UNIVERSIDAD
NACIONAL
DE COLOMBIA

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

Brandon Lotero Londoño

Universidad Nacional de Colombia
Faculty of Engineering and Architecture
Department of Electric, Electronic and Computing Engineering
Manizales, Colombia
2025

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

Brandon Lotero Londoño

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

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

Co-advisor:
Prof. Germán Castellanos-Domínguez, Ph.D.

Academic research group:
Signal Processing and Recognition Group - SPRG

Universidad Nacional de Colombia
Faculty of Engineering and Architecture
Department of Electric, Electronic and Computing Engineering
Manizales, Colombia
2025

Segmentación de imágenes médicas en un contexto de múltiples anotadores: Aplicación al estudio de histopatologías

Brandon Lotero Londoño

Disertación presentada como requisito parcial para recibir el título de:

Magíster en Ingeniería - Automatización Industrial

Director:

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

Codirector:

Prof. Germán Castellanos-Domínguez, Ph.D.

Grupo de investigación:

Grupo de Control y Procesamiento Digital de Señales - GCPDS

Universidad Nacional de Colombia

Facultad de Ingeniería y Arquitecura

Departamento de Ingeniería Eléctrica, Electrónica y Computación

Manizales, Colombia

2025

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to:

My beloved Camila, for her unwavering love, support, and patience throughout this journey.

My family, especially my parents, for their unconditional support during challenging times. I look forward to sharing many more joyful moments together.

My advisors, Prof. Andrés Marino Álvarez-Meza, Prof. Julián Gil, and Prof. Germán Castellanos-Domínguez, for their invaluable guidance and mentorship.

My friends and colleagues at GCPDS, particularly Lucas, Santiago, Marcos, and Rafael, for their support and camaraderie throughout this academic journey.

The Universidad Nacional de Colombia, for providing the resources and facilities that enabled me to pursue this Master's degree. I hope this work contributes meaningfully to the field of early cancer detection.

Brandon Lotero Londoño

2025

ABSTRACT

Keywords: Crowdsourcing, Multiple Annotators, Histopathology, Breast Cancer, Semantic Image Segmentation, Gaussian Processes, Deep Learning, Medical Image Analysis

RESUMEN

Palabras clave: Crowdsourcing, Múltiples Anotadores, Histopatología, Cáncer de Mama, Segmentación Semántica de Imágenes, Procesos Gaussianos, Aprendizaje Profundo, Análisis de Imágenes Médicas

Contents

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

CAD Computer-Aided Diagnosis 2, 5, 6
CCGP Correlated Chained Gaussian Processes 18
CCGPMA Correlated Chained Gaussian Processes for Multiple Annotators 17, 19, 66
CE Cross Entropy 54, 57
CGP Chained Gaussian Processes 18
CNN Convolutional Neural Networks 3, 14, 21, 24, 59, 66
CT Computed Tomography 1, 2, 11

ELBO Evidence Lower Bound 18

GCE Generalized Cross Entropy 54, 57
GCECDL Generalized Cross-Entropy-based Chained Deep Learning 20

ISS Image Semantic Segmentation 2, 3, 6, 11, 13, 21, 23, 24, 35, 65, 66

LF Latent Function 18

MAE Mean Absolute Error 54, 57
MITs Medical Imaging Techniques 1
ML Machine Learning 11, 21
MRI Magnetic Resonance Imaging 1, 2
MV Majority Voting 11, 12

OCR Optical Character Recognition 11

PET Positron Emission Tomography 14

ROI Region of Interest 2, 6, 15, 31

SLFM Semi-Parametric Latent Factor Model 18
SS Semantic segmentation 3
STAPLE Simultaneous Truth and Performance Level Estimation 12–14, 36

WSI Whole Slide Imaging xviii, 1, 5, 6, 8, 15, 29, 41

CHAPTER

ONE

INTRODUCTION

1.1 Motivation

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

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

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

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

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

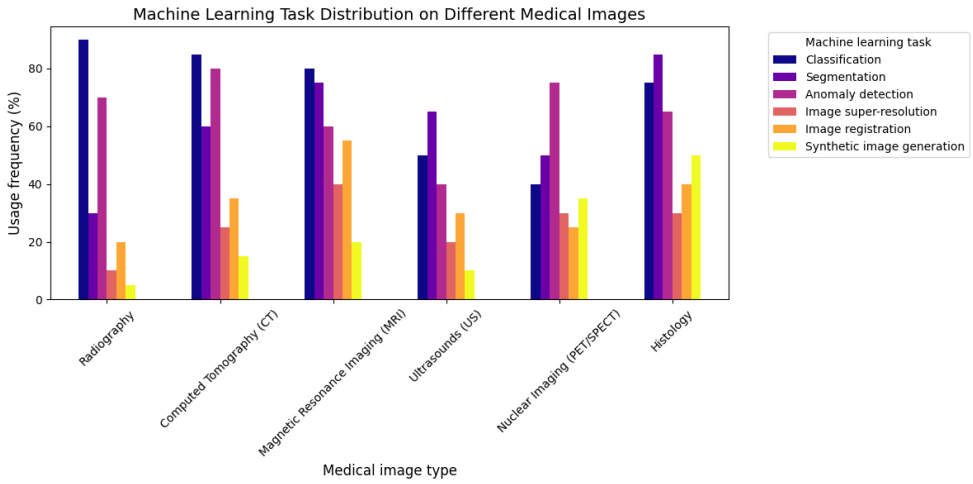

Figure **1-1** Estimation of the popularity of tasks and medical image types based on recent literature review (count of referenced terms).

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

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

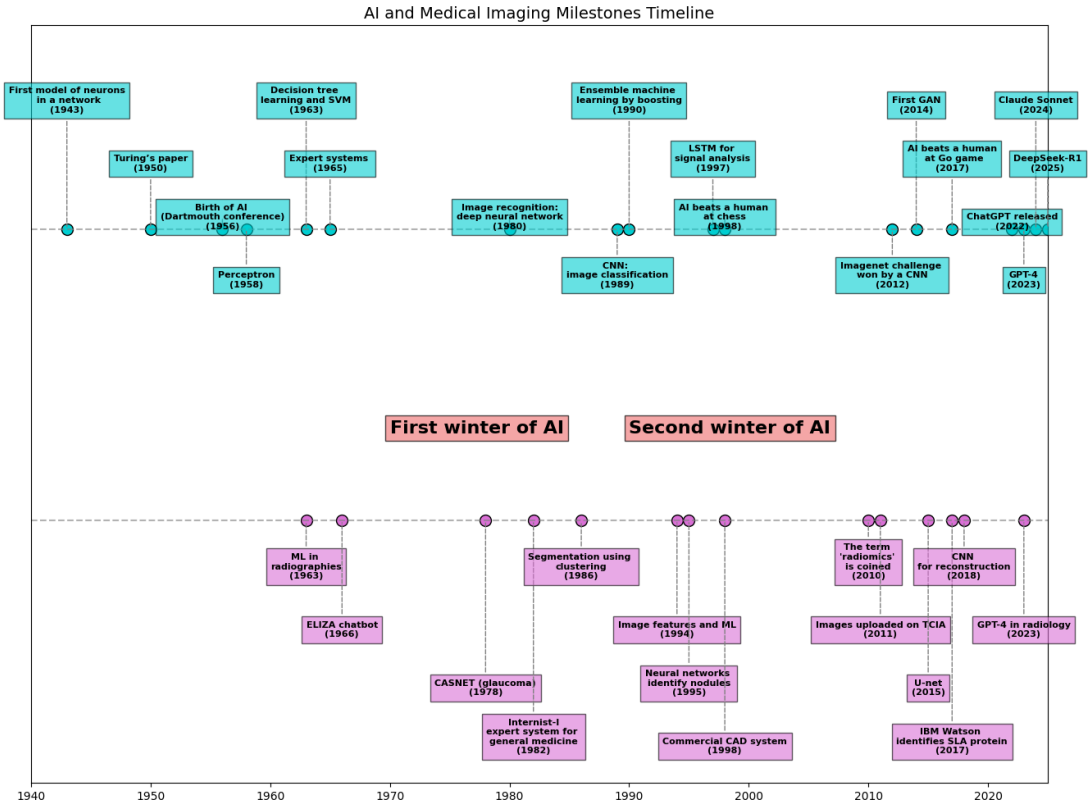

Figure **1-2** AI and machine learning in medical imaging brief timeline.

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

Many medical images datasets however, contain a high variability in class sizes and variations in colors, which is specially noticeable in histopathological images because of the usage of different staining and other factors which can affect the color of the images (see section 2.2.3). This variability can lead to a significant loss of efficiency of machine learning models when using a mixed supervision strategy, as the model can be biased towards the most common classes or colors in the dataset [Shah et al., 2018].

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

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

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

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

1.2 Problem Statement

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

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

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

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

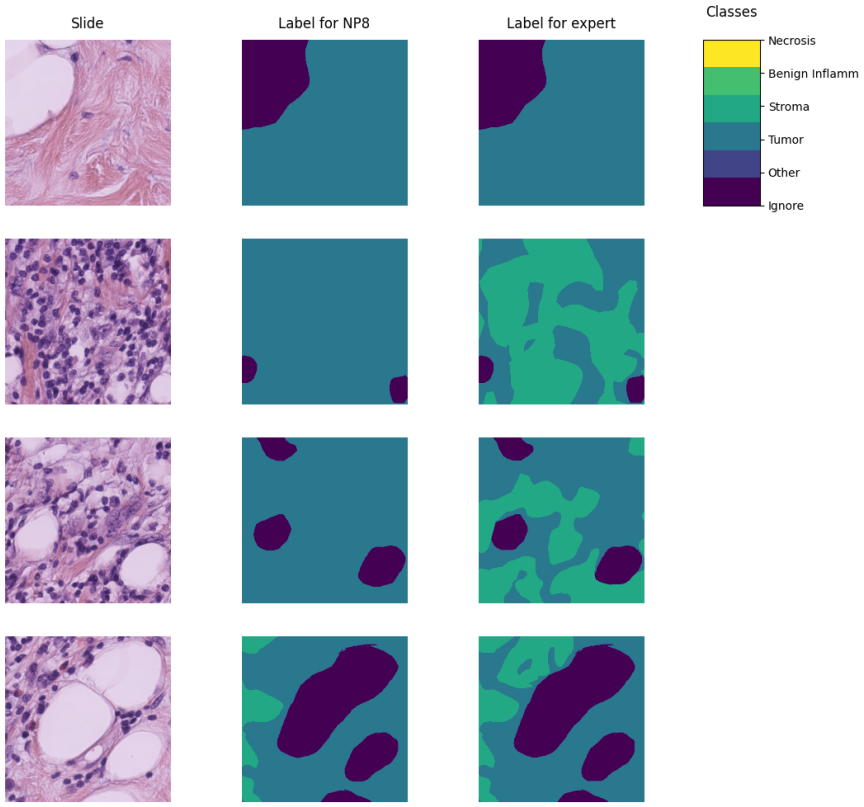

Figure **1-3** Example of a histopathological image segmented by multiple annotators, illustrating variations in label assignment.

