



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

10 **Medical image segmentation in a multiple**
11 **labelers context: Application to the study of**
12 **histopathology**

13 **Brandon Lotero Londoño**

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

16 Advisor:
17 Prof. Andrés Marino Álvarez-Meza, Ph.D.
18 Co-advisor:
19 Prof. Germán Castellanos-Domínguez, Ph.D.

20 Academic research group:
21 Signal Processing and Recognition Group - SPRG

22 Universidad Nacional de Colombia
23 Faculty of Engineering and Architecture
24 Department of Electric, Electronic and Computing Engineering
25 Manizales, Colombia

26 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

44

ACKNOWLEDGEMENTS

45 PENDING

46

Brandon Lotero Londoño

47

2025

48

ABSTRACT

49 PENDING

50 **Keywords:** PENDING

51

RESUMEN

52

53 PENDIENTE

54 **Palabras clave:** PENDIENTE

55

⁵⁶ Contents

⁵⁷ Acknowledgements	vii
⁵⁸ Abstract	ix
⁵⁹ Resumen	xi
⁶⁰ Contents	xv
⁶¹ List of figures	xviii
⁶² List of tables	xix
⁶³ Abbreviations	xxi
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 Facing noisy annotations and low-quality data	19
73 1.4 Aims	21
74 1.4.1 General Aim	22
75 1.4.2 Specific Aims	22

76	1.5 Outline and Contributions	23
77	2 Conceptual preliminaries	25
78	2.1 Modern concept of digital image	25
79	2.1.1 Types of digital images	25
80	2.1.2 Mathematical representations	26
81	2.2 Digital histopathological images	28
82	2.2.1 Whole Slide Imaging (WSI)	28
83	2.2.2 Regions of Interest (ROI)	30
84	2.2.3 Staining Techniques	30
85	2.3 Deep learning fundamentals	31
86	2.3.1 Learning Paradigms	32
87	2.3.2 Architecture and Training	33
88	2.3.3 Challenges and Solutions	33
89	2.3.4 Deep Learning Frameworks	34
90	2.4 Datasets and data sources	34
91	2.4.1 Datasets with emulated noisy annotations	36
92	2.4.2 Real histopathology datasets	39
93	3 Chained Gaussian Processes	43
94	3.1 Gaussian processes	43
95	3.2 Chained Gaussian processes	43
96	3.3 Experimental evaluation	43
97	4 Truncated Generalized Cross Entropy for segmentation	45
98	4.1 Loss functions for multiple annotators	45
99	4.1.1 Generalized Cross Entropy	46
100	4.1.2 Extension to Multiple Annotators	47
101	4.1.3 Reliability Maps and Truncated GCE	48
102	5 Chained deep learning for image segmentation	51
103	5.1 Introduction	51

104	5.2 Using U-NET as a building block	51
105	5.2.1 Backbone Architecture	52
106	5.2.2 UNET Architecture	52
107	5.2.3 Reliability Map Branch	53
108	5.2.4 Integration with TGCE_{SS} Loss	54
109	5.2.5 Training Process	54
110	6 Conclusions	57
111	6.1 Summary	57
112	6.2 Future work	58
113	Bibliography	59

LIST OF FIGURES

115	1-1	Estimation of the tasks and medical image types based on recent literature review (count of referenced terms).	3
116			
117	1-2	AI and machine learning in medical imaging brief timeline.	4
118			
119	1-3	Example of a histopathological image segmented by multiple annotators, illustrating variations in label assignment.	7
120			
121	1-4	Summary diagram for problem Statement	10
122	1-5	Proposed framework for the approach in [López-Pérez et al., 2024]. .	17
123	2-1	Histology evolution timeline. (Image from [Mazzarini et al., 2021]). .	28
124			
125	2-2	(Above) Whole slide imaging system by Omnyx for slide digitization. (Below) Comprehensive digital pathology interface from Omnyx designed to streamline pathologists' diagnostic workflow. (From [Farahani et al., 2015]).	29
126			
127	2-3	Common learning paradigms.	32
128			
129	2-4	Comparative Trends of the top two most popular Deep Learning Frameworks, apparently, tendency was switched to PyTorch since 2022	35
130			
131	2-5	Example of a noisy mask generated by naively introducing random noise into a ground truth mask. Morphological consistency is lost. .	36
132			
133	2-6	Annotations in the Oxford-IIIT Pet data. From left to right: pet image, head bounding box, and trimap segmentation (blue: background region; red: ambiguous region; yellow: foreground region).	38
134			
135			
136			

137	2-7	Noisy mask generated by enhancing the disturbances in the encoder layers weights for the Oxford-IIIT Pet Dataset. Morphological consistency is preserved. From left to right, SNR levels of noise in the encoder layer are 10, 5, 2, 0, -5 dB.	38
138			
139			
140			
141	2-8	Different staining techniques obtained from multi-stain breast cancer dataset [Weitz et al., 2023]. (a) shows H&E, (b) ER, (c) HER2, (d) Ki67 and (e) PGR. (f) shows an example of a Whole Slide Imaging (WSI) that was excluded since it contains multiple tissue sections.	40
142			
143			
144			
145	2-9	Screenshot of the DSA and HistomicsTK web interface while creating the crowdsourced annotations for the dataset presented by [Amgad et al., 2019].	41
146			
147			
148	5-1	Original U-NET architecture.	52
149	5-2	Overview of the proposed model architecture.	53
150	5-3	Training process of the proposed model overview. Estimated segmentation and reliability maps are computed for each input image, and the loss function is computed for the entire batch.	54
151			
152			

LIST OF TABLES

ABBREVIATIONS

- ¹⁵⁵ **CAD** Computer-Aided Diagnosis 2, 5, 6
- ¹⁵⁶ **CCGP** Correlated Chained Gaussian Processes 18
- ¹⁵⁷ **CCGPMA** Correlated Chained Gaussian Processes for Multiple Annotators 17, 19
- ¹⁵⁸ **CE** Cross Entropy 46
- ¹⁵⁹ **CGP** Chained Gaussian Processes 18
- ¹⁶⁰ **CNN** Convolutional Neural Networks 3, 14, 22, 23, 55
- ¹⁶¹ **CT** Computed Tomography 11
- ¹⁶² **ELBO** Evidence Lower Bound 18
- ¹⁶³ **GCE** Generalized Cross Entropy 46
- ¹⁶⁴ **GCECDL** Generalized Cross-Entropy-based Chained Deep Learning 20, 21
- ¹⁶⁵ **ISS** Image Semantic segmentation 2, 3, 6, 11, 13, 21-23, 34
- ¹⁶⁶ **LF** Latent Function 18
- ¹⁶⁷ **MAE** Mean Absolute Error 46, 47
- ¹⁶⁸ **MITs** Medical Imaging Techniques 1
- ¹⁶⁹ **ML** Machine Learning 11, 21
- ¹⁷⁰ **MV** Majority Voting 11, 12
- ¹⁷¹ **OCR** Optical Character Recognition 11
- ¹⁷² **PET** Positron Emission Tomography 14
- ¹⁷³ **ROI** Region of Interest 2, 6, 15, 30
- ¹⁷⁴ **SLFM** Semi-Parametric Latent Factor Model 18
- ¹⁷⁵ **SS** Semantic segmentation 3
- ¹⁷⁶ **STAPLE** Simultaneous Truth and Performance Level Estimation 12-14, 35
- ¹⁷⁷ **WSI** Whole Slide Imaging xviii, 1, 5, 6, 8, 15, 28, 40

178

179

CHAPTER

180

181

ONE

182

INTRODUCTION

183

1.1 Motivation

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

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

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

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

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

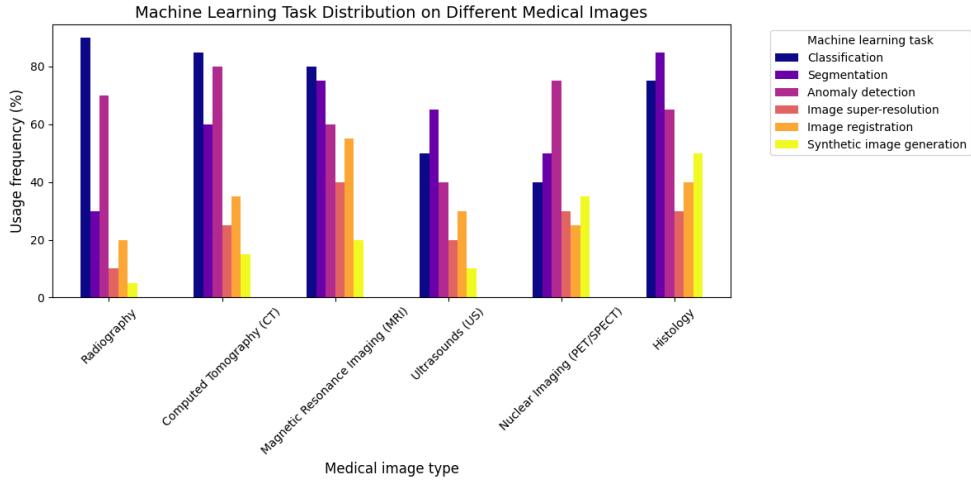


Figure 1-1 Estimation of the tasks and medical image types based on recent literature review (count of referenced terms).

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

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

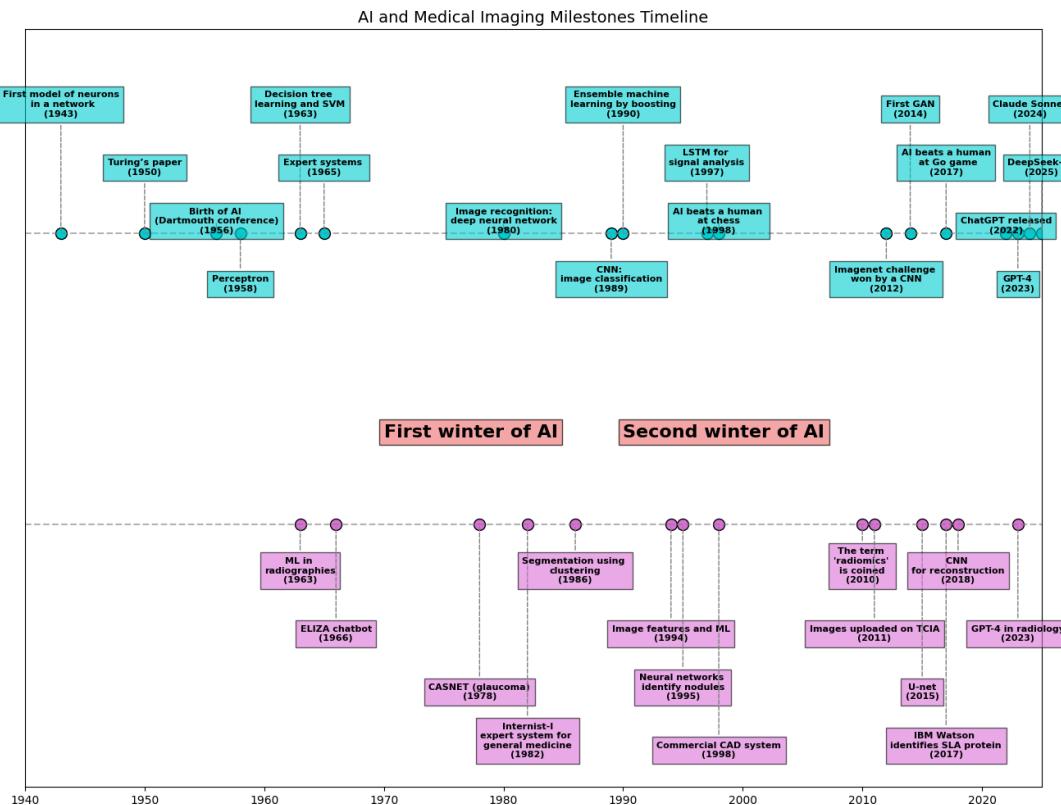


Figure 1-2 AI and machine learning in medical imaging brief timeline.

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

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

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

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

269 A major challenge in histopathological image segmentation arises from the
270 variability in annotations provided by different pathologists. Unlike natural
271 images, where object boundaries are often well-defined, histological structures
272 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 assignment¹. 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 segments².

