



UNIVERSIDAD
NACIONAL
DE COLOMBIA

1 **Medical image segmentation in a multiple**
2 **labelers context: Application to the study of**
3 **histopathology**

4 **Brandon Lotero Londoño**

5 Universidad Nacional de Colombia
6 Faculty of Engineering and Architecture
7 Department of Electric, Electronic and Computing Engineering
8 Manizales, Colombia
9 2023

Medical image segmentation in a multiple labelers context: Application to the study of histopathology

Brandon Lotero Londoño

Dissertation submitted as a partial requirement to receive the grade of:
Master in Engineering - Industrial Automation

Advisor:

Prof. Andrés Marino Álvarez-Meza, Ph.D.

Co-advisor:

Prof. Germán Castellanos-Domínguez, Ph.D.

Academic research group:

Signal Processing and Recognition Group - SPRG

Universidad Nacional de Colombia

Faculty of Engineering and Architecture

Department of Electric, Electronic and Computing Engineering

Manizales, Colombia

2025

27 **Segmentación de imágenes médicas en un**
28 **contexto de múltiples anotadores:**
29 **Aplicación al estudio de histopatologías**

30 **Brandon Lotero Londoño**

31 Disertación presentada como requisito parcial para recibir el título de:

32 **Magíster en Ingeniería - Automatización Industrial**

33 Director:

34 Prof. Andrés Marino Álvarez-Meza, Ph.D.

35 Codirector:

36 Prof. Germán Castellanos-Domínguez, Ph.D.

37 Grupo de investigación:

38 Grupo de Control y Procesamiento Digital de Señales - GCPDS

39 Universidad Nacional de Colombia

40 Facultad de Ingeniería y Arquitectura

41 Departamento de Ingeniería Eléctrica, Electrónica y Computación

42 Manizales, Colombia

43 2023

ACKNOWLEDGEMENTS

45 PENDING

Brandon Lotero Londoño
2025

ABSTRACT

49 PENDING

50 **Keywords:** PENDING

53 PENDIENTE

54 **Palabras clave:** PENDIENTE

56 **Contents**

57	Acknowledgements	vii
58	Abstract	ix
59	Resumen	xi
60	Contents	xiv
61	List of figures	xv
62	List of tables	xvii
63	Abbreviations	xix
64	1 Introduction	1
65	1.1 Motivation	1
66	1.2 Problem Statement	6
67	1.2.1 Variability in Expertise Levels	8
68	1.2.2 Technical Constraints and Image Quality	9
69	1.2.3 Research Question	9
70	1.3 Literature review	11
71	1.3.1 Facing annotation variability in medical images	12
72	1.3.2 Strategies for handling low-quality images	14
73	1.3.3 U-shaped Convolutional Neural Networks (CNN)s	15
74	1.4 Aims	19
75	1.4.1 General Aim	19

76	1.4.2 Specific Aims	20
77	1.5 Outline and Contributions	21
78	Bibliography	22

LIST OF FIGURES

80 **1-1** Estimation of the tasks and medical image types based on recent
81 literature review (count of referenced terms). 3
82 **1-2** AI and machine learning in medical imaging brief timeline. 4
83 **1-3** Example of a histopathological image segmented by multiple
84 annotators, illustrating variations in label assignment. 7
85 **1-4** Summary diagram for problem Statement 10
86 **1-5** Original U-Net architecture. 16
87 **1-6** Proposed framework for the approach in [López-Pérez et al., 2024]. . 19

LIST OF TABLES

- 90 **CAD** Computer-Aided Diagnosis 2, 5, 6
- 91 **CNN** Convolutional Neural Networks xiii, 3, 15, 19, 21
- 92 **CT** Computed Tomography 12
- 93 **ISS** Image Semantic segmentation 2, 3, 6, 11, 13, 19-21
- 94 **MITs** Medical Imaging Techniques 1
- 95 **ML** Machine Learning 11
- 96 **MV** Majority Voting 11, 12
- 97 **OCR** Optical Character Recognition 11
- 98 **PET** Positron Emission Tomography 14
- 99 **ROI** Region of Interest 2, 6
- 100 **SS** Semantic segmentation 3
- 101 **STAPLE** Simultaneous Truth and Performance Level Estimation 12-14
- 102 **WSI** Whole Slide Imaging 1, 5, 6, 8, 14, 17

103

104

105

106

CHAPTER

ONE

107

INTRODUCTION

108

1.1 Motivation

109 Since Roentgen's discovery of X-rays in 1895, medical imaging has advanced
110 significantly, with modalities like radionuclide imaging, ultrasound, CT, MRI, and
111 digital radiography emerging over the past 50 years. Modern imaging extends
112 beyond image production to include processing, display, storage, transmission and
113 analysis. [Zhou et al., 2021]. Other Medical Imaging Techniques (MITs) have arose
114 during the last decades, some of them implying only the examination of certain
115 pieces or tissues instead of complete patients, like histopathological images, which
116 are images of tissue samples obtained from biopsies or surgical resections and are
117 widely used for the diagnosis of diseases like cancer through Whole Slide Imaging
118 (WSI) scanners [Rashmi et al., 2021].

119 Along with the advances in technologies for medical images acquisition,
120 computational technologies on pattern recognition and artificial intelligence have

also emerged, allowing the development of **Computer-Aided Diagnosis (CAD)** systems based on machine learning algorithms. These systems aim to assist physicians in the diagnosis and treatment of diseases, by providing a second opinion or by automating the analysis of medical images. [Panayides et al., 2020]. One of the most used tasks in which machine learning technologies is being used in the universe of medical images is **Image Semantic segmentation (ISS)**, which consists of assigning a label to each pixel in an image according to the object it belongs to. This task is crucial for the development of **CAD** systems, as it allows the identification of **Region of Interest (ROI)** in the images, which can be used to detect and classify diseases [Azad et al., 2024].

