AI-based analysis of abdominal ultrasound images to support medical diagnosis

Hind Dadoun

UNIVERSITÉ CÔTE D'AZUR

ÉCOLE DOCTORALE
SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

THÈSE DE DOCTORAT

Analyse d'images d'échographie abdominale basée sur l'IA pour aider au diagnostic médical

Hind DADOUN

INRIA, Équipe EPIONE

Thèse dirigée par Nicholas AYACHE et co-dirigée par Hervé DELINGETTE et Anne-Laure ROUSSEAU

Soutenue le 06 décembre 2022

Présentée en vue de l'obtention du grade de DOCTEUR EN AUTOMATIQUE, TRAITEMENT DU SIGNAL ET DES IMAGES de l'UNIVERSITÉ CÔTE D'AZUR.

Devant le jury composé de :

ISABELLE BLOCH	Télécom Paris	Présidente et Rapporteur
ALISON NOBLE	University of Oxford	Rapporteur
DANIEL RÜCKERT	Technical University of Munich	Rapporteur
Anne-Laure ROUSSEAU	Hôpital Européen Georges-Pompidou	Co-encadrant
Hervé DELINGETTE	Centre Inria d'Université Côte d'Azur	Co-directeur de thèse
Nicholas AYACHE	Centre Inria d'Université Côte d'Azur	Directeur de thèse

HEALTH DATA HUB · NHANCE · 3iA Côte d'Azur Institut interdisciplinaire d'intelligence artificielle · Inria · ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

Analyse d'images d'échographie abdominale basée sur l'IA pour aider au diagnostic médical

AI-based Analysis of Abdominal Ultrasound Images to Support Medical Diagnosis

Jury

Présidente

Isabelle Bloch	Professeure	Télécom Paris

Rapporteurs

Isabelle Bloch	Professeure	Télécom Paris
Alison Noble	Professeure	University of Oxford
Daniel Rückert	Professeur	Technical University of Munich

Examinateurs

Anne-Laure Rousseau	Praticienne hospitalière	Hôpital Européen Georges-Pompidou
Hervé Delingette	Directeur de recherche	Centre Inria d'Université Côte d'Azur
Nicholas Ayache	Directeur de recherche	Centre Inria d'Université Côte d'Azur

Résumé

L'objectif de notre étude est d'analyser comment les outils d'apprentissage automatique peuvent être adaptés pour être utilisés pour l'interprétation automatique d'images d'échographie abdominale, en prenant en compte une difficulté majeure: l'absence de bases de données d'échographie abdominale propres, annotées et librement accessibles. Dans cette thèse, nous détaillerons ces défis et indiquerons des premiers éléments pour pallier à certains d'entre eux.

Le chapitre 2 décrit la construction d'une grande base de données d'échographie abdominale provenant d'un hôpital universitaire où un total de 8011 examens d'échographie abdominale (120 593 images) de 6482 patients ont été extraits, ainsi que les rapports médicaux correspondants. Nous nous concentrons sur la documentation du jeu de données, y compris ses caractéristiques et la collecte des données, ainsi que sur une évaluation critique des biais du jeu de données et des instances mal étiquetées.

Dans le chapitre 3, nous proposons un logiciel de prétraitement pour les images échographiques qui permet à la fois à une parfaite désidentification des images et la standardisation de leur contenu. La méthode permet la délimitation du cône d'acquisition en échographie et le remplissage des annotations (lignes, caractères) à l'intérieur du cône en combinant une approche probabiliste paramétrique avec un réseau de segmentation, en plus des méthodes de remplissage automatique.

Dans le chapitre 4, nous avons mené une étude pour entraîner et évaluer les performances d'un réseau de neurone profond sur une tâche spécifique autour de l'imagerie échographique, en présence d'une quantité raisonnable de données sans bruit et fortement étiquetées. Ses performances sont ensuite comparées à celle de soignants ayant différents niveaux d'expertise. La détection, la localisation et la caractérisation des lésions focales du foie dans les images échographiques en mode B ont été choisies comme cadre de cette étude: d'abord, car cette tâche a un intérêt clinique bien documenté, et ensuite, car les études précédentes se sont concentrées uniquement sur la caractérisation des lésions en omettant les tâches de détection et de localisation des lésions dans le foie.

Le chapitre 5 explore comment les données non étiquetées peuvent être exploitées pour améliorer les représentation visuelles apprises en utilisant l'apprentissage auto-supervisé et/ou semi-supervisé pour la classification des organes abdominaux. En particulier,

nous proposons d'adapter deux méthodes multi-classes de pointe au contexte de la classification multi-labels: le clustering profond avec PICA, et l'apprentissage semi-supervisé avec FixMatch.

Enfin, nous discutons des défis restants et des orientations futures potentielles.

Mots-clés: imagerie ultrasonore, apprentissage bayésien, détection d'objets, apprentissage auto-supervisé, apprentissage semi-supervisé, traitement du langage naturel.

Abstract

The focus of our study is to analyze how machine learning tools can be used for the automatic interpretation of abdominal ultrasound images, with a major setback : the absence of curated, annotated and openly available abdominal US databases. In this thesis, we will detail those challenges and point out first elements to alleviate some of them.

Chapter 2 describes the construction of a large abdominal ultrasound database where a total of 8011 abdominal ultrasound examinations (120 593 images) from 6482 patients were extracted, along with the corresponding medical reports from a university hospital. We focus on the documentation of the dataset, including its characteristics and data collection, as well as a critical evaluation of dataset biases, and mislabeled instances.

In Chapter 3, we propose a preprocessing pipeline for ultrasound images that results in both a perfect de-identification of the images as well as the standardization of their content. The method allows for the delimitation of the acquisition cone in ultrasound imaging and the inpainting of annotations (lines, characters) inside the cone by combining a parametric probabilistic approach with a segmentation network, in addition to inpainting methods.

In Chapter 4, we conducted a study to train and evaluate the performance of a deep learning-based network on a specific task around ultrasound imaging, when given access to a reasonable amount of strongly labeled noise-free data, and compare its performance to that of caregivers with different levels of expertise. The detection, localization, and characterization of focal liver lesions in B-mode ultrasound images were chosen as the setting for this study : first, because this task has a well documented clinical interest, and second, because previous studies have focused solely on lesion characterization while omitting the tasks of lesion detection and localization in the liver.

Chapter 5 explores how unlabeled data can be leveraged to improve the learned representation of features using either self-supervised and/or semi-supervised learning for abdominal organ classification. In particular we propose to adapt two state-of-the-art multi-class methods to the multi-label classification setting: deep clustering with PICA, and semi-supervised learning with FixMatch.

Finally, we discuss remaining challenges and potential future directions.

Keywords: ultrasound imaging, Bayesian learning, object detection, self-supervised learning, semi-supervised learning, natural language processing.

Remerciements

Il y a trois ans, j'ai contacté le Dr Anne-Laure Rousseau pour lui faire part de mon désir d'être bénévole au sein de son association NHance, à Paris. Ayant du temps libre à la conclusion de mon stage, j'étais désireuse de mettre ce temps à profit en rejoignant un projet à fort impact le temps de quelques semaines. Anne-Laure a du savoir trouver les bons mots, car, à l'issue de cet appel, j'acceptais des entretiens pour une thèse de trois ans à l'autre bout de la France.

C'est au sein de l'équipe EPIONE que j'ai rencontré mes directeurs académiques, Nicholas et Hervé, qui m'ont guidée et conseillée avec bienveillance, même lorsque je me montrais têtue. Nicholas, je te remercie d'avoir toujours été là quand j'avais besoin d'aide, que ce soit pour faire avancer le projet avec les différents acteurs, pour me rediriger vers les pistes à suivre, ou simplement pour ton retour à chaque étape. Hervé, tu m'as donné beaucoup de liberté pour explorer des sujets très différents, et tu as réussi à me guider dans chacun d'entre eux. Je vous remercie tous les deux pour nos discussions constructives et vos idées qui ont façonné cette thèse.

Anne-Laure, ce projet n'existe que grâce à toi, et tu as réussi l'ambitieux défi de le mener à bien. En plus de tes compétences, j'admire sincèrement ton engagement et ton éthique. Tu m'as fait confiance rapidement, ce qui m'a permis de m'y investir et de m'y épanouir. Je te remercie pour ton aide tout au long de cette thèse.

Je tiens également à remercier les membres externes du jury, les professeurs Isabelle Bloch, Alison Noble et Daniel Ruckert pour avoir lu attentivement ce manuscrit et partagé leurs commentaires et conseils. En particulier, je remercie Isabelle pour avoir présidé la soutenance.

Ce fut un plaisir de partager ces trois années avec les permanents et les doctorants de l'équipe, je ne pouvais espérer un meilleur cadre. Je profite de cette occasion pour remercier personnellement Paul, avec qui j'ai partagé un bureau pendant trois ans. Je repars avec un titre, mais aussi un véritable ami.

Pour finir, je suis reconnaissante à toutes ces personnes qui, par leur simple présence et leur soutien indéfectible, me permettent de naviguer sans crainte. Je pense en particulier à Salma, Kenza, Diane et Daniel sur qui je peux toujours compter. Mais aussi Raphaël, qui était au premier rang de cette aventure et grâce à qui, la traversée s'est faite dans la joie. Enfin, merci à mes parents, mon frère et ma famille qui m'ont toujours encouragé dans mes choix sans forcément les comprendre, j'espère que vous êtes fiers aujourd'hui.

Acknowledgement

We thank the French Health Data Hub for providing resources and support, and especially Stephanie Combes and Emmanuel Bacry, Salma Eloualydy, MS in the help provided to construct our database, Christophe Dubois for coordinating the project.

We are grateful for the help provided by the team of the radiology department at Saint Louis Hospital: Eric de Kerviler MD, Fanny Joujou, Kemel Khezzane, Constance De Margerie MD, PhD.

We are grateful to the radiology team from Necker Hospital, and particularly Jean-Michel Correas MD PhD, Anne-Marie Tissier MD, Sylvain Bodard MD.

We thank Deepomatic for making available an annotation platform to the Nance project and especially Thibaut Duguet, Augustin Marty, Alois Brunel, and Vincent Delaitre.

We thank the Clinical Data Warehouse of Greater Paris University Hospitals led by Laure Maillant and especially the imaging team lead by Aurelien Maire.

We thank Remi Rousseau, MS, Guillaume Oules, MS, and Alexandre Dubreucq, MS, both from Ecole Polytechnique, for their help in the design of the Nance projecT.

We greatly appreciated the help that Mariama Bah, MD, and Gregory Khelifi, MD, Albane De Kerautem, MD, and Cecile Poret provided.

This work was supported by the clinical research unit of Saint Louis Hospital: Jerome Lambert, MD, PhD, and Claire Montlahuc, MD, PhD.

We thank Olivier Lucidarme MD, PhD and Charles Doulin from the radiology department at Pitié-Salpêtrière Hospital.

Financial Support

This work has been supported by the French government, through the 3IA Cote d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002. The authors are grateful to the OPAL infrastructure from Universite Cote d'Azur. We thank Deepomatic for making available an annotation platform to the NHance project. We thank the French Health Data Hub for providing resources and support.

Contents

Abbreviations:

Medical

US	ultrasound
HCC	hepatocellular carcinoma
FLLs	focal liver lesions
PACS	picture archiving and communications system
DICOM	digital imaging and communications in medicine

Networks

CNN	convolutional neural network
R-CNN	region-based convolutional neural network
DETR	DEtection TRansformer
PICA	PartItion Confidence mAximisation

Metrics

PPV	positive predictive value
FN	false-negative finding
FP	false-positive finding
TN	true-negative finding
TP	true-positive finding
MCC	Matthews correlation coefficient
AUC	area under the receiver operating characteristic curve

1 Introduction

Contents

This thesis explores how abdominal ultrasound combined with automated image analysis could be used by non-expert caregivers to select individuals that will eventually be directed to a trained sonographer.

1.1 Clinical Context

Access to Imaging Technology in Global Health

In her keynote addressed to the First World Health Organization (WHO) Global Forum on Medical Devices, Dr. Margaret Chan, former WHO Director-General, raised the alarm about the number of people excluded from the benefits of medical devices (e.g., less than one CT scan per million people in low- and middle-income countries, compared to nearly 40 CT scans per million people in high-income countries). Yet these devices are of major importance as they are used to diagnose, monitor or treat medical conditions. She suggested three main reasons for the unequal access to medical imaging services. The most obvious one relates to resources and costs with a considerable gap in annual government spending on health (over $7,000 per capita to less than $10). The second is inherent to the medical device industry largely focusing on financially profitable diseases. Finally, the third reason is lack of training capacity. She goes on to say that the biggest breakthroughs are likely to come from technologies that use "alternative power supplies [...], and can be operated, with no risk to patient safety, by personnel with little specialized training. [...] Or with robust portable machines that extend the advantages of technology beyond the hospital setting or take it from cities to rural areas [Organization, 2011]". Ultrasound offers many of the aforementioned benefits, making it a method of choice for low to mid income countries: either in an emergency, during a patient follow-up visit, or during a public health screening examination. Indeed, it is a non-invasive[1]

[1] There is no known risks from the sound waves used in ultrasound exams. There is no ionizing radiation exposure associated with ultrasound imaging (no potential radiation induced cancer).

imaging modality with no side effects (such as radiation-induced cancers) that allows for cost-effective[2] real-time diagnosis, screening and/or monitoring. Ultrasound is the primary imaging modality recommended for many clinical indications worldwide. More specifically in France, 30 million ultrasound examinations are performed annually in the private sector, with an increase of 1 million examinations each year. This increase is 7 times greater than that of MRI and 6 times greater than that of CT. This represents a total market of 1.5 billion for private practitioners[3].

Ultrasound: a Cost Effective Tool for Real-time Diagnosis, Screening and/or Monitoring

Ultrasound has proven to be a safe and useful tool for assessing the abdomen. In an emergency setting, an observational pilot study [Jang, 2014] showed that US performed by an emergency physician can have a positive impact on decision making and diagnostic workup in patients who did not present with any specific abdominal pain. In another study [Lindelius, 2009], the authors performed a randomised trial to evaluate the diagnostic accuracy of surgeon-performed ultrasound for patients presenting with abdominal pain in the emergency department (ED) and showed that it was significantly higher in the group examined with ultrasound (64.7% vs 56.8%, p = 0.027). Lindelius et al [Lindelius, 2008] also showed that the use of bedside ultrasound results in fewer further requested examinations, and fewer admissions. Finally, [Mjolstad, 2012] showed that adding a close-up ultrasound examination lasting less than 10 minutes to usual care resulted in a corrected diagnosis in one in five patients admitted to a medical department. A variety of diseases can be detected on abdominal ultrasound. For example, it is the primary and often the only imaging modality used to evaluate the gallbladder [Sidhu, 2022]. Specifically, for detection of gallstones and polyps and for diagnosis of acute cholecystitis [Randen, 2008; Zenobii, 2016], it is often recommended as the primary modality in young or thin patients to avoid radiation exposure associated with CT. Another example is the early detection of renal obstruction - visible on ultrasound in the presence of a dilated collecting system - which can prevent permanent renal damage. In addition, pocket-sized ultrasound was shown to be useful for evaluating dilatation of the renal collecting system, especially for ruling out its presence [Kameda, 2018]. Ultrasound is also the most widely used screening and surveillance tool for detecting hepatocellular carcinoma (HCC) worldwide, and is used annually for millions of patients considered to be at high risk of developing HCC [Morgan, 2018]. Ultrasound can also identify splenic abnormalities such as splenomegaly or focal splenic masses [Andrews, 2000; Izranov, 2019].

Barriers in the Use of Ultrasound

Ultrasound imaging requires a qualified sonographer to acquire and interpret ultrasound

[2]The cost of pocket-sized ultrasound machines is much lower than that of standard ones, approximately USD 10,000 vs. USD 49,000 respectively.

[3]https://www.ccomptes.fr/fr/publications/limagerie-medicale

images. Today, the use of ultrasound imaging by the clinician at the bedside of the patient (point of care ultrasound) is relevant to more than 20 different medical and surgical specialties and is no longer limited to radiologists [Moore, 2011]. The legislative framework for the practice of ultrasound by non-radiologists differs considerably among countries, varying from tight legislation to no guidelines at all [Radiology ESR, 2020]. Today, the American Society of Ultrasound provides guidelines to ensure the quality of examinations. Regarding training, the WHO has published - since 2001 - the "Manual of diagnostic ultrasound" which has been reissued and improved since. WHO recommends a minimum of 6 months of full-time training in a specialized center, but it is specified that even in this case, more experience is desirable. Despite these efforts, limitations to its use persist [Schnittke, 2019; Shah, 2015], mainly due to the limited number of trained operators. With appropriate training and the right systems in place, there can be a potential transfer of tasks from trained sonographers and physicians to trained nurses, midwives, and clinical officers, thus enabling the democratization of access to this modality.

1.2 Automated Image Analysis for Ultrasound

Deep learning based strategies were developed around the entire processing chain in US imaging, from ultrasound-specific processing methods [Salvadeo, 2014; Luijten, 2019] to automated interpretation of US images and captioning [Alsharid, 2022]. In the following, we focus on studies around the acquisition and interpretation of ultrasound images.

Clinical Tasks

An abdominal ultrasound examination requires considerable manual effort to obtain standard views of abdominal organs, annotate the views in text, and record measurements of clinically relevant organs. Deep learning methods have the potential to assist in each of these steps, as recently shown in [Matthew, 2022]. A common approach to applying deep learning methods is to help the clinician obtain a more accurate measurement or to acquire the measurement in less time [Lee, 2020; Meng, 2019]. This first approach, although saving the examiner valuable time, still requires the images to be acquired by a highly trained operator, which is is not the goal of this study. Alternatively, some of the methods aim at guiding the user in obtaining the correct standardized plans [Bimbraw, 2020]. This is highly valuable but still requires a human to be trained in the use of an ultrasound machine, and the interpretation of the images to obtain clinically relevant information. To overcome this problem, one study on antenatal care [Heuvel, 2019] proposes to extract information from predefined free-hand sweeps. The main advantage of this approach is that these predefined sweeps can be taught to any healthcare worker, without any ultrasound knowledge, rapidly. Unfortunately it requires the construction of

a specific database (i.e., video clips of exams performed by non-experts) and to define a scanning protocol which is much more complex for the abdomen. Indeed, there are other challenges in ultrasound beyond the prenatal and obstetrical context. Among these difficulties, air in the lungs and digestive tract of an adult body reflects ultrasound and may obscure abdominal organs, stool may mimic tumors that are difficult to recognize, the operator may ask the patient to drink in order to fill the bladder and improve analysis of pelvic organs. Patients who are very short of breath, agitated or tearful may be difficult to examine, obesity and overweight may limit the visibility and analysis of deep structures, and the long course of chronic diseases may render ultrasound analysis more difficult. Finally, some approaches aim at extracting information relevant to the diagnosis from standard planes acquired by experts [George, 2022]. This is not ideal either, as training a model on images acquired by experts does not guarantee a generalization of these methods to potentially lower quality images acquired by non-experts. However, the main benefit is that images can be collected from the hospital retrospectively at almost no cost, providing a first database for the evaluation of deep learning methods. For this reason, we focus on deep learning methods for extracting diagnostically relevant information from standard planes throughout this thesis.

Methodologies and Anatomical Applications

Many studies on ultrasound image analysis focus on fetal ultrasound, where fetal growth and development are monitored to identify potential problems and facilitate diagnosis [Baumgartner, 2017; Zimmer, 2020; Dahdouh, 2015]. Other applications include tumor identification and segmentation in breast ultrasound [Almajalid, 2018], localization of clinically relevant B-line artifacts in lung ultrasound [Baloescu, 2020], accurate identification of cardiac cycle phases (end-diastole (ED) and end-systole (ES)) in echocardiography [Bajaj, 2021], or detection of thyroid nodules [Koundal, 2018], one of the most common nodular lesions in the adult population worldwide. For the abdomen, studies focus on a specific organ of the abdomen, such as the liver or kidneys, and are often disease-specific, such as measuring the severity of hepatic steatosis [Chou, 2021], automatic segmentation of a pancreatic tumor [Iwasa, 2021], or diagnosis of neoplastic polyps of the gallbladder [Jang, 2021]. There are very few studies covering multiple organs of the abdomen, although examples include the study by Xu *et al.* [Xu, 2018] who perform simultaneous view classification and landmark detection for abdominal ultrasound images, and to the best of our knowledge none of them addresses the detection of abnormalities in the complete abdominal ultrasound examination.

Current Challenges in Automated US Image Analysis

Ultrasound images are difficult to acquire because of their low contrast resolution, operator dependence, and patient-related factors that impact image quality. Once acquired, automatic analysis of abdominal ultrasound images poses additional challenges:

1. Lack of a shared database: One of the biggest issue of this topic is the lack of a shared database on which all methods can be compared. Few open source databases are available, and the existing ones focus on a particular organ. Thus, there is a significant need to build a database containing complete abdominal examinations with examples of all organs of interest.

2. Quality assessment: Ultrasound images usually contain informative annotations such as measurements and standard plane specifications placed by the sonographers during the examination. A pre-processing step is needed to prevent the learning algorithm from using this information as a basis for predictions.

3. Non-standardized examinations: Unlike an abdominal CT scan or MRI, the images linked to an abdominal ultrasound examination are not standardized, meaning that the organs may be visualized from different views based on the position of the transducer on the body, with no guarantee that all organs will be visualized in a single examination.

4. Lack of expert annotations: Developing a model capable of detecting an abnormality at the level of the examination (and thus all the organs visualized) would require at least several hundred annotated examples, which is time consuming and costly. Indeed, an expert is needed to annotate the image, and even for an expert, the interpretation of static ultrasound images, taken out of their context, is difficult. Moreover, ultrasound image anotation is very prone to inter-expert variability. It is therefore crucial to explore other ways of learning visual representations capable of separating classes with few to no labeled data. In particular using self-supervised or semi-supervised vision methods, as well as exploring multi-modal learning between text and image as a way to leverage the radiological reports associated with images.

1.3 Objectives and Organization of the Thesis

This thesis is conducted in partnership with NHance, the APHP's *Entrepôt des Données de Santé* (EDS) and the national health data platform, also known as the Health Data Hub (HDH). The objective is to :

1. Build a large database of abdominal ultrasound images that can be easily exploited for automated image analysis.

2. Evaluate the suitability of different learning methods -supervised and/or unsupervised- for computer-aided diagnosis in abdominal ultrasound imaging.

The manuscript is organized as follows, in accordance with the mentioned research objectives:

In Chapter 2, we start by briefly presenting a dataset designed for the detection of a specific pathology on a single organ. Despite the quality of the acquired data and the richness of the annotations, the effort invested cannot be extended to a dataset of more general use under reasonable time and cost constraints. Consequently, we present Abdo-US, a larger, more realistic dataset for abdominal ultrasound image analysis. We detail the study design choices, including: data extraction methods, selection of labels of interest, and the annotation strategy adopted. Next, we assess the variability of the data as well as potential sources of bias and error in the extracted imaging data. Finally, a study is conducted on the suitability of automatic label extraction from electronic radiology records.

In Chapter 3, have developed a statistical method based on a geometrical modeling of the ultrasound area that allows to detect the region of interest and which combined with deep learning becomes faster and more accurate. Finally, using pre-existing processing tools, we manage to remove the annotations present inside the area. This method was made freely available to allow for standardization of ultrasound image content across different databases. We also worked with the *Entrepôt des Données de Santé* and the Health Data Hub teams to integrate this software in their ultrasound image processing workflow and thus accelerate the development of the project. This work was published at the IEEE International Symposium on Biomedical Imaging [Dadoun, 2021].

