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
2023

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

Brandon Lotero Londoño

Dissertation submitted as a partial requirement to receive the grade of:

Master in Engineering - Industrial Automation

Advisor:

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

Co-advisor:

Prof. Germán Castellanos-Domínguez, Ph.D.

Academic research group:

Signal Processing and Recognition Group - SPRG

Universidad Nacional de Colombia

Faculty of Engineering and Architecture

Department of Electric, Electronic and Computing Engineering

Manizales, Colombia

2025

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
2023

ACKNOWLEDGEMENTS

PENDING

Brandon Lotero Londoño
2025

ABSTRACT

PENDING

Keywords: PENDING

RESUMEN

PENDIENTE

Palabras clave: PENDIENTE

Contents

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

CAD Computer-Aided Diagnosis 2, 5, 7

CNN Convolutional Neural Networks 3, 15, 17

CT Computed Tomography 11

ISS Image Semantic segmentation 2, 3, 7, 9, 13, 15–17

MITs Medical Imaging Techniques 1

ML Machine Learning 9, 11

MV Majority Voting 9, 11

OCR Optical Character Recognition 11

PET Positron Emission Tomography 13

ROI Region of Interest 2, 7

SS Semantic segmentation 3

STAPLE Simultaneous Truth and Performance Level Estimation 12–14

WSI Whole Slide Imaging 1, 5, 7, 14

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, CT, 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 «CITE». Histopathology is rapidly gaining prominence, particularly in segmentation and feature extraction, where AI-driven techniques aid in automated cancer diagnosis and tissue structure analysis. The integration of Deep Learning in histological image processing is revolutionizing pathology, enabling more precise and efficient diagnostics. A brief comparison of the tasks and medical image types based on recent literature review, can be seen in Figure **1-1**. [Yu et al., 2025], [Brito-Pacheco et al., 2025], [Ryou et al., 2025], [Hu et al., 2025], [Elhaminia et al., 2025]

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

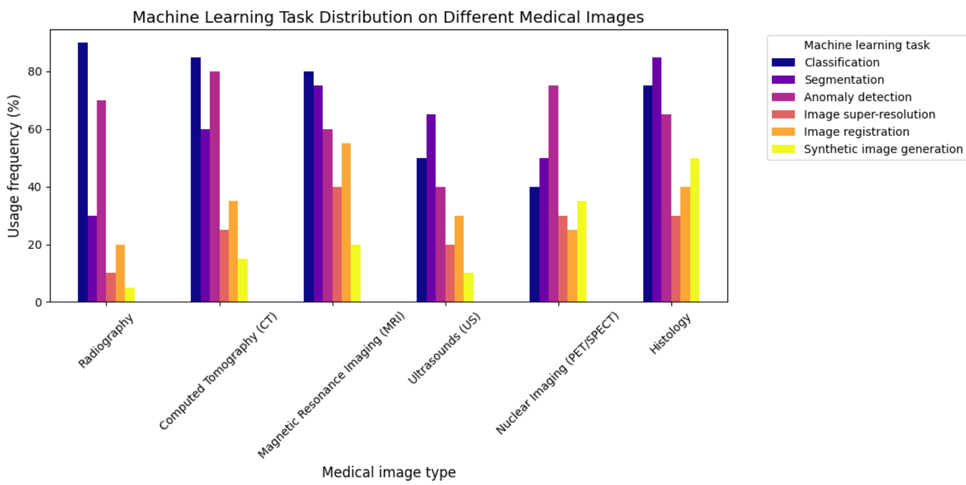

Figure **1-1** Estimation of the tasks and medical image types based on recent literature review (count of referenced terms).