The application of Machine Learning in medical imaging has grown significantly, with key tasks including classification, segmentation, anomaly detection, super-resolution, image registration, and synthetic image generation [Brito-Pacheco et al., 2025]. Among imaging modalities, X-rays and CT scans are widely used for classification and anomaly detection, especially in pulmonary and oncological applications. MRI and ultrasound play a crucial role in segmentation and resolution enhancement, while PET/SPECT imaging is essential for anomaly detection in oncology and neurodegenerative diseases «CITE». Histopathology is rapidly gaining prominence, particularly in segmentation and feature extraction, where AI-driven techniques aid in automated cancer diagnosis and tissue structure analysis. The integration of Deep Learning in histological image processing is revolutionizing pathology, enabling more precise and efficient diagnostics. A brief comparison of the tasks and medical image types based on recent literature review, can be seen in Figure 1-1. [Yu et al., 2025], [Brito-Pacheco et al., 2025], [Ryou et al., 2025], [Hu et al., 2025], [Elhaminia et al., 2025]

For solving the different requirements of tasks in medical images, a variety of computational techniques have been developed [Zhou et al., 2021]. Initially, these needs were covered with simple morphological filters, which implied no training process or elaborated optimization. However, as the complexity of the tasks increased, the need for more sophisticated techniques arose, leading to the

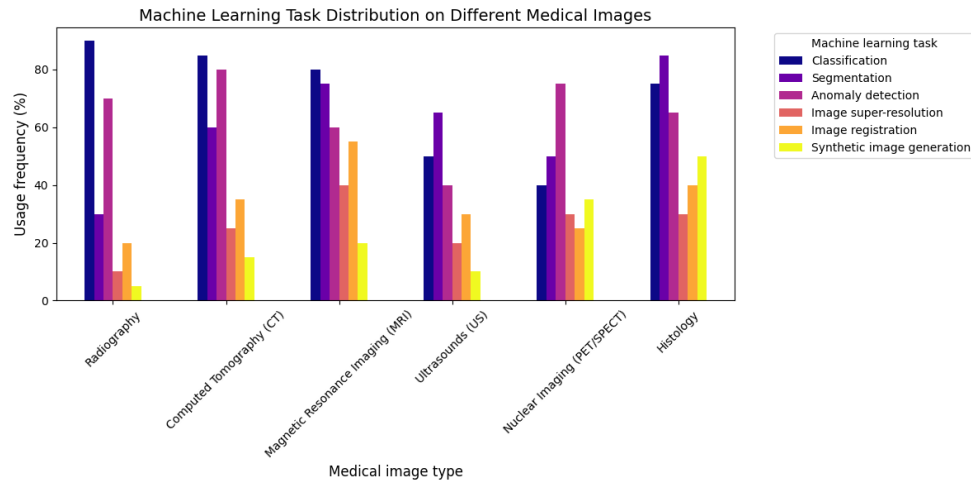


Figure 1-1 Estimation of the tasks and medical image types based on recent literature review (count of referenced terms).

151 application of advanced statistical tools and machine learning algorithms like
 152 Support Vector Machines, Decision Trees, and SGD Neural Networks [Avanzo
 153 et al., 2024]. The coevolution of advances in medical image acquisition,
 154 computational power (i.e. Moore's law) and statistical/mathematical techniques
 155 have led to a convergence for merging state of the art algorithms with medical
 156 imaging [Shalf, 2020]. Figure 1-2 shows a brief timeline of coevolution between
 157 some conspicuous advances in computational pattern recognition and its medical
 158 applications in different scopes (besides medical imaging) [Avanzo et al., 2024].

159 Convolutional Neural Networks (CNN) have been widely used in Semantic
 160 segmentation (SS) tasks, as they have outperformed traditional machine learning
 161 algorithms in this task for both medical and non medical images [Xu et al., 2024]
 162 [Sarvamangala and Kulkarni, 2022]. However, most CNN architectures are deep,
 163 which imply a necessity of a large amount of data to train them. This introduces a
 164 problem since both the acquisition and annotation of medical images are
 165 expensive and time-consuming processes. This is especially true for ISS tasks, as
 166 they require pixel-level annotations, which is taxing in terms of cost, time and
 167 logistics involved [Bhalgat et al., 2018]. Other fashions face this problem through
 168 less expensive annotation strategies like bounding boxes or anatomical landmarks

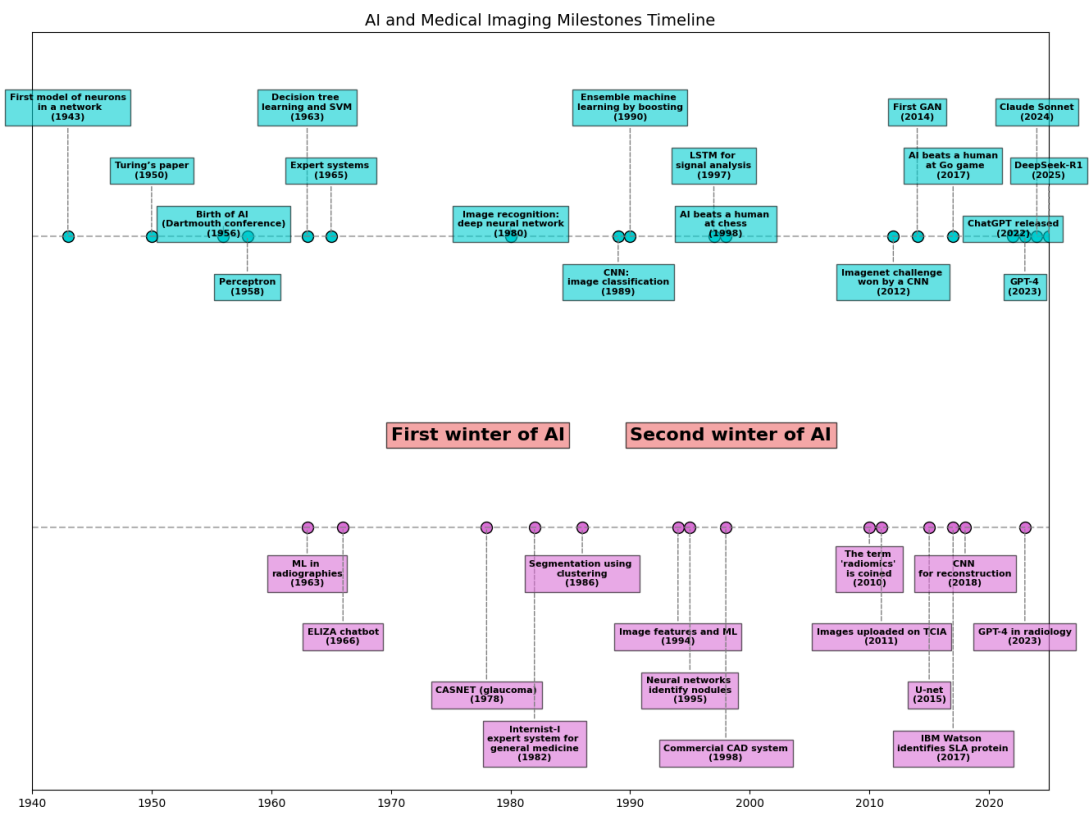


Figure 1-2 AI and machine learning in medical imaging brief timeline.