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

1.2.1 Variability in Expertise Levels

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

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

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

1.2.2 Technical Constraints and Image Quality

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

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

1.2.3 Research Question

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

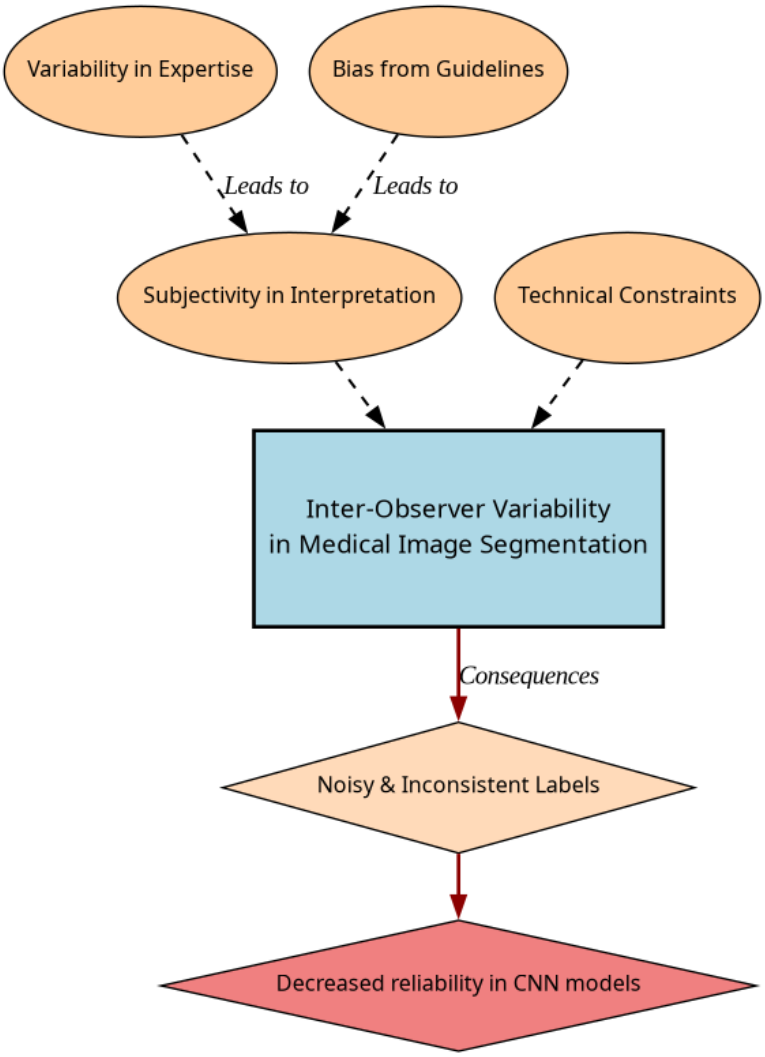

Figure **1-4** Summary diagram for problem Statement

Research Question

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

1.3 Literature review

Certainly, in general Machine Learning (ML) classification tasks [3] where multiple annotators are involved, Majority Voting (MV) is by far the simplest possible approach to implement. This concept was born multiple times and divergently in multiple fields, but it was described as relevant for ML and pattern recognition labeling for classification in [Lam and Suen, 1997], in which the approach is exposed as simple, yet powerful. The authors describe the MV as a method that can be used to improve the accuracy of classification tasks by combining the labels of multiple annotators. The method is based on the assumption that the majority vote of the annotators is more likely to be correct than the vote of a single annotator. The authors also describe the method as a straightforward way to improve the accuracy of classification tasks without the need for complex algorithms or additional data. The authors also prove this method to deliver very similar results to more complicated approaches (Bayesian, logistic regression, fuzzy integral, and neural network) in the particular task of Optical Character Recognition (OCR). Despite its simplicity, modern solutions for delivering accurate medical image segmentation models still rely on Majority Voting at some stage, like [Elnakib et al., 2020], which uses a majority voting strategy for delivering a final output based on the labels of multiple models (VGG16-Segnet, Resnet-18 and Alexnet) in CT images for Liver Tumor Segmentation, or [López-Pérez et al., 2023],

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

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

1.3.1 Facing annotation variability in medical images

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

STAPLE Mechanism

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

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

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

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

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

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

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

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

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

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

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

Finally, this work contemplates STAPLE as useful for 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

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

U-Net based approaches

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

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

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

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

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

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

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

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

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

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

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

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

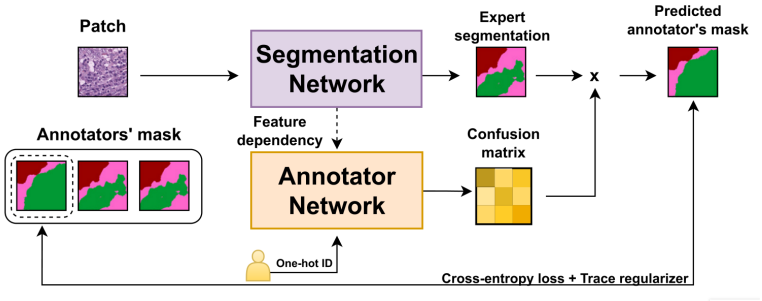

Figure **1-5** Proposed framework for the approach in [López-Pérez et al., 2024].

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

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

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

Bayesian models

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

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

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

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

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

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

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

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

preserving the interdependencies among the annotators, and also outperforming GPC-MV[4], MA-LFC-C[5], MA-DGRL[6], MA-GPC[7], MA-GPCV[8], MA-DL[9], KAAR[10].

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

1.3.2 Facing noisy annotations and low-quality data

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

[4] A GPC using the MV of the labels as the ground truth.
[5] A LRC with constant parameters across the input space.
[6] A multi-labeler approach that considers as latent variables the annotator performance.
[7] A multi-labeler GPC, which is an extension of MA-LFC.
[8] An extension of MA-GPC that includes variational inference and priors over the labelers' parameters.
[9] A Crowd Layer for DL, where the annotators' parameters are constant across the input space.
[10] A kernel-based approach that employs a convex combination of classifiers and codes labelers dependencies.

Loss functions in deep learning models

Loss functions are fundamental components in deep learning models that quantify how well a model's predictions match the ground truth. They serve as the objective function that guides the learning process by measuring the discrepancy between predicted and actual values. In classification tasks, the most common loss functions are Cross-Entropy (CE) and Mean Absolute Error (MAE) [Zhang and Sabuncu, 2018]. CE is particularly effective for classification as it heavily penalizes confident but wrong predictions, though it can be sensitive to noisy labels. MAE, on the other hand, is more robust to outliers and assigns equal weights to all mistakes, but typically requires more training iterations. For image segmentation tasks, specialized loss functions have been developed to handle the unique challenges of pixel-wise classification. The Dice loss, which measures the overlap between predicted and ground truth regions, is widely used in medical image segmentation [Zhao et al., 2020]. More recently, the Generalized Cross Entropy (GCE) loss has emerged as a robust alternative that combines the benefits of both CE and MAE, allowing for better handling of noisy labels through a tunable parameter that controls sensitivity to outliers. However, to the best of our knowledge, none of the current loss functions approaches have focused on the multiple labelers scenario, neither on the per labeler spatial reliability estimation.

Generalized Cross-Entropy for multiple annotators classification

A more recent approach was proposed by [Triana-Martinez et al., 2023], introducing a Generalized Cross-Entropy-based Chained Deep Learning (GCECDL) framework. This method addresses the limitations of traditional label aggregation techniques by modeling each annotator's reliability as a function of the input data. The approach effectively mitigates the impact of noisy labels by using a noise-robust loss function, balancing Mean Absolute Error (MAE) and Categorical Cross-Entropy (CE). Unlike prior approaches, GCECDL accounts for the dependencies among annotators while encoding their non-stationary behavior

across different data samples. Their experiments on multiple datasets demonstrated superior predictive performance compared to state-of-the-art methods, particularly in cases where annotations were highly inconsistent.

The strategy of the authors effectively unlocks the potential of ML models to handle low-quality data and noisy annotations, but it is bounded to classifications tasks only, not being by itself applicable to segmentation tasks. The TGCE equation for handling multiple annotators is defined as:

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

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

1.4 Aims

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

Table **1-1** Summary of state-of-the-art approaches for handling multiple annotators in medical image segmentation

Approach	Key Characteristics	Advantages	Limitations
Majority Voting (MV)	Simple aggregation of multiple annotators' labels	- Simple to implement - No complex algorithms required - Proven effective in many cases	- Assumes equal expertise - Sensitive to correlated errors - May introduce bias in expert tasks
STAPLE	Probabilistic framework estimating true segmentation and rater reliability	- Handles binary and multiclass segmentation - Estimates rater sensitivity/specificity - Robust to correlated errors	- Computationally expensive - Sensitive to low-quality annotations - Tends to over-smooth results - Limited interpretability
U-Net based approaches	Deep learning architecture with encoder-decoder structure	- Captures both global and local information - Effective for complex structures - Works well with limited data	- Requires large amounts of training data - Complex optimization process - May need specialized loss functions
Bayesian Models (CCGPMA)	Gaussian Processes modeling annotator reliability	- Preserves annotator interdependencies - Handles label noise effectively - Works for classification and regression	- Computationally expensive - Not directly applicable to segmentation - Complex optimization process
Generalized Cross-Entropy (GCE)	Loss function balancing MAE and CE	- Robust to noisy labels - Accounts for non-stationary behavior - Handles annotator dependencies	- Limited to classification tasks - Requires parameter tuning - May need adaptation for segmentation

1.4.1 General Aim

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

1.4.2 Specific Aims

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

1.5 Outline and Contributions

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

- A novel loss function for ISS tasks in medical images capable of inferring the best possible segmentation and codifying the reliability of the labelers without needing separate inputs about the labeler performance or multiple networks to be trained.
- A Chained Deep Learning model for ISS tasks in medical images, which can learn from inconsistent labels and improve the segmentation performance outperforming state of the art approaches on a Triple Negative Breast Cancer histopathology dataset.
- A mechanism for noisy annotations emulations by using networks encoding layers weight disturbance for building crowdsourced version of the well known Oxford-IIIT Pet Dataset.
- A python package for using the proposed loss function and architecture in CNN models for ISS tasks in medical images. [11]
- Triple negative breast cancer and Oxford-IIIT Pet Datasets mapping as lazy loaders for the proposed loss function ready to go in the package. [12]
- A public Github repository with the code used in this work. [13]

[11] https://pypi.org/project/seg_tgce/
[12] https://seg-tgce.readthedocs.io/en/latest/experiments.html
[13] https://github.com/blotero/seg_tgce

CHAPTER

TWO

CONCEPTUAL PRELIMINARIES

2.1 Modern concept of digital image

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

2.1.1 Types of digital images

Grayscale images

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

Color images

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

Multispectral images

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

3D images and volumetric data

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

2.1.2 Mathematical representations

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

- **Sampling**: The process of converting a continuous image into a discrete representation. According to the Nyquist-Shannon sampling theorem, the sampling frequency must be at least twice the highest frequency present in the image to avoid aliasing.