In Chapter 4, we develop a framework for the detection, localization, and characterization of focal liver lesions in B-mode ultrasound images. This study was performed on the task-specific dataset. The objective of this study was to compare the performance of expert physicians and non-expert caregivers to state-of-the-art methods with an ideal database. This work has been published in Radiology: Artificial Intelligence [Dadoun, 2022b].

In Chapter 5, we switch back to Abdo-US, a database that more accurately reflects the reality of abdominal ultrasound analysis. We explore self-supervised and semi-supervised learning methods to leverage unlabeled data for abdominal organ classification in the presence of very few labeled data. This work was submitted to a journal [Dadoun, 2022a].

In Chapter 6 the main contributions of this thesis are summarized. Finally, potential future work and perspectives are discussed.

1.4 Publications, software and award

The described contributions led to the following peer-reviewed publications and awards.

Journal Articles

- [Dadoun, 2022b] **Dadoun, H**, Rousseau, A. L., de Kerviler, E., Correas, J. M., Tissier, A. M., Joujou, F., ... & Ayache, N. Deep Learning for the Detection, Localization, and Characterization of Focal Liver Lesions on Abdominal US Images. *Radiology: Artificial Intelligence 4, no. 3 (2022)*

- [Dadoun, 2022a] **Dadoun, H**, Delingette, H., Rousseau, A. L., de Kerviler, E., & Ayache, N. Deep Clustering for Abdominal Organ Classification in US imaging. *Submitted to a journal*

Conference Papers

- [Dadoun, 2021] **Dadoun, H**, Delingette, H., Rousseau, A. L., de Kerviler, E., & Ayache, N. Combining Bayesian and Deep Learning Methods for the Delineation of the Fan in Ultrasound Images. *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) (pp. 743-747). IEEE.*

- [Dadoun, 2023] **Dadoun, H**, Delingette, H., Rousseau, A. L., de Kerviler, E., & Ayache, N. Joint Representation Learning from Radiological Reports and Ultrasound images. *Work in progress for submission to a conference.*

Plug-in or module Software

- EchoFanArea: This software allows the delimitation of the acquisition cone in ultrasound imaging and the inpainting of annotations (lines, characters) inside the cone. It allows both the perfect de-identification of the images but also to standardize the content of the images.

Award

- Won the 2nd prize of the Prix Pierre Laffitte 2O21.

2

AbdoUS: A large dataset for abdominal ultrasound image analysis

Contents

Abstract This chapter describes the construction of a large abdominal ultrasound dataset comprising more than 120,000 images from 6,000 patients. In particular, study design choices are presented, such as: data extraction methods, selection of labels of interest, and the annotation strategy adopted to effectively annotate a subset of the data. Next, we perform an in-depth analysis of the images contained in the dataset as well as the associated annotations (content variation due to different acquisition parameters, label distribution and systematic annotation bias). Finally, we investigate the automatic extraction of labels from patient medical reports, and show that the obtained labels are noisy and cannot be directly used to annotate the images.

2.1 Introduction

The field of automatic visual recognition has grown in recent years with the advent of Convolutional Neural Networks (CNNs) in the mid-1990s, and the availability of large-scale, strongly annotated datasets in the 2000s. More recently, a CNN trained with a set of 129,450 clinical images achieved expert-level performance for skin cancer classification [Esteva, 2017]. More generally, an increasing number of large labeled medical data sets have allowed studies in a variety of medical fields to flourish [Irvin, 2019; Bejnordi, 2017; Menze, 2014]. However, as recently pointed out in [Varoquaux, 2022], the availability of datasets can bias the applications that are considered. This is particularly true in ultrasound imaging, and more specifically in abdominal ultrasound where there is a lack of openly available datasets. For instance, deep learning based ultrasound image analysis studies on the liver and the kidneys, come in seventh and last position, respectively, in a ranking of the number of articles by anatomical structures according to a bibliographic search of all works published until February 1, 2018 and presented in [Liu, 2019a]. Some key factors contribute to the lack of available datasets on abdominal ultrasound and thus (to some extent) to the low number of studies on abdominal ultrasound:

1. Data protection: like any other medical imaging modality, ultrasound is subject to data protection with restricted access on secure servers [Dove, 2018], making it tedious to build a large open-source dataset.

2. The associated data is unstructured and heterogeneous: that is, it is acquired by different machines, different people, and without a strict generalized protocol to perform the examination.

3. Lack of trained operators to acquire ultrasound images: due to its low resolution, ultrasound is a complex modality for which there are few trained operators able to acquire and interpret ultrasound images.

4. Annotation of ultrasound images is challenging: an ultrasound image alone without context is difficult to interpret.

Abdominal ultrasound adds additional complexity as several organs and anatomical regions are involved, namely: the liver, gallbladder and bile ducts, pancreas, spleen, kidneys, aorta and inferior vena cava. It requires to explore different scanning planes to accurately identify each anatomy's specific part viewed from a particular direction. Depending on the depth of the structure to be examined, different transducers can be used- the lower the frequency of the transducer, the greater the penetration; the angle of incidence, the amount of pressure applied and the orientation will also vary accordingly. Images acquired are naturally sensitive to patients' movement, tissue's echogenicity and are very operator-dependent [Strauss, 2007]. Pathologies associated with a given organ may alter its shape, size, contour, position, or textural appearance, resulting in highly variable differences in echographics patterns. Finally, in many cases, the examination is limited to one or multiple areas of the abdomen but not all of them. Readers may refer to [Tempkin, 1999] for detailed presentation of abdominal scanning methods and how images are documented for physician diagnostic interpretation.

Despite these difficulties, ultrasound imaging remains one of the most common techniques for medical diagnosis. It is the only non-invasive imaging modality, with no side effects, such as radiation-induced cancers, that can be used in real-time, making it a method of choice: either in an emergency setting, in consultation for patient follow-up or during a public health screening examination. Therefore, building publicly available datasets will allow training and performance evaluation of deep learning models for diagnostic support on ultrasound images. While essential, the availability of such datasets does not guarantee improved diagnostic accuracy. An important step towards this goal is to critically evaluate dataset biases, mislabeled instances, etc.

In the following, we present two datasets: a task-based and a general-purpose dataset. We start by highlighting the similarities and differences of both datasets. Next, we focus on the second dataset and analyze the associated images and discuss potential challenges related to its use, such as heterogeneous images, class imbalance, and inter-expert variability. Since the images are associated with reports, we explore the use of natural language tools to extract initial information about the dataset.

2.2 Building Two US Datasets : Task-based vs General Purpose

In this section, we present the study framework for the construction of two datasets around ultrasound imaging: a task-based and a general-purpose dataset. The former is a multi-centric dataset designed for the detection of focal liver lesions on ultrasound images(*FLLs dataset*). The associated images were manually selected by expert radiologists only to keep those of high quality and were strongly annotated with localization information in addition to the labels. Despite its many qualities, this small and pathology-specific dataset limits the scope of possible applications. For this reason, we have built a more substantial and general dataset (*US-Abdo dataset*) that better reflects the real conditions of abdominal examinations and that can be applied to several use cases. In total, 8011 abdominal ultrasound examinations (120 593 images) from 6482 patients were extracted (along with the corresponding radiological reports) from the picture archiving and communication system (PACS) of a university hospital.

	FLLs : Task-based Dataset	**US-Abdo: General-purpose Dataset**
Type of Study	Multi-center : Two Centers	Mono-center: Single Center
Data Type	Images	Images and Text
Label Type	Labels + Bounding boxes	Labels
Determination of Ground Truth	Consensus	Majority
Supervision	Fully Supervised	Semi- Supervised
Data Partition	Stratified	Random
Organs of Interest	Liver	Liver, Spleen, Kidneys, Gallbladder
Images	2706	120593
Patients	1074	6482

Tab. 2.1.: Overview of study design choices for each dataset

2.2.1 Study Framework

Data Type: During an ultrasound examination, the examiner performs a complete scan of the area of interest and adjacent structures and takes captures, also known as freeze frames, of the standard scanning plane views and potential visible abnormalities. The freeze-frames along with a textual documentation of the examination form the ultrasound examination report. On average there are 12.5 frames per examination, but this number can vary substantially. Depending on the scope of the study, all or part of the freeze-frames can be included in the final dataset.

Data Collection: International Review Board approval was obtained for this retrospective study, in collaboration with the APHP's *Entrepôt des Données de Santé* registered under the number (no. IRB00011591). Multi-vendor data collected directly from the picture archiving and communication system (PACS) consisted of RGB freeze-frames captured

during ultrasound examination tagged as "abdominal" along with a textual report written by a physician as shown in Figure 2.1.

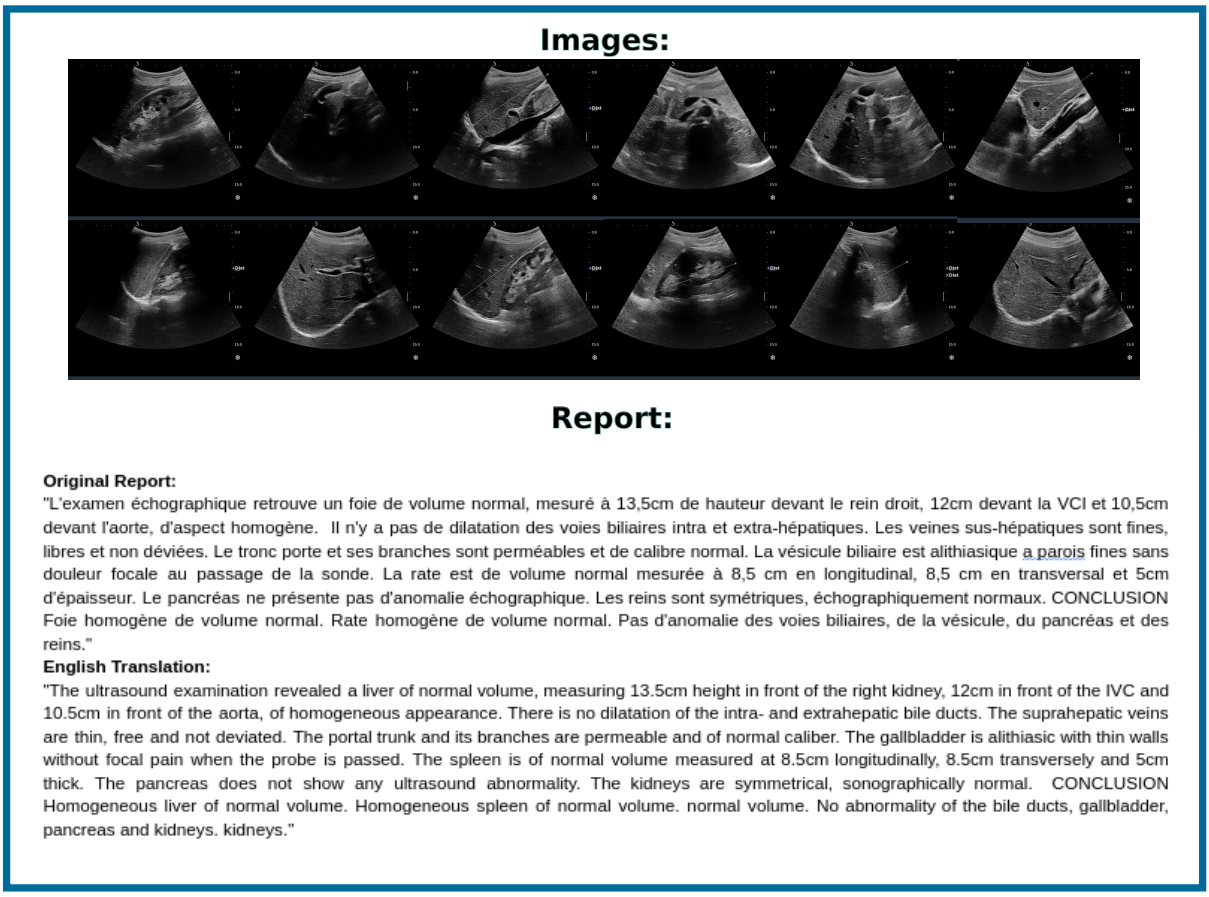


Fig. 2.1.: Example of an abdominal ultrasound examination taken from the picture archiving and communication system (PACS). The examination consists of freeze-frames captured during ultrasound examination along with an unstructured textual report written by a physician and describing the findings of the examination.

Ethics and Legal Compliance: Only adult patients were selected (age $\geq$ 18 years old). All images and clinical reports were de-identified within the centers and no information on the demographics of the study population was retained. A detailed patient's risk re-identification analysis was conducted prior to the data collection and is presented in Appendix A.

Panel of Experts and Annotation Platform: For all studies, a panel of experts was chosen to annotate the images. The panel consisted of four physicians, either radiologists or holders of a national diploma in US imaging, and four sonographers, holders of a national diploma in US imaging from six different health institutions with more than three years of experience. The annotators worked on a tailor-made annotation platform Deepomatic.

2.2.2 Inclusion and Exclusion Criteria

The inclusion and exclusion criteria (at patients, examinations and images level) must be considered before the construction of a dataset. Depending on the research protocol, different criteria may be considered. Naturally, building a study around the detection of focal liver lesions will require a stricter protocol than a more general study on the analysis of abnormalities in the abdomen.

FLLs: For the former, at the patient level, patients with lesions containing liver parenchyma were included if they met the following criteria: 1- lesions were visible on US images (unanimous decision by an arbitration panel), 2 - patients did not receive previous local therapy and 3- a definitive pathological diagnosis of the lesions was obtained. Patients without lesions in the liver parenchyma were selected in case of definitive absence of pathological diagnosis. At the examination level, all abdominal US examinations of patients available in centers 1 and 2 between 2014 and 2019 were selected. For the training and development set, exams performed between 2014 and 2018 were selected. For the test set, examinations performed in 2019 were selected, provided that the corresponding patients had no examinations between 2014 and 2018. At the image level: extremely enlarged images and images obtained in gradient mode simultaneously with CE-US images were excluded.

US-Abdo: For the latter, at the patient level, all adult patients were included. At the examination level, all abdominal US examinations of patients available at Center 1 between 01/07/2015 and 30/06/2018 were selected. At the image level, all images were included in the extracted dataset, regardless of imaging mode and image quality.

2.2.3 Data Partition

In order to distribute the data in the training, development and test sets several strategies are possible. The data can be distributed randomly between the different sets, or in a stratified manner to ensure a similar distribution of classes in each set. In all cases, the sets were constructed so that there was no overlap of patients between sets. The goal of the development set is to choose the parameters of the networks that perform best on images they have never seen.

FLLs: For the focal liver lesion detection database, a total of 1026 patients ($n = 2551$ images) met the inclusion criteria for the training and development set. This set was randomly divided into two subsets containing approximately the same proportions for each class, 80% and 20% for training and development, respectively. A total of 48 new patients ($n = 155$ images) met the inclusion criteria for the test set.

US-Abdo: For the general-purpose abdominal ultrasound database, a total of 6482 patients ($n = 120,593$ images and 8011 abdominal ultrasound exams) met the inclusion criteria for the extracted dataset. 905 exams ($n = 9062$ images) from 875 patients were randomly selected for the labeled training set and 6913 exams ($n = 110,053$ images) from 5417 patients for the unlabeled training set. 47 exams ($n = 503$ images) from 47 patients were randomly selected for the validation set and 146 exams ($n = 975$ images) from 143 patients for the test set.

2.2.4 Label Selection

The selection of labels directly defines the task we are trying to accomplish. When the task is well defined, as is the case for the detection of focal liver lesions, then the labels chosen will be more accurate. Conversely, for a more general task, we must ensure that the labels chosen are general enough to cover the entire scope of the study, but associate them with a well-defined list of diseases to avoid misinterpretation of the general label.

FLLs: For the focal liver lesion database, a hierarchical classification was chosen following the recommendations of the medical expert. The first level describes the liver, and two categories are possible: *homogeneous liver* (i.e., without lesions) or *liver with lesion(s)* with respect to the final diagnosis associated with the image. Selection of lesion(s) led to the second task (i.e., lesion location). Each lesion had to be classified according to the final diagnosis, with two possibilities: *benign lesion(s)* and/or *malignant lesion(s)*. Benign lesions were subdivided into *cyst, angioma, focal nodular hyperplasia or adenoma*. Malignant lesions were subdivided into *metastasis or hepatocellular carcinoma.*

US-Abdo: Regarding the general purpose database, several organs and anatomical regions are involved in a complete abdominal examination, namely: *the liver, gallbladder and bile ducts, pancreas, spleen, kidneys, aorta, and inferior vena cava*. Because some anatomies are more frequently visualized than others, we focused on the four anatomical structures most frequently examined: *liver, kidney, spleen, and gallbladder*. Any difference in ultrasound patterns related to these anatomies is considered abnormal. But because these changes - diffuse or localized - may alter the shape, size, contour, position, or textural appearance of the organ and thus be subject to interpretation by the rater, we have identified a list of diseases associated with each organ, as recommended by medical experts. The complete list is presented in Table 2.2.

Organ	Labels		Sub-categories
Liver	Normal Liver		
	Abnormal Liver	-Liver surface abnormality: surface irregularity, surface nodularity, accentuation of the fissure -Increased echogenicity of the liver parenchyma (ultrasound aspect of diffuse liver steatosis) -Atrophy	-Hepatic parenchyma abnormality (the presence of at least one cyst is considered an abnormality) -Vascular abnormality visible in B-mode : hepatic vein abnormality, portal vein abnormality, portosystemic shunt visible in B-mode -Hepatomegaly
Kidneys	Normal Kidneys		
	Abnormal Kidneys	-Kidney parenchyma abnomality (the presence of at least one cyst is considered an abnormality) -Hydronephrosis : dilatation of the renal pelvis and calyces -Renal atrophy -Nephromegaly	-Stones -Upper ureter abnormality: Pyelitis, Pyelo ureteral junction syndrome
Spleen	Normal Spleen		
	Abnormal Spleen	-Spleen parenchyma abnomality (the presence of at least one cyst is considered an abnormality) -Vascular anomaly visible in B-mode -Splenomegaly	-Splenic atrophy -Presence of an accessory spleen
Gallbladder and Biliary Tract -Vesicular distension	Normal Gallbladder		
	Abnormal Gallbladder	-Stones -Sludge -Other intravesicular abnormalities	-Vascular abnormality visible in B-mode -Vesicular atrophy -Vesicular wall abnormality
	Cholecystectomy Empty gallbladder		
Exclude	If there is a discrepancy between the report and what is visible on the image (e.g., a lesion on an image that is not mentioned in the report). In case of discrepancy within the examination itself (e.g., a fibrosis score F3 or F4 with an ultrasound normal liver).		

Tab. 2.2.: List of diseases associated with each organ

2.2.5 Annotation Strategy and Ground Truth

Annotation of ultrasound images is costly and time consuming because it requires an expert, and even for an expert, interpretation of ultrasound images taken out of context is difficult. This observation implies two factors: 1- The larger the database, the more difficult it will be to annotate all the images, 2- Given the difficulty of the task, annotation errors are to be expected. For this reason, two different annotation strategies were adopted for the two databases.

FLLs: In the task-based dataset, given its small size, all images included in the final dataset were annotated. For each center, the diagnosis associated with each US image was collected by two radiologists with more than 15 years of experience in US image interpretation, by cross-referring with other imaging modalities (contrast-enhanced US (CE-US), CT, MRI, biopsy when available) and clinical files. The final diagnosis was used for the characterization task (liver parenchyma and characterization of FLLs) and did not include the number of lesions in the US image, nor their localization. To determine the ground truth boxes for the localization task (i.e. boxes around the liver and FLLs), an adjudication panel was used as an external standard of reference. Each image was annotated by two experts and at least once by a physician. During this phase, in case of disagreement between two annotators for the localization task, four additional annotators analyzed the questionable image. If the image annotation was not unanimous among the additional annotators, it was excluded from the study.

US-Abdo: Regarding the general-purpose dataset, annotation of the entire dataset would be extremely cost- and time-intensive. Instead, 1065 patients were randomly selected from the 6482 patients in our dataset, with 1098 corresponding examinations and a total of 10,516 images. The remaining unlabeled images will be used in an unsupervised manner. To annotate these images efficiently, we applied different annotation strategies to progressively larger subsets of our dataset: one subset of images is evaluated by multiple raters (triple or double evaluation), while the remaining ones are evaluated by a single rater (single evaluation strategy), as follows:

1. Strategy 1: each examination is assigned to three (or more) randomly selected raters, and the given labels are recorded.

2. Strategy 2: each examination is assigned to two randomly selected raters.

3. Strategy 3: each examination is assigned to one randomly selected rater.

Figure 2.2 illustrates the allocation of annotation strategies in the different sets (training, development, and test). The annotators were asked to select from among the list of labels

presented in Table 2.2 the ones visible on each image. The annotators were not asked to provide any localization information for this task. Multiple factors such as image quality, level of expertise or clarity of the guidelines can play a significant role in the amount of noise potentially induced by a single evaluation strategy. For this reason, the single evaluation strategy is coupled with the double evaluation strategy on the training and development sets. The double evaluation strategy allows to estimate the degree of noise in both sets, and to adapt the training methods accordingly. For the test set, having noise is less admissible. Hence, a triple evaluation was necessary. This way, a majority vote can be used as ground truth. With each image, we keep only the classes that at least two evaluators out of three have selected.

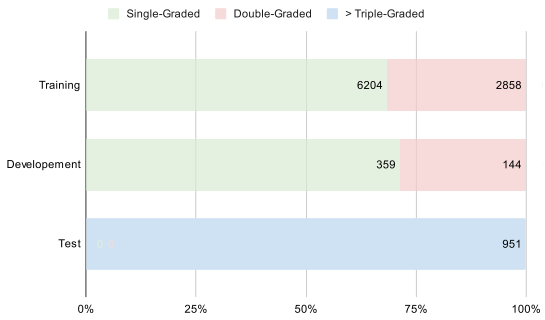


Fig. 2.2.: Allocation of annotation strategies per set.

2.2.6 Overview

Table 2.1 provides an overview of the design choices made to build each dataset. FLLs's dataset was used as is in Chapter 4. The Abdo-US dataset on the other hand was adapted for each chapter. In Chapter 3, a subset of images were randomly chosen to develop the pre-processin tool. In Chapter 5, all images were used but the partition between the labeled and unlabeled sets was different since the study was conducted alongside the annotation process and less labeled images were available at that time. As the focus of the study was the classification of abdominal organs, additional labels were also used, namely the pancreas and bladder. We reiterate in each chapter the characteristics of the adapted datasets for a clear comprehension. Table 2.3 summarizes which dataset was used in the following chapters.

	FLLs	**US-Abdo**
Chapter 3	X (subset)	X (subset)
Chapter 4	X (entire set)	
Chapter 5		X (subset)

Tab. 2.3.: Overview of dataset usage per chapter

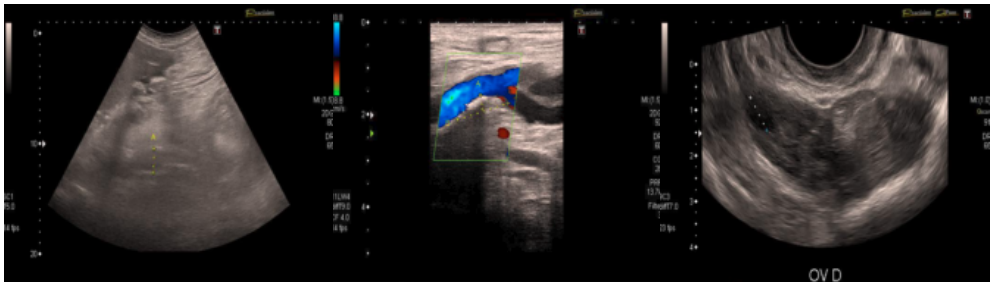


Fig. 2.3.: Example of images from different types of transducers. The image on the left was taken from an "abdominal examination", the image on the center was taken from an "ultrasound for the detection of arteritis" and finally the right image was extracted from a "sus-pubic and endo-vaginal ultrasound".