169 for being used in a semi-supervised strategy [Shah et al., 2018].

170 Many medical images datasets however, contain a high variability in class sizes
171 and variations in colors, which is specially noticeable in histopathological images
172 because of the usage of different staining and other factors which can affect the
173 color of the images. This variability can lead to a significant loss of efficiency of
174 machine learning models when using a mixed supervision strategy, as the model
175 can be biased towards the most common classes or colors in the dataset [Shah
176 et al., 2018].

177 This is where other solutions arise to tackle the problem of the weak image
178 annotation while maintaining low costs. One of these solutions is crowdsourcing
179 strategy, which consists of having multiple annotators labeling the same image,
180 and then combining the labels to obtain a consensus label [Lu et al., 2023]. This
181 strategy can lead to a labeling cost reduction when different levels of expertise are
182 combined, since the crowd may be composed of both experts and laymen, being
183 the latter less expensive to hire [López-Pérez et al., 2023].

184 Recently, diagnosis, prognosis and treatment of cancer have heavily relied on
185 histopathology, where tissue samples are obtained through biopsies or surgical
186 resections and critical information that helps pathologists determine the presence
187 and severity of malignancies [López-Pérez et al., 2024]. The segmentation of
188 histopathological images enables precise identification of structures such as
189 nuclei, glands, and tumors, which are essential for assessing disease progression
190 and treatment response [Rashmi et al., 2021]. Accurate segmentation is
191 particularly crucial in digital pathology, where whole-slide images (WSI) are
192 analyzed using AI-powered CAD systems to support clinical decision-making
193 [López-Pérez et al., 2024].

194 A major challenge in histopathological image segmentation arises from the
195 variability in annotations provided by different pathologists. Unlike natural
196 images, where object boundaries are often well-defined, histological structures
197 may have ambiguous borders, leading to inconsistencies among annotators

[López-Pérez et al., 2023]. Because of this, crowdsourcing labeling is one of the most popular approaches, as illustrated in Figure 1-3, an example of how histopathological images are segmented by multiple experts, showing some variations in label assignment ¹. These discrepancies highlight the need for models that can handle annotation uncertainty effectively. Leveraging crowdsourcing strategies and machine learning techniques that infer annotator reliability can enhance segmentation performance while reducing costs.

1.2 Problem Statement

Throughout the development of medical technology and CAD, the task of ISS has become a crucial step in delivering precise diagnosis and treatment planning [Giri and Bhatia, 2024]. Particularly, in the area of histopathological studies, the usage of Whole Slide Images (WSI) is rather common since this method delivers high quality imaging and allows for the diagnosis of diseases like cancer [Lin et al., 2024].

ISS task consists of assigning a label to each pixel in an image according to the object it belongs to. Accurate segmentation is essential for the development of CAD systems, as it allows the identification of regions of interest (ROI) in the images, which can be used to detect and classify diseases and hence, treatment planning [Sarvamangala and Kulkarni, 2022]. However, modern computational solutions for ISS tasks involve the use of deep learning, which mostly rely large amounts of labeled data to train the models on supervised learning techniques. This means that the model is trained on a dataset with ground-truth labels, which are assumed to be correct and consistent across all samples. In practice, this assumption is often violated due to the high technical complexity of labeling these segments ².

¹obtained from a real world Triple Negative Breast Cancer (TNBC) dataset published in [López-Pérez et al., 2023]

²compared to a more trivial task like image classification on ordinary an well known classes like MNIST

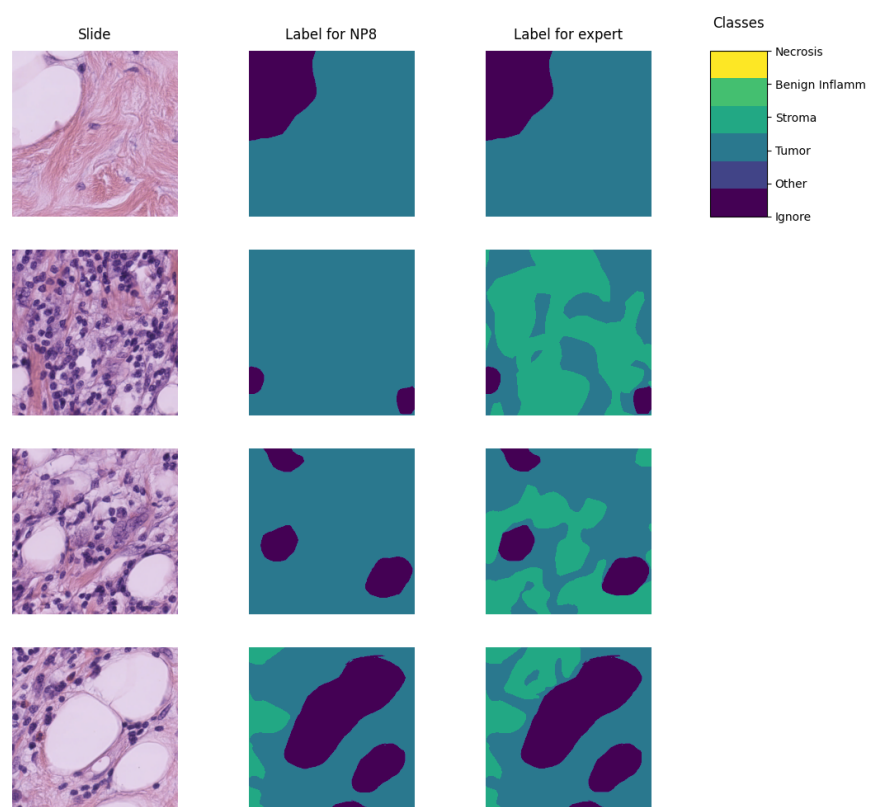


Figure 1-3 Example of a histopathological image segmented by multiple annotators, illustrating variations in label assignment.

The process of labeling medical images is often managed with the help of specialized software tools that allow the annotators to draw the regions, delivering an standard format for the labeled masks [Habis, 2024]. Despite the help of these tools, the labeling process in WSI can have high costs, as it requires long hours of work from specialized personnel. Because of cost constraints in many medical institutions, the labeling processes is often done by multiple labelers with varying levels of expertise, equalizing the cost of the labeling process. However, this strategy can lead to inconsistent labels, as the consensus between the labelers may not be exact due to the diversity in depth of knowledge and experience of the labelers [Xu et al., 2024]. These inconsistencies are mostly represented in the subsections 1.2.1 and 1.2.2.