- **Quantization**: The process of converting continuous intensity values into discrete levels. The number of quantization levels determines the image's bit depth and affects its quality and storage requirements.

- **Resolution**: The number of pixels per unit length in an image, typically measured in pixels per inch (PPI) or dots per inch (DPI).

- **Dynamic range**: The ratio between the maximum and minimum measurable light intensities in an image, often expressed in decibels (dB).

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

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

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

For color images, the representation extends to:

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

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

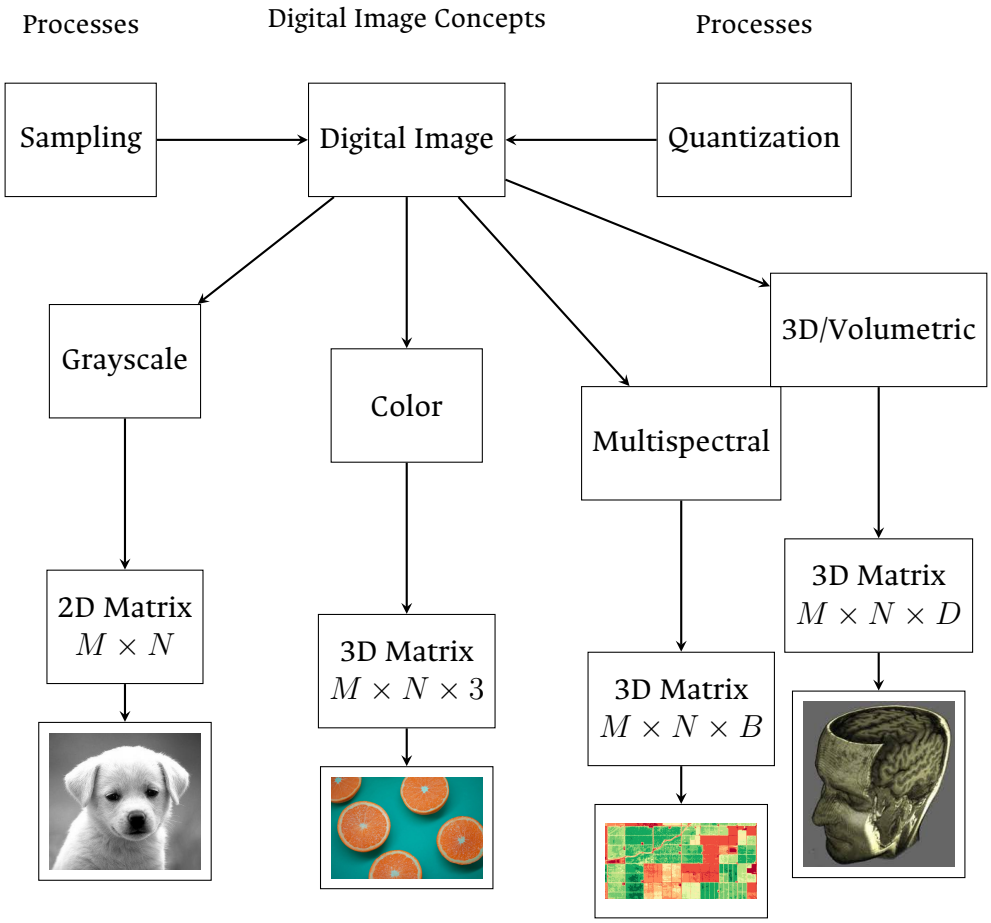

Figure **2-1** Overview of digital image concepts and their mathematical representations. The figure shows the main types of digital images (grayscale, color, multispectral, and volumetric), their mathematical representations, and the fundamental processes of sampling and quantization. Example images are included to illustrate each type.

2.2 Digital histopathological images

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

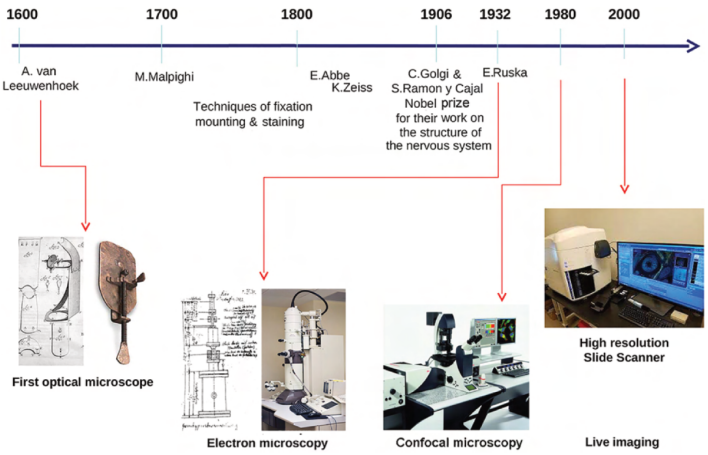

Figure **2-2** Histology evolution timeline. (Image from [Mazzarini et al., 2021]).

2.2.1 Whole Slide Imaging (WSI)

Whole Slide Imaging (WSI) is the process of digitizing entire glass slides at high resolution, creating a digital representation that can be viewed, analyzed, and shared electronically. Modern WSI scanners use sophisticated optical systems that capture multiple fields of view at high magnification, which are then stitched together to create a seamless digital image [Hu et al., 2025]. These systems incorporate high-resolution objectives with magnifications ranging from 20x to 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 2-3 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].

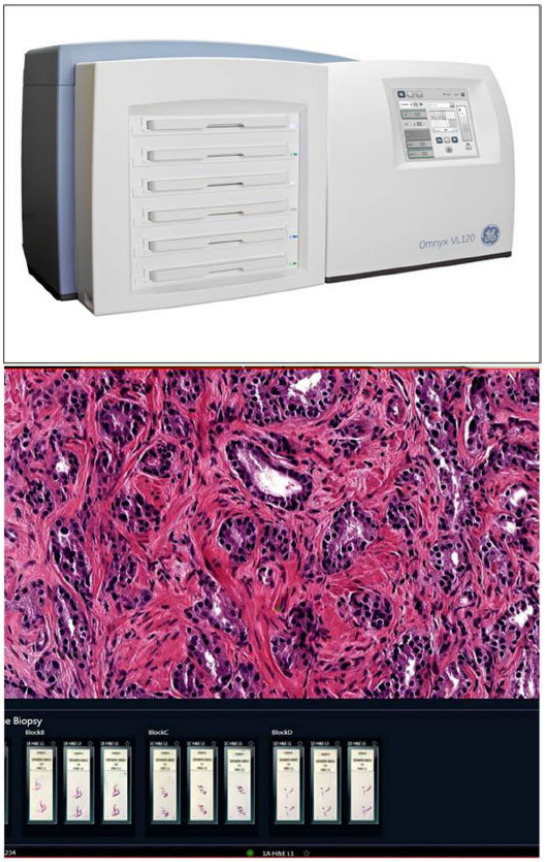

Figure 2-3 (Above) Whole slide imaging system by Omnyx for slide digitization. (Below) Comprehensive digital pathology interface from Omnyx designed to streamline pathologists' diagnostic workflow. (From [Farahani et al., 2015]).

2.2.2 Regions of Interest (ROI)

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

2.2.3 Staining Techniques

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

Hematoxylin and Eosin (H&E)

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

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

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

Special Stains

In addition to H&E, various special stains are used for specific diagnostic purposes. Immunohistochemistry (IHC) uses antibodies to detect specific proteins, playing a crucial role in subtyping breast cancers. Key IHC stains include Estrogen Receptor (ER) staining for detecting estrogen receptors, Progesterone Receptor (PGR) staining for assessing progesterone receptor status, Human Epidermal Growth Factor Receptor 2 (HER2) staining for evaluating HER2 protein expression, and Ki67 staining for measuring cellular proliferation rates. These markers are particularly crucial in breast cancer diagnosis and treatment planning, as they help determine the molecular subtype of the cancer and guide personalized therapeutic approaches. Other specialized stains include Periodic Acid-Schiff (PAS) for highlighting carbohydrates and basement membranes, Masson's Trichrome for distinguishing between collagen and muscle fibers, and silver stains for detecting microorganisms and nerve fibers. These specialized staining techniques complement H&E by providing additional diagnostic information that is crucial for accurate diagnosis and treatment planning [Weitz et al., 2023]. Examples of these staining techniques are shown in Figure 2-9.

2.3 Deep learning fundamentals

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

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

2.3.1 Learning Paradigms

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

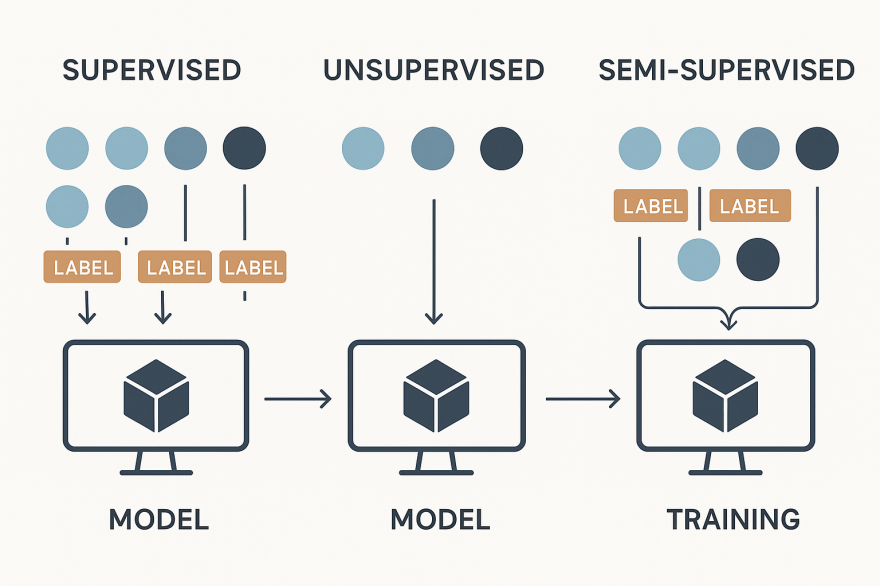

Figure **2-4** Common learning paradigms.