2.3 Image Dataset Analysis

The aim of this dataset is to allow the development of an automated analysis tool for B-mode images obtained with traditional ultrasound systems and focused on four anatomical structures: the liver, the kidney, the spleen and the gallbladder. Unfortunately, abdominal ultrasound examinations performed in the hospital are not restricted to this setting, and as a result, the content of the extracted images is broader. We start by presenting the factors contributing to the high variability of the image content in Section 2.3.1. We then examine the labeled set of our dataset and document the imbalance between classes (Section 2.3.2). In addition, an in-depth analysis of inter-rater reliability is conducted in Section 2.3.3, highlighting potential noise in our labeled training set.

2.3.1 Sources of Image Variability

Several factors contribute to the high variability of the images in our dataset:

1. Different US area shapes: The first element of an ultrasound scanner is the transducer which allows the transmission and reception of ultrasound. Ultrasound transducers come in different shapes, sizes, and have diverse features (see Figure 2.3). During an abdominal examination different transducers can be used. Sector transducers are often used for in-depth examinations but linear transducers can also be used to examine more superficial structures.

2. Multiple US imaging types: The second element is the imaging mode that transforms the delay between emission and reception into an image. The most well-known and routinely used clinically US imaging mode is B-mode imaging, which displays anatomical information. Other types of US imaging modes allow the measurement of functional parameters related to blood flow (Doppler and contrast-

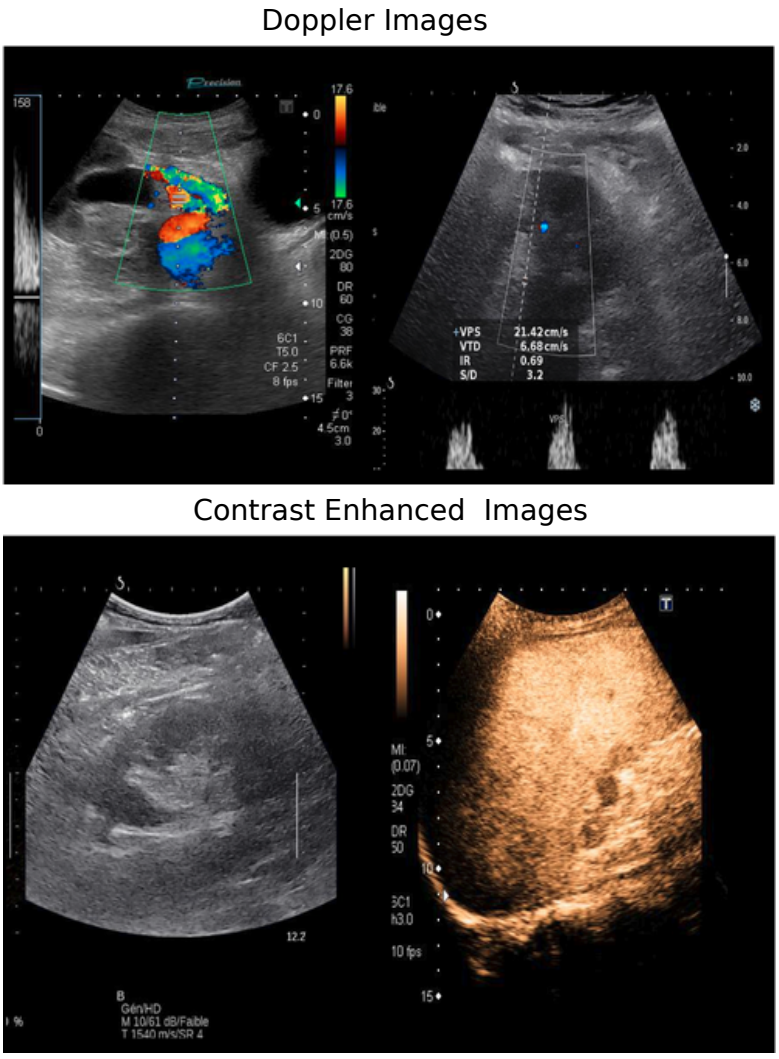


Fig. 2.4.: (Up) Two examples of Doppler images. The image on the left is easily recognizable by simply thresholding the number of colored pixels. The image on the right however, contains very few colored pixels. (Down) Two examples of contrast-enhanced US images. The image on the right is easily identifiable due to its distinct range of color. The image on the left however is very similar to a B-mode image.

enhanced US) and tissue displacement (elastography). In abdominal ultrasound, these US modes are all relevant (eg. Doppler US is indicated as a modality to assess renal perfusion, contrast-enhanced US is often used for the differentiation between benign and malignant focal liver lesions and US elastography for measuring tissue stiffness, a biomarker correlated with liver fibrosis).

3. Highly operator-dependent images: A third element is the operator who acquires images in real-time. As the operator guides the transducer to the correct scan plane, the resulting images are highly operator dependent. Finally, the control console allows the operator to enter various settings and measurements which are manually added to the image during an examination.

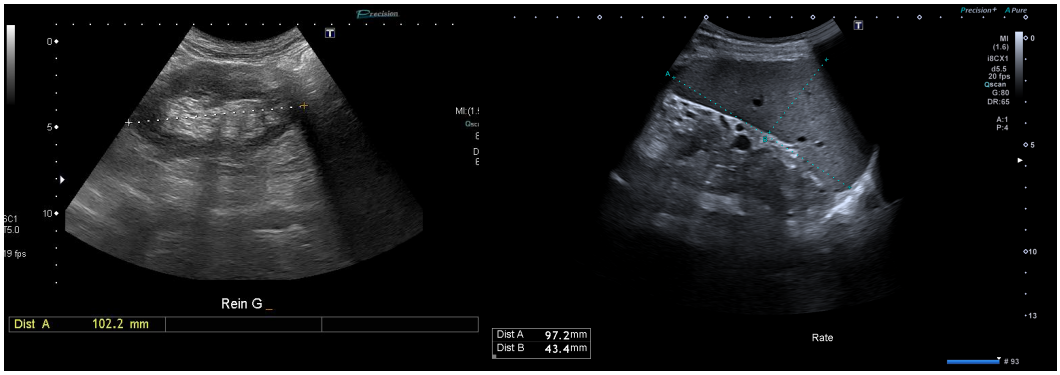
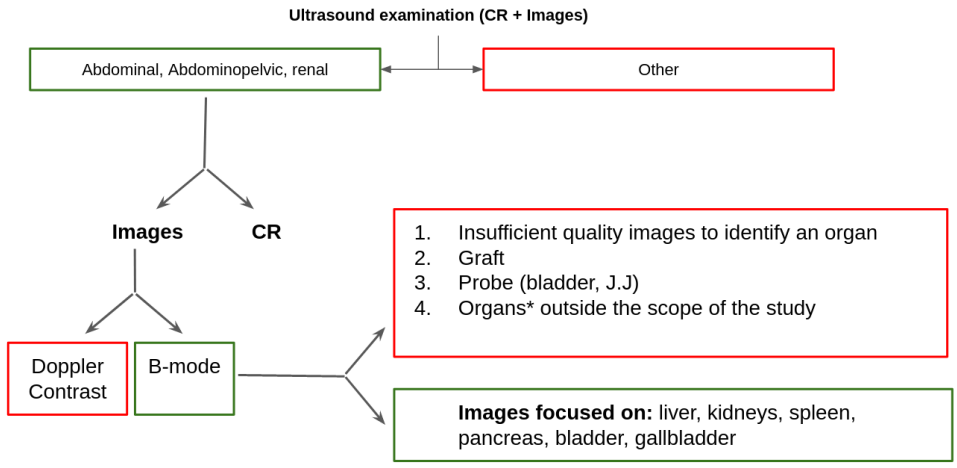


Fig. 2.5.: Image showing several text and graphic elements.



*: Images centered on vessels, ovaries, uterus, prostate, lymph nodes, effusion, location of a biopsy

Fig. 2.6.: A schematic presentation of the ultrasound examinations present in our dataset. Examinations that fall in the "other" category are detailed in Appendix C

4. Different workflows of ultrasound: The use of ultrasound is not limited to monitoring, screening or diagnosis. It can also be used to guide interventional procedures such as biopsies. Depending on the intended use, the images acquired will be very different. Figure 2.6 shows a schematic presentation of the ultrasound examinations present in our dataset. An important task is to automatically filter the dataset to only keep the images of interest.

2.3.2 Label Distribution

We report the distribution of the labels for the different observations in Figure 2.7. Ultrasound is often used for screening, hence most images are normal ("disease-free"), producing a high class imbalance. As several organs can be visible simultaneously in the same image, frequent and infrequent classes can co-occur in the same sample. Therefore

sampling more often minority classes may result in a higher number of majority classes as well. For this reason, the use of more elaborate methods is advised [Liu, 2022].

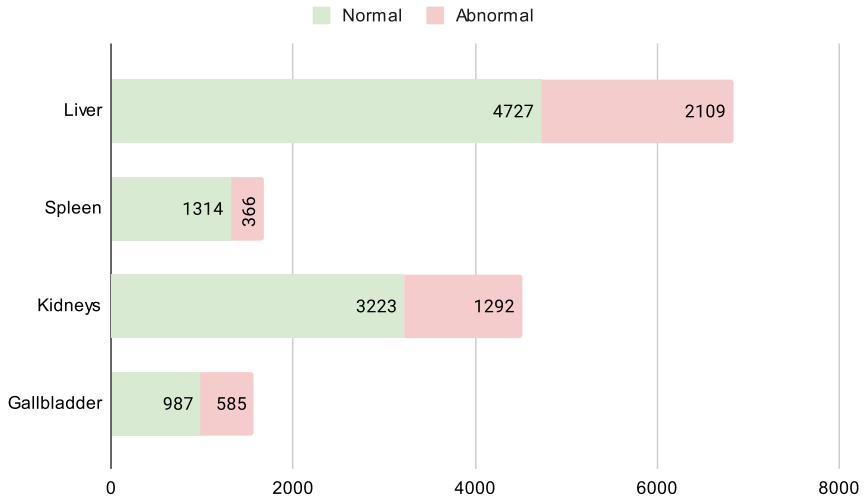


Fig. 2.7.: The AbdoUS dataset consists of 8 labeled observations. We report the number of images which contain these labels in the annotated sets.

2.3.3 Inter-Rater Reliability

	Fully crossed		<———————>				Raters nested within images					
Images/Rater	R1	R2	R1	R2	R3	R4	R1	R2	R3	R4	R5	R6
Image 1	X	X	X	X			X	X				
Image 2	X	X		X	X				X	X		
Image 3	X	X		X	X	X					X	X

Note: This table is strongly inspired by the example figure given in [Putka, 2008].

Tab. 2.4.: Mock example of study designs: left panel represents a crossed design, central panel represent an ill structured design in which raters and images are neither fully crossed nor nested, and right panel represents a nested design.

In ultrasound, any difference in the appearance of the echo pattern may suggest an abnormality in that specific organ, and further examination is often required to confirm a diagnosis. Therefore, image annotation is potentially subject to considerable disagreement among experts, which must be taken into account to converge on a single reference annotation for model learning and evaluation.

When two or more raters independently assign scores to images (e.g., physicians evaluating whether or not lesions are visible on an ultrasound image), Inter-Rater Reliability (IRR) can be used to quantify the degree of agreement between raters. Based on how raters are assigned to images and the type of assessment (e.g., nominal, ordinal, interval, ratio, or categorical data), appropriate statistics are chosen.

In Strategy 1, images are evaluated by three or more raters and are not always assigned to the same set of raters (as described in the center panel of the Table 2.4). These designs -in which raters and images are neither fully crossed nor nested- are referred to as ill-structured measurement designs (ISMDs). In [Putka, 2008], the authors explain that common reliability estimators (e.g., Pearson correlations, intra-class correlations) are problematic when used in ISMDs and propose an alternative reliability estimator between ranging from 0 (worst) to 1 (best) $G(q,k)$:

$$G(q,k) = \frac{\sigma_T^2}{\sigma_T^2 + (q \cdot \sigma_R^2 + \frac{\sigma_{TR,e}^2}{\hat{k}})} \tag{2.1}$$

where $\sigma_T^2, \sigma_R^2, \sigma_{TR,e}^2$ are the variance components after fitting a random effects model with images and raters treated as crossed random factors. T is the ratee main effects, R is the rater main effects, and TR,e is the combination of the Ratee $\times$ Rater interaction and residual effects. q is a multiplier that scales the contribution of variance attributable to rater main effects:

$$q = \frac{1}{\hat{k}} - \frac{\sum_i \sum_{i'} \frac{c_{i,i'}}{k_i \cdot k_i'}}{N_t \cdot (N_t - 1)} \tag{2.2}$$

$\hat{k}$ is the harmonic mean number of raters per image; N_t is the total number of images; $c_{i,i'}$ is the number of raters that each pair of images (i,i') share; and k_i and k_i' are the number of raters who rated images i and i'. To construct the samples for which we wish to compute the reliability score, we consider the labels associated to all images in the test set (Strategy 1). Three labels are allowed per organ: normal, abnormal and absent. Because the absent class is predominant, we select only the images where the organ is considered present by the consensus and compute the reliability score per organ.

Table 2.5 highlights the reliability estimators of raters in ill-structured measurement design for abnormality classification per organ. Rater main effect variance component σ_R^2 is surprisingly low, suggesting that contribution of rater main effects has little impact on observed score variance.

	Variance coponent			Harmonic mean number of raters per image	Multiplier	Reliability score	Support
	σ_T^2	σ_R^2	$\sigma_{TR,e}^2$	$\hat{k}$	q	$G(q,\hat{k})$	n_{images}
Liver	0.120	0.005	0.097	2.814	0.199	0.771	737
Kidneys	0.127	0.001	0.049	2.826	0.195	0.870	405
Spleen	0.090	0.004	0.072	2.202	0.297	0.726	183
Gallbladder	0.160	0.000	0.045	2.259	0.282	0.890	175

Note: $\hat{k}$ is the harmonic mean number of raters per image, $\sigma_T^2, \sigma_R^2, \sigma_{TR,e}^2$ are the variance components for fitting a random effects model with images and raters treated as crossed random factors, q is a multiplier that scales the contribution of variance attributable to rater main effects and $G(q,\hat{k})$ the reliability score.

Tab. 2.5.: Reliability estimator of raters in ill-structured measurement design for abnormality classification per organ in the test set.

Assuming we consider only the subset of images where the number of raters is constant and equal to three, we can compute a second reliability score, the Fleiss kappa coefficient [Fleiss, 1971]. This coefficient is designed specifically for cross designs in which each image is assigned to a constant number n of raters randomly sampled from a larger population. In this case, each image is rated by a different sample of n raters. The advantage of this coefficient, when used appropriately, is that it provides a reliability score per class, which in turn informs more about the distribution of noise per class that occurs in the set of Strategy 3 where each image is randomly assigned to a single rater. Results are shown in Table 2.6. One can see that depending on the organ, the reliability score changes differently per class (i.e., normal, abnormal, or absent). For instance, the gallbladder is the only organ where the class *absent* does not have the highest reliability score. This is not surprising, as the gallbladder can be difficult to visualize when the patient is not on a completely fasted state. As for the liver, the most visualized organ during an abdominal ultrasound, the reliability score is at its lowest for all classes, which may pose a real difficulty during training as roughly six out of ten images in our dataset show this organ.

2.4 Radiological Reports Analysis

The analysis of radiological reports associated with ultrasound examinations is a valuable tool. First, it allows to have a more precise understanding of the attributes of the dataset (e.g., the most frequently examined organs, the most cited conditions, etc.) and thus to better target the research question. Second, it potentially allows to have a global label at the examination level almost at no additionnal cost. The following section is organized into three subsections:

1. Medical Concept Mining: We show how the Unified Medical Language System® (UMLS®) can be used to detect medical concepts in free-text reports and flag the most common ones.

2. Hard-coded Tagging Tool: Given a report and an anatomic region of interest (e.g., liver), we can easily extract a list of sentences that mention that region, and then determine if the region is mentioned in the context of an abnormal finding.

3. Report Model: Rather than using hard-coded rules to classify radiological reports, we analyze whether training a language model can better grasp the complexity of radiological reports.

		Kappa	Z
Liver	Normal	0.498	17.479
	Abnormal	0.531	18.617
	Absent	0.737	25.843
Kidneys	Normal	0.738	25.881
	Abnormal	0.64	22.441
	Absent	0.862	30.244
Spleen	Normal	0.685	24.014
	Abnormal	0.61	21.39
	Absent	0.769	26.959
Gallbladder	Normal	0.646	22.653
	Abnormal	0.81	28.417
	Absent	0.749	26.271

Tab. 2.6.: Fleiss' Kappa and Z-value for 3 Raters with 410 images.

To compare the hard-coded tagging tool to the machine learning report model and assess their performances, we asked the expert's panel presented in Section 2.2.1 to annotate 197 radiological reports according to the labels presented in Figure 2.7.

2.4.1 Medical Concept Mining

Entity	Complete Report (#130616 entities)	Conclusion Report (#20039 entities)
Body Part, Organ, or Organ Component	39.76 %	28.52%
Disease or Syndrome	12.18 %	16.42%
Finding	11.39%	13.59%
Pathologic Function	8.53%	10.34%
Sign or Symptom	5.45%	5.90%
Diagnostic Procedure	5.09%	4.07%
Other*	17.59%	21.15%

Note: Other included the following medical concepts : 'Tissue', 'Body Location or Region', 'Anatomical Abnormality', 'Intellectual Product', 'Congenital Abnormality', 'Body Substance', 'Neoplastic Process', 'Therapeutic or Preventive Procedure', 'Health Care Activity', 'Acquired Abnormality', 'Pharmacologic Substance', 'Mental Process', 'Mental or Behavioral Dysfunction', 'Phenomenon or Process', 'Medical Device', 'Laboratory Procedure', 'Indicator, Reagent, or Diagnostic Aid', 'Body Space or Junction', 'Biomedical Occupation or Discipline', 'Organism Function', 'Biomedical or Dental Material', 'Immunologic Factor', 'Physiologic Function', 'Body System', 'Injury or Poisoning', 'Clinical Attribute', 'Biologically Active Substance', 'Antibiotic', 'Inorganic Chemical', 'Enzyme', 'Classification', 'Hormone', 'Laboratory or Test Result', 'Organic Chemical', 'Research Activity', 'Amino Acid, Peptide, or Protein', 'Hazardous or Poisonous Substance'.

Tab. 2.7.: Distribution of found semantic types (as defined by the Unified Medical Language System) in all radiological reports of our dataset.

The UMLS is a metasaurus that unifies the concepts of several dozen terminologies in the biomedical field. Each concept in the UMLS is assigned a unique concept identifier (CUI), a set of terms (or synonyms), possibly in multiple languages, and a semantic type. We make use of QuickUMLS [Soldaini, 2016], which is a tool for fast, unsupervised biomedical concept extraction from medical text, that works for multiple languages, including French. This tool can be incorporated in a larger framework, medspaCy, a library of tools for performing clinical Natural Language Processing and text processing tasks with the popular spaCy framework. Table 2.7 reports the distribution of found semantic types (as defined by the Unified Medical Language System) in all the radiological reports of our dataset. One can see that the most common semantic type is *Body Part, Organ, or Organ Component*, followed by *Disease or Syndrome* and *Finding* with nearly the same frequency. These semantic types are highly relevant as they provide information regarding the inclusion of a specific anatomical region in the abdominal examination and facilitate the matching of an organ to an abnormal finding (if they are mentioned in the same sentence, for example). Other frequent semantic types include *Pathologic Function, Sign or Symptom, and Diagnostic Procedure*.

2.4.2 Hard-coded Tagging Tool

The mention of an abnormality in the same sentence as an anatomical region does not necessarily imply the existence of an abnormality in that region, and in fact the majority

of mentions of abnormality falls in the context of a negative result. With ultrasound examinations, often used as a screening tool, the physician will most likely say "no evidence of focal lesions in the liver" instead of "homogeneous liver". This shortcoming can be addressed by using approaches that depend on negation and uncertainty words (e.g., "not", "no", "unlikely") to classify mentions as neutral or negative. These words may be different depending on the context, for instance in the medical field, specific words are used to nuance the findings of the report. Hence, it is important to use a tool like medspaCy, which specializes in medical texts. Unfortunately, negation detection on medspaCy is only possible in English. For this reason, we added another component: EDS-NLP that provides a set of spaCy components used to extract information from clinical notes written in French. One of which is a rule-based negation detector, that we use to detect if the concepts recognized by QuickUMLS are referred to in a negative form.

Labelling Function

Given a free text medical report, we first divide it into sentences. For each sentence, we look for a word related to the semantic type "Body part, organ, or organ component" and check whether its unique concept identifier refers to the four organs of interest (i.e., liver, spleen, kidney, and gallbladder). In such case, the sentence is added to the list of sentences assigned to the organ. If a specific organ is not mentioned in the entire report, we consider that the organ was not examined during the ultrasound. Next, for each organ, we search for mentions of abnormalities, and consider as abnormal any concept related to the following semantic types: *Disease or Syndrome, Finding, Pathological Function, Congenital Abnormality, Anatomical Abnormality, Acquired Abnormality, Neoplastic Process, Injury or Poisoning*. If at least one abnormality is listed in non-negative form, the organ is considered abnormal. Table 2.8 shows the most common positive medical concepts detected in sentences linked to the four organs of interest.

Results

Table 2.9 shows the performance of the labeling tool based on the test set. This approach generates several false negatives. In fact, for an organ to be considered abnormal, the abnormality must be mentioned in the same sentence as the one mentioning the organ. As an example, in the following paragraph: "*the liver is homogeneous and of normal volume. We also note the presence of a suspicious image.*", the liver contains a suspicious image, but since the word "liver" is not explicitly mentioned in the second sentence, the labeling tool will classify the liver as normal.

Figure 2.8 displays the output of the labeling tool when run on a medical report sampled from our dataset, one can see that most medical concepts were detected, but few are

Liver

Entity	Count	Percentage
Kyste	206	17.87
Angiomes	148	12.84
Nodule	96	8.33
Steatose hepatique	84	7.29
chronique	83	7.20
Formations	65	5.64
Kyste biliaire	61	5.29
Splenomegale	51	4.42
calcifications	48	4.16
Hypertension portale	42	3.64
Fibrose	31	2.69
DIlatation	31	2.69
cirrhose	28	2.43
abces	23	1.99
Masses	21	1.82
hepatite	17	1.47
Cholestase	16	1.39
cholecystite	13	1.13

Kidneys

Entity	Count	Percentage
Kyste	807	52.75
DIlatation	180	11.76
Formations	120	7.84
angiomyolipomes	69	4.51
Hypotonie	48	3.14
Masses	37	2.42
hydronephrose	35	2.29
calcule	32	2.09
calcifications	31	2.03
insuffisance renale a	18	1.18
Nodule	15	0.98
Splenomegale	13	0.85
obstructif	12	0.78
Syndrome	12	0.78
Pyelonephrites	12	0.78
insuffisante renale chronique	12	0.78
chronique	10	0.65
lithiases	10	0.65

Gallbladder

Entity	Count	Percentage
calcule	85	43.37
Formations	23	11.73
lithiases	21	10.71
cholecystite	18	9.18
DIlatation	11	5.61
polype	9	4.59
obstructif	5	2.55
Masses	4	2.04
Nodule	4	2.04
Syndrome	2	1.02
hepatite	2	1.02
calcifications	2	1.02
signe de Murphy positif	2	1.02
Lithiases vesiculaires	2	1.02
cirrhose	1	0.51
Kyste	1	0.51
carcinome	1	0.51
Splenomegale	1	0.51
Insuffisance hepato	1	0.51
Cholestase	1	0.51

Spleen

Entity	Count	Percentage
Hypertension portale	17	13.49
Splenomegale	22	17.46
Formations	4	3.17
Kyste	29	23.02
Angiomes	4	3.17
Syndrome	4	3.17
Masses	5	3.97
Infarctus	4	3.17
Nodule	12	9.52
calcifications	9	7.14
Cytolyse hepatique	1	0.79
insuffisante renale chronique	1	0.79
Microcalcifications	5	3.97
nodule du foie	1	0.79
maladie	1	0.79
DIlatation	2	1.59
Fibrose	2	1.59
calcule	1	0.79
Hypotonie	1	0.79
polype	1	0.79

Tab. 2.8.: Most common positive medical concepts detected in sentences linked to the four organs of interest

still missing (highlighted in gray). In addition, some concepts are very tricky, such as "absence d'épanchement" which translates to "absence of effusion." Here, the labeling tool identifies "Absence" as a sign or symptom and "effusion" as a pathological function, instead of considering "Absence of effusion" as a whole. This implies that instead of considering only "effusion" as a pathological function that is not present, the world "Absence" is considered independently as a positive sign or symptom by the negation detector, since the word "absence" is used in an affirmative form.