1.2.1 Variability in Expertise Levels

One of the primary sources of inter-observer variability in medical image segmentation is the difference in expertise levels among annotators [López-Pérez et al., 2023]. Experienced radiologists and pathologists tend to produce highly precise annotations, whereas novice labelers may introduce systematic biases due to their limited familiarity with subtle image features. Studies have demonstrated that annotation accuracy tends to improve with experience, yet medical institutions often rely on a mix of annotators to manage costs and workload distribution [Lu et al., 2023].

The training background of annotators and institutional guidelines play a crucial role in shaping labeling practices. Different medical schools and hospitals may adopt distinct segmentation protocols, leading to inconsistencies when datasets are combined from multiple sources [López-Pérez et al., 2023]. For example, some institutions may emphasize conservative delineation of tumor boundaries, while others adopt a more inclusive approach. Such variations contribute to systematic biases in medical image datasets [Banerjee et al., 2025].

Medical images frequently contain structures with ambiguous boundaries, making segmentation inherently subjective. For instance, tumor margins in histopathological slides may not have well-defined edges, leading to variations in how different annotators delineate the regions of interest [Carmo et al., 2025]. These discrepancies arise not only from technical expertise but also from differences in perception and interpretation.

1.2.2 Technical Constraints and Image Quality

Technical constraints in medical imaging, such as resolution differences, noise levels, and contrast variations, can significantly impact segmentation accuracy. Lower-resolution images may obscure fine structures, leading to inconsistencies in boundary delineation [Zhou et al., 2024].

When combined with long sessions, bad images might also increase the cognitive load of the annotators, leading to fatigue and reduced precision in labeling [Kim et al., 2024]. This is particularly relevant in histopathological studies, where the staining process and tissue preparation can introduce color variations and artifacts that affect image quality, even if the same scanning equipment is used [Karthikeyan et al., 2023].

1.2.3 Research Question

Given the challenges posed by inconsistent labels in medical image segmentation, this work aims to address the following research question:

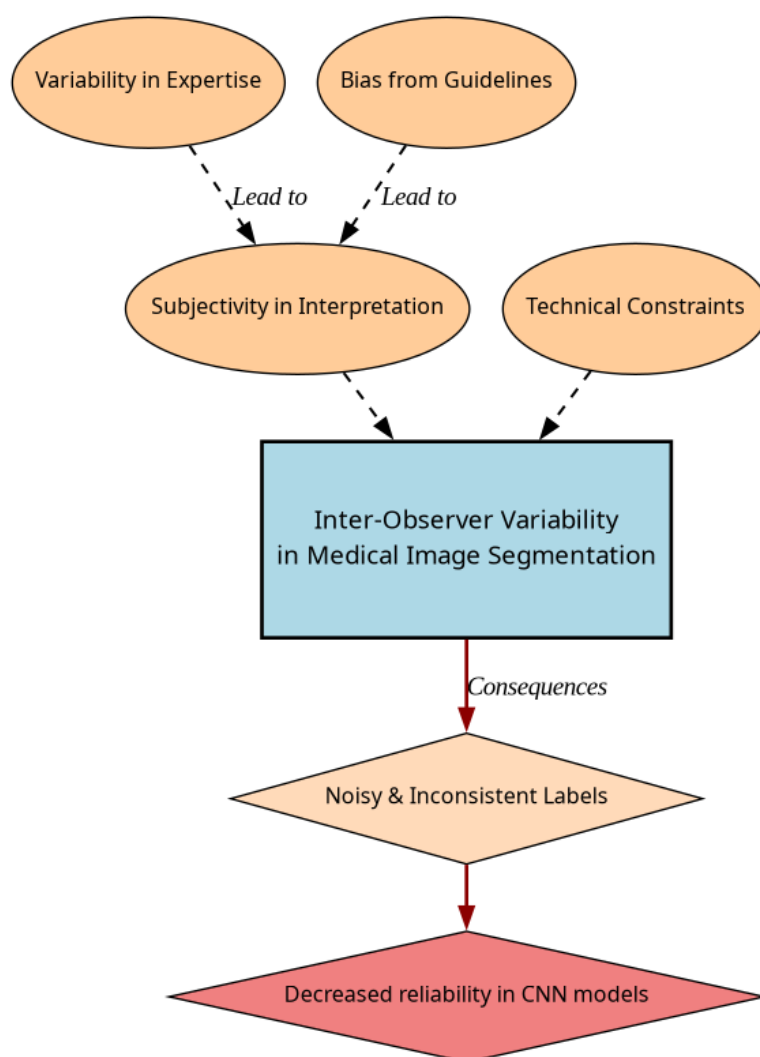


Figure 1-4 Summary diagram for problem Statement

Research Question

How can we develop a learning approach for ISS tasks in medical images that can adapt to inconsistent labels without requiring explicit supervision of labeler performance? Can such approach face problems related to the variability in expertise levels and technical constraints while preserving interpretability, generalization and computational efficiency?

269

1.3 Literature review

270

271 Certainly, in general Machine Learning (ML) classification tasks³ where multiple
272 annotators are involved, Majority Voting (MV) is by far the simplest possible
273 approach to implement. This concept was born multiple times and divergently in
274 multiple fields, but it was described as relevant for ML and pattern recognition
275 labeling for classification in [Lam and Suen, 1997], in which the approach is
276 exposed as simple, yet powerful. The authors describe the MV as a method that
277 can be used to improve the accuracy of classification tasks by combining the labels
278 of multiple annotators. The method is based on the assumption that the majority
279 vote of the annotators is more likely to be correct than the vote of a single
280 annotator. The authors also describe the method as a straightforward way to
281 improve the accuracy of classification tasks without the need for complex
282 algorithms or additional data. The authors also prove this method to deliver very
283 similar results to more complicated approaches (Bayesian, logistic regression,
284 fuzzy integral, and neural network) in the particular task of Optical Character
285 Recognition (OCR). Despite its simplicity, modern solutions for delivering accurate
286 medical image segmentation models still rely on Majority Voting at some stage,
287 like [Elnakib et al., 2020], which uses a majority voting strategy for delivering a
288 final output based on the labels of multiple models (VGG16-Segnet, Resnet-18 and

³In this work, image segmentation is considered as a particular case of classification in which target classes are assigned pixel-wise.