2.3.2 Architecture and Training

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

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

2.3.3 Challenges and Solutions

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

2.3.4 Deep Learning Frameworks

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

These frameworks have democratized deep learning by providing efficient implementations of common operations, automatic differentiation for gradient computation, and GPU acceleration for faster training. They also offer pre-trained models and transfer learning capabilities, along with active communities for support and knowledge sharing. The combination of these frameworks with modern hardware has enabled researchers and practitioners to develop increasingly sophisticated models, pushing the boundaries of what's possible in artificial intelligence. As shown in Figure **2-5**, which presents data from Google Trends over the last five years (as of April 2025), TensorFlow and PyTorch have emerged as the two most prominent frameworks in the deep learning landscape.

2.4 Datasets and data sources

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

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

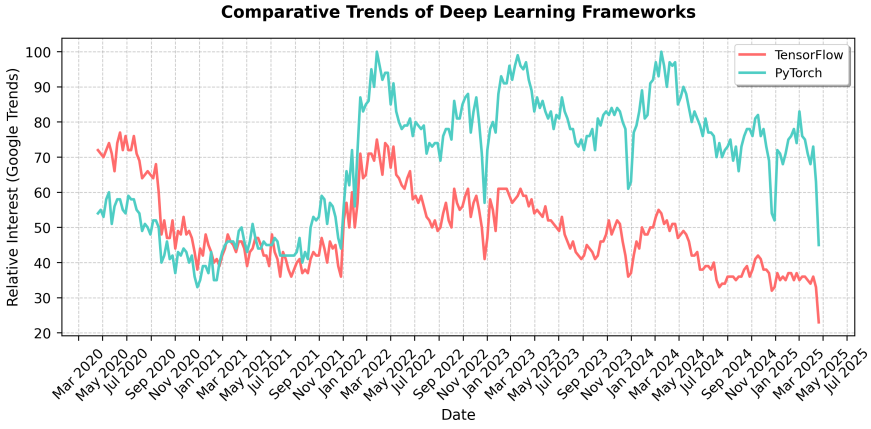

Figure **2-5** Comparative Trends of the top two most popular Deep Learning Frameworks, apparently, tendency was switched to PyTorch since 2022

- Distribution of segmentation labels is not uniform across the image, since some tissues and structures are more common than others.

- Visualization of performance of the models in debug time (like per epoch analysis) is not simple for non experts in the subject, which makes it hard to evaluate whether the model is overfitting or not at a glance.

For these reasons, multiple datasets were created in the pursuit of an initial evaluation of performance of the models against more traditional and familiar images before the focus on histopathology images. Once a decent performance in metrics like Dice coefficient was achieved, the focus was shifted to histopathology images and further tunings on the models were performed if needed.

In any case, both the emulated noisy annotations datasets and the histopathology datasets somehow contained ground truth aggregation, either from the original source (in the case of emulated noisy annotations), the expert annotation (if available) or from the aggregation of multiple labelers [1].

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

2.4.1 Datasets with emulated noisy annotations

A challenge arises for the creation of emulated noisy annotations datasets, since it is expected for images annotations to have some degree of expertise variability, similar to what is expected in real crowdsourced datasets. Simply introducing random noise into the annotated masks does not work, since the original morphological structures from the expected ground truth are far from being preserved. Instead, a "noisy" labeler is expected to produce an annotation which has at least some degree of morphological consistency on itself, even if it shows discrepancies when compared with some metric (like DICE score) against the ground truth annotation.

This has been proven experimentally when introducing random noise into any popular segmentation dataset, in which the resulting mask is just a non coherent map of noise across the image, as shown in Figure 2-6.

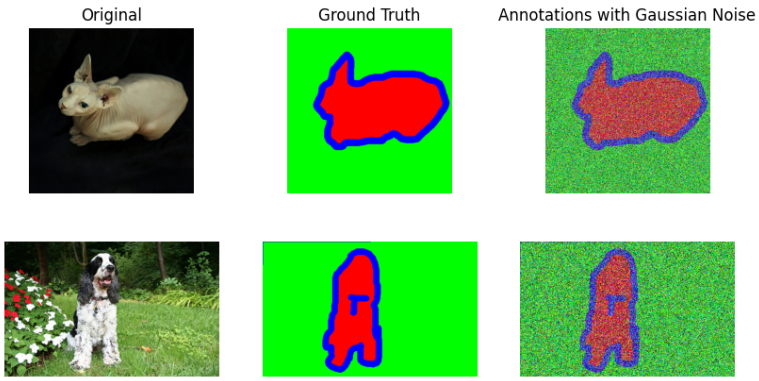

Figure **2-6** Example of a noisy mask generated by naively introducing random noise into a ground truth mask. Morphological consistency is lost.

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

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

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

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

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

Oxford-IIIT Pet Dataset

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

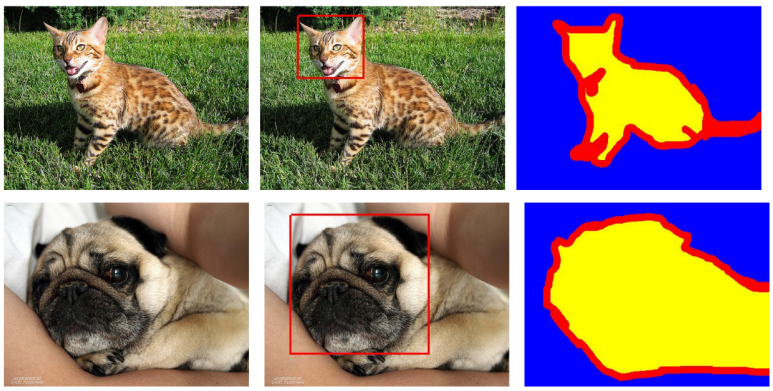

Figure **2-7** Annotations in the Oxford-IIIT Pet data. From left to right: pet image, head bounding box, and trimap segmentation (blue: background region; red: ambiguous region; yellow: foreground region).

annotate the images with a high degree of accuracy, which facilitates model initial debugging. Figure 2-7 shows an example of the annotations in the original dataset.

With the application of the encoder layer weight perturbation technique, the resulting noisy masks are shown in Figure 2-8. It can be seen that the morphological consistency is preserved, even though the resulting masks are far from the ground truth annotations, which goes perfectly well for testing the robustness of the models against noisy annotations in crowdsourcing-like scenarios.

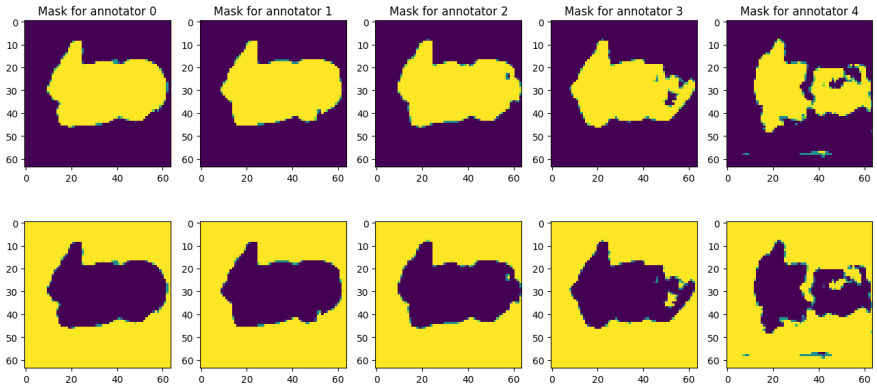

Figure **2-8** Noisy mask generated by enhancing the disturbances in the encoder layers weights for the Oxford-IIIT Pet Dataset. Morphological consistency is preserved. From left to right, SNR levels of noise in the encoder layer are 10, 5, 2, 0, -5 dB.

2.4.2 Real histopathology datasets

Multi-Stain Breast Cancer Histological Dataset

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

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

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

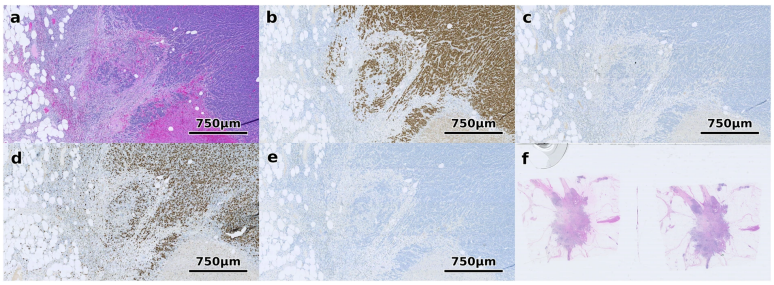

Figure **2-9** Different staining techniques obtained from multi-stain breast cancer dataset [Weitz et al., 2023]. (a) shows H&E, (b) ER, (c) HER2, (d) Ki67 and (e) PGR. (f) shows an example of a WSI that was excluded since it contains multiple tissue sections.

Structured Crowdsourcing Dataset for Histology Images

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

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

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

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

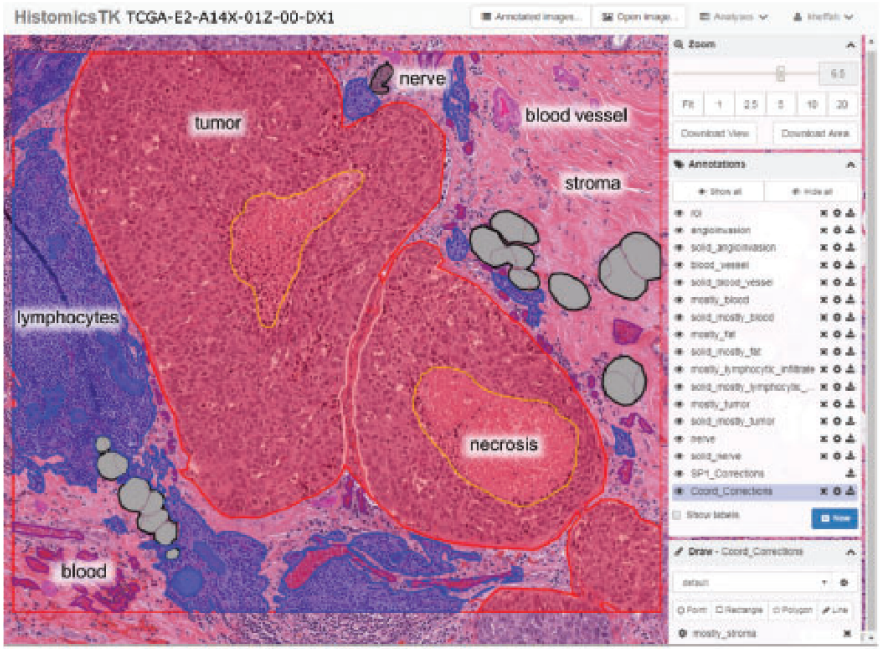

Figure **2-10** Screenshot of the DSA and HistomicsTK web interface while creating the crowdsourced annotations for the dataset presented by [Amgad et al., 2019].