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 and well known classes like MNIST

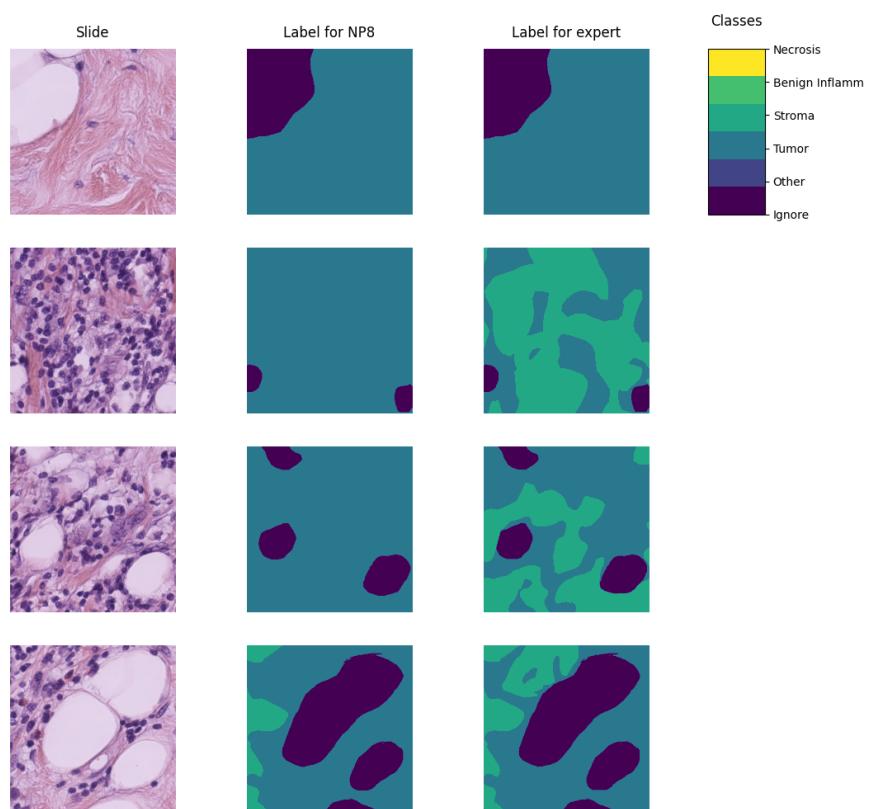


Figure 1-3 Example of a histopathological image segmented by multiple annotators, illustrating variations in label assignment.

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

308 1.2.1 Variability in Expertise Levels

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

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

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

330 1.2.2 Technical Constraints and Image Quality

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

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

341 1.2.3 Research Question

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

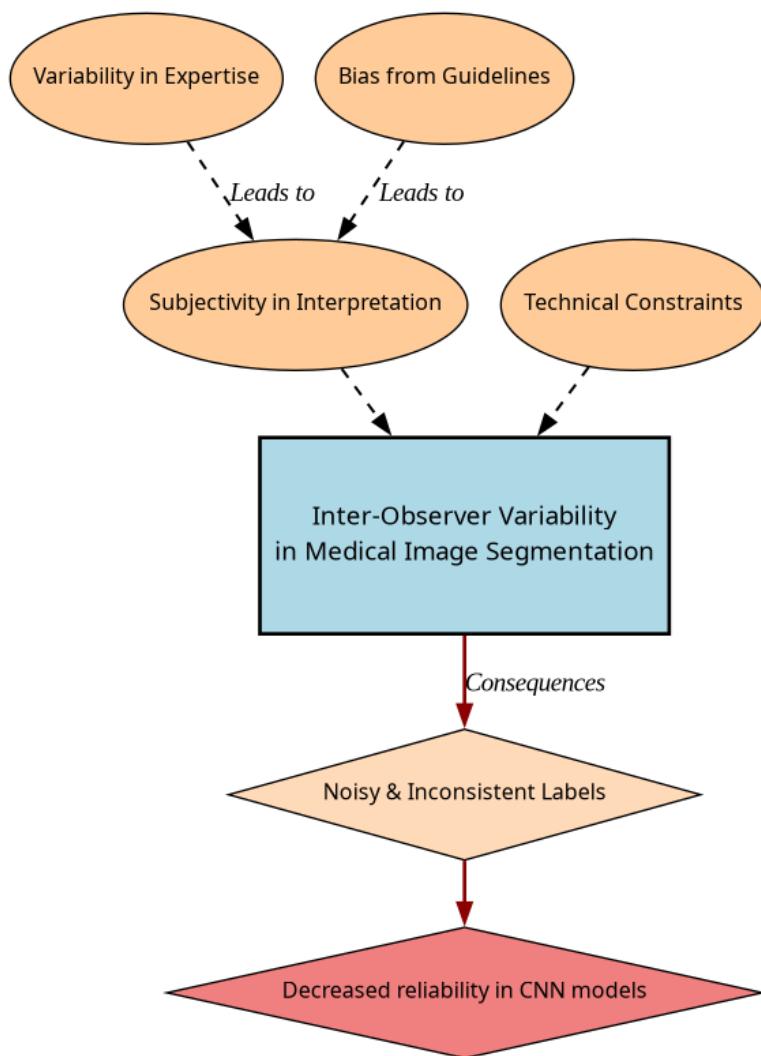


Figure 1-4 Summary diagram for problem Statement

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, while addressing challenges related to variability in expertise levels and technical constraints, and maintaining interpretability, generalization, and computational efficiency?

344

345 1.3 Literature review

346 Certainly, in general Machine Learning (ML) classification tasks ³ where multiple
347 annotators are involved, Majority Voting (MV) is by far the simplest possible
348 approach to implement. This concept was born multiple times and divergently in
349 multiple fields, but it was described as relevant for ML and pattern recognition
350 labeling for classification in [Lam and Suen, 1997], in which the approach is
351 exposed as simple, yet powerful. The authors describe the MV as a method that
352 can be used to improve the accuracy of classification tasks by combining the labels
353 of multiple annotators. The method is based on the assumption that the majority
354 vote of the annotators is more likely to be correct than the vote of a single
355 annotator. The authors also describe the method as a straightforward way to
356 improve the accuracy of classification tasks without the need for complex
357 algorithms or additional data. The authors also prove this method to deliver very
358 similar results to more complicated approaches (Bayesian, logistic regression,
359 fuzzy integral, and neural network) in the particular task of Optical Character
360 Recognition (OCR). Despite its simplicity, modern solutions for delivering accurate
361 medical image segmentation models still rely on Majority Voting at some stage,
362 like [Elnakib et al., 2020], which uses a majority voting strategy for delivering a
363 final output based on the labels of multiple models (VGG16-Segnet, Resnet-18 and
364 Alexnet) in Computed Tomography (CT) images for Liver Tumor Segmentation, or

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

[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

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

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

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

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

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

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

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

405 The popularity and performance of STAPLE has led to its usage in modern
 406 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 label aggregation,hence being a good support for other methods, but not that useful for providing annotations of structures on new and unlabeled images.

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

434 while increasing feature channels through a series of convolutional and
 435 max-pooling layers, capturing high-level semantic information. The decoder path
 436 uses transposed convolutions to gradually recover spatial resolution while
 437 reducing feature channels. Skip connections between corresponding encoder and
 438 decoder layers preserve fine-grained details by concatenating high-resolution
 439 features from the encoder with upsampled features in the decoder, enabling
 440 precise localization of structures.

441 **U-Net based approaches**

442 In [López-Pérez et al., 2024] two networks are trained for delivering a final
 443 segmentation. One network is trained to estimate the annotators reliability and
 444 another one is trained to segment the image. The first network is a deep neural
 445 network that takes as input features of image and the labelers id encoded as
 446 one-hot and outputs a reliability map across the image feature space. This map is
 447 then used to weight the contribution of each annotator to the final segmentation.
 448 The second network is the U-Net used for segmentation.

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

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

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

456 The authors then assume the annotator network to deliver a reliability map
 457 $\{\hat{\mathbf{A}}_\phi^{(r)}(\mathbf{x})\}_{r \in R_n}$ with different dimensions:

⁴⁵⁸ • CR global: a single reliability vector per labeler with dimensions C which
⁴⁵⁹ represent global reliability of the labeler across all input space.

⁴⁶⁰ • CR image: a single reliability vector per image per labeler with dimensions C
⁴⁶¹ which represent local reliability of the labeler across the image.

⁴⁶² • CR pixel: a reliability matrix per image per labeler, with dimensions C which
⁴⁶³ represent local reliability of the labeler across all the pixels in the image.

⁴⁶⁴ These differences in dimensions are determined by the feature extraction space
⁴⁶⁵ from segmentation network which feed the input of the annotator network, which
⁴⁶⁶ the authors vary for experimentation purposes.

⁴⁶⁷ Being $\mathbf{p}_\theta(\mathbf{x}_n)$ the estimation of the latent (ground truth) segmentation delivered by
⁴⁶⁸ the segmentation UNet network, thus, the estimated segmentation probability
⁴⁶⁹ mask for each annotator is given by the product:

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

⁴⁷⁰ where \odot is the element-wise product and ϕ and θ are the parameters of the
⁴⁷¹ annotator network and the segmentation UNet network, respectively, being the
⁴⁷² latter initialized with a ResNet34 backbone pre-trained on ImageNet.

⁴⁷³ The authors propose a loss function involving cross-entropy and a trace based
⁴⁷⁴ regularization on the reliability map, originally proposed in [Zhang et al., 2020]
⁴⁷⁵ which combined, looks like:

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

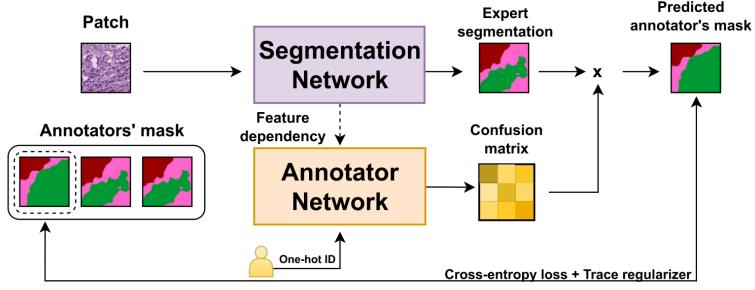


Figure 1-5 Proposed framework for the approach in [López-Pérez et al., 2024].

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

478 When evaluated on a Triple Negative Breast Cancer dataset, this approach achieves
 479 a Dice coefficient of 0.7827, outperforming STAPLE (0.7039) and matching expert-
 480 supervised performance (0.7723). The CR image reliability modeling proved most
 481 effective, as CR pixel, while potentially offering finer-grained reliability estimation,
 482 requires significantly more training data.

483 Despite the decent performance of the approach, solving the problem of multiple
 484 labelers with two networks can be overwhelming for the optimization process,
 485 requiring large amounts of annotated data to properly codify the annotators
 486 spatial reliabilities, which could be managed by a single model with an appropriate
 487 loss function.

488 Bayesian models

489 Bayesian approaches are a good choice for handling label noise and uncertainty in
 490 the labelers. In [Julián and Álvarez Meza Andrés Marino, 2023] the authors
 491 propose a novel approach from Gaussian Processes to model the relationship
 492 between the annotators' reliability and the input data, while also preserving the
 493 interdependencies among the annotators. This is achieved by introducing
 494 Correlated Chained Gaussian Processes for Multiple Annotators (CCGPMA), a

495 framework based on the well known Chained Gaussian Processes (CGP). CGP on
 496 itself cannot consider inter-annotator dependencies, thus, the authors introduce
 497 the Correlated Chained Gaussian Processes (CCGP) to model correlations between
 498 the GP latent functions, which are supposed to be generated from a
 499 Semi-Parametric Latent Factor Model (SLFM):

$$f_j(\mathbf{x}_n) = \sum_{q=1}^Q w_{j,q} \mu_q(\mathbf{x}_n), \quad (1-6)$$

500 where $f_j : \mathcal{X} \rightarrow \mathbb{R}$ is a Latent Function (LF), $\mu_q(\cdot) \sim \mathcal{GP}(0, k_q(\cdot, \cdot))$ with $k_q : \mathcal{X} \times \mathcal{X} \rightarrow$
 501 \mathbb{R} being a kernel function, and $w_{j,q} \in \mathbb{R}$ is a combination coefficient ($Q \in \mathbb{N}$). This
 502 leads to a joint distribution of the form:

$$p(\mathbf{y}, \hat{\mathbf{f}}, u | \mathbf{X}) = p(\mathbf{y} | \boldsymbol{\theta}) \prod_{j=1}^J p(\mathbf{f}_j | \mathbf{u}) p(\mathbf{u}), \quad (1-7)$$

503 where \mathbf{y} is the vector of noisy labels, $\hat{\mathbf{f}}$ is the vector of latent functions, u represents
 504 the inducing points, and \mathbf{X} is the input data.