Alexnet) in **Computed Tomography (CT)** images for Liver Tumor Segmentation, or [López-Pérez et al., 2023], which uses **MV** for combining noisy annotations as an additional annotator to be included in the deep learning solution. Majority voting as a technique for setting a pseudo ground truth label is a powerful approach for its simplicity in many use cases in which the target to be labeled is not tied to an expertise related task, otherwise, the assumption of equal expertise among the labelers can be a source of bias in the final label, which is not desirable in the case of highly technical annotations like medical images. In subsection 1.3.1, we will be reviewing literature which no longer assumes the naive approach of equal expertise among labelers and face the challenge of learning from inconsistent labels.

1.3.1 Facing annotation variability in medical images

Learning from crowds approaches in general face the challenge of not having a ground truth label and hence, an intrinsic difficulty in measuring the real reliability of the labelers annotations. Some approaches assume beforehand a certain level of expertise for each labeler based on experience as an input, like in [TIAN and Zhu, 2015], which introduce the concept of max margin majority voting, using the reliability vector as weights for the weights for the binary and multiclass classifier. The crowdsourcing margin is the minimal difference between the aggregated score of the potential true label and the scores for other alternative labels. Accordingly, the annotators' reliability is estimated as generating the largest margin between the potential true labels and other alternatives. The problem introduced in this approach is assuming an stationary reliability per expert across the whole input space, which is imprecise since annotators performance may change between different tasks or even between different regions of the same image.

STAPLE Mechanism

The **Simultaneous Truth and Performance Level Estimation (STAPLE)** algorithm, introduced in [Warfield et al., 2004] is a probabilistic framework that estimates a

hidden true segmentation from multiple segmentations provided by different raters. It also estimates the reliability of each rater by computing their sensitivity and specificity.

The **STAPLE** algorithm's goal is to maximize the log likelihood function:

$$(\mathbf{p}, \mathbf{q}) = \arg \max_{\mathbf{p}, \mathbf{q}} \ln f(\mathbf{D}, \mathbf{T} \mid \mathbf{p}, \mathbf{q}). \quad (1-1)$$

Where \mathbf{D} is the set of segmentations provided by the raters, \mathbf{T} is the hidden true segmentation, p is the sensitivity and q is the specificity of the raters.

This is achieved by using the Expectation-Maximization algorithm to maximize the log likelihood function in equation, which is done iteratively with step computations:

$$\begin{aligned} (p_j^{(k)}, q_j^{(k)}) = \arg \max_{p_j, q_j} & \sum_{i: D_{ij}=1} W_i^{(k-1)} \ln p_j \\ & + \sum_{i: D_{ij}=1} \left(1 - W_i^{(k-1)}\right) \ln(1 - q_j) \\ & + \sum_{i: D_{ij}=0} W_i^{(k-1)} \ln(1 - p_j) \\ & + \sum_{i: D_{ij}=0} \left(1 - W_i^{(k-1)}\right) \ln q_j. \end{aligned} \quad (1-2)$$

The capacity of STAPLE to accurately estimate the true segmentation, even in the presence of a majority of raters generating correlated errors, was demonstrated, which makes it theoretically a strong choice for setting a ground-truth in binary or multiclass medical **ISS** tasks.

The popularity and performance of **STAPLE** has led to its usage in modern applications medical image, 3d spatial images due to its assumption of decision

space being based on voxel-wise decisions, like the authors in [Grefve et al., 2024] which applied the algorithm on Positron Emission Tomography (PET) images. Other authors still rely heavily on STAPLE for setting a ground truth consensus for histopathological images, like [Qiu et al., 2022].

However, the STAPLE algorithm has some limitations. It assumes independent rater errors, which may not hold in practice, leading to biased estimates. STAPLE is also sensitive to low-quality annotations, potentially degrading final segmentations if the weights are not initialized correctly. The algorithm tends to over-smooth results, blurring fine details, and struggles with multi-class segmentation. Computationally, it is expensive due to its iterative EM approach. Additionally, STAPLE cannot correct systematic biases in annotations and depends on initial estimates, impacting accuracy. Lastly, the estimated performance levels lack interpretability, making it difficult to assess annotator reliability effectively.

Finally, this work contemplates STAPLE as useful for ground truth estimation given the existence of multiple labelers for an input WSI, but not that useful for providing annotations of structures on new and unlabeled images, hence being a good support for other methods.

Chained Gaussian Processes

Other works like [Gil-González et al., 2025] proposed a novel approach from Gaussian Processes to model the relationship between the annotators' reliability and the input data. The authors propose a method to estimate the reliability of each annotator based on their performance on a set of validation images. The reliability is then used to weight the contribution of each annotator to the final segmentation.

1.3.2 Strategies for handling low-quality images

The problem of low-quality images and noisy annotations has been tackled with various strategies. One such approach is the use of deep learning models that

incorporate loss functions designed to mitigate the effects of unreliable labels. Traditional methods such as Majority Voting (MV) or Expectation-Maximization (EM) have been widely used for aggregating multiple annotators' inputs. However, they assume a homogeneous reliability of annotators, which may not hold in real-world scenarios.

A more recent approach was proposed by [Triana-Martinez et al., 2023], introducing a Generalized Cross-Entropy-based Chained Deep Learning (GCECDL) framework. This method addresses the limitations of traditional label aggregation techniques by modeling each annotator's reliability as a function of the input data. The approach effectively mitigates the impact of noisy labels by using a noise-robust loss function, balancing Mean Absolute Error (MAE) and Categorical Cross-Entropy (CE). Unlike prior approaches, GCECDL accounts for the dependencies among annotators while encoding their non-stationary behavior across different image regions. Their experiments on multiple datasets demonstrated superior predictive performance compared to state-of-the-art methods, particularly in cases where annotations were highly inconsistent.

This strategy is especially relevant for handling low-quality medical images, where expert annotations may be inconsistent, and traditional consensus-based approaches fail to account for varying expertise levels. By leveraging deep learning with robust noise-handling loss functions, the reliability of segmentation models can be significantly improved.

1.3.3 U-shaped CNNs

Since the introduction of U-Net [Ronneberger et al., 2015] in 2015 for biomedical image segmentation, U-shaped CNNs have become a prevalent architecture in medical image segmentation tasks. The U-Net's success stems from its ability to capture both global and local information through its contracting and expanding paths, making it particularly effective for complex and heterogeneous structures,

even with limited annotated data. This architecture has been successfully applied to various medical image segmentation tasks, including organ segmentation, tumor segmentation, and brain structure segmentation.