2.4.3 Report Model

The hard-coded tagging tool presented in Section 2.4.2 relies on language rules derived from a medical concept database, and hence no training set is required for this method to work. As a result, the hard-coded tagging tool may be more error-prone and not generalize as well as newer language models trained on hundreds of thousands of

Fig. 2.8.: Output of the labeling tool when run on a report sampled from our dataset. Words highlighted in blue correspond to anatomical regions, words highlighted in green represent entities of interest referenced as negatives, words highlighted in red represent entities of interest referenced as positives, and words highlighted in gray are entities that the labeling tool failed to detect.

unlabeled text corpora. Such models are trained to learn meaningful representations of words/phrases in a corpus. A key benefit is that these same models can be refined for other tasks for which little labeled data is available. In the following, we detail how a deep learning language model can be trained to label radiological reports.

Training Set

Thanks to the annotation of the images presented in Section 2.2.1, we can use the images labels associated to the examination to construct labels at the examination level (as opposed to the image level). If at least one organ in one of the images of the examination is labeled as "abnormal", we assign this as a global label to the examination. Similarly, if the organ has not been detected in any of the images, we assign the label 0 to both concepts related to the organ (i.e., normal/abnormal). A total of 905 examinations forms the training set, with 388 and 416 normal and abnormal examples for the liver, 354 and 184 normal and abnormal examples for the gallbladder, 523 and 350 normal and abnormal examples for the kidneys, and 497 and 149 normal and abnormal examples for

		Precision	Recall	F1-score	support
Liver	Normal	74.59%	90.10%	81.61%	101
	Abnormal	68.00%	30.91%	42.50%	55
Gallbladder	Normal	89.91%	83.05%	86.34%	118
	Abnormal	33.33%	6.25%	10.53%	16
Kidneys	Normal	91.41%	97.39%	94.30%	153
	Abnormal	88.89%	57.14%	69.57%	28
Spleen	Normal	88.65%	99.21%	93.63%	126
	Abnormal	100.00%	15.38%	26.67%	26
micro avg		85.64%	80.42%	82.95%	623
macro avg		79.35%	59.93%	63.14%	623
weighted avg		84.53%	80.42%	79.94%	623
samples avg		83.11%	78.60%	80.09%	623

Tab. 2.9.: Performance of the *labeling tool* based on a set of annotated radiological reports.

the spleen, respectively.

Model

We use the *CamemBERT* [Martin, 2019] model for our report model. *CamemBERT* is a state-of-the-art language model pre-trained on a French corpus *OSCAR*, based on the *RoBERTa* [Liu, 2019b] architecture. *CamemBERT* can be fine-tuned with a multiple-choice classification head on top (a linear layer on top of the pooled output and a softmax) for multi-label classification. The model takes as input the entire report (with a maximum number of tokens set to 200) and outputs a probability vector for the presence or absence of each class. The sentences of the report are then divided into a list of indivisible units-or tokens- (e.g., a sequence of a few characters that constitute all or part of a word). Typically, the partition of these words is optimized on the training corpus and forms the vocabulary of the model, however since this model was pre-trained on generic text, it is essential to consider words that are specific to the domain on which we want to refine the model (abdominal ultrasound radiological reports). To do so, we re-train a word-piece Tokenizer to find the set of words that minimize the number of tokens needed to reconstruct the reports in our training set. We fine-tune the model for 35 epochs, with a learning rate of $1e{-}04$ and batch size of 16 using Adam optimizer. Figure 2.9 provides an overview of the framework.

Results

Table 2.10 shows the performance of the report model based on the test set. Note that the recall has increased with a macro-average of 73% compared to 60% for the labeling tool (Table 2.9) while the precision has decreased from 79.35% to 67%. This suggests that the model reduces the number of false negatives, but increases the number of false

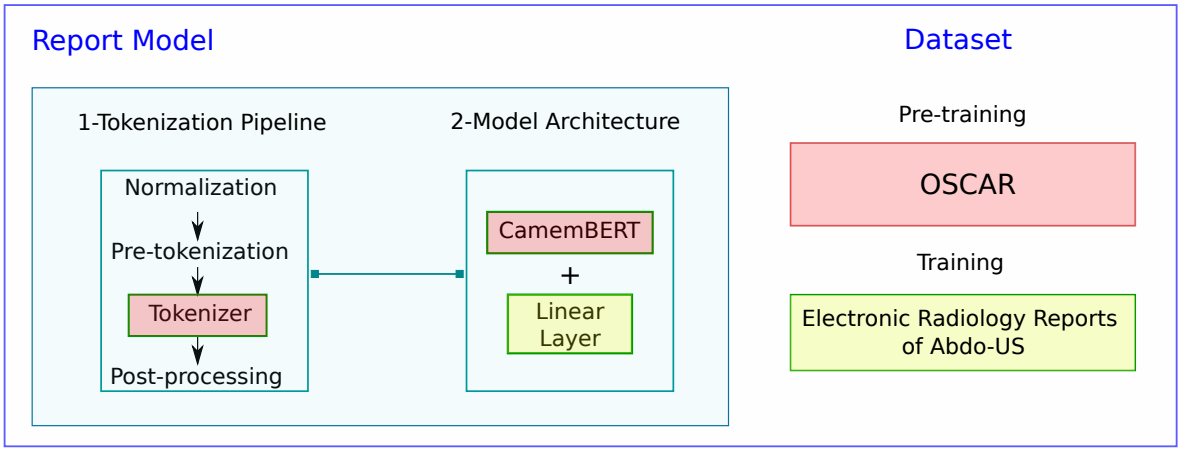


Fig. 2.9.: The report model uses *CamemBERT* a model pre-trained on a French corpus *OSCAR*, based on the *RoBERTa* [Liu, 2019b] architecture. We re-train the word-piece Tokenizer (highlighted in red) on the radiological reports dataset and train the CamemBERT (highlighted in red) for multi-label classification by adding a linear layer (highlighted in yellow) using the training set presented in Section 2.4.3.

		Precision	Recall	F1-score	support
Liver	Normal	54.00%	35.00%	43.00%	101
	Abnormal	39.00%	82.00%	53.00%	55
Gallbladder	Normal	89.00%	86.00%	87.00%	118
	Abnormal	48.00%	81.00%	60.00%	16
Kidneys	Normal	90.00%	90.00%	90.00%	153
	Abnormal	62.00%	75.00%	68.00%	28
Spleen	Normal	80.00%	92.00%	86.00%	126
	Abnormal	75.00%	44.00%	56.00%	26
micro avg		72.00%	77.00%	74.00%	623
macro avg		67.00%	73.00%	68.00%	623
weighted avg		74.00%	77.00%	74.00%	623
samples avg		68.00%	73.00%	69.00%	623

Tab. 2.10.: Performance of the *report model* based on a set of annotated radiological reports.

positives. Overall, the macro-average F1-score is 68% for the reporting model compared to 63.14% for the labeling tool. This metric allows us to obtain an unweighted average of the F1-score for each class and is therefore not sensitive to class imbalance. Similarly, looking at the F1-score helps to assess the performance of the model in terms of precision and recall by taking their harmonic mean.

2.5 Discussion

In this chapter, we detailed the challenges related to the construction of one of the largest abdominal ultrasound datasets. First, we compared study design choices for the

construction of task-based vs general purpose datasets. Specifically, we presented the data extraction methods, the selection of labels of interest, and the annotation strategies adopted to efficiently annotate the datasets. Since no exclusion criteria was used for the general-purpose dataset, an additionnal analysis of the retrieved data was needed. We investigated the two components of an ultrasound examination: the ultrasound images and the reports related to these exams.

For the images, we first showed the high variability of the image contents and the relevance of a pre-processing phase to harmonize their content, which will be then addressed in Chapter 3 [1]. Second, we have shown that there are certain images, and sometimes the entire examination, that are not applicable to the study, and therefore should be removed from the dataset. The relevance of deep clustering for this task will be detailed in Chapter 5 [2]. Finally, we investigated the inter-expert variability of annotators and the challenges this posed for training a machine learning model, in particular the noise that a single annotation implies. Notably, we saw that the variability of annotations for the liver was greater than that of other organs. This motivated the choice made in Chapter 4 [3] on the detection of focal liver lesions, where a clean task-based dataset was used.

For the reports, we investigated the value of text processing tools to provide global labels for examinations whose images were not annotated due to lack of time and resources. Our study showed that the use of hard-coded rules led to many false negatives, which is problematic. Indeed, as pointed out in the earlier sections, the aim of our study is to help non-expert caregivers to sort out patients and to detect those who need an additional examination by experts, so it is necessary to have as little false negatives as possible. The presence of false positives is less problematic, since ultrasound is a non-invasive modality, and performing an additional examination by an expert is safe. On the other hand, training more advanced language models has improved the recall of abnormal classes and the overall macro-average of the F1 score (which gives the same contribution to each class regardless of its prevalence).

These results, although encouraging, do not enable the direct use of the reports to annotate the images, firstly because there are still classification errors that would increase the noise level in the training dataset, and secondly because we can only partially link the reports to the images. An examination where the liver is considered abnormal for example, does not necessarily imply that all the images of the liver are abnormal. In fact, it is uncommon to find such a situation. Often, the abnormality is only visible on a precise plane and localized on a portion of the organ. Similarly, recent multi-modal

[1] Combining Bayesian And Deep Learning Methods For The Delineation Of The Fan In Ultrasound Images
[2] Deep Clustering for Abdominal Organ Classification in US imaging
[3] Detection, Localization, and Characterization of Focal Liver Lesions in Abdominal US with Deep Learning

learning methods [Zhang, 2020; Radford, 2021] do not apply directly to partially paired data.

2.6 Conclusion

In this chapter we described the result of a collaboration between several national organizations, including the Assistance Publique des Hôpitaux de Paris and the Health Data Hub, as well as volunteer caregivers in NHance from several different hospitals. This partnership led to the construction of a large abdominal ultrasound dataset called Abdo-US, which includes labels annotated by expert caregivers, noisy labels produced by language processing models, and finally standard reference evaluation sets labeled by physicians. This dataset forms the basis of the research conducted in this thesis. Thanks to the analysis performed in this chapter, we were able to direct our work towards the challenges that seemed most pressing to us and that will be discussed in the following chapters. However, this dataset can also serve for a more general purpose and contribute to the development and validation of models of abdominal ultrasound interpretation.

3 Combining Bayesian And Deep Learning Methods For The Delineation Of The Fan In Ultrasound Images

Contents

Abstract
Ultrasound (US) images usually contain identifying information outside the ultrasound fan area and manual annotations placed by the sonographers during exams. For those images to be exploitable in a Deep Learning framework, one needs to first delineate the border of the fan which delimits the ultrasound fan area and then remove other annotations inside. We propose a parametric probabilistic approach for the first task. We make use of this method to generate a training data set with segmentation masks of the region of interest (ROI) and train a U-Net to perform the same task in a supervised way, thus considerably reducing computational time of the method, one hundred and sixty times faster. These images are then processed with existing inpainting methods to remove annotations present inside the fan area. To the best of our knowledge, this is the first parametric approach to quickly detect the fan in an ultrasound image without any other information than the image itself. This chapter was published in IEEE 18th International Symposium on Biomedical Imaging (ISBI) [Dadoun, 2021].

3.1 Introduction

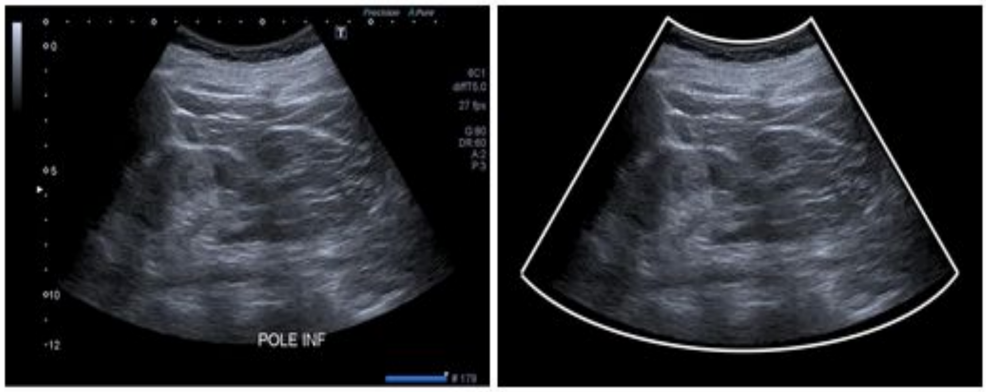


Fig. 3.1.: Left picture shows the original image. The US fan area is limited to a conic section, and several text and graphic elements are present. Right picture shows the result of our pre-processing. The white lines delimiting the US fan area are automatically detected, and all graphic and text elements are removed and replaced by a plausible intensity value.

Ultrasound (US) imaging is one of the most common techniques for medical diagnosis. According to the WHO, two thirds of the world's population do not have access to medical imaging, and ultrasound associated with X-ray could cover 90 % of these needs. The decrease in ultrasound hardware prices allows its diffusion, but limitations persist because acquiring and interpreting an ultrasound image is a difficult examiner-dependent task with few trained operators[Moore, 2011; Liebo, 2011; Choi, 2011]. Hence the importance of developing the research around the entire processing chain

in US imaging [Droste, 2020; Heuvel, 2019; Yap, 2017]. Interested readers may refer to [Sloun, 2019] for an overview of Deep Learning strategies for ultrasound-specific processing methods. Ultrasound imaging provides real-time anatomical information and allows the measurement of functional parameters. These measurements along with other annotations are manually added to the image during an exam to provide a complete report. In order to make them available to the research community following regulatory guidelines, they are usually converted to JPEG/PNG format, and processed to remove all metadata, including acquisition parameters. The presence of biometric measurements and other machine dependent characteristics may be challenging for the task of automated image analysis. First, because there is no guarantee that these annotations do not include sensitive information. Second, because we would like to make sure that they do not induce a bias during the training of a neural network for the task of classification, segmentation or detection. In [Zhang, 2017], the authors "blacked out" the identifying patient information on videos by setting the corresponding intensities of pixels that remained static throughout the entire clip to minimal intensity. Unfortunately, this method only works if we have access to the video sequence of the exam. As for the biometric measurements, [Baumgartner, 2017] removed all the annotations using the inpainting algorithm proposed in [Telea, 2004], but no information was given on how to create the inpainting masks. In this work we propose a pre-processing pipeline for ultrasound images to detect the ultrasound fan area combining Bayesian and Deep Learning methods and show how to apply existing inpainting methods to remove biometric measurements as shown in Fig 3.1.

3.2 Dataset

Data for this pilot study constitute a subset of US-Abdo et FLLs datasets presented in 2.2. We worked closely with volunteer physicians from the NHANCE NGO to construct an heterogeneous dataset that includes images from various manufacturers. The dataset was composed of 1280 images from which 1150 were used to train and validate the method. The 130 remaining images were used to evaluate the method by a trained engineer.

3.3 Detection of the Ultrasound Fan Area

We present a fully parametric method to detect the ultrasound fan area in the image and thus remove most of the annotations present outside. This is achieved by generating segmentation masks of the US region using a probabilistic approach.

3.3.1 Probabilistic Model of US Fan Area

Let I be our image of size $n \cdot m$ where n is the width and m the height of the image. We define $\theta = \{\epsilon_1, ..., \epsilon_{10}\}$ as the set of parameters describing the truncated cone Ω_θ modeling the US fan area (see Fig. 3.2). We seek to optimize θ for every input US image

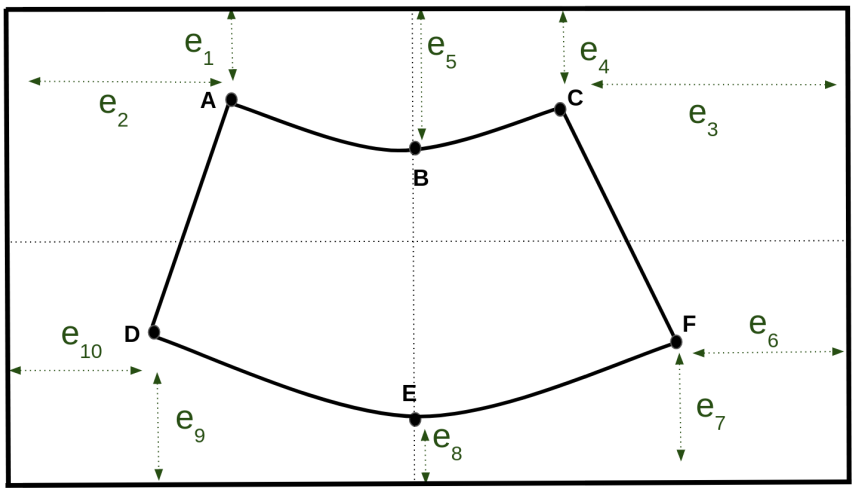


Fig. 3.2.: Parameterisation of the region of interest

via a probabilistic formulation. More precisely, we introduce the hidden binary variable $Z_i \in [0, 1]$, indicating whether voxel i belongs to the fan area Ω_θ. The θ parameters are equipped with a uniform prior and we propose to maximize the marginal log likelihood $p(\theta|I)$ as follows:

$$\log(p(\theta|I)) \propto \log\left(p(\theta)\right) + \log(p(I|\theta)) \tag{3.1}$$

$$\log(p(I|\theta)) = \sum_{i=0}^{n \cdot m} \log\left(p(I_i|Z_i = 1) \cdot p(Z_i = 1|\theta)\right. \tag{3.2}$$

$$\left. + \left(p(I_i|Z_i = 0)\right) \cdot \left(1 - p(Z_i = 1|\theta)\right)\right) \tag{3.3}$$

To model the distributions, we make use of intensity values extracted from bounding boxes of all training images. The distribution of a pixel intensity in the US fan $f_i = P(I_i|Z_i = 1)$ is captured by a mixture of two Gaussians (see Fig 3.3.b) whose parameters are estimated using bounding boxes of size 5×5 located in the center of images. As for the background distribution $b_i = P(I_i|Z_i = 0))$ it is modeled as a uniform distribution whose parameters are estimated using bounding boxes of size 2×2 located in the top left, top right, bottom left and bottom right corners of the image. $P(Z_i = 1|\theta)$ is the probability of voxel i to be inside the fan area Ω_θ which is defined in a closed form manner. Indeed, the truncated cone region Ω_θ is defined analytically as the set of points $\mathbf{x} = (x, y)$ for which $f_i(\mathbf{x}, \theta) \geq 0$ where f_1, f_2 are respectively the equations of the two straight lines AD, CF and f_3, f_4 are respectively the equations of the two parabolas through ABC, DEF as shown in Figure 3.2. From this piecewise analytic description,

we derive a soft implicit definition of the fan shape by summing up the four implicit functions $f_i(\mathbf{x}, \theta)$. The prior label probability is defined as Bernouilli distribution whose parameter depends on the sigmoid of the regularized implicit function:

$$p(Z_i = 1|\theta) = \sigma[\sum_{i=1}^{4} f_i(\mathbf{x}_i, \theta)] \tag{3.4}$$

Where $\mathbf{x}_i$ is the position of voxel i in the image.

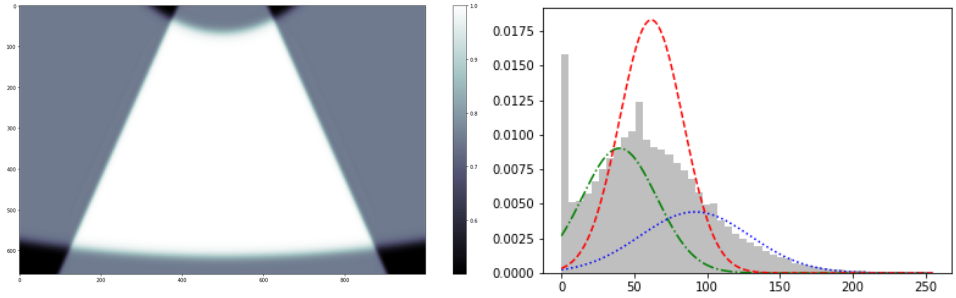


Fig. 3.3.: (Left) Prior label probability $p(Z_i = 1|\theta)$ parameterized by θ; (Right) Normalized histograms of the ROI distribution. In green/blue lines the two Gaussians and in red the mixture of the Gaussians.

3.3.2 Expectation-Maximization (E-M) Algorithm

Maximimizing the log joint probability is not easy since we have the log of sums. Instead, we use a lower bound which is much simpler to compute. More precisely, we replace this maximization:

$$\log p(I, \theta) = \log(p(\theta) + \log(p(I|\theta))$$

With this one :

$$\log p(I, \theta) - D_{\mathrm{KL}}(U||p(Z|I))$$

Where $U = \{U_n\}$ is a surrogate function for the posterior label probability $p(Z|I)$ and D_{KL} is the Kullback–Leibler divergence. This can easily be solved using the E-M algorithm.

- **E**-Step. Compute the posterior label probability for the current value of θ:

$$\begin{aligned} U_i &= p(Z_i = 1|I_i) \\ &= \frac{r_i p(Z_i = 1|\theta)}{r_i p(Z_i = 1|\theta) + (1 - r_i)(1 - p(Z_i = 1|\theta))} \end{aligned}$$

  where $r_i = p(I_i|Z_i = 1)/p(I_i|Z_i = 1) + p(I_i|Z_i = 0)$.

- **M**-Step. Find θ by maximizing the variational lower bound which is equivalent to minimizing the following expression with the current value of U :

$$\begin{aligned}\mathcal{L}(\theta) &= -D_{\mathrm{KL}}(U||p(Z|I)) + \log(p(\theta)) \\ &= \sum_{i=0}^{n \cdot m} -U_i \log(p(Z_i = 1|\theta)) \\ &+ (1 - U_i) \log(1 - p(Z_i = 1|\theta))) \\ &+ \log(p(\theta))\end{aligned}$$

During the M-step we use an optimization algorithm, Limited-memory BFGS algorithm (L-BFGS). This algorithm approximates the Broyden–Fletcher–Goldfarb–Shanno algorithm (BFGS) using a limited amount of computer memory [Byrd, 1995; Zhu, 1997].