505 Combined with inducing-variables based methods for sparse GP approximations,
 506 and maximizing an Evidence Lower Bound (ELBO) for the estimation of the
 507 variational parameters, the authors reach a model whose variational expectations
 508 are not analytically tractable, and hence, the authors derive a Gaussian-Hermite
 509 quadrature approach.

510 Finally, the authors extend this approach for being applied to classification and
 511 regression, reaching the only known approach to involve chained gaussian
 512 processes in multiple annotators classification and regression tasks while

513 preserving the interdependencies among the annotators, and also outperforming
514 GPC-MV⁴, MA-LFC-C⁵, MA-DGRL⁶, MA-GPC⁷, MA-GPCV⁸, MA-DL⁹, KAAR¹⁰.

515 CCGPMA on itself proposes a good approach for handling label noise and
516 uncertainty in the labelers for regression and classification tasks, while also
517 preserving the interdependencies among the annotators, however, it does not face
518 the image segmentation problem, which is the main focus of this works, however,
519 it does not face the image segmentation problem, which is the main focus of this
520 work. Besides, handling so many latent functions during the optimization process
521 is computationally expensive, making it on itself infeasible for large and high
522 resolution datasets.

523 1.3.2 Facing noisy annotations and low-quality data

524 The problem of low-quality data and noisy annotations has been tackled with
525 various strategies. One such approach is the use of deep learning models that
526 incorporate loss functions designed to mitigate the effects of unreliable labels.
527 Traditional methods such as Majority Voting (MV) or Expectation-Maximization
528 (EM) have been widely used for aggregating multiple annotators' inputs. However,
529 they assume a homogeneous reliability of annotators, which may not hold in
530 real-world scenarios.

⁴A GPC using the MV of the labels as the ground truth.

⁵A LRC with constant parameters across the input space.

⁶A multi-labeler approach that considers as latent variables the annotator performance.

⁷A multi-labeler GPC, which is an extension of MA-LFC.

⁸An extension of MA-GPC that includes variational inference and priors over the labelers' parameters.

⁹A Crowd Layer for DL, where the annotators' parameters are constant across the input space.

¹⁰A kernel-based approach that employs a convex combination of classifiers and codes labelers dependencies.

531 **Loss functions in deep learning models**

532 Loss functions are fundamental components in deep learning models that quantify
533 how well a model's predictions match the ground truth. They serve as the
534 objective function that guides the learning process by measuring the discrepancy
535 between predicted and actual values. In classification tasks, the most common
536 loss functions are Cross-Entropy (CE) and Mean Absolute Error (MAE). CE is
537 particularly effective for classification as it heavily penalizes confident but wrong
538 predictions, though it can be sensitive to noisy labels. MAE, on the other hand, is
539 more robust to outliers and assigns equal weights to all mistakes, but typically
540 requires more training iterations. For image segmentation tasks, specialized loss
541 functions have been developed to handle the unique challenges of pixel-wise
542 classification. The Dice loss, which measures the overlap between predicted and
543 ground truth regions, is widely used in medical image segmentation. More
544 recently, the Generalized Cross Entropy (GCE) loss has emerged as a robust
545 alternative that combines the benefits of both CE and MAE, allowing for better
546 handling of noisy labels through a tunable parameter that controls sensitivity to
547 outliers. In multi-annotator scenarios, where multiple experts provide potentially
548 inconsistent segmentations, novel loss functions like the Truncated Generalized
549 Cross Entropy for Semantic Segmentation (TGCE_{SS}) have been developed to
550 account for varying annotator reliability across different image regions. These loss
551 functions are crucial for training accurate segmentation models, especially in
552 medical imaging where precise delineation of anatomical structures is essential for
553 diagnosis and treatment planning.

554 **Generalized Cross-Entropy for multiple annotators classification**

555 A more recent approach was proposed by [Triana-Martinez et al., 2023],
556 introducing a Generalized Cross-Entropy-based Chained Deep Learning (GCECDL)
557 framework. This method addresses the limitations of traditional label aggregation
558 techniques by modeling each annotator's reliability as a function of the input data.

- 559 The approach effectively mitigates the impact of noisy labels by using a
 560 noise-robust loss function, balancing Mean Absolute Error (MAE) and Categorical
 561 Cross-Entropy (CE). Unlike prior approaches, GCECDL accounts for the
 562 dependencies among annotators while encoding their non-stationary behavior
 563 across different data samples. Their experiments on multiple datasets
 564 demonstrated superior predictive performance compared to state-of-the-art
 565 methods, particularly in cases where annotations were highly inconsistent.
- 566 The strategy of the authors effectively unlocks the potential of ML models to handle
 567 low-quality data and noisy annotations, but it is bounded to classifications tasks
 568 only, not being by itself applicable to segmentation tasks. The TGCE equation for
 569 handling multiple annotators is defined as:

$$\text{TGCE}(\mathbf{y}, f(\mathbf{x}); \tilde{\lambda}_x, \tilde{C}) = \tilde{\lambda}_x \frac{1 - (\mathbf{1}^\top (\mathbf{y} \odot f(\mathbf{x})))^q}{q} + (1 - \tilde{\lambda}_x) \frac{1 - (\tilde{C})^q}{q}, \quad (1-8)$$

- 570 where $\tilde{\lambda}_x$ represents the annotator reliability, \tilde{C} is a constant, q is a parameter that
 571 controls the balance between MAE and CE behavior, \mathbf{y} is the annotation vector, and
 572 $f(\mathbf{x})$ is the model prediction. This approach is more deeply discussed in chapter 4.

573 1.4 Aims

- 574 With the mentioned considerations in section 1.3 in mind, this work proposes a
 575 novel approach for ISS tasks in medical images, which aims to train a model whose
 576 learning approach is adaptive to the labeler performance. This is done by
 577 introducing a loss function capable of inferring the best possible segmentation
 578 without needing separate inputs about the labeler performance. This loss function
 579 is designed to implicitly weigh the labelers based on their performance, with the
 580 presence of an intermediate reliability map allowing the model to learn from the

581 most reliable labelers and ignore the noisy labels. This approach differs from
582 existing CNN-based segmentation models, as it does not require explicit
583 supervision of the labeler performance, making it more generalizable and
584 adaptable to different datasets and labelers.

585 **1.4.1 General Aim**

586 The main purpose of this work is to develop a novel approach for ISS tasks in
587 medical images, which can adaptively infer the best possible segmentation without
588 needing separate inputs about the labeler performance. This approach is expected
589 to outperform the segmentation performance of other state of the art approaches,
590 correctly facing the labeler performance inconsistency across the annotators space
591 and the variability of images quality.

592 **1.4.2 Specific Aims**

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

602 1.5 Outline and Contributions

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

- 605 • A python package for using the proposed loss function in CNN models for ISS
606 tasks in medical images. ¹¹
- 607 • Datasets mapping as lazy loaders for the proposed loss function. ¹²
- 608 • A public Github repository with the code used in this work. ¹³

¹¹https://pypi.org/project/seg_tgce/

¹²<https://seg-tgce.readthedocs.io/en/latest/experiments.html>

¹³https://github.com/blotero/seg_tgce

609

610

CHAPTER

611

612

TWO

613

CONCEPTUAL PRELIMINARIES

614

2.1 Modern concept of digital image

615 A digital image is a numerical representation of a visual scene, captured through
616 various imaging devices and stored in a computer. From a mathematical perspective,
617 a digital image can be represented as a function $f(x, y)$ that maps spatial coordinates
618 (x, y) to intensity values. In the discrete domain, this function is sampled at regular
619 intervals, creating a matrix of values known as pixels (picture elements).

620

2.1.1 Types of digital images

621

Grayscale images

622 Grayscale images are the simplest form of digital images, where each pixel
623 represents a single intensity value. Mathematically, a grayscale image can be
624 represented as a 2D matrix I of size $M \times N$, where each element $I(i, j)$ represents
625 the intensity at position (i, j) . The intensity values typically range from 0 (black)
626 to 255 (white) in 8-bit images, or from 0 to 65535 in 16-bit images.

627 Color images

628 Color images extend the grayscale concept by representing each pixel with multiple
629 channels, typically Red, Green, and Blue (RGB). A color image can be represented
630 as a 3D matrix I of size $M \times N \times 3$, where $I(i, j, k)$ represents the intensity of the
631 k -th color channel at position (i, j) . Other color spaces like HSV (Hue, Saturation,
632 Value) or CMYK (Cyan, Magenta, Yellow, Key) are also commonly used in different
633 applications.

634 Multispectral images

635 Multispectral images capture information across multiple wavelength bands
636 beyond the visible spectrum. These images can be represented as a 3D matrix I of
637 size $M \times N \times B$, where B is the number of spectral bands. Each band $I(i, j, b)$
638 represents the intensity at position (i, j) for the b -th spectral band. This
639 representation is particularly useful in medical imaging, remote sensing, and
640 scientific applications.

641 3D images and volumetric data

642 Three-dimensional images extend the concept of pixels to voxels (volume elements).
643 A 3D image can be represented as a 3D matrix V of size $M \times N \times D$, where D
644 represents the depth dimension. Each voxel $V(i, j, k)$ represents the intensity at
645 position (i, j, k) in the 3D space. This representation is fundamental in medical
646 imaging (CT, MRI), scientific visualization, and computer graphics.

647 2.1.2 Mathematical representations

648 The mathematical foundation of digital images relies on several key concepts:

- 649 • **Sampling:** The process of converting a continuous image into a discrete
650 representation. According to the Nyquist-Shannon sampling theorem, the
651 sampling frequency must be at least twice the highest frequency present in
652 the image to avoid aliasing.
- 653 • **Quantization:** The process of converting continuous intensity values into
654 discrete levels. The number of quantization levels determines the image's
655 bit depth and affects its quality and storage requirements.
- 656 • **Resolution:** The number of pixels per unit length in an image, typically
657 measured in pixels per inch (PPI) or dots per inch (DPI).
- 658 • **Dynamic range:** The ratio between the maximum and minimum measurable
659 light intensities in an image, often expressed in decibels (dB).

660 The mathematical representation of a digital image can be expressed as:

$$I(x, y) = \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} f(i, j) \cdot \delta(x - i, y - j) \quad (2-1)$$

661 where $I(x, y)$ is the digital image, $f(i, j)$ represents the intensity values, and $\delta(x -$
662 $i, y - j)$ is the Kronecker delta function.

663 For color images, the representation extends to:

$$I(x, y) = \begin{bmatrix} I_R(x, y) \\ I_G(x, y) \\ I_B(x, y) \end{bmatrix} \quad (2-2)$$

664 where I_R , I_G , and I_B represent the red, green, and blue channels respectively.

665 2.2 Digital histopathological images

666 Digital histopathology represents a significant advancement in medical imaging,
 667 where traditional glass slides containing tissue samples are digitized using
 668 specialized scanning devices. This transformation has revolutionized the field of
 669 pathology by enabling remote diagnosis, computer-aided analysis, and digital
 670 archiving of tissue samples [Amgad et al., 2019]. This process has evolved
 671 significantly over the past few decades, as shown in Figure 2-1.

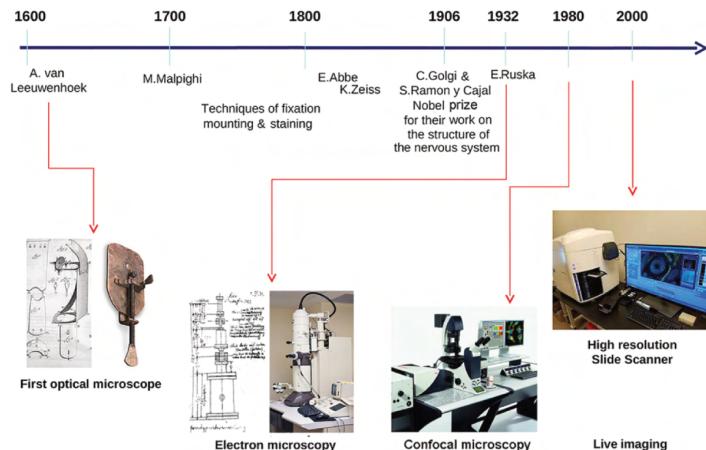


Figure 2-1 Histology evolution timeline. (Image from [Mazzarini et al., 2021]).