The U-Net architecture consists of a symmetric encoder-decoder structure with skip connections. The encoder path progressively reduces spatial dimensions while increasing feature channels through a series of convolutional and max-pooling layers, capturing high-level semantic information. The decoder path uses transposed convolutions to gradually recover spatial resolution while reducing feature channels. Skip connections between corresponding encoder and decoder layers preserve fine-grained details by concatenating high-resolution features from the encoder with upsampled features in the decoder, enabling precise localization of structures. The architecture overview can be seen in figure 1-5.

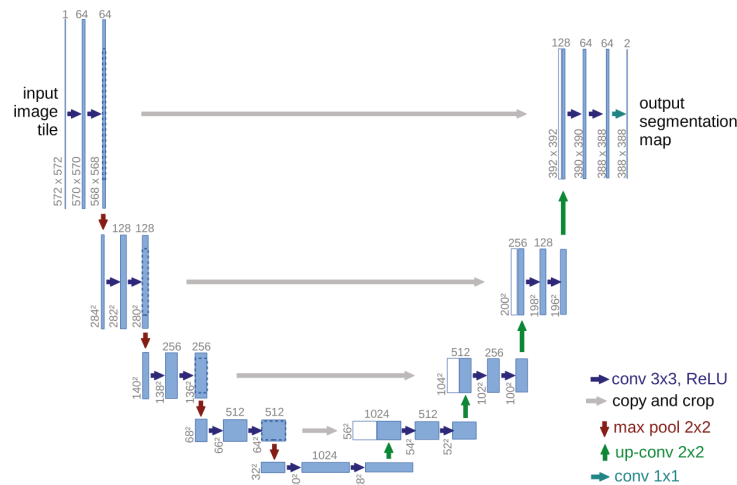


Figure 1-5 Original U-Net architecture.

U-Net based approaches

In [López-Pérez et al., 2024] two networks are trained for delivering a final segmentation. One network is trained to aggregate the annotations of the labelers

and another one is trained to segment the image. The first network is a U-Net that takes as input the image and the annotations of the labelers and outputs a reliability map. This map is then used to weight the contribution of each annotator to the final segmentation. The second network is the U-Net used for segmentation.

In this approach, it is assumed that the images are labeled for at least one labeler and not all of them, which is closer to a real world scenario, in which it is common to have images with variability in the amount of annotations, per patch. Hence, the input data can be modeled as:

$$\mathcal{D} = (\mathbf{X}, \tilde{\mathbf{Y}}) = \{(\mathbf{x}_n, \tilde{\mathbf{y}}_n^r) : n = 1, \dots, N; r \in R_n\}, \quad (1-3)$$

Where every \mathbf{x}_n is an input patch from a ROI in one WSI, $\tilde{\mathbf{y}}_n$ is the noisy annotation from the r labeler, N is the number of patches in the dataset and $R_n \subset \{1, \dots, R\}$ is the set of labelers that annotated the image \mathbf{x}_n .

The authors then assume the annotator network to deliver a reliability map $\{\hat{\mathbf{A}}_\phi^{(r)}(\mathbf{x})\}_{r \in R_n}$ with different dimensions:

- CR global: a single reliability vector per labeler with dimensions C which represent global reliability of the labeler across all input space.
- CR image: a single reliability vector per image per labeler with dimensions C which represent local reliability of the labeler across the image.
- CR pixel: a reliability matrix per image per labeler, with dimensions C which represent local reliability of the labeler across all the pixels in the image.

Being $\mathbf{p}_\theta(\mathbf{x}_n)$ the estimation of the latent (ground truth) segmentation delivered by the segmentation UNet network, then, the estimated segmentation probability mask for each annotator is given by the product:

$$\mathbf{p}_{\theta, \phi}^{(r)}(\mathbf{x}_n) := \mathbf{A}_{\phi}^{(r)}(\mathbf{x}) \odot \mathbf{p}(\mathbf{x}_n), \quad (1-4)$$

where \odot is the element-wise product and θ and ϕ are the parameters of the annotator network and the segmentation UNet network, respectively, being the latter initialized with a ResNet34 backbone pre-trained on ImageNet.

The authors propose a loss function involving cross-entropy and a trace based regularization on the reliability map, originally proposed in [Zhang et al., 2020] which combined, looks like:

$$\mathcal{L}(\theta, \phi) := \sum_{n=1}^N \sum_{r=1}^R \mathbb{I}(\tilde{\mathbf{y}}_n^{(r)} \in R_n) \cdot \left[\text{CE} \left(\mathbf{A}_{\phi}^{(r)}(\mathbf{x}_n) \cdot \mathbf{p}_{\theta}(\mathbf{x}_n), \tilde{\mathbf{y}}_n^{(r)} \right) + \lambda \cdot \text{tr} \left(\mathbf{A}_{\phi}^{(r)}(\mathbf{x}_n) \right) \right] \quad (1-5)$$

Being \mathbb{I} the indicator function, CE the cross-entropy loss, and λ the regularization parameter.

When trained against a Triple Negative Breast Cancer images dataset, this approach achieves a Dice coefficient of 0.7827, outperforming the well-known STAPLE (0.7039) and being competitive with the supervised method with expert labels (0.7723). The reliability modeling approach which gave the best results was the CR image, the authors explain that CR pixel could have a deeper understanding of the labelers reliability per image but requires a much larger amount of data to be trained, hence CR image offers a balanced solution.

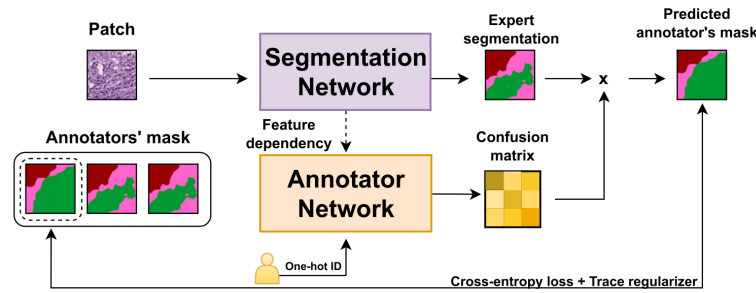


Figure 1-6 Proposed framework for the approach in [López-Pérez et al., 2024].