CHAPTER

THREE

CHAINED GAUSSIAN PROCESSES

3.1 Background and Related Methods

In recent years, several approaches have been proposed to handle supervised learning problems in the context of multiple annotators. This section reviews key methods that form the foundation for understanding chained Gaussian processes and their application to multiple annotator scenarios in regression and classification tasks.

3.1.1 Kernel Alignment-Based Annotator Relevance Analysis (KAAR)

The Kernel Alignment-Based Annotator Relevance Analysis (KAAR) method, introduced in [Julián and Álvarez Meza Andrés Marino, 2023], addresses the challenge of estimating annotator expertise in scenarios where the ground truth is

unavailable. The key innovation of KAAR lies in its use of Centered Kernel Alignment (CKA) to measure the similarity between input features and annotator labels.

Given a dataset with input features $\boldsymbol{X}$ and labels from multiple annotators $\boldsymbol{Y}$, KAAR computes a kernel matrix $\boldsymbol{K_\nu}$ as a convex combination of R basis kernels:

$$\boldsymbol{K_\nu} = \sum_{r=1}^{R} \nu_r \boldsymbol{K}_r, \tag{3-1}$$

where $\boldsymbol{K}_r$ is computed using a kernel function over the input features, and ν_r represents the relevance weight for the r-th annotator. The weights are optimized by maximizing the CKA between $\boldsymbol{K_\nu}$ and a target kernel $\boldsymbol{F}$ computed from the input features:

$$\rho(\boldsymbol{K_\nu}, \boldsymbol{F}) = \frac{\langle \bar{\boldsymbol{K}}_{\boldsymbol{\nu}}, \bar{\boldsymbol{F}} \rangle_{\mathsf{F}}}{|| \bar{\boldsymbol{K}}_{\boldsymbol{\nu}} ||_{\mathsf{F}} || \bar{\boldsymbol{F}} ||_{\mathsf{F}}}, \tag{3-2}$$

A key advantage of KAAR is its ability to capture dependencies between annotators through the kernel framework, without requiring explicit parametric modeling of these relationships.

3.1.2 Localized Kernel Alignment-Based Annotator Relevance Analysis (LKAAR)

LKAAR extends KAAR by introducing locality in the kernel alignment framework. Rather than assuming constant annotator performance across the input space,

LKAAR models the annotator relevance as a function of the input features. The kernel combination in LKAAR takes the form:

$$\boldsymbol{K_q} = \sum_{r=1}^{R} \boldsymbol{Q}_r \boldsymbol{K}_r \boldsymbol{Q}_r, \tag{3-3}$$

where $\boldsymbol{Q}_r$ is a diagonal matrix whose elements represent the local relevance of annotator r. These elements are computed as:

$$q_r(\boldsymbol{x}_n) = \begin{cases} \beta_0^{(r)} + \sum_{n'=1}^{N} \beta_{n'}^{(r)} \kappa_\beta(\boldsymbol{x}_n, \boldsymbol{x}_{n'}), & \text{if } n \in N_r \\ 0, & \text{Otherwise} \end{cases} \tag{3-4}$$

This localized approach allows LKAAR to better model inconsistent annotators whose performance varies across different regions of the input space.

3.1.3 Regularized Chained Deep Neural Network (RCDNN)

RCDNN represents a deep learning approach to multiple annotator learning, inspired by the chained Gaussian processes model. The method models both the ground truth estimation and annotator reliability through a deep neural network architecture. The likelihood function in RCDNN takes the form:

$$p(\boldsymbol{Y}|\boldsymbol{\theta}) = \prod_{n=1}^{N} \prod_{r \in R_n} \left(\prod_{k=1}^{K} \zeta_{n,k}^{\delta(y_n^{(r)} - k)} \right)^{\lambda_n^{(r)}} \left(\frac{1}{K} \right)^{1-\lambda_n^{(r)}}, \tag{3-5}$$

where $\lambda_n^{(r)}$ represents the reliability of annotator r for instance n, and $\zeta_{n,k}$ is the estimated probability of class k for instance n.

The RCDNN architecture is built upon a sophisticated multi-layer structure that combines several key innovations in deep learning. At its core, it employs multiple dense layers with carefully chosen activation functions, each designed to capture different aspects of the annotator-data relationship. To prevent overfitting, a comprehensive regularization strategy is implemented, combining both l1 and l2 norms with dropout mechanisms. The network architecture splits into specialized branches: one dedicated to estimating the ground truth labels, and another focused on modeling annotator reliability. This dual-branch approach allows the network to simultaneously learn both the true underlying patterns in the data and the individual characteristics of each annotator. Furthermore, the implementation of Monte Carlo dropout during prediction provides a mechanism for uncertainty estimation, offering not just predictions but also confidence measures for those predictions.

3.1.4 Chained Gaussian Processes Model

The Chained Gaussian Processes (CGP) model provides a probabilistic framework for modeling multiple outputs through a chain of Gaussian processes. In the context of multiple annotators, CGP can be used to model both the ground truth and annotator behavior. The key idea is to model a likelihood function with J parameters as in [**?**]:

$$p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}) = \prod_{n=1}^{N} p(y_n|\theta_1(\boldsymbol{x}_n), \ldots, \theta_J(\boldsymbol{x}_n)), \tag{3-6}$$

where each parameter $\theta_j(\boldsymbol{x})$ is modeled using a separate Gaussian process. This framework provides remarkable flexibility in modeling complex relationships

within the data. The use of Gaussian processes enables natural uncertainty quantification through posterior distributions, allowing the model to express varying degrees of confidence in its predictions. The framework inherently handles missing data through its probabilistic nature, and the automatic relevance determination property of GPs helps identify the most important features for prediction.

The CGP framework becomes particularly powerful when integrated with concepts from KAAR and LKAAR. By incorporating kernel alignment principles, the model can make more informed choices about kernel functions, better capturing the underlying structure of the data. The framework can be extended to include input-dependent lengthscales, allowing for varying degrees of smoothness across the input space. Furthermore, through multi-output GP formulations, the model can capture complex dependencies between different annotators, providing a rich representation of how different experts' opinions relate to each other.

However, implementing CGP in practice presents several significant challenges. The computational complexity of the model scales cubically with the dataset size, making it potentially prohibitive for large-scale applications. The need for approximate inference techniques in non-Gaussian likelihood scenarios can introduce additional complexity and potential approximation errors. When dealing with high-dimensional inputs, the model may struggle with the curse of dimensionality, requiring careful feature selection or dimensionality reduction preprocessing. Additionally, the selection of appropriate kernel functions remains a critical challenge, as it significantly impacts the model's performance and its ability to capture relevant patterns in the data.

3.2 Tooling and Implementation

The implementation of Chained Gaussian Processes for multiple annotator learning requires robust and flexible software tools that can handle both the

probabilistic nature of Gaussian processes and the complexity of deep architectures. In our implementation, we leverage two key frameworks: GPflow and GPflux. These tools provide the foundation for building scalable and efficient implementations of our models while maintaining the theoretical rigor necessary for proper uncertainty quantification.

The choice of these frameworks is motivated by several factors. First, both are built on top of TensorFlow, providing access to automatic differentiation and GPU acceleration capabilities. Second, they offer a modular design that allows for easy extension and modification, which is crucial when implementing novel model architectures. Third, they provide robust implementations of various Gaussian process models and inference methods, allowing us to focus on the specific challenges of multiple annotator learning rather than reimplementing basic GP functionality.

3.2.1 GPflow Overview

GPflow is a Python package for building Gaussian process models in TensorFlow. Developed as a modern and scalable alternative to GPy, it leverages TensorFlow's computational capabilities to provide efficient implementation of Gaussian process models. The framework is designed with four key principles in mind: ease of use, computational efficiency, extensibility, and automatic differentiation. This work was originally published in [de G. Matthews et al., 2016].

At its core, GPflow provides implementations of various Gaussian process models including GPR (Gaussian Process Regression), GPMC (Gaussian Process Monte Carlo), SGPR (Sparse Gaussian Process Regression), and SGPMC (Sparse Gaussian Process Monte Carlo). These implementations are built on top of TensorFlow's computational graph framework, allowing for automatic differentiation and GPU acceleration where available.

The framework's architecture is modular, with clear separation between model components such as kernels, likelihoods, and optimization routines. This modularity makes it straightforward to extend existing models or create new ones. GPflow includes a comprehensive suite of kernel functions, ranging from the basic RBF (Radial Basis Function) kernel to more sophisticated ones like the Spectral Mixture kernel.

One of GPflow's distinguishing features is its robust handling of model parameters. It provides facilities for parameter transformation (ensuring parameters remain in valid ranges during optimization), prior specification, and flexible optimization strategies. The framework also includes utilities for model evaluation and prediction, with built-in support for uncertainty estimation.

3.2.2 GPflux Framework

GPflux extends GPflow to enable the construction and training of Deep Gaussian Process (DGP) models using modern deep learning practices. Built on top of both GPflow and TensorFlow, GPflux provides a flexible framework for implementing deep probabilistic models that combine the expressiveness of deep neural networks with the uncertainty quantification capabilities of Gaussian processes. This work was originally published in [Dutordoir et al., 2021].

The framework implements Deep Gaussian Processes as a composition of Gaussian process layers, where each layer transforms its inputs through a Gaussian process mapping. This hierarchical structure allows the model to capture complex, non-stationary patterns in data while maintaining uncertainty estimates throughout the entire processing pipeline. GPflux handles this complexity through a carefully designed architecture that separates the concerns of model specification, inference, and prediction.

A key innovation in GPflux is its integration with Keras-style layers, allowing practitioners to combine deterministic neural network layers with Gaussian

process layers in a single model. This hybrid approach enables the construction of models that can leverage both the structural assumptions encoded in neural network architectures and the flexibility of Gaussian processes.

For inference, GPflux implements various approximate methods, including stochastic variational inference with inducing points, which makes it possible to train deep Gaussian process models on large datasets. The framework provides built-in support for minibatch training, making it practical to work with datasets that don't fit in memory.