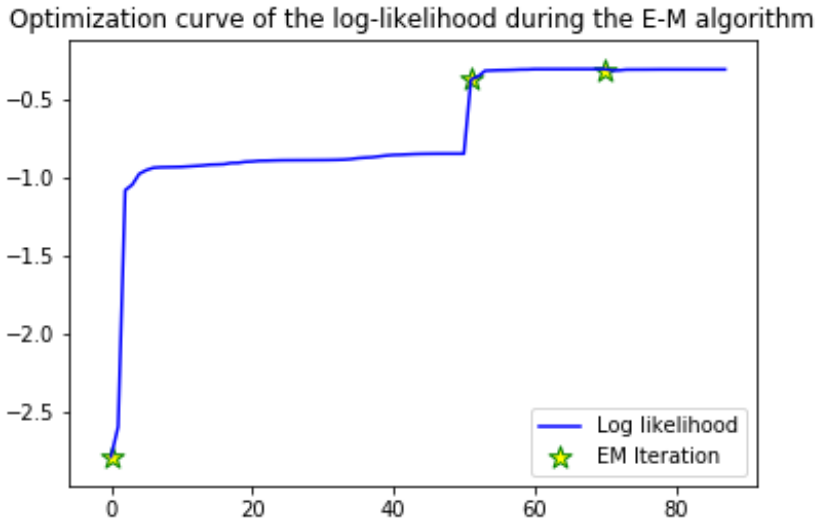


Fig. 3.4.: Log-likelihood optimization during the EM algorithm.

3.3.3 Reducing the Computational Time of the Method using Deep Learning

We use the segmentation masks generated by the method to train a neural network for the same task. This is done by training on CPU a simple U-Net on 70 % of our dataset. We validate the method on 20 % of our dataset and test it on the remaining 10 % . We use the Binary Cross Entropy (BCE) with logits loss. In inference, the processing of one frame takes approximately 0.45 seconds which is 160 times faster than the Bayesian method.

3.4 Implementation and Results

In this section we explicit the implementation details of our method and show the quantitative and qualitative results.

3.4.1 Implementation of the E-M Algorithm

We only optimize the Kullback-Leibler term during this step since we use a uniform prior as $p(\theta)$. Indeed, we do not have access to a preferred range of values for θ. We show in Figure 3.4 the optimization curve of the log-likelihood during the E-M algorithm when the prior on θ is far from the ground truth. We can see in Figure 3.5 that the algorithm is robust to the change of shape in the ultrasound fan area.

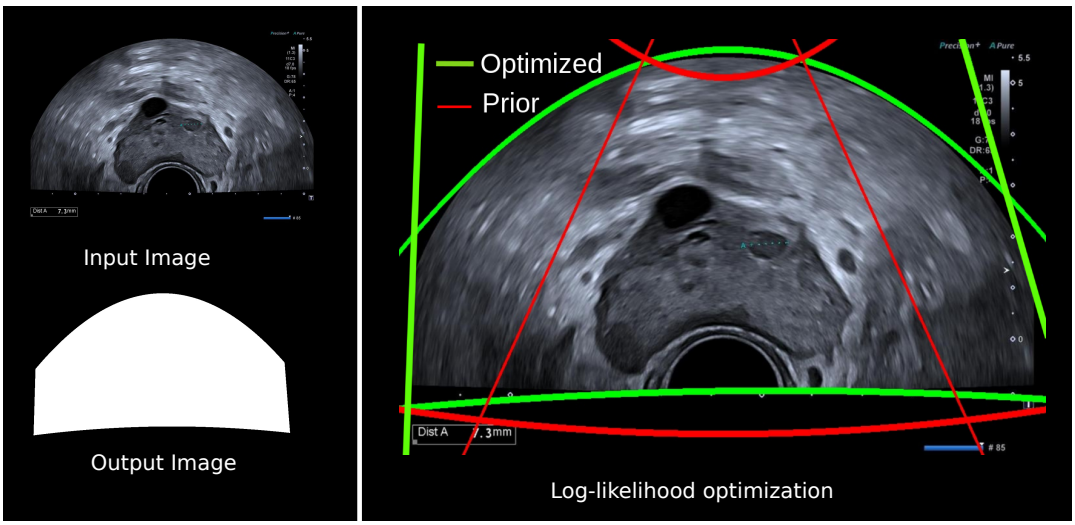


Fig. 3.5.: Example of a generated mask when the prior on the ultrasound fan area is very different from the truth.

3.4.2 Training Details of the U-Net

We use Adam Optimizer [Kingma, 2014] with learning rate 0.001 and a batch size of 1. After 10 epochs, we achieve a validation loss of 0.062 and Dice loss of 0.017.

3.4.3 Bayesian Method Compared to the U-Net

The Bayesian method developed in Section 3.3 is constrained by our piecewise analytic description. It provides a good approximation of the ground truth mask, but results in a rigid delineation of the ultrasound fan area. Whereas the masks generated by the U-Net

are less regularized and can therefore capture more information on the ultrasound fan area. An example of the two masks is shown in Figure 3.6.

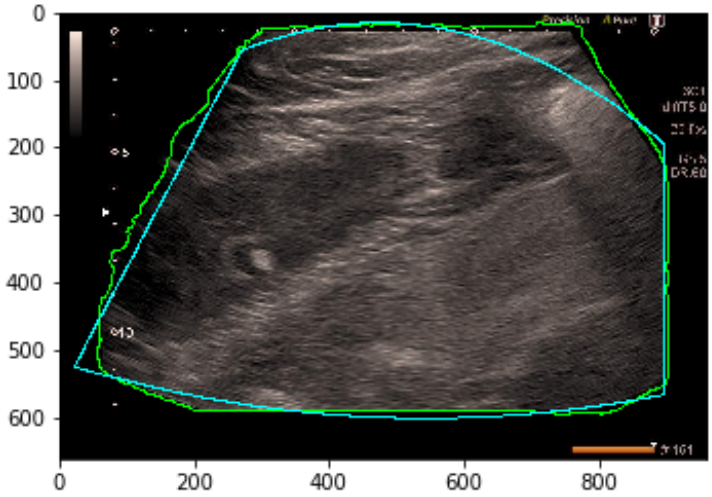


Fig. 3.6.: (Blue) Mask generated by the Bayesian method. (Green) Mask generated by the U-Net.

3.4.4 Evaluation of the Method

For the remaining 10 % of the dataset, which we had never used before, we asked a trained engineer to evaluate the results of the method using 3 labels:

- Perfect match between the ultrasound fan area and the detected area.
- Good detection when the area of the missing part is less than 1% of the image.
- Poor detection when the area of the missing part is more than 1% of the image.

We see in Table 3.1 that 90 images were labeled as a perfect match, 37 images were labeled as good detection, with mean area of mismatch <0.15 %. This part corresponds to the corners of the fan that were slightly cropped due to the parametric definition of the fan area. Yet those tiny errors on the fan margin have no impact on the interpretation of the image content. Finally 3 images were labeled as missing a relatively large part of the detected ultrasound fan area. This happens when a part of the ultrasound fan area is totally dark,the method then mixes up the background with the foreground. Examples are shown in Fig 3.7.

Label	# Images	Mean mismatch area
Perfect detection	90 (69.2 %)	0.00%
Good detection	37 (28.5 %)	0.15%
Total	127 (97.7 %)	0.05%
Poor detection	3 (2.3 %)	5.0%

Tab. 3.1.: Evaluation of the detection method on 130 images

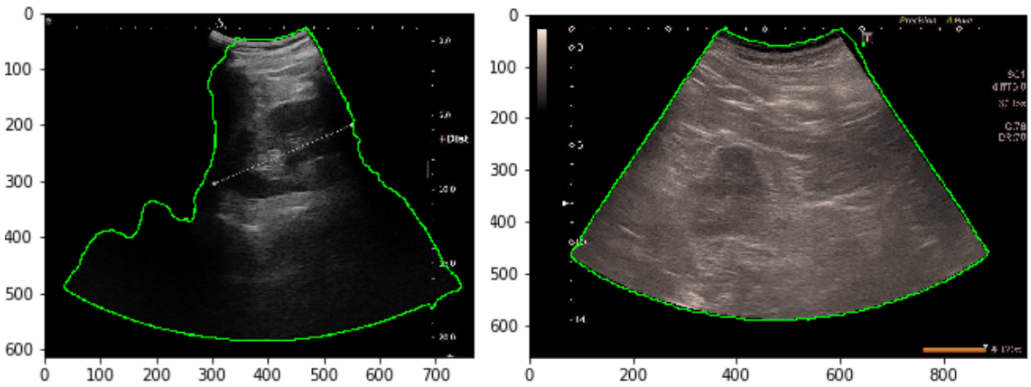


Fig. 3.7.: (Left) Example of the label 'Poor detection', a part of the fan is not detected because it is filled with low intensity pixels. (Right) Example of the label 'Good detection', the missing area corresponding to the cropped corners is barely visible to the naked eye.

3.5 Inpainting Images with Annotations Inside the Ultrasound Fan Area

Here we show how to use open CV's module inpaint to replace segments and annotations present on the cone by pixels of the background. More precisely we use in-paint Telea which is based on [Telea, 2004]. Values of pixels of the region to be inpainted are replaced by a weighted sum of neighboring pixels starting from the boundary. The challenge is to generate in a fully unsupervised way a mask of the region to be in-painted. This is done by maximizing the contrast of the image and masking all pixels below a threshold value. We also replace all colored pixels in the image with random shades of gray so that the inpainting algorithm doesn't use colored pixels present in the boundary. Finally we denoise the resulting image using non-local-means filtering [Buades, 2005]. The method uses small patches centered on pixels. The patch of interest is compared to other patches. Then it replaces the value of the pixel by the average intensity of pixels that have patches close to the current patch. An example is shown in Fig 3.8.

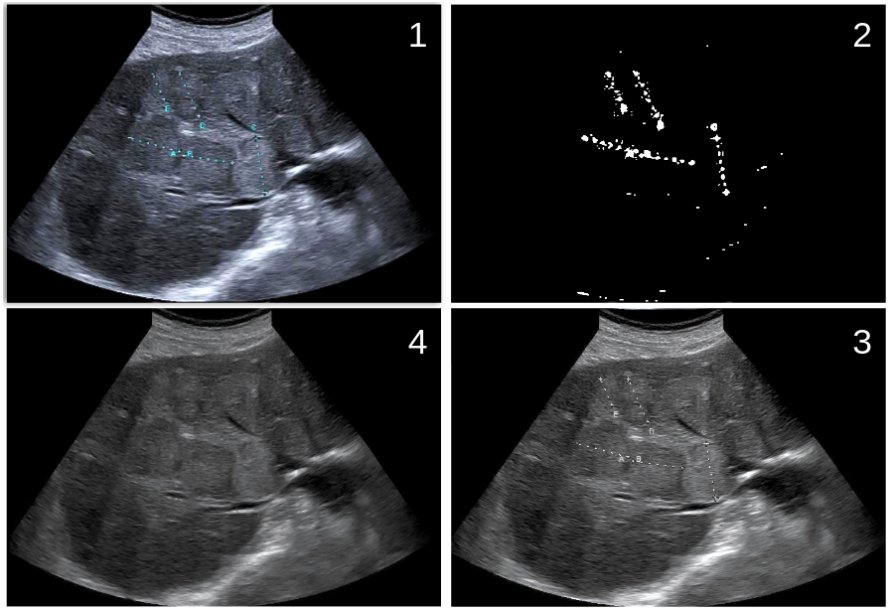


Fig. 3.8.: Pipeline to generate masks for the inpainting algorithm. We maximize the contrast of the input image (1) and mask all pixels below a threshold value (2). We also replace all colored pixels in the image with random shades of gray (3) so that the inpainting algorithm doesn't use colored pixels present in the boundary. Finally we apply the inpaiting algorithm and denoise the resulting image using non-local-means filtering(4).

3.6 Conclusion

We achieved our primary objective, namely the construction of an automated pipeline for ultrasound fan area detection. We have shown that this method is scalable by evaluating it on 130 varied images obtained from different machines and various shapes of the ultrasound fan area. The next step is to work closely with the Clinical Data Warehouse of Greater Paris University Hospitals, to assess the method on a larger dataset. The primary novelty of this method is the use of Bayesian statistics to generate training data for a Deep Learning application. We believe that this work is an important step for a larger and better exploitation of ultrasound images in the area of medical image analysis using Deep Learning. A possible improvement of the method is to replace the uninformative uniform $p(\theta)$ with a distribution estimated on the training set and include it in the M-step of the EM algorithm. The method could be further improved by adding a regularization term in the U-Net loss function. This would allow for more regular approximations of the ultrasound fan area. The tools developed in the framework of this project are available under an open-source license.

4

Detection, Localization, and Characterization of Focal Liver Lesions in Abdominal US with Deep Learning

Contents

Abstract

Through a retrospective, multicenter, institutional review board-approved study, two object detectors, Faster Region based Convolutional Neural Network (Faster-RCNN) and DEtection vision TRansformer (DETR) were fine-tuned on a dataset of 1026 patients (n = 2551 B-mode abdominal US images between 2014 and 2018) to detect liver parenchymal lesions on abdominal US images, localize focal liver lesions (FLs), and characterize them. Their performance was analyzed on a test set of 48 new patients (n = 155 B-mode abdominal US images between 2018 and 2019) and compared with three caregivers, one non-expert and two experts, blinded to clinical history. A sign test was used to statistically compare accuracy, specificity, sensitivity, and Positive Predictive Value between all raters. Results indicated that:

1. The vision transformer network, DEtection TRansformer, DETR showed higher performance for all tasks compared to the recurrent convolutional network Faster RCNN.

2. For the detection of lesions in the liver parenchyma, DETR met or exceeded the performances of two experts, with a specificity of 90% (95% CI : 75, 100) and a sensitivity of 97% (95% CI: 97, 97).

3. For the localization of focal liver lesions (FLLs) DETR showed comparable performances to that of two experts, with a positive predictive value (PPV) of 77% (95% CI: 70, 84) and a sensitivity of 84% (95% CI: 77, 89).

4. For the characterization of focal liver lesions (benign vs. malignant), DETR achieved a higher performance compared to all raters, with a specificity value of 81% (95% CI : 67-91) and a sensitivity value of 82% (95% CI: 62, 100).

This chapter was published in Radiology: Artificial Intelligence (2022) [Dadoun, 2022b].

4.1 Introduction

The accurate detection and assessment of focal liver lesions (FLLs) is a critical public health issue due to the incidence increase of primary hepatic malignant lesions. Liver cancer is the third leading cause of cancer-related deaths worldwide [Bray, 2018] with

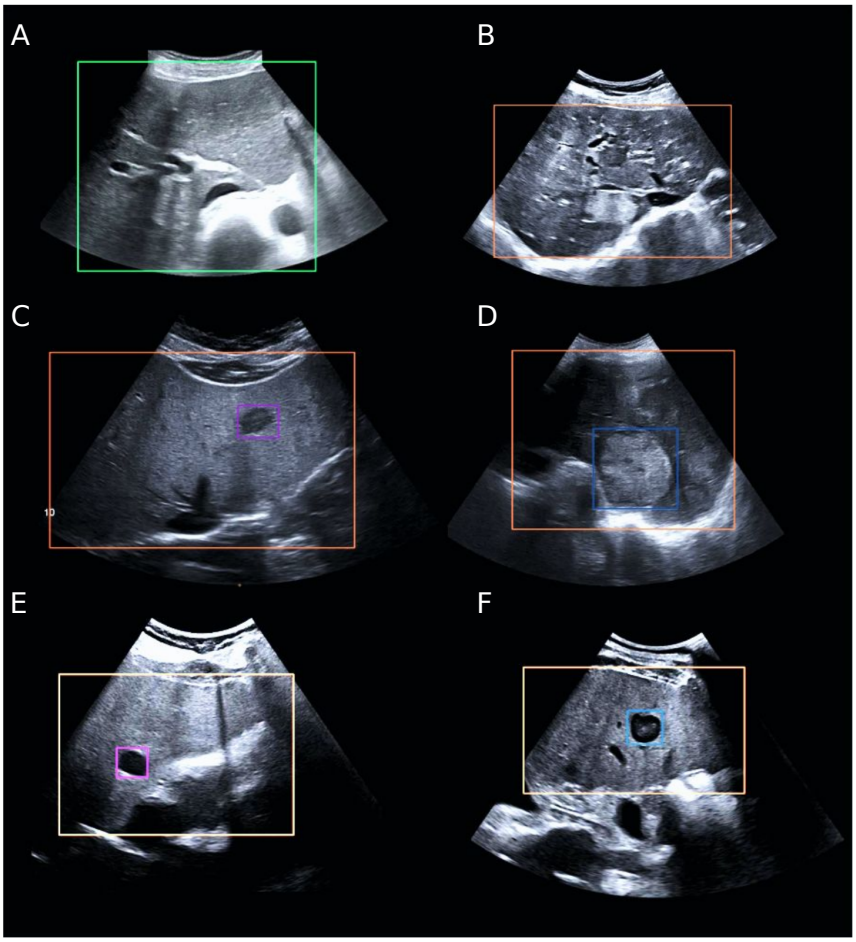


Fig. 4.1.: Upper left picture (A) shows a liver without lesions (highlighted by a green square), upper right picture (B) shows a liver with lesions (highlighted by an orange square). Middle left picture (C) shows a benign lesion - focal nodular hyperplasia- (highlighted by a purple small square) and on the right (D) a malignant lesion- hepatocellular carcinoma- (blue small square). In this example benign and malignant lesions have different texture and size. Bottom left picture (E) shows a benign lesion -cyst- (highlighted by a purple square). It has a circular shape and dark pixel intensities. Right picture (F) shows a malignant lesion -metastasis- (highlighted by a blue square) with similar characteristics. These images highlight the difficulty of malignant versus benign discrimination.

hepatocellular carcinoma (HCC) being the primary type affecting adults [Craig, 2020]. These lesions are usually discovered in patients with third-stage liver failure or other cancers such as colorectal cancer. They can also be found incidentally during abdominal imaging studies [Marrero, 2014]. Non-contrast enhanced US is one of the most commonly used modalities to investigate FLLs during the screening stage for high-risk patients. However, acquiring and interpreting an ultrasound image is a difficult and examiner-dependent task with a limited number of trained operators [Marrero, 2014]. This is particularly true in developing countries where healthcare providers identify lack of training as the main limitation to US use [Shah, 2015]. A computer-aided assistance tool could allow more early-stage malignant lesions to be detected, increase differential diag-

nosis and enable an efficient and cost-effective treatment [Cadier, 2017; Trinchet, 2009]. Overall, it could assist non-expert caregivers in performing an adequate assessment of the liver. Machine learning methods have shown promising results in earlier studies [Ta, 2018; Yao, 2018; Yang, 2020; Schmauch, 2019] for the diagnosis of FLLs in US images. A previous study classified malignant and benign lesions in 95 three-minute cine clips using automatically extracted B-mode and contrast-specific features on a support vector machine classifier [Ta, 2018]. It showed results comparable to those of experts with more than 15 years of experience. Another study showed improved performance when using sparse representation-based feature extraction methods on multi-modal US images on 111 patients [Yao, 2018]. A pre-trained convolutional network with an added attention module of the region of interest (ROI) was evaluated in 367 two-dimensional B-mode US images but assumed that only one subtype of lesions could be present in the liver [Schmauch, 2019]. The network achieved an area under the receiver operating characteristic (ROC) (AUC) score of 0.94 for FLL detection and 0.916 for FLL characterization over three-fold cross validations. Finally, Yang et al. [Yang, 2020] constructed a large multi-centric dataset of 24,343 B-mode US images along with radiomics signatures derived from FLLs and liver, ultrasonic features including posterior acoustic enhancement, echogenicity, shape of the fan, and clinical parameters. Diagnostic performances were verified using external validation and were compared with that of contrast-enhanced CT/ MRI and radiologists, with an AUC of 0.924 for classification of malignant from benign FLLs. Nonetheless, most of these previous studies used small datasets, with a large class imbalance for benign and malignant lesions (80% and 20%, respectively in [Yang, 2020]). To the best of our knowledge, none of these methods localized the lesions in the liver and gave a specific characterization for each lesion. Therefore, our study explores the use of deep learning-based networks for the detection, localization and characterization of focal liver lesions in non-contrast enhanced US imaging, to assist non-expert caregivers in performing a proper assessment of the screening test.

4.2 Methods

4.2.1 Study Design

IRB approval (blinded) was obtained for this retrospective, multicenter study, and informed consent was waived. Data for this pilot study were obtained in collaboration with a Clinical Data Warehouse. US exams were extracted from the picture archiving and communication system (PACS) of two university hospitals, Center 1 (Necker Hospital) and Center 2 (Saint Louis Hospital). Only adult patients were selected (age $\geq$ 18 years old). Patients with liver parenchyma containing lesions were included if they met the following criteria: 1- lesions were visible on the US images (decision made unanimously by an adjudication panel), 2 - patients did not receive previous local therapy

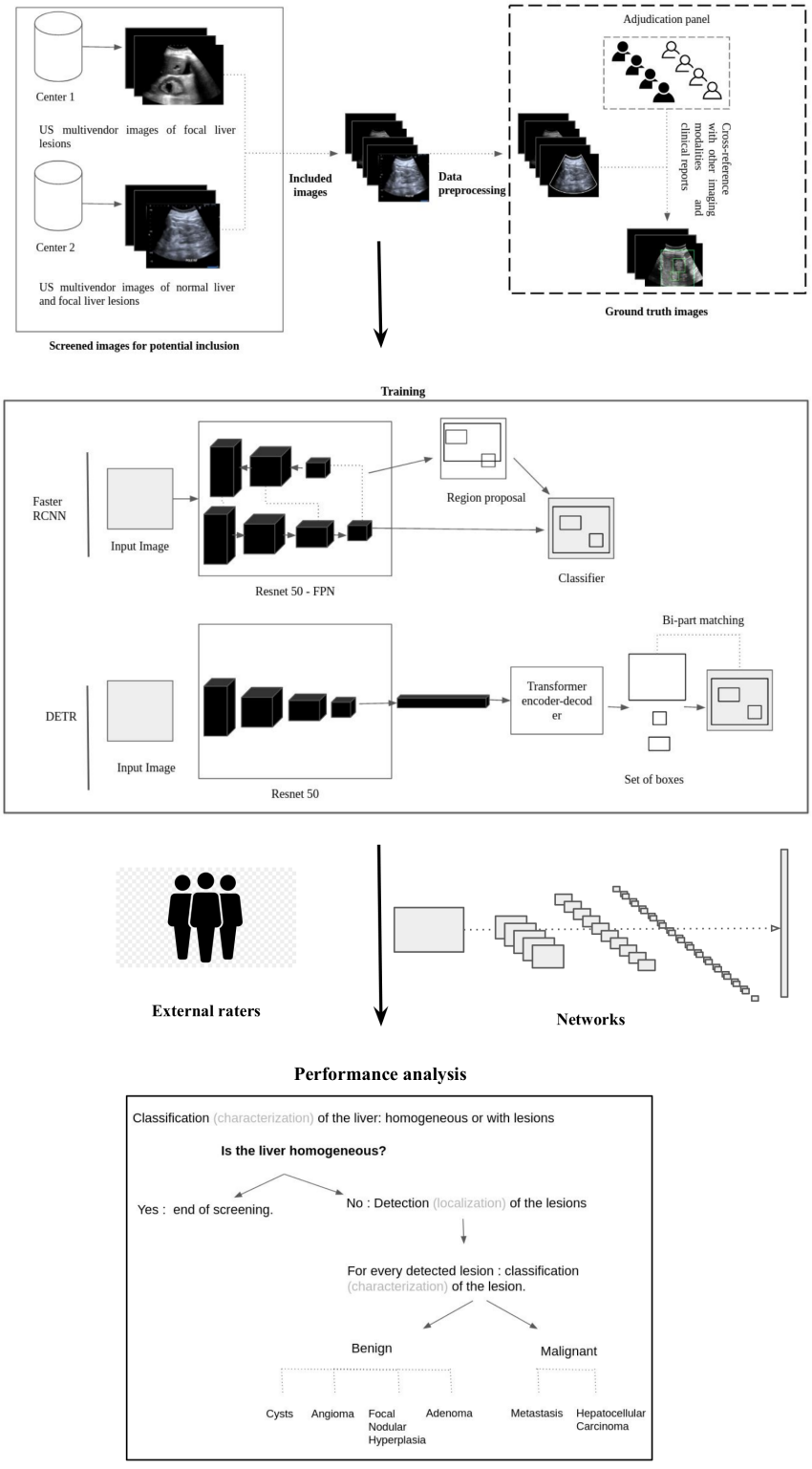


Fig. 4.2.: Workflow of the study. Note.—RCNN = recurrent convolutional neural network, DETR = Detection Transformer, FPN = Feature Pyramid Network.

and 3- a definite pathological diagnosis of lesions was obtained (“cyst,” “angioma,” “focal nodular hyperplasia”, “adenoma” , “metastasis” or “hepatocellular carcinoma.”). Patients

without lesions in the liver parenchyma were selected in case of a definite absence of pathological diagnosis. All patients' abdominal US examinations available in centers 1 and 2 between 2014 and 2019 were selected. For the training and development set, examinations completed between 2014 and 2018 were selected. For the test set, examinations completed in 2019 were selected, provided that the corresponding patients did not undergo any examination between 2014 and 2018. At an image level, extremely magnified images and images obtained with degraded mode simultaneously with CE-US images were excluded.