1.4 Aims

With the mentioned considerations in section 1.3 in mind, this work proposes a novel approach for ISS tasks in medical images, which aims to train a model whose learning approach is adaptive to the labeler performance. This is done by introducing a loss function capable of inferring the best possible segmentation without needing separate inputs about the labeler performance. This loss function is designed to implicitly weigh the labelers based on their performance, with the presence of an intermediate reliability map allowing the model to learn from the most reliable labelers and ignore the noisy labels. This approach differs from existing CNN-based segmentation models, as it does not require explicit supervision of the labeler performance, making it more generalizable and adaptable to different datasets and labelers.

1.4.1 General Aim

The main purpose of this work is to develop a novel approach for ISS tasks in medical images, which can adaptively infer the best possible segmentation without needing separate inputs about the labeler performance. This approach is expected to outperform the segmentation performance of other state of the art approaches, eliminate the need for explicit labeler supervision, and enhance automation in medical image analysis.

1.4.2 Specific Aims

- To develop a novel loss function for ISS tasks in medical images, capable of inferring the best possible segmentation without needing separate inputs about the labeler performance.
- Introducing a tensor map which codifies the reliability of each labeler, allowing the model to implicitly weigh the labelers based on their performance across the mask and classes space.
- To develop and test a deep learning model for ISS tasks in medical images, which can learn from inconsistent labels and improve the segmentation performance compared to other solutions in state of the art.

1.5 Outline and Contributions

As an output of this work, some contributions were made to the field of ISS in medical images. The main contributions are:

- A python package for using the proposed loss function in CNN models for ISS tasks in medical images. ⁴
- Datasets mapping as lazy loaders for the proposed loss function. ⁵
- A public Github repository with the code used in this work. ⁶

⁴https://pypi.org/project/seg_tgce/

⁵<https://seg-tgce.readthedocs.io/en/latest/experiments.html>

⁶https://github.com/blotero/seg_tgce

BIBLIOGRAPHY

- 476 [Avanzo et al., 2024] Avanzo, M., Stancanella, J., Pirrone, G., Drigo, A., and Retico,
477 A. (2024). The evolution of artificial intelligence in medical imaging: From
478 computer science to machine and deep learning. *Cancers (Basel)*, 16(21):3702.
479 Author Joseph Stancanella is employed by Elekta SA. The remaining authors
480 declare no commercial or financial conflicts of interest. (page 3)
- 481 [Azad et al., 2024] Azad, R., Aghdam, E. K., Rauland, A., Jia, Y., Avval, A. H.,
482 Bozorgpour, A., Karimijafarbigloo, S., Cohen, J. P., Adeli, E., and Merhof, D.
483 (2024). Medical image segmentation review: The success of u-net. *IEEE*
484 *Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10076–10095.
485 (page 2)
- 486 [Banerjee et al., 2025] Banerjee, A., Shan, H., and Feng, R. (2025). Editorial:
487 Artificial intelligence applications for cancer diagnosis in radiology. *Frontiers in*
488 *Radiology*, 5. (page 8)
- 489 [Bhalgat et al., 2018] Bhalgat, Y., Shah, M. P., and Awate, S. P. (2018). Annotation-
490 cost minimization for medical image segmentation using suggestive mixed
491 supervision fully convolutional networks. CoRR, abs/1812.11302. (page 3)
- 492 [Brito-Pacheco et al., 2025] Brito-Pacheco, D., Giannopoulos, P., and Reyes-
493 Aldasoro, C. C. (2025). Persistent homology in medical image processing: A
494 literature review. (page 2)

- 495 [Carmo et al., 2025] Carmo, D. S., Pezzulo, A. A., Villacreses, R. A., Eisenbeisz,
496 M. L., Anderson, R. L., Van Dorin, S. E., Rittner, L., Lotufo, R. A., Gerard, S. E.,
497 Reinhardt, J. M., and Comellas, A. P. (2025). Manual segmentation of opacities
498 and consolidations on ct of long covid patients from multiple annotators. *Scientific*
499 *Data*, 12(1):402. (page 9)
- 500 [Elhaminia et al., 2025] Elhaminia, B., Alsalemi, A., Nasir, E., Jahanifar, M., Awan,
501 R., Young, L. S., Rajpoot, N. M., Minhas, F., and Raza, S. E. A. (2025). From
502 traditional to deep learning approaches in whole slide image registration: A
503 methodological review. (page 2)
- 504 [Elnakib et al., 2020] Elnakib, A., Elmenabawy, N., and S Moustafa, H. (2020).
505 Automated deep system for joint liver and tumor segmentation using majority
506 voting. *MEJ-Mansoura Engineering Journal*, 45(4):30–36. (page 11)
- 507 [Gil-González et al., 2025] Gil-González, J., Daza-Santacoloma, G., Cárdenas-Peña,
508 D., Orozco-Gutiérrez, A., and Álvarez Meza, A. (2025). Generalized cross-entropy
509 for learning from crowds based on correlated chained gaussian processes. *Results*
510 *in Engineering*, 25:103863. (page 14)
- 511 [Giri and Bhatia, 2024] Giri, K. and Bhatia, S. (2024). Artificial intelligence in
512 nephrology- its applications from bench to bedside. *International Journal of*
513 *Advances in Nephrology Research*, 7(1):90–97. (page 6)
- 514 [Grefve et al., 2024] Grefve, J., Söderkvist, K., Gunnlaugsson, A., Sandgren, K.,
515 Jonsson, J., Keeratijarut Lindberg, A., Nilsson, E., Axelsson, J., Bergh,
516 A., Zackrisson, B., Moreau, M., Thellenberg Karlsson, C., Olsson, L.,
517 Widmark, A., Riklund, K., Blomqvist, L., Berg Loegager, V., Strandberg,
518 S. N., and Nyholm, T. (2024). Histopathology-validated gross tumor
519 volume delineations of intraprostatic lesions using psma-positron emission
520 tomography/multiparametric magnetic resonance imaging. *Physics and Imaging in*
521 *Radiation Oncology*, 31:100633. (page 14)