3.2.3 Implementation Considerations

When implementing Chained Gaussian Processes for multiple annotator learning using these tools, several key considerations must be addressed. First, the model architecture must be carefully designed to capture both the individual annotator characteristics and their interdependencies. This requires extending the basic GP layers provided by GPflux to handle multiple output streams - one for each annotator's reliability model.

Second, the inference procedure must be adapted to handle the unique challenges of multiple annotator data, such as missing labels and varying levels of annotator expertise across different regions of the input space. The variational inference capabilities of GPflux provide a solid foundation for this, but careful attention must be paid to the design of the variational approximation to ensure it captures the relevant uncertainty in both the ground truth and annotator reliability estimates.

Finally, computational efficiency becomes a critical concern when scaling to large datasets with multiple annotators. The minibatch training capabilities of both frameworks help address this, but additional considerations such as appropriate inducing point selection and batch size tuning become important for achieving good performance in practice.

3.3 Experimental Setup

Synthetic datasets were used to evaluate the proposed methods in classification and regression tasks, following very similar procedures to the ones used in [**?**], which are as follows:

3.3.1 Classification

A fully synthetic data for a 1-D ($P = 1$) multiclass classification problem with $K = 3$ classes is generated. The input feature matrix $\mathbf{X}$ is constructed by randomly sampling $N = 100$ points from a uniform distribution over the interval $[0, 1]$. For each sample n, the true label is determined by computing three target values:

$$
\begin{aligned}
t_{n,1} &= \sin(2\pi x_n), \\
t_{n,2} &= -\sin(2\pi x_n), \\
t_{n,3} &= -\sin(2\pi(x_n + 0.25)) + 0.5,
\end{aligned}
$$

and assigning the class label as the index $i \in \{1, 2, 3\}$ that maximizes $t_{n,i}$. This construction ensures a nontrivial, nonlinear decision boundary between the classes.

For evaluation, the test set is generated by extracting 200 equally spaced samples from the interval $[0, 1]$, providing a dense coverage of the input space to assess generalization performance.

Note that at this point, the fully synthetic dataset does not contain real annotator labels. Therefore, it is necessary to simulate annotator labels as corrupted versions of the hidden ground truth. In these experiments, the simulation is designed to account for: 1) dependencies among annotators, and 2) labeler performance that varies as a function of the input features.

To achieve this, an SLFM-based approach (termed SLFM-C) is employed, which generates annotator labels as follows:

1. **Definition of deterministic functions:** Define Q deterministic functions $\hat{\mu}_q : \mathcal{X} \to \mathbb{R}$ and their combination parameters $\hat{w}_{l,r,q} \in \mathbb{R}$, for all $r \in R$, $n \in N$.

2. **Annotator performance computation:** Compute $\hat{f}_{l,r,n} = \sum_{q=1}^{Q} \hat{w}_{l,r,q} \hat{\mu}_q(\hat{x}_n)$, where $\hat{x}_n \in \mathbb{R}$ is the nth component of $\hat{\mathbf{x}} \in \mathbb{R}^N$. Here, $\hat{\mathbf{x}}$ is the 1-D representation of the input features in **X**, obtained using the t-distributed Stochastic Neighbor Embedding (t-SNE) approach.

3. **Reliability calculation:** Calculate $\hat{\lambda}_n^r = \varsigma(\hat{f}_{l,r,n})$, where $\varsigma(\cdot) \in [0, 1]$ is the sigmoid function.

4. **Label assignment:** The rth annotator's label for the nth sample is given by

$$y_n^r = \begin{cases} y_n, & \text{if } \hat{\lambda}_n^r \geq 0.5 \\ \tilde{y}_n, & \text{if } \hat{\lambda}_n^r < 0.5 \end{cases}$$

   where y_n is the true label and $\tilde{y}_n$ is a flipped version of the actual label y_n.

This procedure ensures that the simulated annotator labels exhibit both inter-annotator dependencies and input-dependent reliability.

3.3.2 Regression

A fully synthetic dataset is also generated for a 1-D regression problem, where the ground truth for the nth sample is defined as:

$$y_n = \sin(2\pi x_n) \sin(6\pi x_n)$$

Here, the input matrix **X** is constructed by randomly sampling 100 points uniformly from the interval $[0, 1]$. The test set is created by extracting equally spaced samples

from the same interval, ensuring a clear separation between training and testing data.

To simulate the presence of multiple annotators, the ground truth is corrupted with annotator-specific noise. For each annotator r, the observed label for the nth sample is given by:

$$y_n^r = y_n + \epsilon_n^r$$

where $\epsilon_n^r \sim \mathcal{N}(0, v_n^r)$ is Gaussian noise, and v_n^r represents the error variance for the rth annotator, which may depend on the input features. This setup allows us to model both the bias and the reliability of each annotator, as well as potential dependencies between annotators.

To further increase realism, the error variance v_n^r can be modeled as a function of the input features, capturing scenarios where annotator reliability varies across the input space, thus:

$$\epsilon_n^r \sim \mathcal{N}(0, v_n^r)$$

being v_n^r, the rth annotator's error variance for the nth sample.

Again, for modeling the error as a function of the input features, an SLFM-based approach (termed SLFM-R) is employed to build the noisy labels, as follows:

1. **Definition of deterministic functions:** Define Q deterministic functions $\hat{\mu}_q : \mathcal{X} \to \mathbb{R}$ and their combination parameters $\hat{w}_{l,r,q} \in \mathbb{R}$, for all $r \in R$, $n \in N$.
2. **Annotator performance computation:** Compute $\hat{f}_{l,r,n} = \sum_{q=1}^{Q} \hat{w}_{l,r,q} \hat{\mu}_q(\hat{x}_n)$, where $\hat{x}_n \in \mathbb{R}$ is the nth component of $\hat{\mathbf{x}} \in \mathbb{R}^N$. Here, $\hat{\mathbf{x}}$ is the 1-D representation of the input features in $\mathbf{X}$, obtained using the t-distributed Stochastic Neighbor Embedding (t-SNE) approach.
3. **Compute variance :** Finally, determine $v_n^r = \exp(\hat{f}_{l,r,n})$.

3.4 Summary

3.4.1 Comparative Analysis

The evolution of multiple annotator learning methods represents a fascinating progression in how we approach the challenge of learning from diverse expert opinions. KAAR established a foundation with its non-parametric approach through kernel alignment, though it operates under the assumption of constant annotator performance across the input space. LKAAR advanced this concept by introducing input-dependent annotator behavior, recognizing that expert performance often varies across different types of inputs. RCDNN brought the power and flexibility of deep learning to the problem, offering excellent scalability but requiring careful attention to regularization and architecture design. CGP represents perhaps the most theoretically elegant approach, providing a principled probabilistic framework, though it faces significant computational challenges in practical applications.

The selection of an appropriate method for a given application requires careful consideration of several factors. The size and dimensionality of the dataset play a crucial role, as some methods scale better than others with increasing data volume. The need for uncertainty quantification may favor probabilistic approaches like CGP, while computational constraints might push towards more efficient methods like KAAR. The presence of input-dependent annotator behavior could make LKAAR or RCDNN more appropriate choices. Additionally, when interpretability is a key requirement, as in many medical applications, methods like KAAR and LKAAR might be preferred over the more complex deep learning approaches. Table 3-1 provides a comprehensive comparison of the key characteristics and capabilities of each method discussed.

Table **3-1** Comparison of Multiple Annotator Learning Methods

Characteristic	**KAAR**	**LKAAR**	**RCDNN**	**CGP**
Input-dependent annotator behavior	No	Yes	Yes	Yes
Annotator dependencies modeling	Yes	Yes	Yes	Yes
Uncertainty quantification	No	No	Partial	Yes
Computational scalability	High	Medium	High	Low
Missing data handling	Yes	Yes	Yes	Yes
Non-linear relationships	Yes	Yes	Yes	Yes
Training complexity	Low	Medium	High	High
Hyperparameter sensitivity	Low	Medium	High	Medium
Interpretability	High	Medium	Low	Medium
Memory requirements	Low	Medium	High	High

Note: Partial uncertainty quantification in RCDNN refers to the use of Monte Carlo dropout for approximate uncertainty estimation.

3.5 Conclusion

In this chapter, we have presented a comprehensive overview of the Chained Gaussian Processes model and its application to multiple annotator learning. We have discussed the theoretical underpinnings of the model, its implementation using GPflow and GPflux, and its evaluation on synthetic datasets. It was shown in [**?**] that CCGPMA outperforms other method in state of the art methods in terms of accuracy and uncertainty quantification (see 1.3.1).

CHAPTER

FOUR

TRUNCATED GENERALIZED CROSS ENTROPY FOR SEGMENTATION

4.1 Loss functions for multiple annotators

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

4.1.1 Generalized Cross Entropy

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

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

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

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, $\mathbf{1}^\top \mathbf{y} = \mathbf{1}^\top f(\mathbf{x}) = 1$, $\mathbf{1} \in \{1\}^K$ being an all-ones vector. In addition, the MAE loss can be rewritten for softmax outputs, yielding:

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

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

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

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

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

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

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

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

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

4.1.2 Extension to Multiple Annotators

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

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

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

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

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

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

4.1.3 Reliability Maps and Truncated GCE

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

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

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

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

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

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

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

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

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

$$\mathscr{L}\left(\mathbf{Y}_r[n], f(\mathbf{X}[n];\theta)|_r(\mathbf{X}[n];\theta)\right) = \frac{1}{N}\sum_{n}^{N} TGCE_{SS}(\mathbf{Y}_r[n], f(\mathbf{X}[n];\theta)|_r(\mathbf{X}[n];\theta)) \tag{4-11}$$