672 2.2.1 Whole Slide Imaging (WSI)

673 Whole Slide Imaging (WSI) is the process of digitizing entire glass slides at high
 674 resolution, creating a digital representation that can be viewed, analyzed, and
 675 shared electronically. Modern WSI scanners use sophisticated optical systems that
 676 capture multiple fields of view at high magnification, which are then stitched
 677 together to create a seamless digital image [Hu et al., 2025]. These systems
 678 incorporate high-resolution objectives with magnifications ranging from 20x to
 679 40x, precise motorized stages for accurate slide positioning, automated focus

systems to maintain image quality, and high-quality cameras equipped with large sensor arrays. The resulting digital slides can reach sizes of several gigabytes, containing billions of pixels that capture the microscopic details of tissue samples [Hu et al., 2025]. Figure 2-2 shows a whole slide imaging system by Omnyx for slide digitization and a comprehensive digital pathology interface from Omnyx designed to streamline pathologists' diagnostic workflow [Farahani et al., 2015].

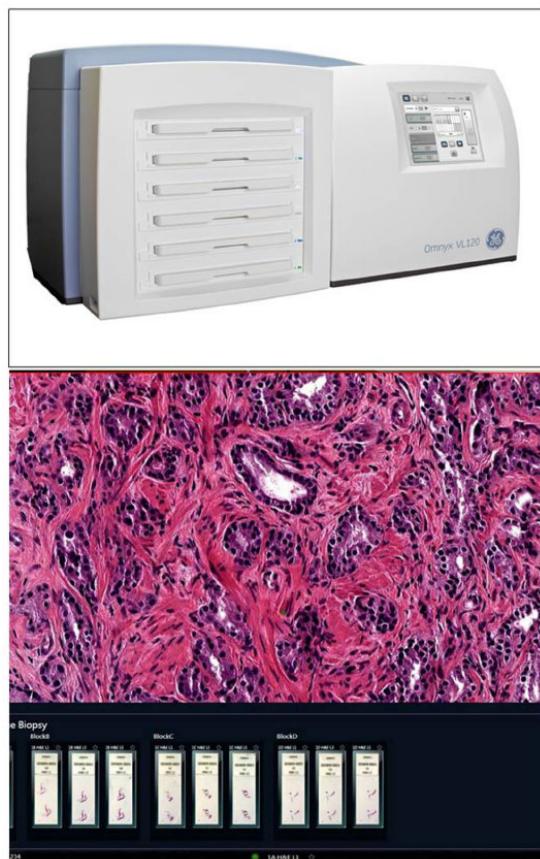


Figure 2-2 (Above) Whole slide imaging system by Omnyx for slide digitization. (Below) Comprehensive digital pathology interface from Omnyx designed to streamline pathologists' diagnostic workflow. (From [Farahani et al., 2015]).

686 2.2.2 Regions of Interest (ROI)

687 In digital histopathology, Region of Interests (ROIs) are specific areas within a
688 whole slide image that contain diagnostically relevant information. These regions
689 can be manually annotated by pathologists, automatically detected using
690 computer vision algorithms, or defined based on specific tissue characteristics or
691 abnormalities. The importance of ROIs lies in their ability to focus computational
692 analysis on relevant areas, reduce computational complexity in automated
693 systems, facilitate targeted diagnosis and research, and enable efficient storage and
694 transmission of critical information.

695 2.2.3 Staining Techniques

696 Histopathological analysis relies heavily on various staining techniques to enhance
697 the visibility of different tissue components and cellular structures. The choice of
698 staining method depends on the specific diagnostic requirements and the type of
699 tissue being examined.

700 Hematoxylin and Eosin (H&E)

701 Hematoxylin and Eosin (H&E) staining is the most widely used technique in
702 histopathology, particularly in breast cancer diagnosis [Pan et al., 2021]. This
703 staining method provides essential visualization through two components:
704 hematoxylin, which stains cell nuclei blue/purple to highlight nuclear morphology,
705 and eosin, which stains cytoplasm and extracellular matrix pink/red to reveal
706 tissue architecture.

707 The popularity of H&E staining in breast cancer histopathology stems from its
708 ability to clearly visualize tumor architecture and growth patterns, distinguish
709 between different types of breast cancer, identify important diagnostic features

710 like nuclear pleomorphism, and assess tumor grade and stage. Beyond breast
711 cancer, H&E staining finds extensive application across various medical specialties
712 including general pathology, dermatology, gastroenterology, neurology, and
713 oncology.

714 **Special Stains**

715 In addition to H&E, various special stains are used for specific diagnostic purposes.
716 Immunohistochemistry (IHC) uses antibodies to detect specific proteins, playing a
717 crucial role in subtyping breast cancers. Key IHC stains include Estrogen Receptor
718 (ER) staining for detecting estrogen receptors, Progesterone Receptor (PGR)
719 staining for assessing progesterone receptor status, Human Epidermal Growth
720 Factor Receptor 2 (HER2) staining for evaluating HER2 protein expression, and
721 Ki67 staining for measuring cellular proliferation rates. These markers are
722 particularly crucial in breast cancer diagnosis and treatment planning, as they help
723 determine the molecular subtype of the cancer and guide personalized therapeutic
724 approaches. Other specialized stains include Periodic Acid-Schiff (PAS) for
725 highlighting carbohydrates and basement membranes, Masson's Trichrome for
726 distinguishing between collagen and muscle fibers, and silver stains for detecting
727 microorganisms and nerve fibers. These specialized staining techniques
728 complement H&E by providing additional diagnostic information that is crucial for
729 accurate diagnosis and treatment planning [Weitz et al., 2023]. Examples of these
730 staining techniques are shown in Figure 2-8.

731 **2.3 Deep learning fundamentals**

732 Deep learning has emerged as a powerful subset of machine learning,
733 revolutionizing the field of artificial intelligence. Its roots can be traced back to the
734 early development of artificial neural networks in the 1940s and 1950s, with
735 significant milestones including the perceptron in 1958 and the backpropagation

⁷³⁶ algorithm in the 1980s. However, it wasn't until the early 21st century, with the
⁷³⁷ advent of more powerful computational resources and the availability of large
⁷³⁸ datasets, that deep learning truly began to flourish.

⁷³⁹ 2.3.1 Learning Paradigms

⁷⁴⁰ Deep learning systems can be categorized into three main learning paradigms. The
⁷⁴¹ most common approach is supervised learning, where models learn from labeled
⁷⁴² data by mapping inputs to known outputs. This paradigm requires a large amount
⁷⁴³ of labeled training data, which can be expensive and time-consuming to acquire.
⁷⁴⁴ Semi-supervised learning offers a hybrid approach that leverages both labeled and
⁷⁴⁵ unlabeled data, proving particularly useful when labeled data is scarce but
⁷⁴⁶ unlabeled data is abundant. Finally, unsupervised learning enables models to
⁷⁴⁷ discover patterns and structures from unlabeled data without explicit guidance,
⁷⁴⁸ making it valuable for tasks like clustering and dimensionality reduction.

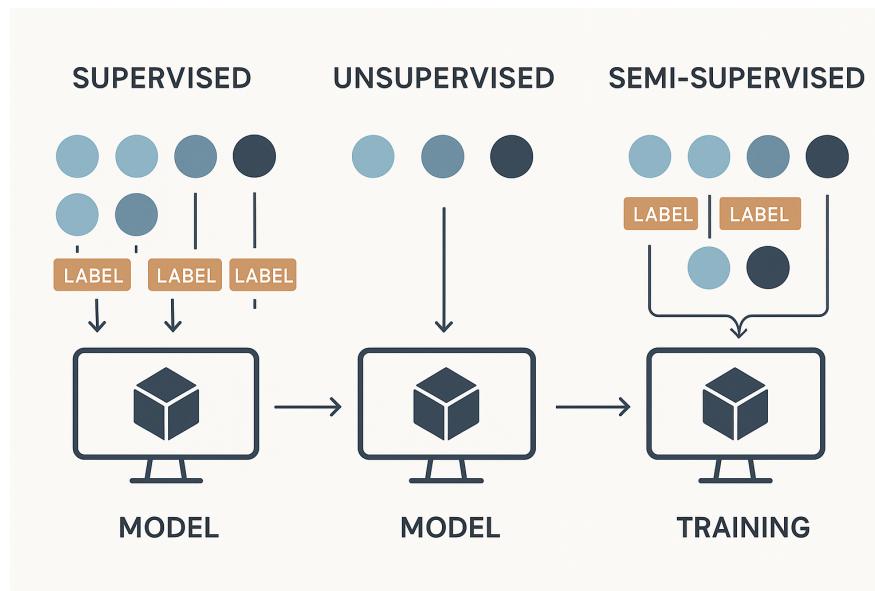


Figure 2-3 Common learning paradigms.

749 2.3.2 Architecture and Training

750 Deep learning architectures are characterized by their layered structure, where
751 each layer progressively extracts and transforms features from the input data. The
752 early layers typically focus on low-level feature extraction, such as edges, textures,
753 and basic patterns in the case of image processing. As information flows through
754 the network, middle layers combine these basic features into more complex
755 representations. The final layers perform high-level reasoning and make the
756 ultimate predictions or classifications.

757 The training process relies heavily on the gradient descent algorithm, which
758 iteratively adjusts the model's parameters to minimize a loss function. This loss
759 function serves as a crucial component of the learning process, quantifying how
760 well the model's predictions match the actual targets. By providing a measure of
761 the model's performance, the loss function guides the optimization process,
762 enabling the network to learn meaningful patterns from the training data.

763 2.3.3 Challenges and Solutions

764 Despite their power, deep learning systems face several significant challenges. One
765 of the most prominent issues is overfitting, where models may memorize training
766 data instead of learning generalizable patterns. This challenge is typically
767 addressed through various regularization techniques such as dropout, L1/L2
768 regularization, and early stopping. Another critical challenge is the substantial
769 data requirements; deep learning models often need massive amounts of training
770 data to achieve good performance, which can be a limiting factor in many
771 applications. Additionally, the complex, layered nature of deep learning models
772 makes them difficult to interpret, often referred to as "black boxes." This lack of
773 transparency can be particularly problematic in critical applications where
774 understanding the decision-making process is essential.

775 2.3.4 Deep Learning Frameworks

776 The development of powerful open-source frameworks has significantly
777 accelerated deep learning research and applications. TensorFlow, developed by
778 Google, provides a comprehensive ecosystem for building and deploying machine
779 learning models [Abadi et al., 2016]. PyTorch, created by Facebook’s AI Research
780 lab, offers dynamic computation graphs and has become particularly popular in
781 research settings. Caffe, known for its speed and modularity, is widely used in
782 computer vision applications.

783 These frameworks have democratized deep learning by providing efficient
784 implementations of common operations, automatic differentiation for gradient
785 computation, and GPU acceleration for faster training. They also offer pre-trained
786 models and transfer learning capabilities, along with active communities for
787 support and knowledge sharing. The combination of these frameworks with
788 modern hardware has enabled researchers and practitioners to develop
789 increasingly sophisticated models, pushing the boundaries of what’s possible in
790 artificial intelligence. As shown in Figure 2-4, which presents data from Google
791 Trends over the last five years (as of April 2025), TensorFlow and PyTorch have
792 emerged as the two most prominent frameworks in the deep learning landscape.

793 2.4 Datasets and data sources

794 Throughout the development of this work, multiple datasets were used for
795 evaluation of ISS models. The common elements of all these datasets are that they
796 contain RGB images and are crowdsourced with multiple labelers, where not
797 necessarily all labeler label all images.