- [Habis, 2024] Habis, A. A. (2024). *Developing interactive artificial intelligence tools to assist pathologists with histology annotation*. Theses, Institut Polytechnique de Paris. (page 8)
- [Hu et al., 2025] Hu, D., Jiang, Z., Shi, J., Xie, F., Wu, K., Tang, K., Cao, M., Huai, J., and Zheng, Y. (2025). Pathology report generation from whole slide images with knowledge retrieval and multi-level regional feature selection. *Computer Methods and Programs in Biomedicine*, 263:108677. (page 2)
- [Karthikeyan et al., 2023] Karthikeyan, R., McDonald, A., and Mehta, R. (2023). What’s in a label? annotation differences in forecasting mental fatigue using ecg data and seq2seq architectures. (page 9)
- [Kim et al., 2024] Kim, Y., Lee, E., Lee, Y., and Oh, U. (2024). Understanding novice’s annotation process for 3d semantic segmentation task with human-in-the-loop. In *Proceedings of the 29th International Conference on Intelligent User Interfaces, IUI ’24*, page 444–454, New York, NY, USA. Association for Computing Machinery. (page 9)
- [Lam and Suen, 1997] Lam, L. and Suen, S. (1997). Application of majority voting to pattern recognition: an analysis of its behavior and performance. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 27(5):553–568. (page 11)
- [Lin et al., 2024] Lin, Y., Lian, A., Liao, M., and Yuan, S. (2024). Bcdnet: A fast residual neural network for invasive ductal carcinoma detection. (page 6)
- [López-Pérez et al., 2023] López-Pérez, M., Morales-Álvarez, P., Cooper, L. A. D., Molina, R., and Katsaggelos, A. K. (2023). Crowdsourcing segmentation of histopathological images using annotations provided by medical students. In Juarez, J. M., Marcos, M., Stiglic, G., and Tucker, A., editors, *Artificial Intelligence in Medicine*, pages 245–249, Cham. Springer Nature Switzerland. (pages 5, 6, 8, and 12)

- [Lu et al., 2023] Lu, X., Ratcliffe, D., Kao, T.-T., Tikhonov, A., Litchfield, L., Rodger, C., and Wang, K. (2023). Rethinking quality assurance for crowdsourced multi-roi image segmentation. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 11(1):103–114. (pages 5 and 8)
- [López-Pérez et al., 2024] López-Pérez, M., Morales-Álvarez, P., Cooper, L. A., Felicelli, C., Goldstein, J., Vadasz, B., Molina, R., and Katsaggelos, A. K. (2024). Learning from crowds for automated histopathological image segmentation. *Computerized Medical Imaging and Graphics*, 112:102327. (pages xv, 5, 16, and 19)
- [Panayides et al., 2020] Panayides, A. S., Amini, A., Filipovic, N. D., Sharma, A., Tsaftaris, S. A., Young, A., Foran, D., Do, N., 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. (page 2)
- [Qiu et al., 2022] Qiu, Y., Hu, Y., Kong, P., Xie, H., Zhang, X., Cao, J., Wang, T., and Lei, B. (2022). Automatic prostate gleason grading using pyramid semantic parsing network in digital histopathology. *Frontiers in Oncology*, 12. (page 14)
- [Rashmi et al., 2021] Rashmi, R., Prasad, K., and Udupa, C. B. K. (2021). Breast histopathological image analysis using image processing techniques for diagnostic purposes: A methodological review. *Journal of Medical Systems*, 46(1):7. (pages 1 and 5)
- [Ronneberger et al., 2015] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F., editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham. Springer International Publishing. (page 15)
- [Ryou et al., 2025] Ryou, H., Thomas, E., Wojciechowska, M., Harding, L., Tam, K. H., Wang, R., Hu, X., Rittscher, J., Cooper, R., and Royston, D. (2025). Reticulin-free quantitation of bone marrow fibrosis in mpns: Utility and applications. *eJHaem*, 6(2):e70005. (page 2)

- [Sarvamangala and Kulkarni, 2022] Sarvamangala, D. R. and Kulkarni, R. V. (2022). Convolutional neural networks in medical image understanding: a survey. *Evolutionary Intelligence*, 15(1):1–22. (pages 3 and 6)
- [Shah et al., 2018] Shah, M. P., Merchant, S. N., and Awate, S. P. (2018). Ms-net: Mixed-supervision fully-convolutional networks for full-resolution segmentation. In Frangi, A. F., Schnabel, J. A., Davatzikos, C., Alberola-López, C., and Fichtinger, G., editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 379–387, Cham. Springer International Publishing. (page 5)
- [Shalf, 2020] Shalf, J. (2020). The future of computing beyond moore’s law. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 378(2166):20190061. (page 3)
- [TIAN and Zhu, 2015] TIAN, T. and Zhu, J. (2015). Max-margin majority voting for learning from crowds. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc. (page 12)
- [Triana-Martinez et al., 2023] Triana-Martinez, J. C., Gil-González, J., Fernandez-Gallego, J. A., Álvarez Meza, A. M., and Castellanos-Dominguez, C. G. (2023). Chained deep learning using generalized cross-entropy for multiple annotators classification. *Sensors*, 23(7). (page 15)
- [Warfield et al., 2004] Warfield, S., Zou, K., and Wells, W. (2004). Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, 23(7):903–921. (page 12)
- [Xu et al., 2024] Xu, Y., Quan, R., Xu, W., Huang, Y., Chen, X., and Liu, F. (2024). Advances in medical image segmentation: A comprehensive review of traditional, deep learning and hybrid approaches. *Bioengineering*, 11(10). (pages 3 and 8)

- 605 [Yu et al., 2025] Yu, J., Li, B., Pan, X., Shi, Z., Wang, H., Lan, R., and Luo, X. (2025).
606 Semi-supervised gland segmentation via feature-enhanced contrastive learning
607 and dual-consistency strategy. *IEEE Journal of Biomedical and Health Informatics*,
608 pages 1–11. (page 2)
- 609 [Zhang et al., 2020] Zhang, L., Tanno, R., Xu, M.-C., Jin, C., Jacob, J., Cicarrelli, O.,
610 Barkhof, F., and Alexander, D. (2020). Disentangling human error from ground
611 truth in segmentation of medical images. In Larochelle, H., Ranzato, M., Hadsell,
612 R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*,
613 volume 33, pages 15750–15762. Curran Associates, Inc. (page 18)
- 614 [Zhou et al., 2021] Zhou, S. K., Greenspan, H., Davatzikos, C., Duncan, J. S.,
615 Van Ginneken, B., Madabhushi, A., Prince, J. L., Rueckert, D., and Summers, R. M.
616 (2021). A review of deep learning in medical imaging: Imaging traits, technology
617 trends, case studies with progress highlights, and future promises. *Proceedings of*
618 *the IEEE*, 109(5):820–838. (pages 1 and 2)
- 619 [Zhou et al., 2024] Zhou, Z., Gong, H., Hsieh, S., McCollough, C. H., and Yu, L.
620 (2024). Image quality evaluation in deep-learning-based ct noise reduction using
621 virtual imaging trial methods: Contrast-dependent spatial resolution. *Medical*
622 *Physics*, 51(8):5399–5413. (page 9)