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

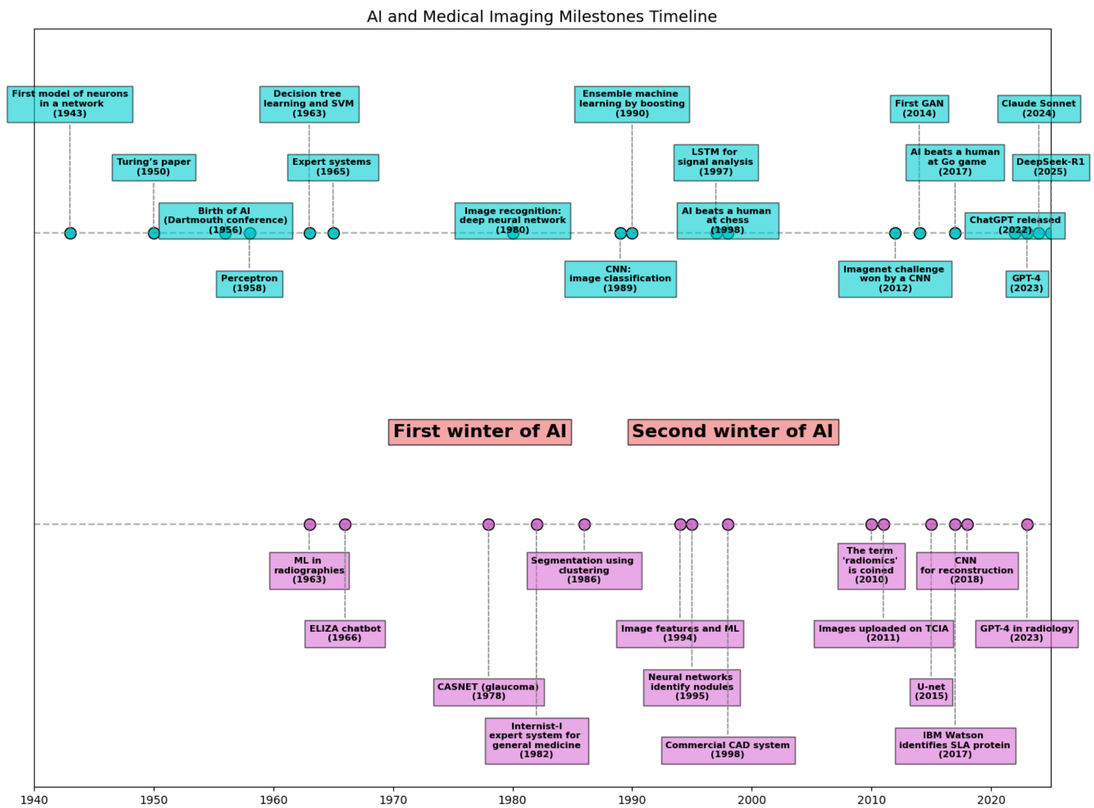

Figure **1-2** AI and machine learning in medical imaging brief timeline.

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. 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 mainlining 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

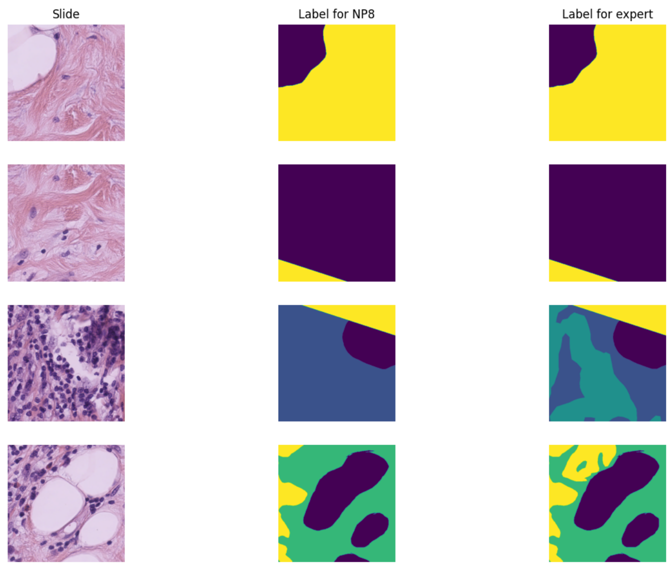

Figure **1-3** Example of a histopathological image segmented by multiple annotators, illustrating variations in label assignment.

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

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].

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

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

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:

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?