4.2.2 Data Acquisition

Data collected retrospectively in Center 1 consisted exclusively of multi-vendor US images with FLLs (four vendors). Data collected retrospectively in Center 2 consisted of multi-vendor images of healthy liver parenchyma and FLLs (two vendors).

4.2.3 Data Preprocessing

The Digital Imaging and Communications in Medicine (DICOM) red, green, blue images extracted from the PACS were converted to Joint Photographic Experts Group (JPEG) format. They were processed using a de-identification tool to remove identifying data and metadata. This tool uses the Dicom Ultrasound Attributes to remove the upper band of the image before converting it to JPEG format. A parametric method was used to detect the US fan area and remove annotations and other machine dependent characteristics outside this region. In addition, biometric measurements present inside the region were removed using existing inpainting methods (Figure 3.1) [Dadoun, 2021]. This enables networks to be trained on raw data with no manually added annotations by examiners and therefore may be more suitable for a real-time use when the examiner is not an expert.

4.2.4 Determination of the Ground Truth

For each center, the diagnosis associated with each US image was collected by two radiologists with more than 15 years of experience in US image interpretation, by cross-referring with other imaging modalities (contrast-enhanced US (CE-US), CT, MRI, biopsy when available) and clinical files. The final diagnosis was used for the characterization task (liver parenchyma and characterization of FLLs) and did not include the number of lesions in the US image, nor their localization. To determine the ground truth boxes for the localization task (i.e. boxes around the liver and FLLs), an adjudication panel was used as an external standard of reference. The panel consisted of four physicians, either radiologists or holders of a national diploma in US imaging, and four sonographers

holders of a national diploma in US imaging from six different health institutions with more than three years of experience. The annotators, who worked on a tailor-made annotation platform, were first asked to localize the liver, and classify it as "homogeneous liver" (i.e., without lesions) or as "liver with lesion (s)" with respect to the final diagnosis associated with the image. The selection of "liver with lesion (s)" led to the second task (i.e., localization of lesions). Each lesion had to be classified in accordance with the final diagnosis, with two possibilities: "benign lesion (s)" and/or "malignant lesion (s)." Benign lesions were further subclassified as "cyst," "angioma," "focal nodular hyperplasia" or "adenoma." Malignant lesions were further subclassified as "metastasis" or "hepatocellular carcinoma." Examples of such annotations are shown in Figure 4.1. Each image was annotated by two experts and at least once by a physician. During this phase, in case of a disagreement between two annotators for the localization task, four additional annotators analyzed the questionable picture. If the annotation of the image was not unanimous between the additional annotators, it was excluded from the study.

4.2.5 Data Partitions

A total of 1026 patients (n = 2551 images) met the inclusion criteria for the training and development set. This set was randomly split into two subsets containing roughly the same proportions for each class, 80% and 20% respectively for training and development. The objective of the development set is to choose the settings of the networks that achieve the best performance on images it has never seen. A total of 48 new patients (n = 155 images) met the inclusion criteria for the test set. All images and clinical reports were de-identified within the centers and no information on the demographics of the study population was retained.

4.2.6 Models

Two object detection networks were used. The former, DEtection TRansformer (DETR), is an end-to-end object detection vision transformer network [Dosovitskiy, 2020] and is more suitable for a real-time application [Carion, 2020]. The latter, Faster Recurrent Convolutional Neural Network (Faster-RCNN), is a two-stage object detection network that showed more robust performance on natural image datasets [Ren, 2015]. Figure 4.2 compares the workflows of both networks. The networks were trained for 100 epochs, after which the validation loss stopped decreasing and the networks started overfitting. Stochastic gradient descent with Nesterov momentum [Goyal, 2017] was used for optimization (Supplementary Appendix A.1), with a learning rate of 0.0048, a batch size of 4, a momentum of 0.9 and a weight decay of 0.0001. For the first epoch, a warm up schedule was applied [Sutskever, 2013]. For the rest of the training, the learning rate of

each parameter group was decreased by 0.9 once the number of epochs reached 75 and 90 epochs.

4.2.7 Data Augmentation

To improve the performances of the networks, a data augmentation strategy based on Google's Brain Team's bounding box augmentation policies was applied [Zoph, 2020]. The data augmentation scheme is based on a learned set of specific sub-policies that were proven to improve generalization performance (Supplementary Appendix A.2). These sub-policies consist of a list of intensity, geometric and bounding box operations - Rotation, Cutout, Sharpness, Translation, Contrast, Brightness, Solarization, Shear - applied to the image or to a specific object inside the bounding box. During training, one of those policies is randomly selected and then applied to the current image as presented in [Zoph, 2020]. The transformations described include those applied to the image with or without altering the bounding boxes coordinates and those applied to a specific object inside the bounding box. For each transformation, one can specify the probability of applying the transformation and the magnitude of the transformation. We used the set of policies described in version 3 of Google's Brain Team's bounding box augmentation code, except for policies altering with the pixel's intensity, hue and value.

4.2.8 Objective function

To overcome the limitations of the loss function originally used in Faster CNN, a different loss function was used, namely a ranking based loss introduced by [Oksuz, 2018] and based on a paper by the same authors [Oksuz, 2020] defining a new performance metric for object detection, the Localization Recall Precision (LRP) The LRP metric aims to assign an error value in the range [0, 1] considering the localization, recall and precision, which are implicitly included in the formulation and weighted by the number of predictions and annotations for the recall and precision, respectively. The loss function used in DETR is similar to the original loss in Faster CNN, with the exception that it uses an optimal bipartite matching between predictions and ground truths and a different localization loss (a linear combination of the L1 loss and the generalized Intersection over Union (IoU) loss [Rezatofighi, 2019]. DETR was designed to remove the need for many hand-designed components, thus the loss function was kept unchanged.

4.2.9 Model Evaluation

Performances of all raters (expert/non-expert caregivers and networks) were compared against the ground truth. The non-expert, an emergency doctor, had five years of experience with US imaging. The experts, a radiologist (Expert 1, S.B) and an advanced

practice sonographer (Expert 2, F.J), had a national degree in US imaging with nine and eight years of experience, respectively. All three caregivers and networks were blinded to clinical history and reports. Performances were evaluated in detail with regards to the capability to: 1- Detect liver parenchyma with FLLs, 2- Localize FLLs using bounding boxes, regardless of the label assigned to them, 3- Characterize the FLLs correctly localized by all raters into benign and malignant and 4-Characterize FLLs into benign and malignant subcategories. For each of these tasks and for all raters, the positive predictive value (PPV) (1) or the specificity (3), and the sensitivity (2) are reported. For the localization task, the F1-score (5) is reported as well. For binary classification (i.e Task 1 and 2), the accuracy (4) and the Matthews correlation coefficient (MCC) (6) metrics are reported as well. For multiclass classification, the macro specificity, sensitivity and accuracy are used instead. The formulas of all metrics are defined as follows:

1) PPV	$\frac{TP}{TP+FP}$	4) Accuracy	$\frac{TP+TN}{TP+FP+TN+FN}$
2) Sensitivity	$\frac{TP}{TP+FN}$	5) F1 score	$2 \cdot \frac{\text{PPV}\cdot\text{Sensitivity}}{\text{PPV}+\text{Sensitivity}}$
3) Specificity	$\frac{TN}{TN+FP}$	6) MCC	$\frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$

To report all these metrics a class-specific threshold was set to select the probabilities output by the networks, the threshold was defined as the value that maximizes the F1-score of each class during the validation step. In this setting, a prediction (box + label) is considered a true positive if it has an IoU score with a ground truth box higher than 0.3 and the same label as the ground truth box and false positive otherwise. The ground truth boxes with no intersections with predicted boxes are false negatives (FN).

4.2.10 Statistical Analysis

Because images from the same patient are not independent, bootstrapping experiments were used, as previously described in [Martin Bland, 1995]. The test set is resampled to contain only one randomly selected image per patient and the inference is repeated 1000 times. At each iteration the resampled test set had the same size as the patient population size (npatients= nimages = 48). For accuracy, specificity, sensitivity, PPV, F1 score and MCC values we report the mean across the observations, and 95% Confidence Intervals computed using the 2.5 and 97.5 percentiles of the ranked observations. The sign test is used to report the significant results on all metrics. Given multiple comparisons, the Bonferroni correction method was used to adjust p-values (19), $P < .05$ was considered to indicate a significant difference.

4.2.11 Data Availability

All code used for this study is publicly available.

4.3 Results

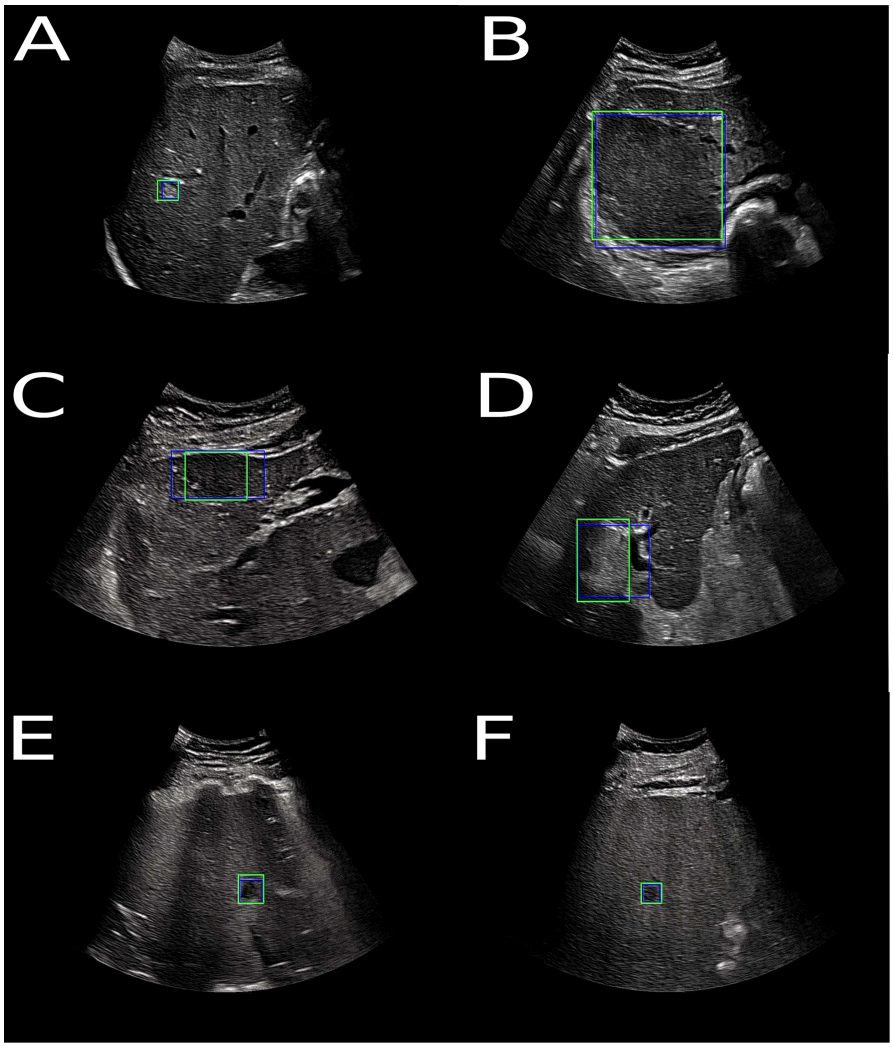


Fig. 4.3.: True positive findings- malignant and benign lesions correctly identified by the Transformer based network DETR for the FLL characterization task. Blue boxes correspond to the ground truth, green boxes are determined by the Transformer based network DETR. Upper left picture (A) shows a benign lesion - angioma-, upper right picture (B) shows a benign lesion - adenoma. Middle left picture (B) shows a benign lesion - focal nodular hyperplasia- and on the right (D) a malignant lesion- hepatocellular carcinoma-.Bottom left picture (E) shows a benign lesion -cyst- and bottom right (F) picture shows a malignant lesion -metastasis-. Note.—RCNN = recurrent convolutional neural network, DETR = Detection Transformer

	Accuracy	Specificity	Sensitivity	Matthews Correlation Coefficient
Non-Expert	78*†(70, 86)	80 (58, 100)	77*†(68, 84)	0.51*†(0.31, 0.68)
Expert 1	80*†(74, 86)	74 (50, 92)	82*†(76, 87)	0.51*†(0.31 − 0.68)
Expert 2	99 (98, 100)	98 (92, 100)	100 (100, 100)	0.98(0.95, 1.00)
Faster RCNN	93 (88, 98)	71 (50, 92)	100 (100, 100)	0.81(0.66, 0.95)
DETR	96 (92, 98)	90(75, 100)	97 (97, 97)	0.88(0.77, 0.95)

Notes-Detection of liver parenchyma with (37 patients : $n_{\text{images}} = 100$) or without focal liver lesion (11 patients : $n_{\text{images}} = 55$). Reported results were computed on the bootstrapped test set. Each subset contained only one image per patient. Intervals in parentheses are 95% Confidence Intervals. ⋆P $\leq .05$ in comparison with DETR. † $P \leq .05$ in comparison with Expert 2. RCNN = recurrent convolutional neural network, DETR = Detection TRansformer.

Tab. 4.1.: Detection of Liver Parenchyma With or Without Focal Liver Lesions (FLLs) Performance

4.3.1 Detection of Lesions in the Liver Parenchyma

- Table 4.1 shows detection performance of liver parenchyma with or without FLLs in terms of accuracy, specificity, sensitivity, and MCC. The non-expert caregiver achieved an accuracy score of 78% (95% CI: 70, 86), Expert 1 performed slightly higher with a score of 80% (95% CI: 74, 86) and Expert 2 had the best accuracy score overall, with 99% (95% CI: 98, 100). Faster-RCNN achieved an accuracy of 93% (95% CI: 88, 98) and DETR an accuracy of 96% (95% CI: 92, 98). Overall, the networks were consistent in their predictions with the MCC of 81% (95 % CI : 66, 95) for FasterRCNN and 88% (95% CI: 77, 95) for DETR. Overall, we found no evidence of a difference between Expert 2 and DETR in terms of accuracy, specificity and sensitivity. In addition, both Expert 2 and DETR showed higher accuracy and sensitivity compared to the non-expert caregiver and Expert 1 (P $\leq .001$).

4.3.2 Localization of lesions

Table 4.2 shows the localization of FLLs' performance in terms of PPV, sensitivity, and F1-score. As there are no true negatives for this task, the accuracy and specificity are not reported. Expert 2 and networks achieved comparable results for this task, with a mean PPV of 76% (95% CI: 72, 79) and mean sensitivity of 78% (95% CI: 74, 83). Expert 1 had a PPV of 73% (95% CI: 63, 84) and sensitivity of 69% (95% CI: 62, 77). This result was consistent with the detection accuracy of liver with lesions, where Expert 1 had lower performance compared to Expert 2 and networks.

4.3.3 Characterization of lesions

To report the characterization performance, a subset of lesions - that were detected and localized both by the networks and the experts- was selected. Table 4.3 shows

	IoU	PPV	Sensitivity	F1-Score
Expert 1	0.72 ± 0.14	73 (63, 84)	69 (62, 77)	71 (64, 79)
Expert 2	0.68 ± 0.12	80 (71, 89)	78 (71,85)	79 (72, 86)
Faster RCNN	0.71 ± 0.10	72 (66, 79)	0.73 (0.66, 0.8)	73 (67, 78)
DETR	0.69 ± 0.12	77 (70, 84)	84 (77, 89)	80 (74, 85)

Notes- Analysis based on 37 patients (214 lesions). Reported results were computed on the bootstrapped test set. Each subset contained only one image per patient. - Dispersion index is shown as ± Standard Deviation - Intervals in parentheses are 95% Confidence Intervals. IoU = intersection over union, PPV = positive predictive value, RCNN = recurrent convolutional neural network, DETR = Detection TRansformer.

Tab. 4.2.: Localization of Focal Liver Lesions (FLLs) Performance

characterization of FLLs' performance in terms of accuracy, specificity, sensitivity, and MCC. DETR achieves the highest accuracy with 81% (95% CI: 68, 94) against 61% (95% CI: 50, 71) for the best performing expert (Expert 2). Expert 2 accuracy was significantly different with all other raters ($P < .001$) except with Expert 1 ($P = .25$). DETR was significantly different with all other raters except with FasterRCNN ($P = .18$). Expert 2 showed the highest sensitivity 87% (95% CI: 73, 100) among all raters and the lowest specificity 33% (95% CI: 25, 44). DETR showed the second highest sensitivity 82% (95% CI: 62, 100) and the highest specificity 81% (95% CI: 67, 91) among all raters. Both networks also had a higher MCC with the true labels: 63% (95% CI: 37, 88) for DETR and 25% (95% CI: 4, 50) for Expert 2.

4.3.4 Sub-characterization

The sub-characterization performance is also reported in Table 3 on the subset of lesions that were detected and localized both by the networks and the experts. DETR achieved the highest accuracy with 76% (95% CI: 62, 91) against 52% (95% CI: 36, 70) for the best performing expert (Expert 1). All raters (experts and networks) showed a comparable specificity for the sub-characterization task with a mean value of 91± 3 and sensitivity with a mean value of 59 ± 10. It should be noted that pairwise comparisons for this task did not show significant differences between all raters (experts and networks).
In summary, for the detection of lesions in the liver parenchyma no significant difference was found between the best performing network (DETR) and the best performing expert (Expert 2) in the small test set; for the localization of FLLs, all raters achieved comparable results; for the characterization of FLLs detected and localized by all raters, both networks achieved higher performance in comparison with experts; for the sub-characterization of benign and malignant FLLs, pairwise comparisons between raters were not significantly different. More details on the discrimination performance of the networks are reported in (Supplementary Appendix A.3, A.4 and A.5).

	Accuracy	Specificity	Sensitivity	Matthews Correlation Coefficient
Characterization of focal liver lesions				
Expert 1	$59^{\star}(47, 70)$	$79\dagger(62, 92)$	$40^{\star}\dagger(22, 55)$	$0.20^{\star}\dagger$ $(-0.04, -0.44)$
Expert 2	$61^{\star}(50, 71)$	$33^{\star}(25, 44)$	$87(73, 100)$	$0.25^{\star}$ $(0.04, -0.50)$
Faster RCNN	$76\dagger(64, 90)$	$72^{\star}\dagger(55, 90)$	$81(60, 100)$	$0.53\dagger(0.27, - 0.81)$
DETR	$81\dagger(68, 94)$	$81(67, 91)$	$82(62, 100)$	$0.63\dagger$ $(0.37, -0.88)$
Sub-Characterization of focal liver lesions into six classes				
Expert 1	$52\mp$ (36, 70)	91 ‡(88, 94)	48‡(38, 61)	-
Expert 2	$50\mp$(36, 67)	87 ‡(84, 92)	54‡(40, 68)	-
Faster RCNN	$72\mp$(54, 91)	0.93 ‡(88, 98)	70‡ (45, 93)	-
DETR	$76\mp$(62, 91)	94 ‡(90, 98)	65‡(50, 80)	-

Notes- Characterization of focal liver lesions into benign (24 patients: $n_{benign} = 48$) or malignant (13 patients: $n_{malignant} = 74$). Sub-Characterzation of focal liver lesions into six classes (37 patients: $n_{lesions}$ = 122); The six classes are: cyst, angioma, focal nodular hyperplasia, adenoma,metastasis and hepatocellular carcinoma. Results reported on the subset of lesions that were localized by all raters on the bootstrapped test set. Each subset contained only one image per patient. Intervals in parentheses are 95% Confidence Intervals. $\star$P ≤ 0.05 in comparison with DETR. $\dagger$ $P \leq 0.05$ in comparison with Expert 2 - $\mp$ Overall Accuracy. ‡ Macro Average, average accuracy at the class level. RCNN = recurrent convolutional neural network, DETR = Detection TRansformer.

Tab. 4.3.: Characterization and Sub-characterization of Focal Liver Lesions' (FLLs') Performance

4.3.5 Discrimination Performance of the Networks

To determine the discrimination performance of the networks with regards to the thresholds levels applied to the outputs, the precision-recall curve, and the area under the precision-recall curve, commonly called average precision (AP) metric, were used. These are conventional measures in computer vision for object detection and localization tasks (19, 20). Figure 4.4 shows precision-recall curves and their AP for each network and for the characterization tasks - Liver parenchyma with or without lesions and benign or malignant lesions. The curves are consistent with the reported results above.

4.3.6 False Positive Findings for the FLL Characterization Task

The 3 images in Figure 4.5 correspond to benign lesions identified as malignant by the transformer-based DETR network.

Images mischaracterized as benign only by the Transformer based network DETR- Image A shows a benign lesion (cyst) with hypogenic rather than anechogenic content.

Images mischaracterized as benign by the Transformer based network DETR and Expert 2- Image B and C- Image B shows a benign lesion (angioma), with unclear boundaries and heterogeneous content, which is not characteristic of an angioma. Image C shows a benign lesion (FNH), this image is also not typical and should be discreetly hypoechoic or isoechoic with a discreetly hyperechoic central scar. Both images (B and C) were identified as metastases by Expert 2 and correctly characterized by Expert 1.

