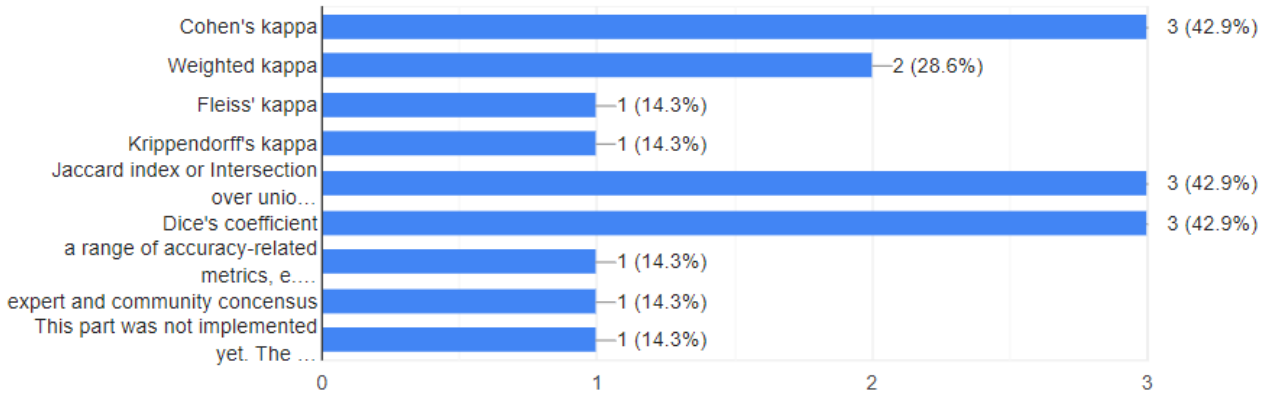
For Diabetic Retinopathy and other conditions, annotation is usually provided by an ophthalmologist's diagnosis assigned to the image.

Manual

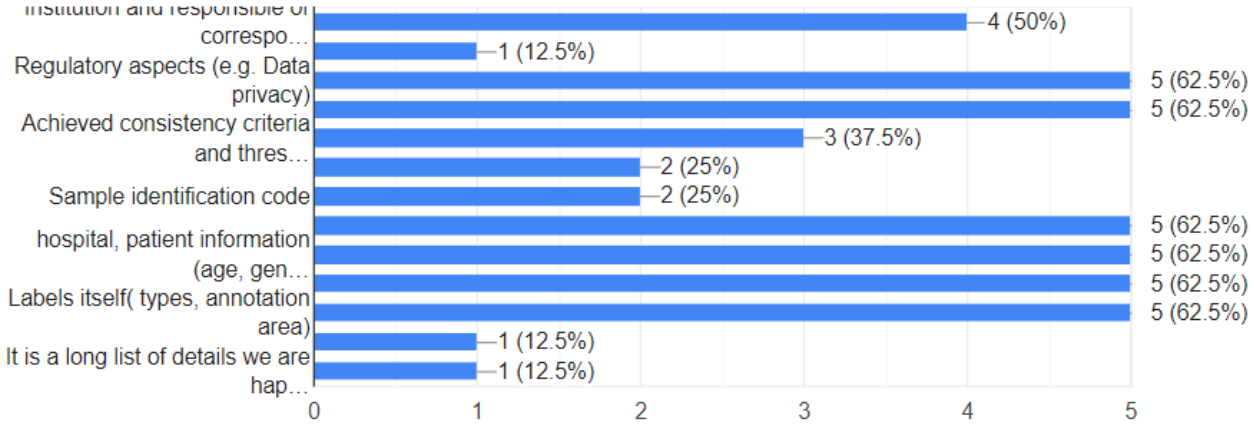
Structured data are used, thus simple R programming is used to recode structured data to required standardized labels.

Most companies have their own annotation tool. For the group we are developing a new case-creation tool where doctors enter the symptoms and expected conditions in a semantically structured way.

8. What annotation quality criterions are currently used in your topic group?



9. What kind of metadata do you consider relevant for data annotation?



10. What type of ontology are you using? (If any)

-

Medical ICD10 classification of diseases.

N/A

not finally decided yet, but most likely SnomedCT - maybe also HPO, ICD10 and we will in any case need to filter and also extend them. We will also need new ontologies e.g. for triage levels.

11. Which additional information do you need to encode the actual meaning of the annotation?

ideally also dental history and tooth specific clinical findings

-

The following paper provides additional indicators: "Guidelines for Data Acquisition, Quality and Curation for Observational Research Designs (DAQCORD)" (https://t.co/brhboFYI54?amp=1)

N/A

Annex B

Examples of endoscopic image metadata

There is an example of metadata of an endoscopic image. The annotation results include:

Attributes	Examples
json version	–
folder number	Trial_Date_Hospital_Batch
filename	PatientNum_Uniquecode.png
file path	/Data/Endoscopic/Trial_Date_Hospital_Batch/ PatientNum_Uniquecode.png
source	Hospital Equipment
size: width/height/depth	1280 × 720

segmented object: name, coordinates	<pre>"label_classification": [
	{
	"name": "lesion",
	"value": "polyp"
	},
	{
	"name": "modality",
	"value": "colonoscopy"
	}
	],
	"label_segmentation": [
	{
	"name": "lesion",
	"type": "2d_mask",
	"label": [
	{
	"name": "lesion",
	"value": "polyp"
	}
	],
	"value":
	"/Data/Endoscopic/Trial_Date_Hospital_Batch/SegNum
	_PatientNum_Uniquecode.nii.gz"
	}
	],
	"label_detection": [
	{
	"name": "lesion",
	"type": "2d_boundingbox",
	"label": [
	{
	"name": "lesion",
	"value": "polyp"
	}
	],
	"value": [123, 223, 40, 50]
	}
	]</pre>

Bibliography

NOTE – From [b-Sangkuhl et al., 2020] onwards are references on several data annotation tools.

[b-Bushberg et al., 2012] Bushberg, J. T., Seibert, J. A., Leidholdt E. M. Jr., Boone, J. M. (2012), *The Essential Physics of Medical Imaging*. Philadelphia, PA, USA: Lippincott Williams & Wilkins. https://pubmed.ncbi.nlm.nih.gov/28524933/

[b-Dou et al., 2016] Dou, Q., Chen, H., Yu, L., Zhao, L., Qin, J., Wang, D., Mok, V. C. T., Shi, L., Heng, P-A. (2016), *Automatic detection of cerebral microbleeds from MR images via 3D convolutional neural networks*. IEEE Transactions on Medical Imaging., vol. 35, no. 5, pp. 118-1195. https://ieeexplore.ieee.org/document/7403984/authors#authors

[b-Ding et al., 2017] Ding, J., Li, A., Hu, Z., and Wang, L. (2017), *Accurate pulmonary nodule detection in computed tomography images using deep convolutional neural networks*. https://arxiv.org/abs/1706.04303

[b-Hsieh 2009] Hsieh Jiang. (2009), *Computed Tomography Principles, Design, Artifacts, and Recent Advances*, 2nd Edition. https://www.wiley.com/en-us/Computed+Tomography+Principles%2C+Design%2C+Artifacts%2C+and+Recent+Advances%2C+2nd+Edition-p-9780470563533

[b-Korolev et al., 2017] Korolev, S., Safiullin, A., Belyaev, M. and Dodonova, Y. (2017), *Residual and plain convolutional neural networks for 3D brain MRI classification*. https://arxiv.org/abs/1701.06643

[b-Nie et al., 2016] Nie, D., Zhang, H., Adeli, E., Liu, Luyan, and Shen, D. (2016), *3D deep learning for multi-modal imaging-guided survival time prediction of brain tumor patients*. pp. 212-220. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5278791/

[b-Panayides et al., 2020] Panayides, A. S., Amini, A., Filipovic, N. D., Sharma, A., Tsaftaris, S. A., Young, A., Foran, D., Do, Nhan, Golemati, S., Kurc, T., Huang, K., Nikita, K. S., Veasey, B. P., Zervakis, M., Saltz, J. H., and Pattichis, C. S. (2020), *AI in Medical Imaging Informatics: Current Challenges and Future Directions*. IEEE Journal of Biomedical and Health Informatics, 24(7): 1837-1857. https://pubmed.ncbi.nlm.nih.gov/32609615/

[b-Sangkuhl et al., 2020] Sangkuhl, K., Whirl-Carrillo, M., Whaley, R. M., Woon, M., Lavertu, A., Altman, R.B., Carter, L., Verma, A., Ritchie M. D., and Klein, T. E. (2020), *Pharmacogenomics Clinical Annotation Tool (PharmCAT)*. Clinical Pharmacology & Therapeutics, 107(1): 203-210. https://pubmed.ncbi.nlm.nih.gov/31306493/

[Diete et al., 2017] Diete, A., Sztyler, T., Stuckenschmidt, H. (2017), *A smart data annotation tool for multi-sensor activity recognition*. IEEE Xplore. https://ieeexplore.ieee.org/document/7917542

[Ferreira et al., 2012] Ferreira, P. M., Mendonça, T., Rozeira, J., Rocha, P. (2012), *An annotation tool for dermoscopic image segmentation*. Proceedings of the 1st International Workshop on Visual Interfaces for Ground Truth Collection in Computer Vision Applications. https://dl.acm.org/doi/10.1145/2304496.2304501

[Hollink et al., 2003] Hollink, L., Schreiber, G., Wielinga, B. J. (2003), *Semantic annotation of image collections*. https://www.researchgate.net/publication/2939637_Semantic_Annotation_of_Image_Collections

[Hou et al., 2019] Hou, R., Denisenko, E., Forrest, A. R. R. (2019), *scMatch: a single-cell gene expression profile annotation tool using reference datasets*. Bioinformatics, 35(22):4688-4695. https://pubmed.ncbi.nlm.nih.gov/31028376/

[Iakovidis et al., 2014] Iakovidis, D. K., Goudas, T., Smailis, C., Maglogiannis, I. (2014), *Ratsnake: A versatile image annotation tool with application to computer-aided diagnosis*. The Scientific World Journal. https://www.hindawi.com/journals/tswj/2014/286856/