798 As it has been mentioned in Chapter 1, the main goal of this work is mainly
799 focused on crowdsourced histopathology images semantic segmentation, however,
800 these datasets present the following challenges:

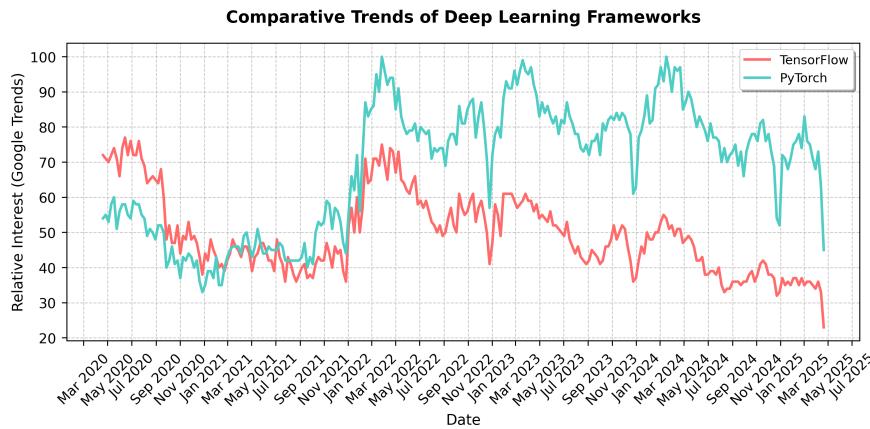


Figure 2-4 Comparative Trends of the top two most popular Deep Learning Frameworks, apparently, tendency was switched to PyTorch since 2022

- 801 • Distribution of segmentation labels is not uniform across the image, since
802 some tissues and structures are more common than others.
- 803 • Visualization of performance of the models in debug time (like per epoch
804 analysis) is not simple for non experts in the subject, which makes it hard to
805 evaluate whether the model is overfitting or not at a glance.
- 806 For these reasons, multiple datasets were created in the pursuit of an initial
807 evaluation of performance of the models against more traditional and familiar
808 images before the focus on histopathology images. Once a decent performance in
809 metrics like Dice coefficient was achieved, the focus was shifted to histopathology
810 images and further tunings on the models were performed if needed.
- 811 In any case, both the emulated noisy annotations datasets and the histopathology
812 datasets somehow contained ground truth aggregation, either from the original
813 source (in the case of emulated noisy annotations), the expert annotation (if
814 available) or from the aggregation of multiple labelers ¹.

¹STAPLE in the case of histopathology datasets with no expert annotations available

815 2.4.1 Datasets with emulated noisy annotations

- 816 A challenge arises for the creation of emulated noisy annotations datasets, since it
 817 is expected for images annotations to have some degree of expertise variability,
 818 similar to what is expected in real crowdsourced datasets. Simply introducing
 819 random noise into the annotated masks does not work, since the original
 820 morphological structures from the expected ground truth are far from being
 821 preserved. Instead, a “noisy” labeler is expected to produce an annotation which
 822 has at least some degree of morphological consistency on itself, even if it shows
 823 discrepancies when compared with some metric (like DICE score) against the
 824 ground truth annotation.
- 825 This has been proven experimentally when introducing random noise into any
 826 popular segmentation dataset, in which the resulting mask is just a non coherent
 827 map of noise across the image, as shown in Figure 2-5.

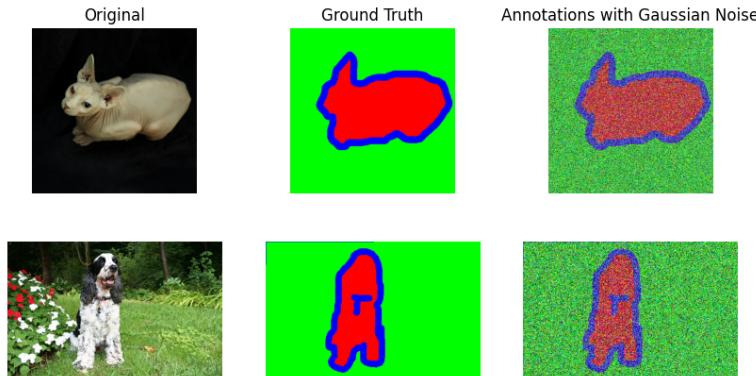


Figure 2-5 Example of a noisy mask generated by naively introducing random noise into a ground truth mask. Morphological consistency is lost.

- 828 For this reason, a more sophisticated approach was needed to create datasets with
 829 emulated noisy annotations. The approach used here was to use a pre-trained
 830 U-Net model to produce a good enough segmentation of the image, which was
 831 then used to produce a “noisy” annotation by introducing random noise into the
 832 last encoder layers of the model, thus preserving the morphological consistency of

833 the original annotation. This strategy works since slight modifications introduced
834 into the encoder layers weights, somehow resemble conceptual disturbances with
835 respect to the original ground truth label, which kind of emulates the cultural
836 behavior of having a different interpretation of the image, either for having a
837 different level of expertise or for having a different point of view or school of
838 thought. The first encoding layers were not modified, since it is expected human
839 labelers would agree on the most fundamental structures (analog to extracted
840 features from initial convolutional layers) in a similar way, thus preserving the
841 morphological consistency of the original annotation.

842 In this way, the level of “disturbance” with respect to the ground truth for an
843 emulated annotator can be controlled by the level of noise introduced into the
844 weights of the last encoder layers, thus:

$$\mathbf{W}_{noisy} = \mathbf{W}_{original} + \mathcal{N}(0, \sigma^2) \quad (2-3)$$

845 where $\mathbf{W}_{original}$ represents the original weights of the last encoder layers, $\mathcal{N}(0, \sigma^2)$
846 is a Gaussian distribution with mean 0 and variance σ^2 , and \mathbf{W}_{noisy} are the resulting
847 noisy weights. The variance σ^2 controls the level of noise introduced, and thus the
848 degree of disturbance in the resulting segmentation masks.

849 Oxford-IIIT Pet Dataset

850 Using the techniques described above, the Oxford-IIIT Pet Dataset [Parkhi et al.,
851 2012] was used to create a dataset with emulated noisy annotations. The almost
852 perfectly uniform distribution of the dataset classes makes it an ideal playground
853 dataset for testing segmentation models with a high degree of confidence in the
854 ground truth annotations, at the same time that cats and dogs are a common sight
855 in the daily life of most people, making it easier to find a labeler that is able to

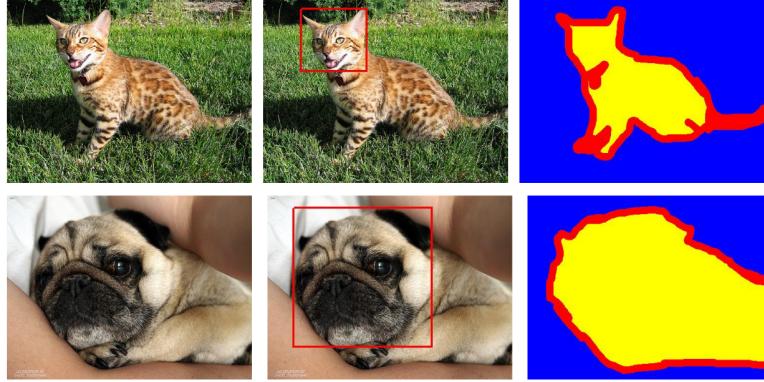


Figure 2-6 Annotations in the Oxford-IIIT Pet data. From left to right: pet image, head bounding box, and trimap segmentation (blue: background region; red: ambiguous region; yellow: foreground region).

856 annotate the images with a high degree of accuracy, which facilitates model initial
857 debugging. Figure 2-6 shows an example of the annotations in the original dataset.

858 With the application of the encoder layer weight perturbation technique, the
859 resulting noisy masks are shown in Figure 2-7. It can be seen that the
860 morphological consistency is preserved, even though the resulting masks are far
861 from the ground truth annotations, which goes perfectly well for testing the
862 robustness of the models against noisy annotations in crowdsourcing-like
scenarios.

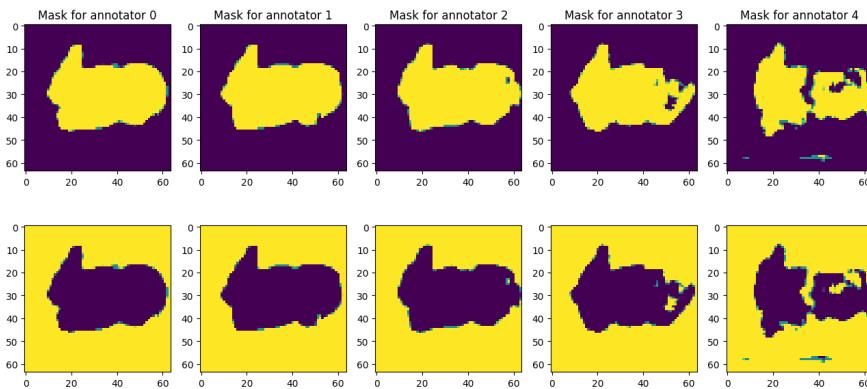


Figure 2-7 Noisy mask generated by enhancing the disturbances in the encoder layers weights for the Oxford-IIIT Pet Dataset. Morphological consistency is preserved. From left to right, SNR levels of noise in the encoder layer are 10, 5, 2, 0, -5 dB.

864 **2.4.2 Real histopathology datasets**

865 **Multi-Stain Breast Cancer Histological Dataset**

866 The Multi-Stain Breast Cancer Histological Dataset [Weitz et al., 2023] represents
867 one of the largest publicly available collections of whole slide images (WSIs) from
868 surgical resection specimens of primary breast cancer patients. This dataset is
869 particularly valuable for our work because it contains matched pairs of H&E and
870 IHC-stained tissue sections from the same tumor, with a total of 4,212 WSIs from
871 1,153 patients. The IHC stains include important biomarkers such as ER, PGR,
872 HER2, and KI67, which are routinely used in breast cancer diagnosis and
873 treatment planning (more on staining techniques in Section 2.2.3).

874 The dataset's relevance to our work stems from several key aspects. The matched
875 H&E and IHC stains allow for studying the consistency of segmentation across
876 different staining modalities, which is crucial for understanding how different
877 visualization methods affect annotation quality. With 1,153 patients, the dataset
878 provides a robust foundation for training and evaluating segmentation models in a
879 real-world clinical setting. The inclusion of routine diagnostic cases makes the
880 dataset representative of actual clinical practice, where variations in staining
881 quality and tissue preparation are common. Furthermore, the multiple biomarker
882 stains (ER, PGR, HER2, KI67) enable the study of how different tissue
883 characteristics affect segmentation performance and annotator agreement.

884 This dataset serves as an ideal testbed for our crowdsourced segmentation
885 approach. It allows us to evaluate how different staining modalities affect
886 annotator performance and agreement, while also providing insights into the
887 relationship between tissue characteristics and segmentation difficulty. The
888 dataset's comprehensive nature enables validation of our models' performance
889 across different biomarker expressions and assessment of the generalizability of
890 our approach to real-world clinical data.

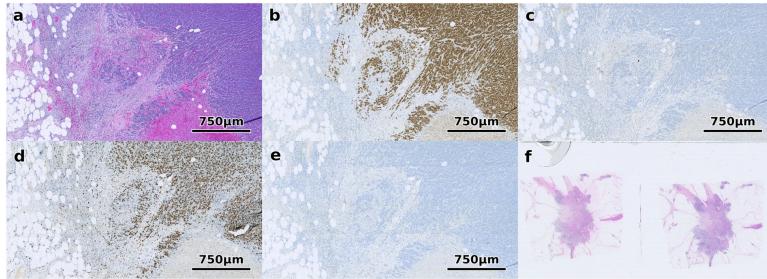


Figure 2-8 Different staining techniques obtained from multi-stain breast cancer dataset [Weitz et al., 2023]. (a) shows H&E, (b) ER, (c) HER2, (d) Ki67 and (e) PGR. (f) shows an example of a WSI that was excluded since it contains multiple tissue sections.

891 Structured Crowdsourcing Dataset for Histology Images

892 The dataset presented by [Amgad et al., 2019] is particularly relevant to our work
 893 as it represents one of the first systematic studies of crowdsourced annotations in
 894 histopathology. The authors recruited 25 participants with varying levels of
 895 expertise (from senior pathologists to medical students) to delineate tissue regions
 896 in 151 breast cancer slides using the Digital Slide Archive platform, resulting in
 897 over 20,000 annotated tissue regions.