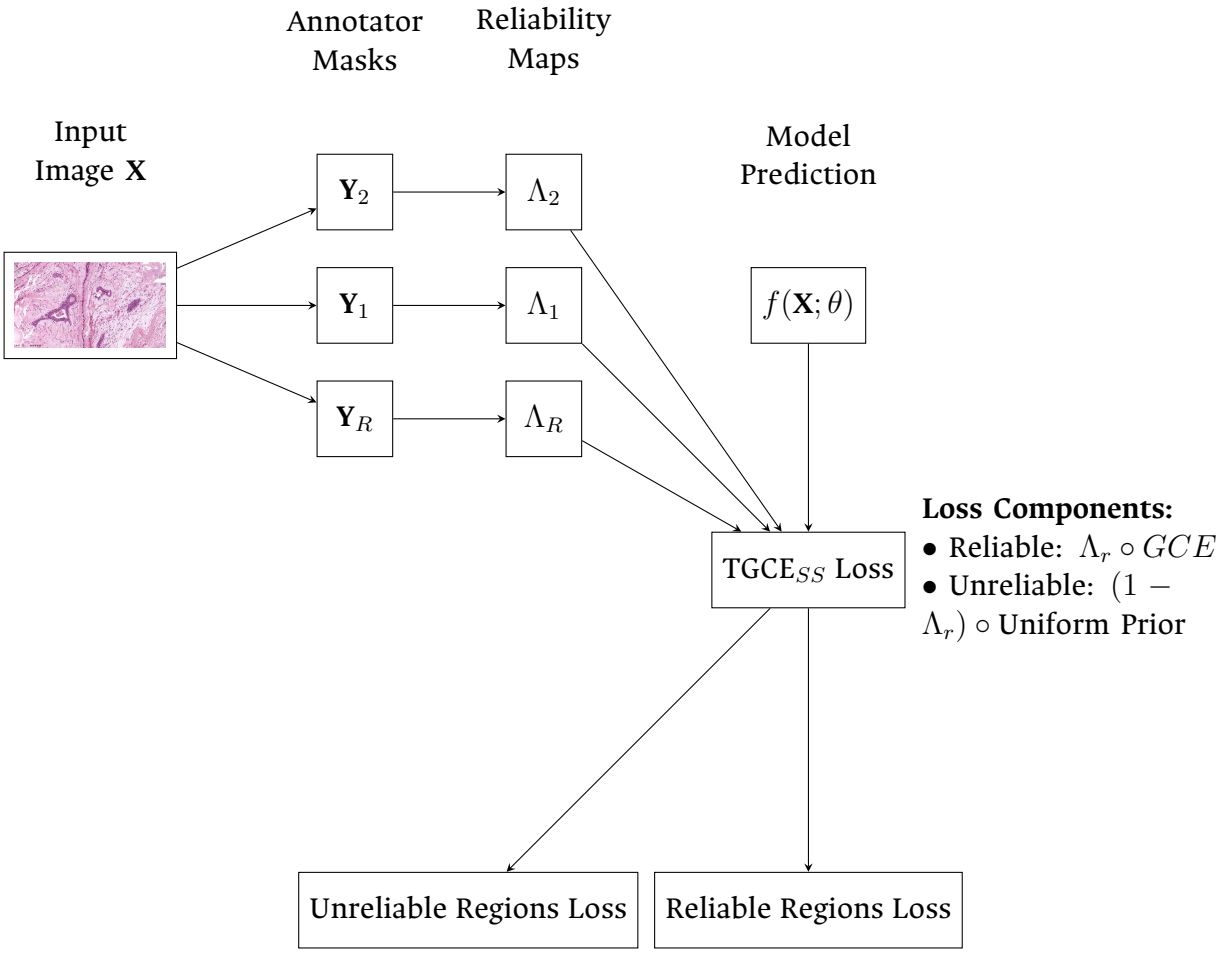

Figure **4-1** Working mechanism of the proposed Truncated Generalized Cross Entropy for Semantic Segmentation (TGCE$_{SS}$) loss function. The loss combines reliability maps Λ_r with the model's predictions $f(\mathbf{X};\theta)$ to compute a weighted loss that accounts for annotator reliability across different image regions.

CHAPTER

FIVE

CHAINED DEEP LEARNING FOR IMAGE SEGMENTATION

5.1 Introduction

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.

5.2 Using U-NET as a building block

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.

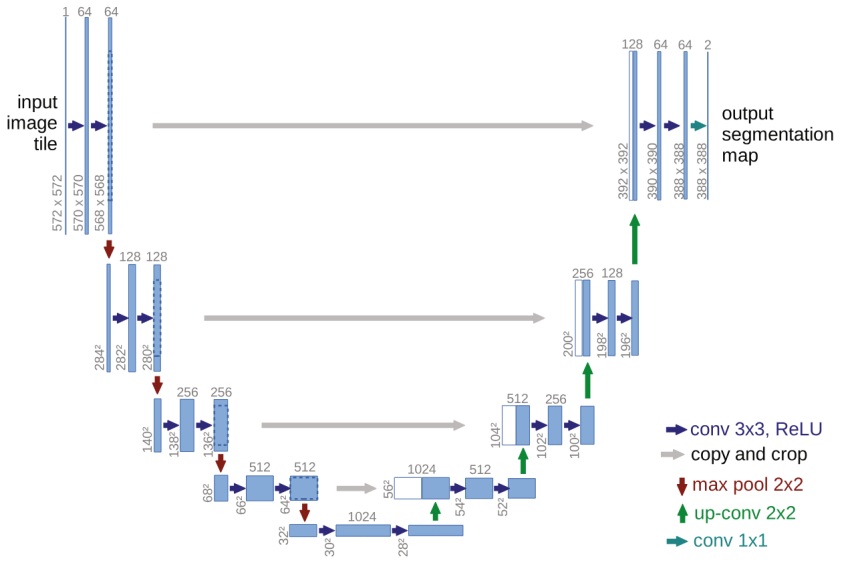

Figure **5-1** Original U-NET architecture.

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

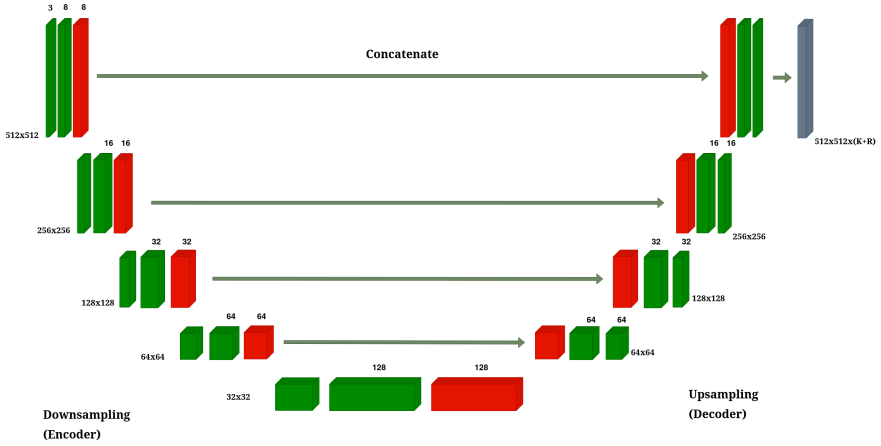

Figure **5-2** Overview of the proposed model architecture.

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

5.2.3 Reliability Map Branch

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

5.2.4 Integration with TGCE$_{SS}$ Loss

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

5.2.5 Training Process

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

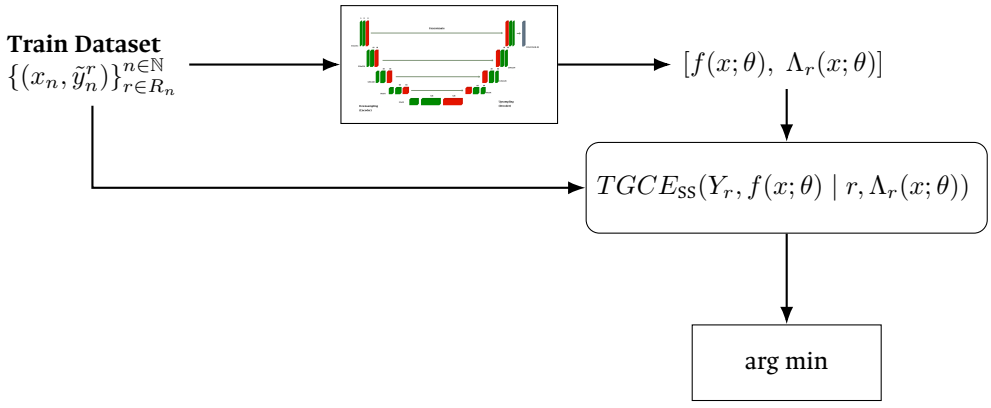

Figure **5-3** Training process of the proposed model overview. Estimated segmentation and reliability maps are computed for each input image, and the loss function is computed for the entire batch.

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

CHAPTER

SIX

CONCLUSIONS

6.1 Summary

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

- 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.

- Existing loss functions in the state of the art were insufficient to model the annotators' performance across the input space, from which the need for a new loss function which fulfils the requirements of the task was identified.

- A novel loss function was proposed and evaluated, showing to be a good alternative to the existing loss functions in the state of the art, and providing an efficient way to optimize parameters for reaching capability to model the annotators' performance across the input space.

- The use of U-Shaped Deep Learning models as a building block for the proposed solution was a valid approach, as it allows for the model to learn the spatial relationships between the pixels in the image, and the annotations, and to use this information to make predictions about the annotations of new images.

- Chained Deep Learning approaches in combination with the proposed loss function were the best approach for the task, outperforming the state of the art in the task of ISS of histopathology images.

6.2 Future work

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

[1] Including GPFlow, and GPFlux.

BIBLIOGRAPHY

[Abadi et al., 2016] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P. A., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., and Zhang, X. (2016). Tensorflow: A system for large-scale machine learning. *CoRR*, abs/1605.08695. (page 35)

[Amgad et al., 2019] Amgad, M., Elfandy, H., Hussein, H., Atteya, L. A., Elsebaie, M. A. T., Abo Elnasr, L. S., Sakr, R. A., Salem, H. S. E., Ismail, A. F., Saad, A. M., Ahmed, J., Elsebaie, M. A. T., Rahman, M., Ruhban, I. A., Elgazar, N. M., Alagha, Y., Osman, M. H., Alhusseiny, A. M., Khalaf, M. M., Younes, A.-A. F., Abdulkarim, A., Younes, D. M., Gadallah, A. M., Elkashash, A. M., Fala, S. Y., Zaki, B. M., Beezley, J., Chittajallu, D. R., Manthey, D., Gutman, D. A., and Cooper, L. A. D. (2019). Structured crowdsourcing enables convolutional segmentation of histology images. *Bioinformatics*, 35(18):3461–3467. (pages xviii, 29, 41, and 42)

[Avanzo et al., 2024] Avanzo, M., Stancanello, J., Pirrone, G., Drigo, A., and Retico, A. (2024). The evolution of artificial intelligence in medical imaging: From computer science to machine and deep learning. *Cancers (Basel)*, 16(21):3702. Author Joseph Stancanello is employed by Elekta SA. The remaining authors declare no commercial or financial conflicts of interest. (page 3)

[Azad et al., 2024] Azad, R., Aghdam, E. K., Rauland, A., Jia, Y., Avval, A. H., Bozorgpour, A., Karimijafarbigloo, S., Cohen, J. P., Adeli, E., and Merhof, D. (2024). Medical image segmentation review: The success of u-net. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10076–10095. (page 2)

[Banerjee et al., 2025] Banerjee, A., Shan, H., and Feng, R. (2025). Editorial: Artificial intelligence applications for cancer diagnosis in radiology. *Frontiers in Radiology*, 5. (page 8)