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

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

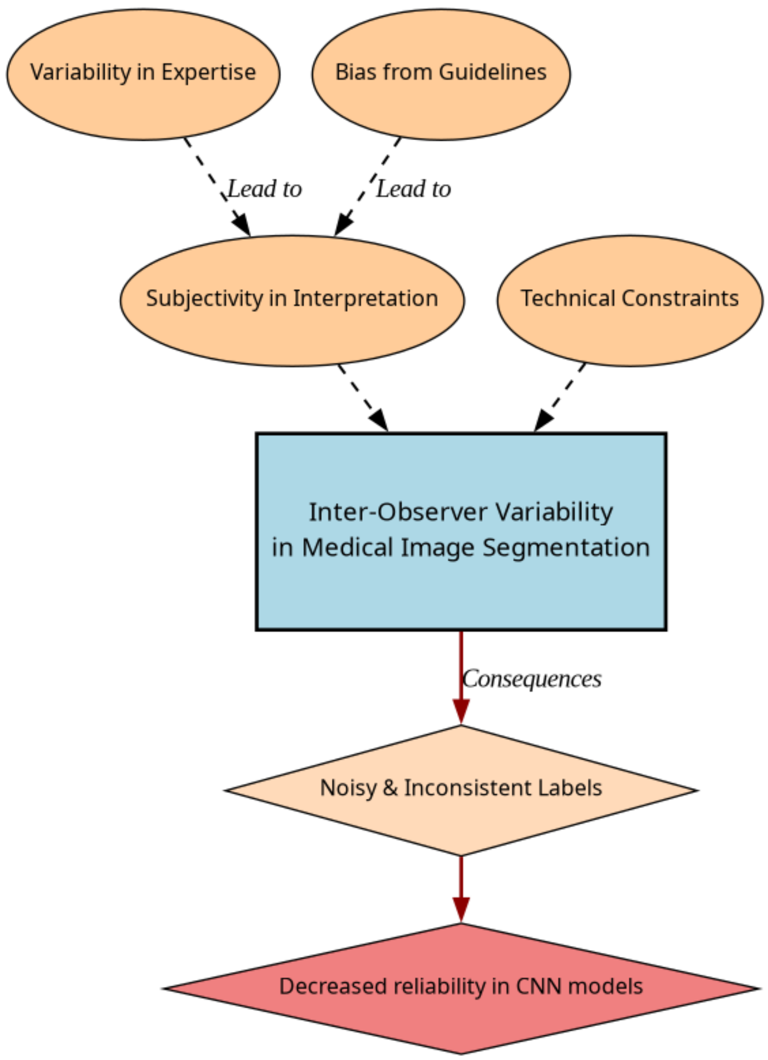

Figure **1-4** Summary diagram for problem Statement approach to implement. This concept was born multiple times and divergently in multiple fields, but it was described as relevant for ML and pattern recognition labeling for classification in [Lam and Suen, 1997], in which the approach is exposed as 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 Computed Tomography (CT) images for Liver Tumor Segmentation, or [López-Pérez et al., 2023], which uses MV for combining noisy annotations as an additional annotator to be included in the deep learning solution. Majority voting as a technique for setting a pseudo ground truth label is a powerful approach for its simplicity in many use cases in which the target to be labeled is not tied to an expertise related task, otherwise, the assumption of equal expertise among the labelers can be a source of bias in the final label, which is not desirable in the case of highly technical annotations like medical images. In subsection 1.3.1, we will be reviewing literature which no longer assumes the naive approach of equal expertise among labelers and face the challenge of learning from inconsistent labels.

1.3.1 Facing annotation variability in medical images

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

STAPLE Mechanism

The Simultaneous Truth and Performance Level Estimation (STAPLE) algorithm, introduced in [Warfield et al., 2004] is a probabilistic framework that estimates a hidden true segmentation from multiple segmentations provided by different raters. It also estimates the reliability of each rater by computing their sensitivity and specificity.

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

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

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

Chained Gaussian Processes

Other works like [Gil-González et al., 2025] proposed a novel approach

1.3.2 Strategies for handling low-quality images

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

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

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

1.4 Aims

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

1.4.1 General Aim

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

1.4.2 Specific Aims

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

1.5 Outline and Contributions

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

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

[4] https://pypi.org/project/seg_tgce/
[5] https://seg-tgce.readthedocs.io/en/latest/experiments.html
[6] https://github.com/blotero/seg_tgce

BIBLIOGRAPHY

[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 8)

[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)

[Gil-González et al., 2025] Gil-González, J., Daza-Santacoloma, G., Cárdenas-Peña, D., Orozco-Gutiérrez, A., and Álvarez Meza, A. (2025). Generalized cross-entropy for learning from crowds based on correlated chained gaussian processes. *Results in Engineering*, 25:103863. (page 14)

[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 7)

[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 13)

[Habis, 2024] Habis, A. A. (2024). *Developing interactive artificial intelligence tools to assist pathologists with histology annotation*. Theses, Institut Polytechnique de Paris. (page 7)

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

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

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

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

[Lin et al., 2024] Lin, Y., Lian, A., Liao, M., and Yuan, S. (2024). Bcdnet: A fast residual neural network for invasive ductal carcinoma detection. (page 7)

[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. (page 5)

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

[Qiu et al., 2022] Qiu, Y., Hu, Y., Kong, P., Xie, H., Zhang, X., Cao, J., Wang, T., and Lei, B. (2022). Automatic prostate gleason grading using pyramid semantic parsing network in digital histopathology. *Frontiers in Oncology*, 12. (page 13)

[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)

[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 7)

[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 14)

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

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

[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)

[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)