898 Key aspects of this dataset make it valuable for our work. The systematic
 899 evaluation of inter-participant discordance revealed varying levels of agreement
 900 across different tissue classes, with low discordance for tumor and stroma, and
 901 higher discordance for more subjectively defined or rare tissue classes. The
 902 inclusion of feedback from senior participants helped in curating high-quality
 903 annotations, demonstrating that fully convolutional networks trained on these
 904 crowdsourced annotations can achieve high accuracy (mean AUC=0.945). The
 905 dataset also provides evidence that the scale of annotation data significantly
 906 improves image classification accuracy.

907 This dataset is particularly valuable for our work because it provides crucial
 908 insights into how annotator expertise affects segmentation quality. It
 909 demonstrates the feasibility of using crowdsourced annotations for training
 910 accurate segmentation models, showing that even with varying levels of expertise,

911 aggregated annotations can produce reliable ground truth. The dataset includes a
912 systematic analysis of inter-annotator agreement, which is crucial for
913 understanding the challenges in crowdsourced histopathology segmentation and
914 informing the development of more robust segmentation approaches.

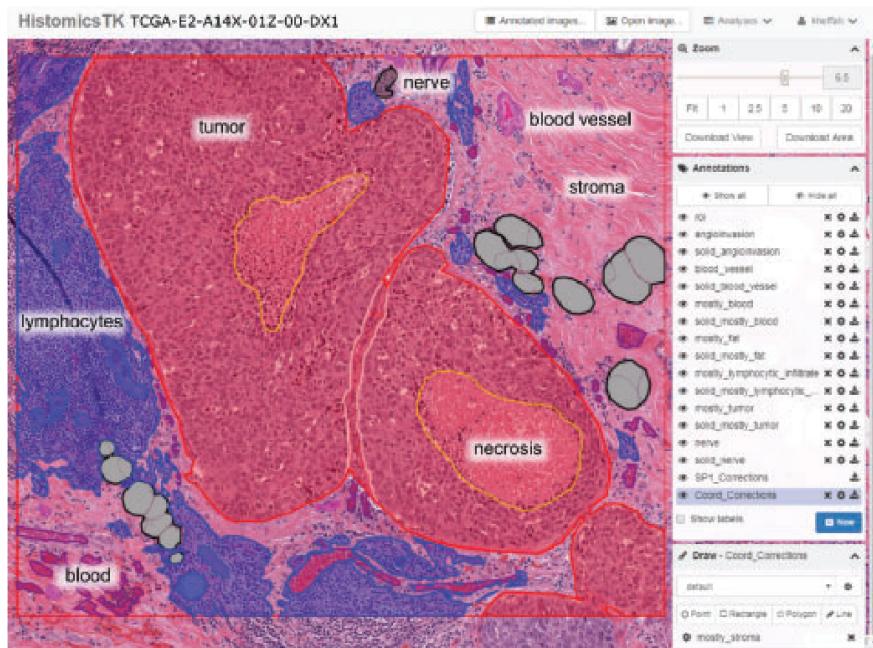


Figure 2-9 Screenshot of the DSA and HistomicsTK web interface while creating the crowdsourced annotations for the dataset presented by [Amgad et al., 2019].

915

916

CHAPTER

917

THREE

918

919

CHAINED GAUSSIAN PROCESSES

920

3.1 Gaussian processes

921

3.2 Chained Gaussian processes

922

3.3 Experimental evaluation

923

924

CHAPTER

925

926

FOUR

927

928

TRUNCATED GENERALIZED CROSS ENTROPY FOR SEGMENTATION

929

4.1 Loss functions for multiple annotators

930 As mentioned in Section 2.3.2, a loss function is a key element for defining the
931 objective function of a deep learning model. The categorical cross-entropy loss is a
932 common loss function for classification tasks. However, in the case of multiple
933 annotators, the categorical cross-entropy loss is not able to handle the varying
934 reliability of the annotators. In this section, we will propose a loss function that is
935 able to handle multiple annotators' segmentation masks while accounting for their
936 varying reliability across different regions of the image.

937 4.1.1 Generalized Cross Entropy

938 The Generalized Cross Entropy (GCE) loss function was first introduced by [Zhang
 939 and Sabuncu, 2018] as a robust alternative to the standard cross-entropy loss,
 940 particularly effective in handling noisy labels. Let us first consider the Cross
 941 Entropy (CE) and Mean Absolute Error (MAE) loss functions:

$$MAE(\mathbf{y}, f(\mathbf{x})) = \|\mathbf{y} - f(\mathbf{x})\|_1 \quad (4-1)$$

$$CE(\mathbf{y}, f(\mathbf{x})) = \sum_{k=1}^K y_k \log(f_k(\mathbf{x})) \quad (4-2)$$

942 where $y_k \in \mathbf{y}$, $f_k(\mathbf{x}) \in f(\mathbf{x})$, and $\|\cdot\|_1$ stands for the l_1 -norm. Of note, $\mathbf{1}^\top \mathbf{y} =$
 943 $\mathbf{1}^\top f(\mathbf{x}) = 1$, $\mathbf{1} \in \{1\}^K$ being an all-ones vector. In addition, the MAE loss can be
 944 rewritten for softmax outputs, yielding:

$$MAE(\mathbf{y}, f(\mathbf{x})) = 2(1 - \mathbf{1}^\top (\mathbf{y} \odot f(\mathbf{x}))) \quad (4-3)$$

945 where \odot stands for the element-wise product.

946 The CE loss function exhibits several distinct characteristics that make it
 947 particularly sensitive to noisy labels. It is unbounded from above and heavily
 948 penalizes confident but wrong predictions, which can lead to instability in the
 949 presence of label noise. In contrast, the MAE loss offers a more robust alternative
 950 with its bounded nature and symmetric properties in softmax-based
 951 representations. While the MAE loss is more resilient to noisy labels, it tends to

952 train more slowly due to its equal weighting of all mistakes regardless of
 953 confidence.

954 The GCE loss function, as defined by the authors in [Zhang and Sabuncu, 2018],
 955 provides a flexible framework that bridges these two approaches:

$$GCE(\mathbf{y}, f(\mathbf{x})) = 2 \frac{1 - (\mathbf{1}^\top (\mathbf{y} \odot f(\mathbf{x})))^q}{q}, \quad (4-4)$$

956 with $q \in (0, 1]$. Remarkably, the limiting case for $q \rightarrow 0$ in GCE is equivalent to the
 957 CE expression, and when $q = 1$, it equals the MAE loss. In addition, the GCE holds
 958 the following gradient with regard to θ :

$$\frac{\partial GCE(\mathbf{y}, f(\mathbf{x}; \theta) | k)}{\partial \theta} = -f_k(\mathbf{x}; \theta)^{q-1} \nabla_\theta f_k(\mathbf{x}; \theta). \quad (4-5)$$

959 The GCE loss combines the best of both worlds, offering robustness to label noise
 960 while maintaining the convexity property necessary for optimization. The
 961 truncation parameter q provides a mechanism to control the sensitivity to outliers,
 962 allowing for fine-tuning of the loss function's behavior based on the specific
 963 characteristics of the dataset.

964 4.1.2 Extension to Multiple Annotators

965 In the context of multiple annotators, we need to consider the varying reliability
 966 of each annotator across different regions of the image. Let's consider a k -class
 967 multiple annotators segmentation problem with the following data representation:

$$\mathbf{X} \in \mathbb{R}^{W \times H}, \{\mathbf{Y}_r \in \{0, 1\}^{W \times H \times K}\}_{r=1}^R; \quad \mathbf{Y} \in [0, 1]^{W \times H \times K} = f(\mathbf{X}) \quad (4-6)$$

968 where the segmentation mask function maps the input to output as:

$$f : \mathbb{R}^{W \times H} \rightarrow [0, 1]^{W \times H \times K} \quad (4-7)$$

969 The segmentation masks \mathbf{Y}_r satisfy the following condition for being a softmax-like
970 representation:

$$\mathbf{Y}_r[w, h, :] \mathbf{1}_k^\top = 1; \quad w \in W, h \in H \quad (4-8)$$

971 4.1.3 Reliability Maps and Truncated GCE

972 The key innovation in our approach is the introduction of reliability maps Λ_r for
973 each annotator:

$$\left\{ \Lambda_r(\mathbf{X}; \theta) \in [0, 1]^{W \times H} \right\}_{r=1}^R \quad (4-9)$$

974 These reliability maps serve as a sophisticated mechanism to estimate the
975 confidence of each annotator at every spatial location (w, h) in the image. By
976 learning these maps jointly with the segmentation model, the network gains the
977 ability to adapt to the varying levels of expertise across different regions of the
978 image. This approach allows for dynamic weighting of each annotator's

979 contribution based on their estimated reliability in specific areas, effectively
 980 handling cases where annotators might demonstrate different levels of expertise in
 981 different parts of the image.

982 The proposed Truncated Generalized Cross Entropy for Semantic Segmentation
 983 ($TGCE_{SS}$) combines the robustness of GCE with the flexibility of reliability maps:

$$TGCE_{SS}(\mathbf{Y}_r, f(\mathbf{X}; \theta)|_r(\mathbf{X}; \theta)) = \mathbb{E}_r \left\{ \mathbb{E}_{w,h} \left\{ \Lambda_r(\mathbf{X}; \theta) \circ \mathbb{E}_k \left\{ \mathbf{Y}_r \circ \left(\frac{\mathbf{1}_{W \times H \times K} - f(\mathbf{X}; \theta)^{\circ q}}{q} \right); k \in K \right\} + \right. \right. \right. \\ \left. \left. \left. (\mathbf{1}_{W \times H} - \Lambda_r(\mathbf{X}; \theta)) \circ \left(\frac{\mathbf{1}_{W \times H} - (\frac{1}{k} \mathbf{1}_{W \times H})^{\circ q}}{q} \right); w \in W, h \in H \right\}; r \in R \right\} \right. \quad (4-10)$$

984 where $q \in (0, 1)$ controls the MAE or CE level in the same way as in the GCE loss
 985 function. The loss function consists of two main components:

- 986 • The first term weighted by Λ_r represents the GCE loss for regions where the
 987 annotator is considered reliable
- 988 • The second term weighted by $(1 - \Lambda_r)$ provides a uniform prior for regions
 989 where the annotator is considered unreliable

990 For a batch containing N samples, the total loss is computed as:

$$\mathcal{L}(\mathbf{Y}_r[n], f(\mathbf{X}[n]; \theta)|_r(\mathbf{X}[n]; \theta)) = \frac{1}{N} \sum_n^{N} TGCE_{SS}(\mathbf{Y}_r[n], f(\mathbf{X}[n]; \theta)|_r(\mathbf{X}[n]; \theta)) \quad (4-11)$$

991

992

CHAPTER

993

FIVE

994

995

CHAINED DEEP LEARNING FOR IMAGE SEGMENTATION

996

997

5.1 Introduction

998

As mentioned in Chapter 1, U-shaped CNNs have been proven a good solution for segmentation tasks in medical images, due to their ability to capture both global and local information with relatively low datasets.

999

1000

1001

5.2 Using U-NET as a building block

1002

Our proposed model architecture combines the strengths of UNET with a ResNet-34 backbone, specifically designed to work with the $TGCE_{SS}$ loss function. The architecture is illustrated in Figure 5-2.

1003

1004

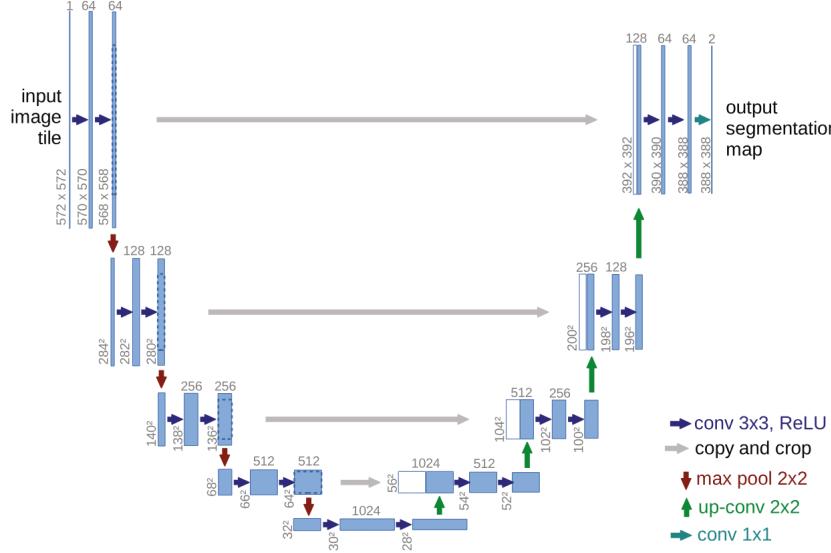


Figure 5-1 Original U-NET architecture.