[Bhalgat et al., 2018] Bhalgat, Y., Shah, M. P., and Awate, S. P. (2018). Annotation-cost minimization for medical image segmentation using suggestive mixed supervision fully convolutional networks. *CoRR*, abs/1812.11302. (page 3)

[Brito-Pacheco et al., 2025] Brito-Pacheco, D., Giannopoulos, P., and Reyes-Aldasoro, C. C. (2025). Persistent homology in medical image processing: A literature review. (page 2)

[Carmo et al., 2025] Carmo, D. S., Pezzulo, A. A., Villacreses, R. A., Eisenbeisz, M. L., Anderson, R. L., Van Dorin, S. E., Rittner, L., Lotufo, R. A., Gerard, S. E., Reinhardt, J. M., and Comellas, A. P. (2025). Manual segmentation of opacities and consolidations on ct of long covid patients from multiple annotators. *Scientific Data*, 12(1):402. (page 9)

[de G. Matthews et al., 2016] de G. Matthews, A. G., van der Wilk, M., Nickson, T., Fujii, K., Boukouvalas, A., León-Villagrá, P., Ghahramani, Z., and Hensman, J. (2016). Gpflow: A gaussian process library using tensorflow. (page 48)

[Dutordoir et al., 2021] Dutordoir, V., Salimbeni, H., Hambro, E., McLeod, J., Leibfried, F., Artemev, A., van der Wilk, M., Hensman, J., Deisenroth, M. P., and John, S. (2021). Gpflux: A library for deep gaussian processes. (page 49)

[Elhaminia et al., 2025] Elhaminia, B., Alsalemi, A., Nasir, E., Jahanifar, M., Awan, R., Young, L. S., Rajpoot, N. M., Minhas, F., and Raza, S. E. A. (2025). From traditional to deep learning approaches in whole slide image registration: A methodological review. (page 2)

[Elnakib et al., 2020] Elnakib, A., Elmenabawy, N., and S Moustafa, H. (2020). Automated deep system for joint liver and tumor segmentation using majority voting. *MEJ-Mansoura Engineering Journal*, 45(4):30–36. (page 11)

[Farahani et al., 2015] Farahani, N., Parwani, A. V., and Pantanowitz, L. (2015). Whole slide imaging in pathology: advantages, limitations, and emerging perspectives. *Pathology and Laboratory Medicine International*, 7:23–33. (pages xvii and 30)

[Giri and Bhatia, 2024] Giri, K. and Bhatia, S. (2024). Artificial intelligence in nephrology- its applications from bench to bedside. *International Journal of Advances in Nephrology Research*, 7(1):90–97. (page 6)

[Grefve et al., 2024] Grefve, J., Söderkvist, K., Gunnlaugsson, A., Sandgren, K., Jonsson, J., Keeratijarut Lindberg, A., Nilsson, E., Axelsson, J., Bergh, A., Zackrisson, B., Moreau, M., Thellenberg Karlsson, C., Olsson, L., Widmark, A., Riklund, K., Blomqvist, L., Berg Loegager, V., Strandberg, S. N., and Nyholm, T. (2024). Histopathology-validated gross tumor volume delineations of intraprostatic lesions using psma-positron emission tomography/multiparametric magnetic resonance imaging. *Physics and Imaging in Radiation Oncology*, 31:100633. (page 14)

[Habis, 2024] Habis, A. A. (2024). *Developing interactive artificial intelligence tools to assist pathologists with histology annotation*. Theses, Institut Polytechnique de Paris. (page 8)

[Hu et al., 2025] Hu, D., Jiang, Z., Shi, J., Xie, F., Wu, K., Tang, K., Cao, M., Huai, J., and Zheng, Y. (2025). Pathology report generation from whole slide images with knowledge retrieval and multi-level regional feature selection. *Computer Methods and Programs in Biomedicine*, 263:108677. (pages 2, 29, and 30)

[Julián and Álvarez Meza Andrés Marino, 2023] Julián, G. G. and Álvarez Meza Andrés Marino (2023). A supervised learning framework in the context of multiple annotators. (pages 17 and 43)

[Karthikeyan et al., 2023] Karthikeyan, R., McDonald, A., and Mehta, R. (2023). What's in a label? annotation differences in forecasting mental fatigue using ecg data and seq2seq architectures. (page 9)

[Kim et al., 2024] Kim, Y., Lee, E., Lee, Y., and Oh, U. (2024). Understanding novice's annotation process for 3d semantic segmentation task with human-in-the-loop. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, IUI '24, page 444–454, New York, NY, USA. Association for Computing Machinery. (page 9)

[Lam and Suen, 1997] Lam, L. and Suen, S. (1997). Application of majority voting to pattern recognition: an analysis of its behavior and performance. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 27(5):553–568. (page 11)

[Lin et al., 2024] Lin, Y., Lian, A., Liao, M., and Yuan, S. (2024). Bcdnet: A fast residual neural network for invasive ductal carcinoma detection. (page 6)

[López-Pérez et al., 2023] López-Pérez, M., Morales-Álvarez, P., Cooper, L. A. D., Molina, R., and Katsaggelos, A. K. (2023). Crowdsourcing segmentation of histopathological images using annotations provided by medical students. In Juarez, J. M., Marcos, M., Stiglic, G., and Tucker, A., editors, *Artificial Intelligence in Medicine*, pages 245–249, Cham. Springer Nature Switzerland. (pages 5, 6, 8, and 11)

[Lu et al., 2023] Lu, X., Ratcliffe, D., Kao, T.-T., Tikhonov, A., Litchfield, L., Rodger, C., and Wang, K. (2023). Rethinking quality assurance for crowdsourced multi-roi image segmentation. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 11(1):103–114. (pages 5 and 8)

[López-Pérez et al., 2024] López-Pérez, M., Morales-Álvarez, P., Cooper, L. A., Felicelli, C., Goldstein, J., Vadasz, B., Molina, R., and Katsaggelos, A. K. (2024). Learning from crowds for automated histopathological image segmentation. *Computerized Medical Imaging and Graphics*, 112:102327. (pages xvii, 5, 15, and 17)

[Mazzarini et al., 2021] Mazzarini, M., Falchi, M., Bani, D., and Migliaccio, A. R. (2021). Evolution and new frontiers of histology in bio-medical research. *Microscopy Research and Technique*, 84(2):217–237. (pages xvii and 29)

[Pan et al., 2021] Pan, X., Lu, Y., Lan, R., Liu, Z., Qin, Z., Wang, H., and Liu, Z. (2021). Mitosis detection techniques in h&e stained breast cancer pathological images: A comprehensive review. *Computers & Electrical Engineering*, 91:107038. (page 31)

[Panayides et al., 2020] Panayides, A. S., Amini, A., Filipovic, N. D., Sharma, A., Tsaftaris, S. A., Young, A., Foran, D., Do, N., Golemati, S., Kurc, T., Huang, K., Nikita, K. S., Veasey, B. P., Zervakis, M., Saltz, J. H., and Pattichis, C. S. (2020). Ai in medical imaging informatics: Current challenges and future directions. *IEEE Journal of Biomedical and Health Informatics*, 24(7):1837–1857. (page 2)

[Parkhi et al., 2012] Parkhi, O. M., Vedaldi, A., Zisserman, A., and Jawahar, C. V. (2012). Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*. (page 38)

[Qiu et al., 2022] Qiu, Y., Hu, Y., Kong, P., Xie, H., Zhang, X., Cao, J., Wang, T., and Lei, B. (2022). Automatic prostate gleason grading using pyramid semantic parsing network in digital histopathology. *Frontiers in Oncology*, 12. (page 14)

[Rashmi et al., 2021] Rashmi, R., Prasad, K., and Udupa, C. B. K. (2021). Breast histopathological image analysis using image processing techniques for diagnostic purposes: A methodological review. *Journal of Medical Systems*, 46(1):7. (pages 1 and 5)

[Ronneberger et al., 2015] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F., editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham. Springer International Publishing. (page 14)

[Ryou et al., 2025] Ryou, H., Thomas, E., Wojciechowska, M., Harding, L., Tam, K. H., Wang, R., Hu, X., Rittscher, J., Cooper, R., and Royston, D. (2025). Reticulin-free quantitation of bone marrow fibrosis in mpns: Utility and applications. *eJHaem*, 6(2):e70005. (page 2)

[Sarvamangala and Kulkarni, 2022] Sarvamangala, D. R. and Kulkarni, R. V. (2022). Convolutional neural networks in medical image understanding: a survey. *Evolutionary Intelligence*, 15(1):1–22. (pages 3 and 6)

[Shah et al., 2018] Shah, M. P., Merchant, S. N., and Awate, S. P. (2018). Ms-net: Mixed-supervision fully-convolutional networks for full-resolution segmentation. In Frangi, A. F., Schnabel, J. A., Davatzikos, C., Alberola-López, C., and Fichtinger, G., editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 379–387, Cham. Springer International Publishing. (page 5)

[Shalf, 2020] Shalf, J. (2020). The future of computing beyond moore's law. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 378(2166):20190061. (page 3)

[TIAN and Zhu, 2015] TIAN, T. and Zhu, J. (2015). Max-margin majority voting for learning from crowds. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc. (page 12)

[Triana-Martinez et al., 2023] Triana-Martinez, J. C., Gil-González, J., Fernandez-Gallego, J. A., Álvarez Meza, A. M., and Castellanos-Dominguez, C. G. (2023). Chained deep learning using generalized cross-entropy for multiple annotators classification. *Sensors*, 23(7). (page 20)

[Warfield et al., 2004] Warfield, S., Zou, K., and Wells, W. (2004). Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, 23(7):903–921. (page 12)

[Weitz et al., 2023] Weitz, P., Valkonen, M., Solorzano, L., Carr, C., Kartasalo, K., Boissin, C., Koivukoski, S., Kuusela, A., Rasic, D., Feng, Y., Sinius Pouplier, S., Sharma, A., Ledesma Eriksson, K., Latonen, L., Laenkholm, A.-V., Hartman, J., Ruusuvuori, P., and Rantalainen, M. (2023). A multi-stain breast cancer histological whole-slide-image data set from routine diagnostics. *Scientific Data*, 10(1):562. (pages xviii, 32, 40, and 41)

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

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

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

[Zhang and Sabuncu, 2018] Zhang, Z. and Sabuncu, M. R. (2018). Generalized cross entropy loss for training deep neural networks with noisy labels. (pages 20, 58, and 59)

[Zhao et al., 2020] Zhao, R., Qian, B., Zhang, X., Li, Y., Wei, R., Liu, Y., and Pan, Y. (2020). Rethinking dice loss for medical image segmentation. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 851–860. (page 20)

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

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