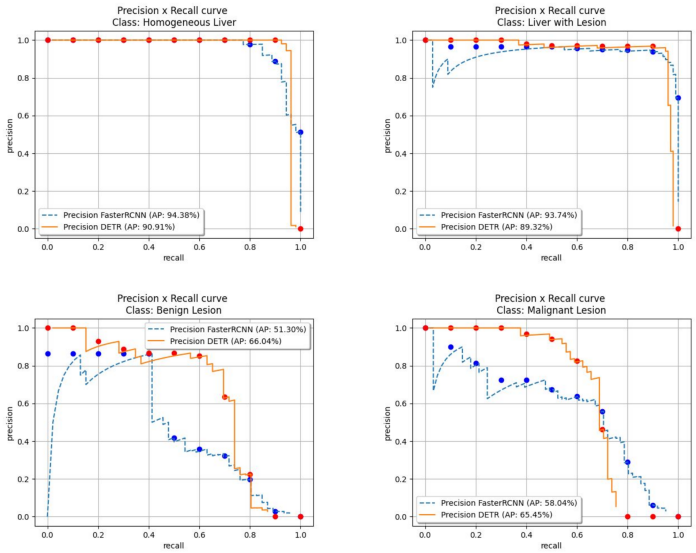


Fig. 4.4.: Precision-Recall curves and their Average Precision (AP). We use 11-point interpolated average precision (represented by dots in the graphs) to compute the average precision. It averages the precisions at each point in a set of 11 recall values (0,0. 1,...,1). Note.—RCNN = recurrent convolutional neural network, DETR = Detection Transformer

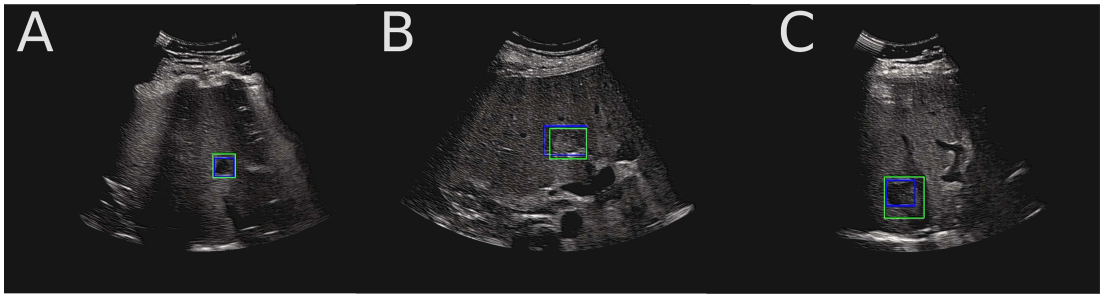


Fig. 4.5.: False positive findings- benign lesions identified as malignant- by the Transformer based network DETR for the FLL characterization task. Blue boxes correspond to the ground truth, green boxes are determined by the Transformer based network DETR. Note.—RCNN = recurrent convolutional neural network, DETR = Detection Transformer

4.3.7 False Negative Findings for the FLL Characterization Task

The 6 images in Figure 4.6 correspond to malignant lesions, identified as benign by the Transformer based network DETR.

Images mischaracterized as benign only by the Transformer based network DETR - Images A and B - show malignant lesions (HCC) and were mischaracterized as benign (FNH) by the Transformer based network DETR, they show liver surface nodularity

and heterogeneity of the hepatic architecture that would normally raise suspicion for malignancy. Both experts correctly characterized these two images as malignant and further sub-characterized Image 1 as HCC and Image 2 as metastasis.

Images mischaracterized as benign by all raters - Images C and E - were identified as benign (angioma) by both experts, they show unique round, hyperechoic lesions with sharp edges whose ultrasound semiology would correspond to an angioma. Indeed, hyperechoic metastases are difficult to differentiate from angiomas. The sensitivity of ultrasound for metastasis detection is reported in the literature as low and variable, ranging from 50%-76% (29) .

Images mischaracterized as benign by the Transformer based network DETR and Expert 1- Image 4 and 6- Image D shows a malignant lesion (metastasis), hypoechoic with a smooth border, and was identified as benign (cyst) by Expert 1 and malignant by Expert 2 (mischaracterized as HCC) because of the lack of posterior acoustic enhancement. Image F, shows a malignant lesion (HCC), heterogeneous and poorly limited, and was identified as benign (FNH) by Expert 1 and malignant by Expert 2 (mischaracterized as metastasis) because of the liver parenchyma that shows areas of different echogenicity.

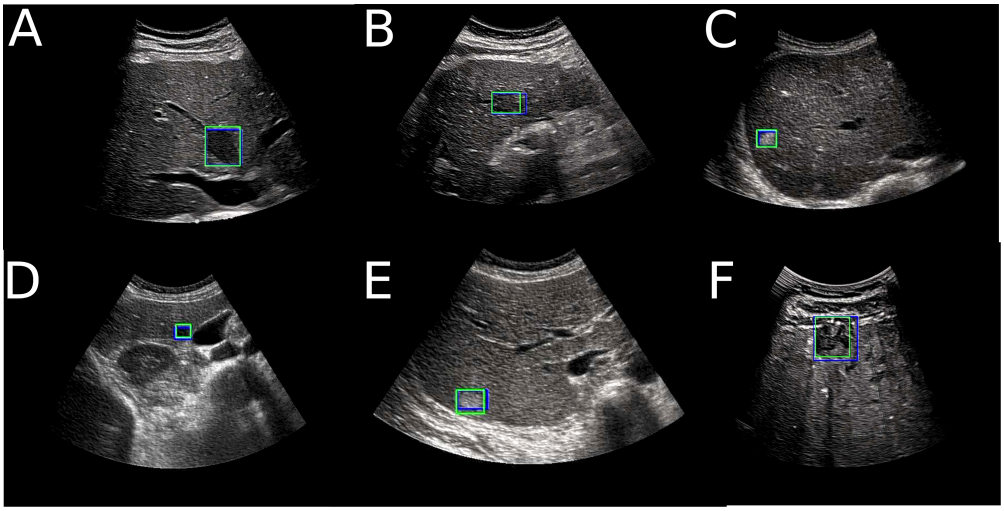


Fig. 4.6.: False negative findings - malignant lesions, identified as benign- by the Transformer based network DETR for the FLL characterization task. Blue boxes correspond to the ground truth, green boxes are determined by the Transformer based network DETR. Note.—RCNN = recurrent convolutional neural network, DETR = Detection Transformer

4.4 Discussion

In this research, we presented a framework for the detection, localization, and characterization of FLLs in B-mode US images. We began by investigating the detection accuracy of FLLs. Next, we included the localization of FLLs' task with the aim of drawing the

examiner's attention to a region of interest. Finally, with FLLs' localization, we could characterize the malignancy of each lesion in the image. The network DETR had a specificity value of 90% (95% CI : 75-100) and a sensitivity value of 97% (95% CI: 97, 97) for the detection of lesions in the liver parenchyma on the test set. It was able to correctly localize 80% of the lesions, and among the lesions correctly localized by all raters (experts and networks), had a specificity of 81% (95% CI: 67, 91) and a sensitivity of 82% (95% CI: 62, 100) for FLL characterization (benign vs. malignant). A previous study showed high performance for FLL detection (ROC-AUC scores of 0.935) in B-mode US images, using repeated random cross-validation [Schmauch, 2019] with roughly the same proportion (70%) of liver parenchyma with lesions as in our study. Our contribution for this task lies in the comparison of screening performance to non-expert and expert caregivers. Our analysis led to the conclusion that the network DETR indeed matches the performances of experts for such a task. We believe this is the first attempt to automatically localize FLLs in B-mode US images. This step enables us to characterize each lesion individually in addition to the global class assigned to the liver (with or without lesions). As our main objective aims for the detections to occur in the areas of interest, we chose an IoU threshold of 0.3. This way, slightly delocalised correct detections were still considered as true positives. We demonstrated that both networks met the performances of experts for this task.

Previous studies investigated the use of deep learning networks for the characterization of lesions in US imaging [Schmauch, 2019; Yang, 2020; Yao, 2018; Ta, 2018; Park, 2013]. Focusing only on the characterization task implies that either the whole liver is classified as benign or malignant, or that the lesions are first localized by an expert and then classified. This is particularly relevant for the diagnosis task of FLLs, where additional external clinical information or features from multiphase imaging with the administration of contrast material (i.e. CE-US, MRI, or US) are often used to reach a final diagnosis. Yang et al. study follows this objective and showed diagnostic performance comparable to that of CE-US using solely B-mode ultrasound data, segmented regions of interest and clinical information. As for the screening of FLLs, it is particularly important to localize the lesions during the examination in order to draw the attention of the examiner when needed and enable him or her to make a decision regarding follow-up based on their observations and the characterization predicted by the network. Both our trained networks showed higher performance for this task, compared to the experts.

Our results for the sub characterization of benign and malignant lesions stay below the reported performances in CE-US [Yasaka, 2018; Park, 2013] and show that the sub-characterization task is much more challenging in non-CE US, which correlates well with the literature [Hanna, 2016]. A previous study showed high performance for the sub-characterization task in B-mode US images at an image level (as opposed to lesion level) [Schmauch, 2019], and while this may be confusing in practice, it still gives insight as to how we can improve performance. Indeed, we think these results could be further improved by adding more weight to the liver features when classifying the lesions, as many experts look at patterns in the liver when making a primary diagnosis based on

US [Harvey, 2001]. This also applies to the use of US clip-videos instead of static US images. Another interesting approach is to investigate the use of self-supervised and semi-supervised learning, as we have access to thousands of unlabeled images. Previous studies showed that these approaches could close the gap between strong and weak supervision, for instance in digital pathology [Dehaene, 2020].

Our study has several limitations. First, the number of images in the test set is limited and the study is retrospective. To ensure the generalizability of the network, we need further investigations with a larger multi-centric and multi-vendor prospective cohort. Second, the reference standard used for localization was based on a unanimous annotation by an adjudication panel to avoid inter-expert variability. While this enables for a more robust learning, it makes it difficult to construct a larger dataset and create bias. A less stringent criteria would be the majority vote and could be used instead during the annotation of the test set. In addition, excluding images of poor quality or deemed uninterpretable likely creates bias and can potentially impact future predictions on images of the same type. Given that ultrasound images are operator dependent, their quality may vary considerably between operators and also during the same examination (the operator may capture images in several planes to best describe the abnormality, some of which become difficult to interpret a posteriori without context). One way to address this issue is use of Image Quality Assessment networks (with access to references that are considered to be of good quality and poor quality images) to filter images before applying the screening of FLLs network. Future studies are needed to analyze how these networks can be included in the method. Third, the experts and networks were blinded to clinical information that is normally available and that can help for screening of FLLs (e.g at-risk patients with known cirrhosis or chronic hepatitis B virus infection). The decision to remove clinical information from our study was to analyze solely the information that can be extracted from ultrasound images. Finally, given the size of our test set and the imbalance in sub-categories, we were not able to draw a comparative analysis study between our method and the experts for this task. Nonetheless, it is important to underline that the network performed poorly compared to experts for the detection of cysts. We believe this is due to the lack of training examples for this class, but also to the fact that, for a network which has only seen focal liver lesions, cysts can be confused with the portal vein or the inferior vena cava in the sagittal plane. A potential improvement to avoid this confusion would be to add annotated examples of the different anatomical structures visible in abdominal ultrasound during training. In brief, this study provides new information on the use of deep learning networks for the detection, localization, and characterization of focal liver lesions in B-mode US images. Experiments on a test set and comparison to experts showed that the vision transformer network DETR can be a companion to the examiner's visual attention by helping focus his or her efforts on areas of interest. This can also provide insights to non-expert caregivers and facilitate screening of FLLs with a potential to increase early detection of HCC in the future.

5 Deep Clustering for Abdominal Organ Classification in US imaging

Contents

Abstract The use of ultrasound (US) imaging has developed considerably in several medical specialties recently. In particular, abdominal pain accounts for a significant part of medical consultations. In this context, ultrasound is the only non-invasive and non-ionizing imaging modality that allows real-time medical exploration of a specific body part. However, acquiring and interpreting US images remains a difficult and examiner-dependent task, with a limited number of trained operators. For abdominal organs, ultrasound images are even more difficult to interpret because some of the organs of interest are located deep inside the body and patient-related factors, such as the presence of fatty tissue, can hinder the reading. In this work, we present a simple framework for abdominal organ clustering using unlabeled ultrasound images. This method can serve as a tool to preprocess large uncurated databases, reducing the need for annotation in abdominal ultrasound studies. When few labeled examples are available, we explore how unlabeled data can be leveraged to improve the performance of multi-label classification as opposed to the traditional transfer learning approach. In particular, we show that for supervised fine-tuning, deep clustering is an effective pre-training method, with performance matching that of ImageNet pre-training using five times less labeled data. Finally, we combine this pre-training method with semi-supervised learning and report the performances. This chapter was submitted to a journal [Dadoun, 2022a].

5.1 Introduction

Several studies have sought to classify abdominal organs on ultrasound images. Due to the lack of freely available databases, most of them use in-house data-sets of very different sizes ranging from 4094 [Cheng, 2017] to 187,219 [Xu, 2018] labeled images. They all share a common approach: the use of transfer learning[1] to initialize their classification models, combined with supervised learning methods. Transfer learning allows the reuse of deep networks trained on large scale annotated databases such as ImageNet, to learn specific tasks for which fewer labeled examples are available, while leveraging the learned general-purpose features of the original network. This applies to ultrasound images where labeled examples are hard to obtain due to the limited availability of expert annotations.

In a study, the authors evaluate transfer learning (and more specifically fine-tuning) with deep CNNs for the classification of abdominal ultrasound images [Cheng, 2017]. Classes were chosen based on 11 categories specified on the images, which would correspond to standard plane views acquired during an abdominal examination. Images that fell outside

[1] either by re-training the complete model, or by tuning the last layers of the model only.

these standard plane views were excluded, in addition to color or spectral Doppler, and images with very limited or unrecognizable anatomy. The study shows that using a large VGGNet network trained on 4094 images yields 77.9% accuracy on a test set of 1423 images. On a slightly different setting, the authors propose a multi-task learning framework for both the classification of views (including a class for *other* views), and a landmark detection for each relevant view [Xu, 2018]. A total of 187,219 ultrasound images from 706 patients were collected, and 20% of the dataset was used for testing. For the classification task, the study reports a 4.07% improvement compared to single-task learning, suggesting that the classification task benefits from sharing the low level features with the landmark detection task. Finally, Li *et al* [Li, 2021] use a public dataset of 360 ultrasound images to propose a classification method that combines the deep learning techniques and k-Nearest-Neighbor (k-NN) classification for the multi-label classification of six anatomical structures (bladder, bowel, gallbladder, kidney, liver, and spleen), making again the assumption that each image contains a single organ. In particular they show that for various classification models (ResNet-101, ResNet-152, DenseNet-121, DenseNet-169, and DenseNet-201) the use of a k-NN classification on top of fixed features outperforms the fine-tuning of a fully connected layer. However they also show that this does not apply when using a Resnet50 model, which is the most commonly used model in this setting.

In summary, the available literature on the subject focuses on supervised methods using transfer learning to overcome the lack of labeled data. These methods all use a multi-class classification models, either considering one organ per image, which is rarely valid in practice, or considering classes as standard views (containing several organs), which forces the model to classify only known standard views.

In this study we consider a different angle, and explore how unlabeled data can be used for the task of abdominal organ classification in ultrasound images in the multi-label setting (non-mutually-exclusive classes), as multiple organs may be visible simultaneously on the same image in abdominal ultrasound. This is achieved by considering three different approaches to exploit unlabelled data: i) self-supervised followed by supervised learning, ii) semi-supervised learning or iii) combining self-supervised pre-training followed by semi-supervised learning. Fig. 5.1 provides a schematic overview of these three different learning methods, in addition to the transfer learning approach.

More specifically, we explore different questions in this context:

1. Are the features learned in a self-supervised manner more useful for downstream tasks on the same domain compared to features learned on a different domain database such as ImageNet (Fig. 5.1: Strategy 2 vs 1)?

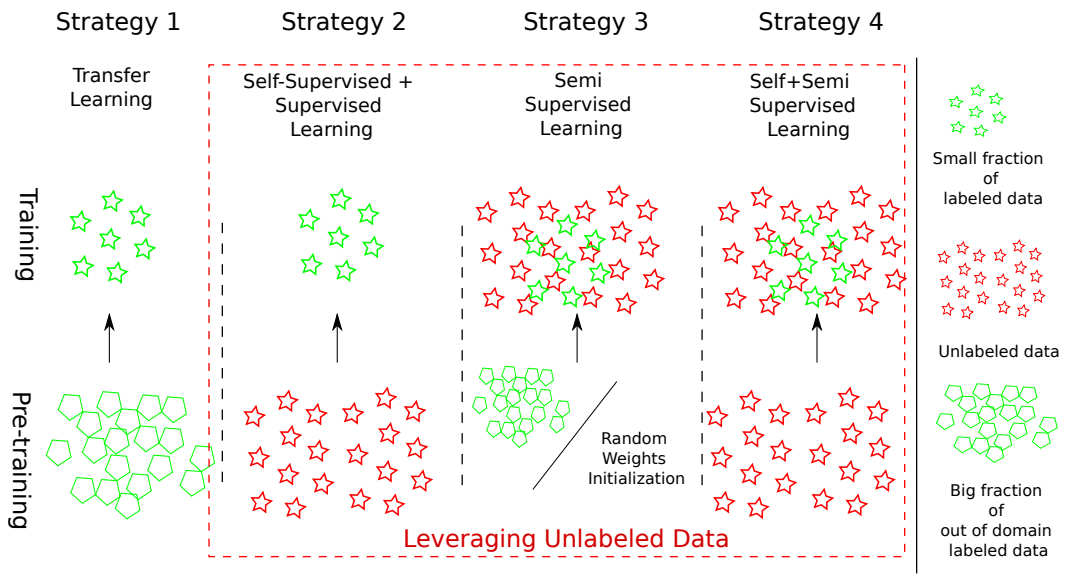


Fig. 5.1.: Overview of three different scenarios where unlabeled data (represented by red stars) can be leveraged with few labeled examples (represented by green stars): during pre-training in a self-supervised manner (Strategy 2), during training in a semi-supervised manner (Strategy 3), and during both stages (Strategy 4). Transfer learning (Strategy 1) is presented as a baseline method that does not require unlabeled data but rather a large amount of *out-of-domain* labeled data (represented by green polygons).

2. Is there any advantage to leverage labeled and unlabeled data simultaneously during training through semi-supervised learning rather than consecutively as proposed by self-supervised methods (Fig. 5.1: Strategy 3 vs 2)?

3. Is there any relevance in combining the two paradigms (Fig. 5.1: Strategy 4)?

5.1.1 Self-supervised Learning

Self-supervised learning methods are used to train networks on large scale unlabeled *in-domain* data using surrogate tasks before transferring those learned representations to more specific tasks on few labeled data. Following the taxonomy presented in Jing *et al* [Jing, 2020], the surrogate tasks can be summarized in four categories:

- Generation based methods (e.g image inpainting [Pathak, 2016]).

- Semantic-free label-based methods (e.g contour detection [Ren, 2018], temporal order correction [Jiao, 2020]).

- Cross modal-based methods (e.g audio and video correspondence [Korbar, 2018; Bonmati, 2021; Bardes, 2021]).

- Context based methods (e.g clustering [Caron, 2018]).

All of the above tasks share the same hypothesis, (1) Visual features are needed to solve the task and (2) the Convolutional Neural Networks (CNNs) can capture those features by solving the surrogate task. Previous works explored the use of such method in ultrasound imaging. For instance, Jiao *et al.* [Jiao, 2020] used cross-modal contrastive learning in multi-modal fetal ultrasound video and audio to learn strong representations and transfer them to a supervised task of fetal standard plane detection, demonstrating higher performance than with features whose weights were initialized with ImageNet. These results were obtained using 90 scans of 55,000 frames each for training, which in practice represents a large amount of labeled data.

5.1.2 Semi-supervised Learning

Semi-supervised learning methods, on the other hand, utilize both labeled and unlabeled data during training to learn feature representations specific to the learning task. In this case, the networks are trained with an objective function composed of two terms: a supervised loss applied to labeled data and an unsupervised loss applied to unlabeled data. The latter is derived either using consistency regularization [Sajjadi, 2016; Laine, 2016] or pseudo-labeling [McLachlan, 1975; Lee, 2013]. Most of these methods are evaluated on curated data sets where the data distribution is close to uniform and unlabeled data contains no novel class. Unfortunately, the strong performance of such methods does not always translate to non-curated data-sets [Su, 2021] that violate assumptions implicit to the semi-supervised learning approaches. These assumptions [Chapelle, 2009] include smoothness, ("If two points x_1,x_2 in a high-density region are close, then so should be the corresponding outputs y_1,y_2"), and cluster assumption, ("If points are in the same cluster, they are likely to be of the same class"). In fetal 2D ultrasound imaging, a study [Tan, 2019] explored the use of consistency regularization with unsupervised data augmentation for anatomy classification to test whether the reported results in semi-supervised learning translate to non-ideal data. In particular, the authors show that the inclusion of challenging classes in the unlabeled set can harm the performance, compared to a supervised setting.

5.1.3 Deep Clustering

As described previously, several surrogate tasks can be considered to analyse the self-supervised learning framework. One of which is deep clustering, which we propose to use because it has the advantage of being closely related to the classification task. Furthermore, using the cluster assignments on a small development set allows to validate the cluster assumption needed for the semi-supervised methods. Traditionally, clustering

methods such as K-Means aim to learn cluster assignments based on fixed feature representations. For high dimensional imaging data, in particular, such fixed features are generally not sufficiently discriminating and need to be learned during clustering. In a recent effort to address this shortcoming, Caron *et al*. [Caron, 2018], proposed a deep clustering framework for the unsupervised learning of CNNs trained in an end-to-end fashion. The proposed method iterates between clustering the learned CNN features using K-Means (i.e clustering) and updating the CNN weights by predicting the cluster assignments as pseudo-labels (i.e representation learning). However, the process of alternating between these two learning objectives may be prone to error-propagation, which motivated other studies to propose simultaneous learning methods [Huang, 2020; Van Gansbeke, 2020]. More recently, on a related application domain, Kart *et al* [Kart, 2021] adapted the framework of DeepCluster [Caron, 2018] to categorize uncurated large-scale cardiac MRI images, and used normalized mutual information (NMI) and cluster purity (CP) to evaluate the clustering quality which gives little information on the relevance and usefulness of the grouped semantic categories.

5.1.4 Contributions

Based on these observations, and keeping the above-mentioned questions in mind we propose a framework for the multi-label classification of abdominal organs in ultrasound images based on a large database of 6951 ultrasound examinations (89 830 images) from 5788 patients with very few labeled examples. Our contributions can be listed as follows:

1. We propose to adapt two state-of-the-art multi-class methods to the multi-label classification setting: deep clustering with PICA [Huang, 2020] (Fig. 5.1: Strategy 2), and semi-supervised learning with FixMatch [Sohn, 2020] (Fig. 5.1: Strategy 3).

2. We evaluate the use of deep clustering in self-supervised learning, and show that the learned features transfer better to the classification task, with performance higher than that of ImageNet initialization (Fig. 5.1: Strategy 2 vs 1).

3. We show that combining deep clustering pre-training with semi-supervised learning (Fig. 5.1: Strategy 4) yields robust results, even when the number of labelled examples is extremely limited (less than 275 images).

5.2 Methodology

5.2.1 Problem Definition

We wish to assign to each ultrasound image a set of *non-mutually-exclusive* target labels corresponding to the C organs of interest: *liver, kidney, gallbladder, pancreas, spleen* and *bladder* or *other* if none of these organs are present in the image. This corresponds to the setting of multi-label classification, where the target label is a binary vector representing the absence or presence of each organ: $\mathbf{v} \in \{0,1\}^{C+1}$. We assume that we have access to a large amount of unlabeled data and very little labeled data. In this case, unlabeled data can be leveraged to improve classification performance either during pre-training and/or directly during training as presented in Fig. 5.1. In the remainder, we refer to the sets of unlabeled and labeled images respectively sampled during training at each iteration as $\mathbf{U} = \{I_1, ..., I_{N_u}\}$ and $\mathbf{S} = \{I_1, ..., I_{N_s}\}$ for the sake of clarity. We also introduce a development set $\mathbf{D} = [(I_1, v_1), ..., (I_d, v_d)]$ whose objective is to choose the network settings that achieve the best performance on images the network has never seen. All the models presented hereafter share a common architecture composed of two parts:

1. A feature extractor $f(.)$ that maps an image $\mathbf{I}$ into a vector representation $\mathbf{x} = f(\mathbf{I})$.

2. A single and/or a multi-label classification head: $g(.)$ that assigns each feature representation $\mathbf{x}$ with a class membership distribution.