5.2.1 Backbone Architecture

The model employs a pre-trained ResNet-34 as its encoder backbone, leveraging its deep residual learning framework for efficient feature extraction. The choice of ResNet-34 provides several key advantages: efficient feature extraction through residual connections, pre-trained weights that capture rich visual representations, and stable gradient flow during training. We modify the ResNet-34 backbone to serve as the encoder in our UNET architecture by removing the final fully connected layer and utilizing the feature maps from different stages of the network for skip connections.

5.2.2 UNET Architecture

The UNET architecture follows a traditional encoder-decoder structure with skip connections, where the encoder path implements the ResNet-34 structure. The decoder path employs transposed convolutions for upsampling, creating a

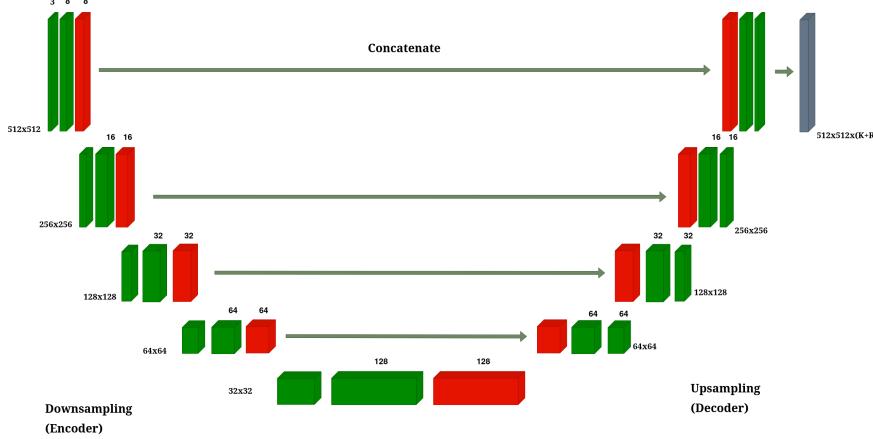


Figure 5-2 Overview of the proposed model architecture.

1018 symmetrical architecture that effectively captures both high-level and low-level
 1019 features. The architecture incorporates four downsampling stages in the encoder,
 1020 corresponding to the ResNet-34 blocks, and four upsampling stages in the decoder.
 1021 These stages are connected through skip connections that bridge corresponding
 1022 encoder and decoder stages, allowing the network to preserve fine-grained details.
 1023 Each convolution operation is followed by batch normalization and ReLU
 1024 activation to ensure stable training and effective feature learning.

1025 5.2.3 Reliability Map Branch

1026 A key innovation in our architecture is the parallel branch dedicated to estimating
 1027 reliability maps. This branch processes the same encoder features as the main
 1028 segmentation path but focuses on learning the confidence of each annotator.
 1029 Through a series of 1×1 convolutions, the branch reduces channel dimensions
 1030 while maintaining spatial information. The final output consists of R reliability
 1031 maps Λ_r , one for each annotator, with values constrained to the $[0, 1]$ range
 1032 through a sigmoid activation function. This design allows the network to learn and
 1033 adapt to the varying reliability of different annotators across different regions of
 1034 the image.

1035 5.2.4 Integration with TGCE_{SS} Loss

1036 The model produces two distinct outputs: segmentation masks $\hat{\mathbf{Y}} = f(\mathbf{X}; \theta)$ and
1037 reliability maps $\{\Lambda_r(\mathbf{X}; \theta)\}_{r=1}^R$. These outputs work in tandem with the TGCE_{SS}
1038 loss function described in Section ???. The loss function simultaneously guides the
1039 learning of both the segmentation masks and reliability maps, ensuring that the
1040 model learns to balance the contributions of different annotators based on their
1041 estimated reliability.

1042 5.2.5 Training Process

1043 The training process begins with the initialization of the ResNet-34 backbone
1044 using pre-trained weights, providing a strong foundation for feature extraction.
1045 The entire network is then trained end-to-end using the Adam optimizer with a
1046 learning rate of 10^{-4} . The TGCE_{SS} loss function plays a crucial role in updating
1047 both the segmentation and reliability branches, ensuring that the model learns to
1048 effectively handle multiple annotators' inputs while accounting for their varying
 reliability.

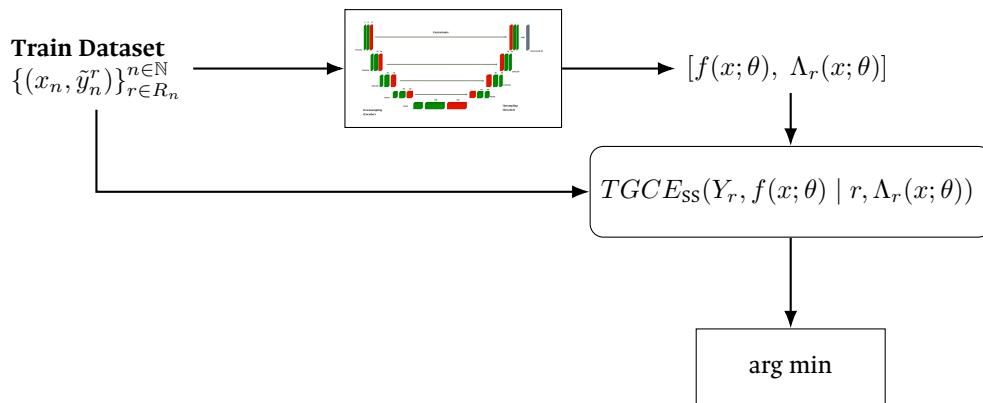


Figure 5-3 Training process of the proposed model overview. Estimated segmentation and reliability maps are computed for each input image, and the loss function is computed for the entire batch.

1049

1050 The model's architecture is specifically designed to address the challenges of
1051 multi-annotator segmentation. Through the ResNet-34 backbone, it learns robust
1052 segmentation features that capture high-level patterns in the data. The UNET's
1053 skip connections enable the preservation of fine-grained details, while the parallel
1054 reliability branch allows the model to adapt to annotator-specific characteristics.
1055 This comprehensive design enables the model to effectively handle multiple
1056 annotators' inputs while maintaining high segmentation accuracy and reliability
1057 estimation.

1058

1059

CHAPTER

1060

SIX

1061

1062

CONCLUSIONS

1063

6.1 Summary

1064 Throughout this work, several techniques have been explored to develop a model
1065 that can face ISS of real world histopathology images in crowdsourcing scenarios.
1066 Different approaches were studied and evaluated to reach an optimal solution
1067 which could model the annotators' performance from the input space. The
1068 following remarks are based on the results obtained from the experiments
1069 conducted and reviewed literature:

1070

- Chained Gaussian Processes approaches are a robust approach for developing classification and regression tasks, however, their computational powerful requirements make them unsuitable for image segmentation tasks, specially since deep learning approaches have shown to be more efficient in the task of extracting features from images.

1071

1072

1073

1074

- 1075 • Existing loss functions in the state of the art were insufficient to model the
1076 annotators' performance across the input space, from which the need for a
1077 new loss function which fulfils the requirements of the task was identified.
- 1078 • A novel loss function was proposed and evaluated, showing to be a good
1079 alternative to the existing loss functions in the state of the art, and providing
1080 an efficient way to optimize parameters for reaching capability to model the
1081 annotators' performance across the input space.
- 1082 • The use of U-Shaped Deep Learning models as a building block for the
1083 proposed solution was a valid approach, as it allows for the model to learn
1084 the spatial relationships between the pixels in the image, and the
1085 annotations, and to use this information to make predictions about the
1086 annotations of new images.
- 1087 • Chained Deep Learning approaches in combination with the proposed loss
1088 function were the best approach for the task, outperforming the state of the
1089 art in the task of **ISS** of histopathology images.

1090 6.2 Future work

1091 As tools and frameworks for bayesian approaches like Chained Gaussian Processes
1092 ¹ evolve, it is expected to see a decrease in computational requirements and hence,
1093 making mixed models more appealing for the task of **ISS** of histopathology images.
1094 In that sense, the use of **CCGPMA** as a building block for the proposed solution could
1095 be a good approach, as it is a powerful model that can take advantage of the inter-
1096 dependencies between the annotators, and the powerfulness of **CNNs** to extract
1097 features from the images.

¹Including GPFlow, and GPFlux.

BIBLIOGRAPHY

- 1099 [Abadi et al., 2016] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J.,
1100 Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga,
1101 R., Moore, S., Murray, D. G., Steiner, B., Tucker, P. A., Vasudevan, V., Warden,
1102 P., Wicke, M., Yu, Y., and Zhang, X. (2016). Tensorflow: A system for large-scale
1103 machine learning. CoRR, abs/1605.08695. (page 34)
- 1104 [Amgad et al., 2019] Amgad, M., Elfandy, H., Hussein, H., Atteya, L. A., Elsebaie,
1105 M. A. T., Abo Elnasr, L. S., Sakr, R. A., Salem, H. S. E., Ismail, A. F., Saad,
1106 A. M., Ahmed, J., Elsebaie, M. A. T., Rahman, M., Ruhban, I. A., Elgazar, N. M.,
1107 Alagha, Y., Osman, M. H., Alhusseiny, A. M., Khalaf, M. M., Younes, A.-A. F.,
1108 Abdulkarim, A., Younes, D. M., Gadallah, A. M., Elkashash, A. M., Fala, S. Y., Zaki,
1109 B. M., Beezley, J., Chittajallu, D. R., Manthey, D., Gutman, D. A., and Cooper, L.
1110 A. D. (2019). Structured crowdsourcing enables convolutional segmentation of
1111 histology images. *Bioinformatics*, 35(18):3461–3467. (pages xviii, 28, 40, and 41)
- 1112 [Avanzo et al., 2024] Avanzo, M., Stancanello, J., Pirrone, G., Drigo, A., and Retico,
1113 A. (2024). The evolution of artificial intelligence in medical imaging: From
1114 computer science to machine and deep learning. *Cancers (Basel)*, 16(21):3702.
1115 Author Joseph Stancanello is employed by Elekta SA. The remaining authors
1116 declare no commercial or financial conflicts of interest. (page 3)

- 1117 [Azad et al., 2024] Azad, R., Aghdam, E. K., Rauland, A., Jia, Y., Avval, A. H.,
1118 Bozorgpour, A., Karimijafarbigloo, S., Cohen, J. P., Adeli, E., and Merhof, D.
1119 (2024). Medical image segmentation review: The success of u-net. *IEEE
1120 Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10076–10095.
1121
(page 2)
- 1122 [Banerjee et al., 2025] Banerjee, A., Shan, H., and Feng, R. (2025). Editorial:
1123 Artificial intelligence applications for cancer diagnosis in radiology. *Frontiers in
1124 Radiology*, 5.
(page 8)
- 1125 [Bhalgat et al., 2018] Bhalgat, Y., Shah, M. P., and Awate, S. P. (2018). Annotation-
1126 cost minimization for medical image segmentation using suggestive mixed
1127 supervision fully convolutional networks. *CoRR*, abs/1812.11302.
(page 3)
- 1128 [Brito-Pacheco et al., 2025] Brito-Pacheco, D., Giannopoulos, P., and Reyes-
1129 Aldasoro, C. C. (2025). Persistent homology in medical image processing: A
1130 literature review.
(page 2)
- 1131 [Carmo et al., 2025] Carmo, D. S., Pezzulo, A. A., Villacreses, R. A., Eisenbeisz,
1132 M. L., Anderson, R. L., Van Dorin, S. E., Rittner, L., Lotufo, R. A., Gerard, S. E.,
1133 Reinhardt, J. M., and Comellas, A. P. (2025). Manual segmentation of opacities
1134 and consolidations on ct of long covid patients from multiple annotators. *Scientific
1135 Data*, 12(1):402.
(page 9)
- 1136 [Elhaminia et al., 2025] Elhaminia, B., Alsalemi, A., Nasir, E., Jahanifar, M., Awan,
1137 R., Young, L. S., Rajpoot, N. M., Minhas, F., and Raza, S. E. A. (2025). From
1138 traditional to deep learning approaches in whole slide image registration: A
1139 methodological review.
(page 2)
- 1140 [Elnakib et al., 2020] Elnakib, A., Elmenabawy, N., and S Moustafa, H. (2020).
1141 Automated deep system for joint liver and tumor segmentation using majority
1142 voting. *MEJ-Mansoura Engineering Journal*, 45(4):30–36.
(page 11)