[b-Kipp 2001] Kipp Michael. (2001), *Anvil – A Generic Annotation Tool for Multimodal Dialogue*. 7th European Conference on Speech Communication and Technology. https://www.researchgate.net/publication/221487883_ANVIL_-_A_Generic_Annotation_Tool_for_Multimodal_Dialogue

[Kumar et al., 2007] Kumar, A., Jawahar, C. V. (2007), *Content-level annotation of large collection of printed document images*. Ninth International Conference on Document Analysis and Recognition (ICDAR 2007). IEEE, pp. 799-803. https://ieeexplore.ieee.org/document/4377025/authors#authors

[Karlgren et al., 2007] Karlgren, K., Dahlström, A., Lonka, K., and Ponzer, S. (2007), *A new educational annotation tool for supporting medical teams to improve their teamwork and communication*. Paper presented at the ICEM/ILE 2007 – The International Council for Educational Media. September 20-22, Nicosia, Cyprus.

[Karlgren et al., 2008] Karlgren, K., Dahlström, A., and Ponzer, S. (2008), *Design of an annotation tool to support simulation training of medical teams*. European Conference on Technology Enhanced Learning. Springer, Berlin, Heidelberg, pp. 179-184. https://link.springer.com/chapter/10.1007/978-3-540-87605-2_20#citeas

[Kim et al., 2011] Kim, E., Huang, X., Tan, G. (2011), *Markup SVG – An Online Content-Aware Image Abstraction and Annotation Tool*. IEEE Transactions on Multimedia, 13(5) pp. 993-1006. https://ieeexplore.ieee.org/document/5940230

[Kraljevic et al., 2019] Kraljevic, Z., Bean, D., Mascio, A., Roguski, L., Folarin, A., Roberts, A., Bendayan, R., and Dobson, R. (2019), *MedCAT – Medical Concept Annotation Tool*. (2019), arXiv:1912.10166. https://arxiv.org/abs/1912.10166

[Oronoz et al., 2013] Oronoz, M., Casillas, A., Gojenola, K., and Perez, A. (2013), *Automatic annotation of medical records in Spanish with disease, drug and substance names.* Iberoamerican Congress on Pattern Recognition. Springer, Berlin, Heidelberg. https://link.springer.com/chapter/10.1007/978-3-642-41827-3_67#citeas

[Philbrick et al., 2019] Philbrick, K. A., Weston, A. D., Akkus, Z., Kline, T. L., Panagiotis, K., Sakinis, T., Kostandy, P., Boonrod A., Zeinoddini, A., Takahashi, N., and Erickson, B. J. (2019), *RIL-Contour: a Medical Imaging Dataset Annotation Tool for and with Deep Learning*. Journal of Digital Imaging, 32, 571-581. https://link.springer.com/article/10.1007/s10278-019-00232-0#citeas

[Russell et al., 2007] Russell B. C., Torralba, A., Murphy, K. P., and Freeman, W. T. (2007), *LabelMe: A database and web-based tool for image annotation*. International Journal of Computer Vision, 77, 157-173. https://link.springer.com/article/10.1007/s11263-007-0090-8

[Rubin et al., 2008] Rubin, D., Mongkolwat, P., Kleper, V., Supekar, K., and Channin, D. (2008), *Medical Imaging on the Semantic Web: Annotation and Image Markup*. Association for the Advancement of Artificial Intelligence. https://aaai.org/papers/0019-medical-imaging-on-the-semantic-web-annotation-and-image-markup/

[Rubin et al., 2009] Rubin, D. L., Mongkolwat, P., Kleper, V., Supekar, K., and Channin, D. S. (2009), *Annotation and image markup: accessing and interoperating with the semantic content in medical imaging*. IEEE Intelligent Systems, 24(1) pp. 57-65. https://ieeexplore.ieee.org/document/4763656

[Ramos et al., 2015] Ramos, A. H., Lichtenstein, L., Gupta, M., Lawrence, M. S., Pugh, T. J., Saksena, G. Meyerson, M., Getz, G. (2015), *Oncotator: cancer variant annotation tool*. Human mutation, 36(4):E2423-E2429. https://pubmed.ncbi.nlm.nih.gov/25703262/

[Seifert et al., 2010] Seifert, S., Kelm, M., Moeller, M., Mukherjee, S., Cavallaro, A., Huber, M, Comaniciu, D. (2010), *Semantic annotation of medical images*. Medical Imaging 2010: Advanced PACS-based Imaging Informatics and Therapeutic Applications. https://www.spiedigitallibrary.org/conference-proceedings-of-spie/7628/1/Semantic-annotation-of-medical-images/10.1117/12.844207.short?SSO=1

[b-US FDA] U.S. Food & Drug Administration. *Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)*. https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf

[Yao et al., 2007] Yao, B., Yang, X., and Zhu, S-C. (2007), *Introduction to a large-scale general purpose ground truth database: methodology, annotation tool and benchmarks*. International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition. Springer, Berlin, Heidelberg, pp. 169-183. https://link.springer.com/chapter/10.1007/978-3-540-74198-5_14