- 1143 [Farahani et al., 2015] Farahani, N., Parwani, A. V., and Pantanowitz, L.
1144 (2015). Whole slide imaging in pathology: advantages, limitations, and
1145 emerging perspectives. *Pathology and Laboratory Medicine International*, 7:23–33.
1146 (pages xvii and 29)
- 1147 [Giri and Bhatia, 2024] Giri, K. and Bhatia, S. (2024). Artificial intelligence in
1148 nephrology- its applications from bench to bedside. *International Journal of
1149 Advances in Nephrology Research*, 7(1):90–97. (page 6)
- 1150 [Grefve et al., 2024] Grefve, J., Söderkvist, K., Gunnlaugsson, A., Sandgren, K.,
1151 Jonsson, J., Keeratijarut Lindberg, A., Nilsson, E., Axelsson, J., Bergh,
1152 A., Zackrisson, B., Moreau, M., Thellenberg Karlsson, C., Olsson, L.,
1153 Widmark, A., Riklund, K., Blomqvist, L., Berg Loegager, V., Strandberg,
1154 S. N., and Nyholm, T. (2024). Histopathology-validated gross tumor
1155 volume delineations of intraprostatic lesions using psma-positron emission
1156 tomography/multiparametric magnetic resonance imaging. *Physics and Imaging in
1157 Radiation Oncology*, 31:100633. (page 14)
- 1158 [Habis, 2024] Habis, A. A. (2024). Developing interactive artificial intelligence tools to
1159 assist pathologists with histology annotation. Theses, Institut Polytechnique de Paris.
1160 (page 8)
- 1161 [Hu et al., 2025] Hu, D., Jiang, Z., Shi, J., Xie, F., Wu, K., Tang, K., Cao, M., Huai, J.,
1162 and Zheng, Y. (2025). Pathology report generation from whole slide images with
1163 knowledge retrieval and multi-level regional feature selection. *Computer Methods
1164 and Programs in Biomedicine*, 263:108677. (pages 2, 28, and 29)
- 1165 [Julián and Álvarez Meza Andrés Marino, 2023] Julián, G. G. and Álvarez Meza
1166 Andrés Marino (2023). A supervised learning framework in the context of
1167 multiple annotators. (page 17)
- 1168 [Karthikeyan et al., 2023] Karthikeyan, R., McDonald, A., and Mehta, R. (2023).
1169 What's in a label? annotation differences in forecasting mental fatigue using ecg
1170 data and seq2seq architectures. (page 9)

- 1171 [Kim et al., 2024] Kim, Y., Lee, E., Lee, Y., and Oh, U. (2024). Understanding
1172 novice's annotation process for 3d semantic segmentation task with human-
1173 in-the-loop. In *Proceedings of the 29th International Conference on Intelligent User
1174 Interfaces, IUI '24*, page 444–454, New York, NY, USA. Association for Computing
1175 Machinery. (page 9)
- 1176 [Lam and Suen, 1997] Lam, L. and Suen, S. (1997). Application of majority
1177 voting to pattern recognition: an analysis of its behavior and performance.
1178 *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*,
1179 27(5):553–568. (page 11)
- 1180 [Lin et al., 2024] Lin, Y., Lian, A., Liao, M., and Yuan, S. (2024). Bcdnet: A fast
1181 residual neural network for invasive ductal carcinoma detection. (page 6)
- 1182 [López-Pérez et al., 2023] López-Pérez, M., Morales-Álvarez, P., Cooper, L. A. D.,
1183 Molina, R., and Katsaggelos, A. K. (2023). Crowdsourcing segmentation
1184 of histopathological images using annotations provided by medical students.
1185 In Juarez, J. M., Marcos, M., Stiglic, G., and Tucker, A., editors, *Artificial
1186 Intelligence in Medicine*, pages 245–249, Cham. Springer Nature Switzerland.
1187 (pages 5, 6, 8, and 12)
- 1188 [Lu et al., 2023] Lu, X., Ratcliffe, D., Kao, T.-T., Tikhonov, A., Litchfield, L., Rodger,
1189 C., and Wang, K. (2023). Rethinking quality assurance for crowdsourced multi-
1190 roi image segmentation. *Proceedings of the AAAI Conference on Human Computation
and Crowdsourcing*, 11(1):103–114. (pages 5 and 8)
- 1192 [López-Pérez et al., 2024] López-Pérez, M., Morales-Álvarez, P., Cooper, L. A.,
1193 Felicelli, C., Goldstein, J., Vadasz, B., Molina, R., and Katsaggelos, A. K. (2024).
1194 Learning from crowds for automated histopathological image segmentation.
1195 *Computerized Medical Imaging and Graphics*, 112:102327. (pages xvii, 5, 15, and 17)
- 1196 [Mazzarini et al., 2021] Mazzarini, M., Falchi, M., Bani, D., and Migliaccio, A. R.
1197 (2021). Evolution and new frontiers of histology in bio-medical research.
1198 *Microscopy Research and Technique*, 84(2):217–237. (pages xvii and 28)

- 1199 [Pan et al., 2021] Pan, X., Lu, Y., Lan, R., Liu, Z., Qin, Z., Wang, H., and Liu, Z. (2021).
1200 Mitosis detection techniques in h&e stained breast cancer pathological images:
1201 A comprehensive review. *Computers & Electrical Engineering*, 91:107038. (page 30)
- 1202 [Panayides et al., 2020] Panayides, A. S., Amini, A., Filipovic, N. D., Sharma, A.,
1203 Tsaftaris, S. A., Young, A., Foran, D., Do, N., Golemati, S., Kurc, T., Huang, K.,
1204 Nikita, K. S., Veasey, B. P., Zervakis, M., Saltz, J. H., and Pattichis, C. S. (2020). Ai
1205 in medical imaging informatics: Current challenges and future directions. *IEEE*
1206 *Journal of Biomedical and Health Informatics*, 24(7):1837–1857. (page 2)
- 1207 [Parkhi et al., 2012] Parkhi, O. M., Vedaldi, A., Zisserman, A., and Jawahar, C. V.
1208 (2012). Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*.
1209 (page 37)
- 1210 [Qiu et al., 2022] Qiu, Y., Hu, Y., Kong, P., Xie, H., Zhang, X., Cao, J., Wang, T.,
1211 and Lei, B. (2022). Automatic prostate gleason grading using pyramid semantic
1212 parsing network in digital histopathology. *Frontiers in Oncology*, 12. (page 14)
- 1213 [Rashmi et al., 2021] Rashmi, R., Prasad, K., and Udupa, C. B. K. (2021).
1214 Breast histopathological image analysis using image processing techniques for
1215 diagnostic purposes: A methodological review. *Journal of Medical Systems*, 46(1):7.
1216 (pages 1 and 5)
- 1217 [Ronneberger et al., 2015] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-
1218 net: Convolutional networks for biomedical image segmentation. In Navab, N.,
1219 Hornegger, J., Wells, W. M., and Frangi, A. F., editors, *Medical Image Computing*
1220 and *Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham. Springer
1221 International Publishing. (page 14)
- 1222 [Ryou et al., 2025] Ryou, H., Thomas, E., Wojciechowska, M., Harding, L., Tam,
1223 K. H., Wang, R., Hu, X., Rittscher, J., Cooper, R., and Royston, D. (2025). Reticulin-
1224 free quantitation of bone marrow fibrosis in mpns: Utility and applications.
1225 *eJHaem*, 6(2):e70005. (page 2)

- 1226 [Sarvamangala and Kulkarni, 2022] Sarvamangala, D. R. and Kulkarni, R. V. (2022).
1227 Convolutional neural networks in medical image understanding: a survey.
1228 *Evolutionary Intelligence*, 15(1):1-22. (pages 3 and 6)
- 1229 [Shah et al., 2018] Shah, M. P., Merchant, S. N., and Awate, S. P. (2018).
1230 Ms-net: Mixed-supervision fully-convolutional networks for full-resolution
1231 segmentation. In Frangi, A. F., Schnabel, J. A., Davatzikos, C., Alberola-
1232 López, C., and Fichtinger, G., editors, *Medical Image Computing and Computer*
1233 *Assisted Intervention - MICCAI 2018*, pages 379–387, Cham. Springer International
1234 Publishing. (page 5)
- 1235 [Shalf, 2020] Shalf, J. (2020). The future of computing beyond moore’s law.
1236 *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering*
1237 *Sciences*, 378(2166):20190061. (page 3)
- 1238 [TIAN and Zhu, 2015] TIAN, T. and Zhu, J. (2015). Max-margin majority voting for
1239 learning from crowds. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and
1240 Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 28.
1241 Curran Associates, Inc. (page 12)
- 1242 [Triana-Martinez et al., 2023] Triana-Martinez, J. C., Gil-González, J., Fernandez-
1243 Gallego, J. A., Álvarez Meza, A. M., and Castellanos-Dominguez, C. G. (2023).
1244 Chained deep learning using generalized cross-entropy for multiple annotators
1245 classification. *Sensors*, 23(7). (page 20)
- 1246 [Warfield et al., 2004] Warfield, S., Zou, K., and Wells, W. (2004). Simultaneous
1247 truth and performance level estimation (staple): an algorithm for the validation
1248 of image segmentation. *IEEE Transactions on Medical Imaging*, 23(7):903–921.
1249 (page 12)
- 1250 [Weitz et al., 2023] Weitz, P., Valkonen, M., Solorzano, L., Carr, C., Kartasalo, K.,
1251 Boissin, C., Koivukoski, S., Kuusela, A., Rasic, D., Feng, Y., Sinius Pouplier,
1252 S., Sharma, A., Ledesma Eriksson, K., Latonen, L., Laenholm, A.-V., Hartman,
1253 J., Ruusuvuori, P., and Rantalainen, M. (2023). A multi-stain breast cancer
1254 histological whole-slide-image data set from routine diagnostics. *Scientific Data*,
1255 10(1):562. (pages xviii, 31, 39, and 40)

- 1256 [Xu et al., 2024] Xu, Y., Quan, R., Xu, W., Huang, Y., Chen, X., and Liu, F. (2024).
1257 Advances in medical image segmentation: A comprehensive review of traditional,
1258 deep learning and hybrid approaches. *Bioengineering*, 11(10). (pages 3 and 8)

1259 [Yu et al., 2025] Yu, J., Li, B., Pan, X., Shi, Z., Wang, H., Lan, R., and Luo, X. (2025).
1260 Semi-supervised gland segmentation via feature-enhanced contrastive learning
1261 and dual-consistency strategy. *IEEE Journal of Biomedical and Health Informatics*,
1262 pages 1–11. (page 2)

1263 [Zhang et al., 2020] Zhang, L., Tanno, R., Xu, M.-C., Jin, C., Jacob, J., Cicarrelli, O.,
1264 Barkhof, F., and Alexander, D. (2020). Disentangling human error from ground
1265 truth in segmentation of medical images. In Larochelle, H., Ranzato, M., Hadsell,
1266 R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*,
1267 volume 33, pages 15750–15762. Curran Associates, Inc. (page 16)

1268 [Zhang and Sabuncu, 2018] Zhang, Z. and Sabuncu, M. R. (2018). Generalized
1269 cross entropy loss for training deep neural networks with noisy labels.
1270 (pages 46 and 47)

1271 [Zhou et al., 2021] Zhou, S. K., Greenspan, H., Davatzikos, C., Duncan, J. S.,
1272 Van Ginneken, B., Madabhushi, A., Prince, J. L., Rueckert, D., and Summers, R. M.
1273 (2021). A review of deep learning in medical imaging: Imaging traits, technology
1274 trends, case studies with progress highlights, and future promises. *Proceedings of
1275 the IEEE*, 109(5):820–838. (pages 1 and 2)

1276 [Zhou et al., 2024] Zhou, Z., Gong, H., Hsieh, S., McCollough, C. H., and Yu, L.
1277 (2024). Image quality evaluation in deep-learning-based ct noise reduction using
1278 virtual imaging trial methods: Contrast-dependent spatial resolution. *Medical
1279 Physics*, 51(8):5399–5413. (page 9)