5.2.2 Deep Clustering

Starting from a well-established deep clustering method (PICA [Huang, 2020]), we add a loss term to take into account the variability of the cluster size distribution at each iteration. To compare performance of both methods, we propose to use a small fraction of labeled data to assign each cluster to a relevant semantic multi-label category. The feature extractor of the best performing method can then be used to fine tune a supervised multi-label classification model.

PICA

Our method builds upon the framework of PICA presented in Fig 5.2, where two different augmentation schemes $(\tilde{.})$ and $(\hat{.})$ are applied to all input images before passing through the CNN composed of a feature extractor $f(.)$ and a classification head $g_\tau(\mathbf{x}) = p = \{p_1, ..., p_K\}$ where K is the number of clusters. We denote by $\mathbf{P} \in \mathbb{R}^{N_u, K}$ the cluster prediction matrix of N_u images in $\mathbf{U}$. Then a Partition Uncertainty Index (PUI) is introduced as the cosine similarity set of all the cluster pairs:

$$\mathcal{M}_{PUI}(j_1, j_2) = cos(\hat{q}_{j_1}, \tilde{q}_{j_2}) \quad \forall (j_1, j_2) \in [0, ..., K] \tag{5.1}$$

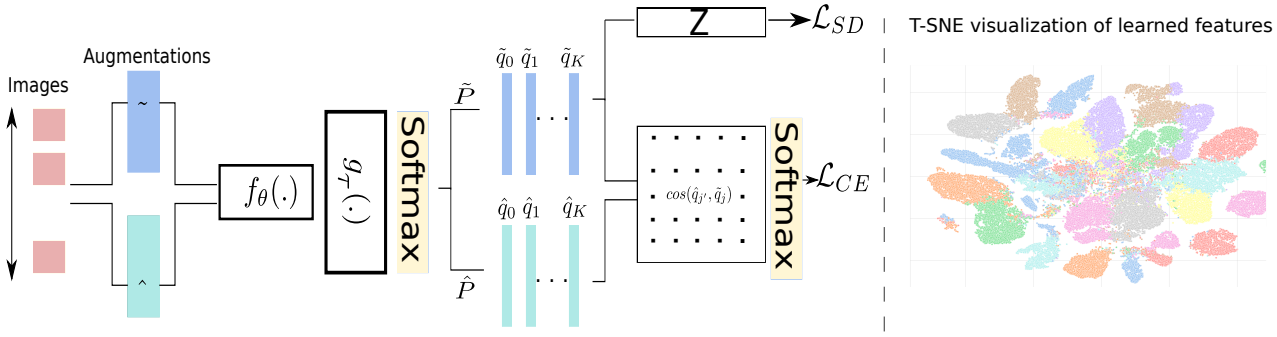


Fig. 5.2.: Overview of the deep clustering framework.Two different augmentation schemes $(\tilde{.})$ and $(\hat{.})$ are applied to all input images before passing through the CNN composed of a feature extractor $f(.)$ and a classification head $g_\tau(\mathbf{x}) = p = \{p_1, ..., p_K\}$. This framework can serve as an unsupervised classification model when no labeled examples are available. $\mathbf{P}$ is the cluster prediction matrix of all images in $\mathbf{U}$ and Z is the cluster size distribution. A T-SNE visualization of the learned features and their assigned clusters represented by different colors is shown on the right.

Where $\tilde{q}_j$ and $\hat{q}_j$ are the j-th rows of $\tilde{\mathbf{P}}$ and $\hat{\mathbf{P}}$ respectively.
To obtain the most confident predictions for each cluster without using any particular distance metric between samples, the authors propose to :

1. Force $\tilde{q}_{j_1}$ and $\hat{q}_{j_2}$ to be orthogonal when $j_1 \neq j_2$, which also means that their cosine similarity is equal to zero.

2. Force $\tilde{q}_j$ and $\hat{q}_j$ to be equal, meaning that predictions on two augmented views of the same image should be equal in which case the cosine similarity is equal to one.

To derive an objective function, the authors propose to apply a softmax operation as self-attention to each cluster j:

$$m_{j,j'} = \frac{\exp \mathcal{M}_{PUI}(j, j')}{\sum_{k=0}^{K} \exp \mathcal{M}_{PUI}(j, k)} \quad \forall j' \in [0, ..., K] \tag{5.2}$$

Yielding the following differentiable objective function:

$$\mathcal{L}_{CE} = \frac{1}{K} \sum_{j=0}^{K} -\log(m_{j,j}) \tag{5.3}$$

In addition, a constraint on the cluster size distribution Z is introduced to avoid trivial solutions:

$$\mathcal{L}_{NE} = \log(K) - H(Z) \tag{5.4}$$

with $z_i = \frac{\tilde{q}_i}{\sum_{k=0}^{K} \tilde{q}_k}$ and H is the entropy function.

The final objective function of PICA is formulated as:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda \mathcal{L}_{NE}$$

with λ a hyper parameter.

Cluster size distribution
By minimising the negative entropy of the cluster size distribution $\mathcal{L}_{NE}$ introduced in (5.4), PICA forces $\mathbf{Z}$ to follow a uniform distribution at each iteration, meaning that all clusters have approximately the same size at each iteration. However, if the whole target data space is unbalanced, as it is the case in ultrasound imaging, then at each iteration the cluster size distribution needs to be also unbalanced. As we do not have any prior information to favor one cluster over any other, we set $\mathbf{Z}$ to follow a Symmetric Dirichlet distribution:

$$f\left(z_1, \ldots, z_K; \boldsymbol{\alpha}\right) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{i=1}^{K} z_i^{\alpha-1} \tag{5.5}$$

with B the Beta function, $\sum_{i=1}^{K} z_i = 1$, $z_i \geq 0$ for all $i \in \{1, \ldots, K\}$, and $\boldsymbol{\alpha} = [\alpha, ..., \alpha] \in \mathbb{R}^K$such that $\alpha \geq 0$.

Doing so, the variance of the Dirichlet is captured by the magnitude of the chosen value (α) for the parameters:

- when $\alpha \gg 1$: the entropy of a Dirichlet draw is high, upper bounded by $\log(K)$ if $\alpha \to \infty$
- when $\alpha \ll 1$: the entropy of a Dirichlet draw approaches a delta peak on a random entry, which would correspond to an entropy equal to one.

In other words, (1) setting a large value for α would correspond to the case where at each iteration we force a uniform distribution (similar to using the negative entropy loss $\mathcal{L}_{NE}$), and (2) setting a small value for α would correspond to the trivial solution where at each iteration, all images are assigned to a single cluster. By taking α slightly higher than 1, we obtain a trade-off between both scenarios.

Therefore, we maximize the likelihood of $\mathbf{z}$ being a Symmetric Dirichlet distribution with $\alpha = 1 + \epsilon$:

$$\mathcal{L}_{SD} = -\frac{1}{log(K)} \sum_{k=0}^{K} (\alpha - 1) log(z_k) \tag{5.6}$$

The final objective function becomes:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda \mathcal{L}_{SD}(\alpha) \tag{5.7}$$

Matching clusters with multi-label targets

The objective of deep clustering is to separate the set of images into mutually exclusive clusters. Each cluster can then be matched with the dominant class and all images assigned to this cluster will have the same label. However, in our application, an image can contain several labels. As such, we consider each different set of labels that exist in the multi-label dataset as a single label. By utilizing label priors, the number of plausible and most common combinations is set to $K < 2^C + 1$, where C corresponds to the number of classes. Doing so, the target label becomes a one hot encoding label ($\mathbf{w} \in \{0,1\}^K$) where the dimension K corresponds to the number of label combinations, leading to:

$$g_\tau(\mathbf{x}) = p = \{p_1, ..., p_K\} \tag{5.8}$$

To evaluate the relevance of the grouped semantic categories in the context of multi-label classification, instead of using directly the output probabilities p, we allocate to each cluster k a soft assignment. This is done by averaging the ground-truth multi-class labels $\mathbf{v} \in \{0,1\}^{C+1}$ of all images $\mathbf{I}$ in the development set $\mathbf{D}$ assigned to cluster k as shown in Algorithm 5.3: $\hat{v}_k = \frac{1}{|k|} \sum_{i \in k} v_i$.

Algorithm 1: Matching clusters with multi-label targets

Input

- Development set, $\mathbf{D} = [(I_1, v_1), ..., (I_d, v_d)]$.
- The cluster prediction matrix $\mathbf{P}$ of all images in $\mathbf{D}$.
- The target cluster number, K.
- The target class number, C.

Initialisation ;
$\hat{v}_k = [0, ..., 0] \in \mathbb{R}^{C+1}$
for $i \in \mathbf{D}$ **: do**
  $\hat{v}_{\arg\max(p_i)} += v_i$
end
for $k \in K$ **: do**
  $\hat{v}_k = \frac{1}{|k|} \times \hat{v}_k$
end
Output Cluster matching with dominating target class

Fig. 5.3.: Matching clusters with multi-label targets.

Fine-tuning for supervised learning

The weights learned during the deep clustering phase can be used to initialize a multi-label supervised classification model. This implies using the same feature extractor f_θ but discarding entirely the single-label clustering head $g_\tau(\mathbf{x}) = p = \{p_1, ..., p_K\}$ and replacing it by a multi-label classification head $g_\Phi(\mathbf{x}) = p = \{p_0, ..., p_C\}$. During training, we use the supervised objective function defined in Equation (5.9), where the target $\mathbf{v} \in \{0, 1\}^{C+1}$ is a vector representing the absence or presence of each organ. Model weights θ (feature extractor) and Φ (classification head) are learned jointly.

$$\mathcal{L}_{BCE} = \sum_{i \in \mathbf{S}} \frac{1}{C+1} \sum_{c=0}^{C} \mathbf{v}_c^{(i)} \log p_c^{(i)} + (1 - \mathbf{v}_c^{(i)})(1 - \log p_c^{(i)}) \tag{5.9}$$

5.2.3 Semi-Supervised Classification

Another common way of using unlabeled data is semi-supervised learning where unlabeled data is utilized simultaneously with labeled data during training using an unsupervised objective function for unlabeled examples. In the following, we present a state-of-the-art semi-supervised model for single-label classification. We then show how this model can be adapted to the multi-label classification setting. The feature extractor of this model can also be initialized with the deep clustering model, allowing to combine both methods.

FixMatch

The objective function $\mathcal{L}_u$ on the unlabeled set combines two lines of work:

- Pseudo-labeling: The objective is to use the model's prediction as pseudo-labels when the corresponding class probabilities fall above a certain threshold.

$$\mathcal{L}_u = \sum_{i \in \mathbf{U}} 1_{max(\tilde{p}^{(i)}) \gg \tau} \cdot \mathbf{H}(\tilde{y}^{(i)}, \tilde{p}^{(i)}) \tag{5.10}$$

  Where $\mathbf{H}$ is the cross-entropy and $\tilde{y}_c^{(i)} = 1_{\tilde{p}_c^{(i)} \geq \max_c(\tilde{p}_c^{(i)})}$ in the single-label setting.

- Consistency regularization: The objective is to force the model to output similar predictions when fed with perturbed versions of the same image.

We denote by $\hat{(.)}$ and $\tilde{(.)}$ the strong and weak augmentations applied to the input images, the loss function in this case is:

$$\mathcal{L}_u = \sum_{i \in \mathbf{U}} \left\| \tilde{p}^{(i)} - \hat{p}^{(i)} \right\|^2 \tag{5.11}$$

Both objective functions can be combined, as described in FixMatch [Sohn, 2020] by enforcing the pseudo-label $\tilde{y}^{(i)}$ obtained from the weak augmented version of the image against the model's output probabilities for the strongly augmented version of the image $\hat{p}^{(i)}$ as follows:

$$\mathcal{L}_u = \sum_{i \in \mathbf{U}} 1_{max(\tilde{p}^{(i)}) \gg \tau} \cdot \mathbf{H}(\tilde{y}^{(i)}, \hat{p}^{(i)}) \tag{5.12}$$

Yielding the following global loss function in the semi-supervised setting:

$$\mathcal{L} = \sum_{j \in \mathbf{S}} \mathbf{H}(\mathbf{v}^{(j)}, p^{(j)}) + \sum_{i \in \mathbf{U}} \lambda \mathcal{L}_u(\tilde{p}^{(i)}, \hat{p}^{(i)}) \tag{5.13}$$

FixMatch for multi-label classification

The objective function derived in FixMatch supposes a single-label classification problem, where each pseudo-label is a one-hot encoding vector. Fig. 5.4 describes two ways of using FixMatch objective function in multi-label classification: i) using a multi-label objective function for both pseudo-labels and true labels by redefining the pseudo-label term (One-Head model) or ii) Using a single-label objective function for pseudo-labels and a multi-label objective function for true labels through the use of an additional classification head (Two-Head model).

1. One-Head model: One way to use the setting of FixMatch directly is to consider a different formulation for the pseudo-label. In this case, a Sigmoid activation function is used instead of a Softmax function to ensure that the probability of each class is considered independently. Instead of taking the maximum probability of a class as the target label, we can threshold each probability class to 0.5: [Rizve, 2021]

   $$\tilde{y}_c^{(i)} = 1_{\tilde{p}_c^{(i)} \geq 0.5}$$

   If the maximum probability of at least one class is greater than a threshold τ, then the derived pseudo-label is considered as a target label during training.

2. Two-Head model: Another way is to use a network with two classifiers $g_\phi(\mathbf{x}) = \{p_1, ..., p_K\}$ and $g_\kappa(\mathbf{x}) = \{\bar{p}_1, ..., \bar{p}_C\}$. Doing so we can compute $\mathcal{L}$ on the outputs of g_ϕ and add a supervised loss (Eq 5.9) on the output of g_κ. The semi-supervised objective function is slightly modified as follows:

$$\mathcal{L} = \sum_{j \in \mathbf{S}} \mathbf{H}(\mathbf{w}^{(j)}, p^{(j)}) + \mathcal{L}_{BCE}(\mathbf{v}^{(j)}, \bar{p}^{(j)}) + \lambda \sum_{i \in \mathbf{U}} \mathcal{L}_u(\tilde{p}^{(i)}, \hat{p}^{(i)})$$

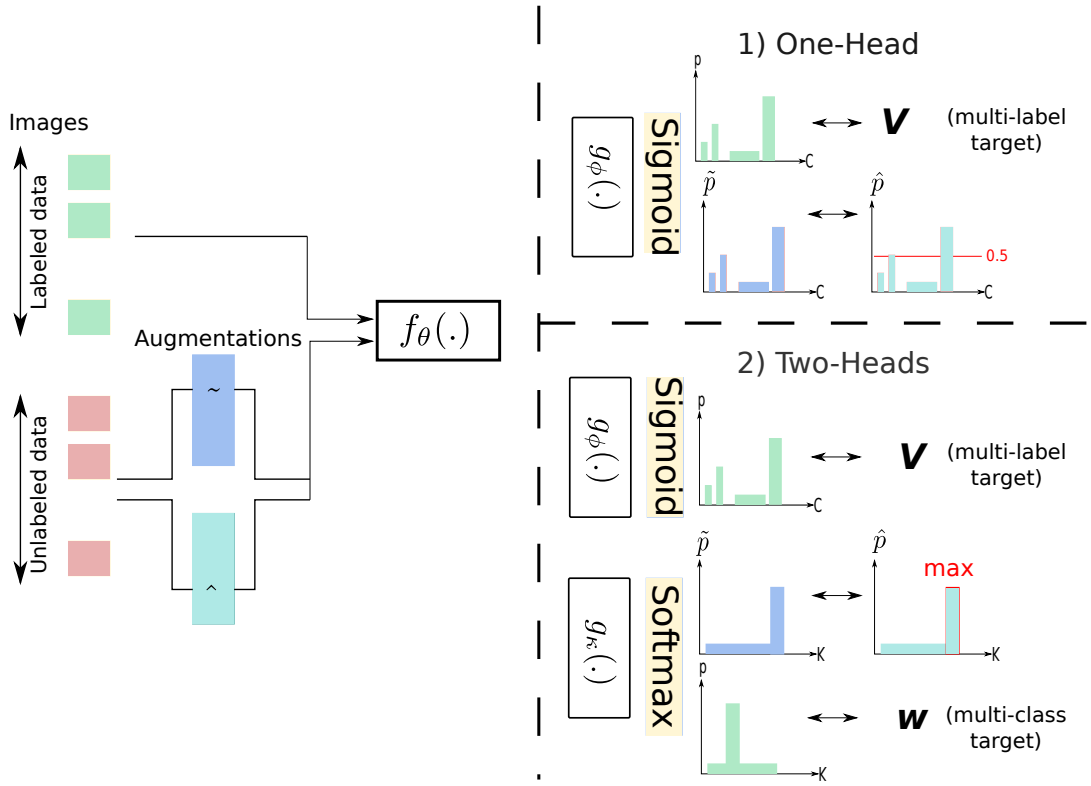


Fig. 5.4.: Semi-supervised learning: Two approaches to using FixMatch objective function in multi-label classification: i) Using a multi-label objective function for both pseudo-labels and true labels or ii) Using a single-label objective function for pseudo-labels and a multi-label objective function for true labels.

Self and Semi-Supervised Learning The feature extractor $f_\theta(\boldsymbol{x})$ of the semi-supervised learning model can be initialized either randomly (without pre-training) or from a pre-trained model (usually trained on a different database with a large size of labeled examples). It can also be initialized by the weights learned during the deep clustering phase, thus unifying these two approaches. The key distinctions with the fine-tuning method presented in 5.2.2 are the classification head(s) and the objective function, which includes the unlabeled data during training as well.

5.3 Experiments

		Liver		Gallbladder		Other		Kidney		Pancreas		Spleen		Bladder		weighted avg	
		Best	Mean	Best	Mean	Best	Mean	Best	Mean	Best	Mean	Best	Mean	Best	Mean	Best	Mean
F1-score	LSD	72.08	**70.71**	50.0	47.44	65.23	**60.41**	71.5	**68.94**	65.08	**54.08**	69.8	59.35	74.75	**68.09**	66.95	**64.73**
	PICA	69.94	67.36	59.9	**50.05**	58.33	49.98	73.59	64.47	58.21	48.23	68.75	**62.74**	64.66	57.9	65.99	60.38
Precision	LSD	72.8	**67.66**	79.25	61.68	63.89	**60.05**	78.71	**64.04**	76.09	**58.03**	74.17	56.21	91.43	**75.42**	70.29	**63.82**
	PICA	70.51	60.47	77.78	**61.74**	56.76	51.18	69.27	59.04	62.9	40.49	72.73	**59.82**	90.62	69.97	63.82	57.78
Recall	LSD	81.8	74.7	54.78	**42.96**	71.43	**61.84**	86.08	**76.2**	68.06	54.72	71.11	65.19	82.76	**66.55**	73.24	**68.05**
	PICA	91.89	**77.77**	53.91	42.78	62.22	51.24	83.23	72.59	76.39	**66.39**	71.11	**66.67**	74.14	53.1	70.11	66.42
n_{labels}		555		115		315		316		72		135		58		1566	

Tab. 5.1.: Abdominal organ classification results for the unsupervised model using deep clustering trained on 84967 unlabeled images. Performance is obtained on a test set of 101 abdominal exams ($n_{images} = 1242, n_{labels} = 1566$) with 5 trials:the average and best results are reported separately.

5.3.1 Data set

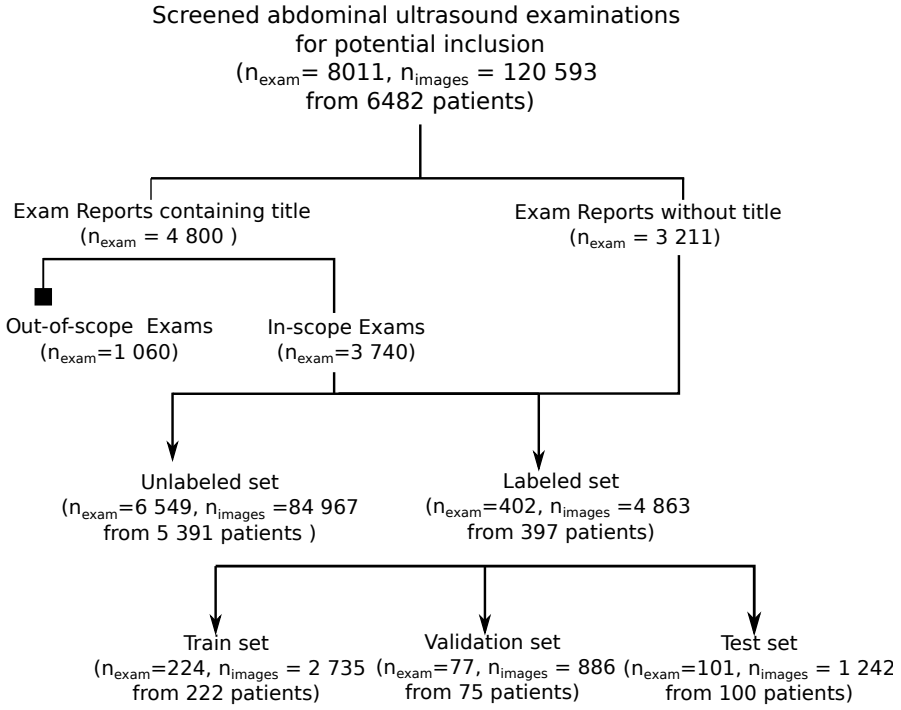


Fig. 5.5.: Flowchart describing the distribution of US examinations.

International Review Board approval was obtained for this retrospective study, in collaboration with the Clinical Data Warehouse registered under the number (no. IRB00011591). Multi-vendor data collected directly from the picture archiving and communication system (PACS) consisted of RGB freeze frames captured during ultrasound examination tagged as "abdominal" along with a textual report written by a physician. In total, 8011 abdominal ultrasound examinations (120 593 images) from 6482 patients were extracted. Abdominal exams are performed to look for abnormalities in the abdomen, but also to determine if blood is circulating at a normal rate and level or study the distribution of vessels in a structure (i.e., Color Doppler Mode), or provide guidance during a biopsy

procedure. In some cases, ultrasound examinations of the pelvis (supra-pubic, vaginal or endorectal scanning) are performed in addition and combined with the abdominal examination. However, in this study, we are only interested in gray-scale ultrasound images (i.e., Brightness Mode) obtained with traditional ultrasound systems and focused on six anatomical structures, which would correspond to the following examination report titles: *abdominal, renal, bladder, liver*, or *urinary tract* ultrasound. For 60% of the reports, a title was available, which allowed us to exclude 1060 out-of-scope examinations that did not fit the five aforementioned titles (conforming to the medical experts' recommendations). Because this information was not always available, we decided to create an additional class *other* when none of the organs of interest were present in the image. The final dataset consisted of 6951 ultrasound examinations (89 830 images) from 5788 patients. 224 examinations (2735 images) from 222 patients were randomly selected for the labeled training set and 6549 examinations (84967 images) from 5391 patients for the unlabeled training set. 77 examinations (886 images) from 75 patients were randomly selected for the validation set and 101 examinations (1242 images) from 100 patients for the test set. The sets were constructed to ensure that there is no overlap of patients between sets. Data partition is summarized in Fig. 5.5. An adjudication panel was used as an external standard of reference. The panel consisted of four physicians, either radiologists or holders of a French national diploma in US imaging, and four sonographers holders of a French national diploma in US imaging from six different health institutions with more than three years of experience. The annotators, who worked on a tailor-made annotation platform, were asked to activate the tags corresponding to the organs present in each image. Each image was annotated once by one of the experts, while images for the test set were annotated twice.

5.3.2 Experimental Settings

Training was conducted with four GeForce GTX 1080 Ti GPUs using Pytorch with a computation time ranging from few hours to 24 hours depending on the experiment. During training, and for all experiments, at each iteration we randomly sample a subset t of exams, instead of sampling a set of images, to make sure that almost all the classes are represented in a batch since standard abdominal exams include views of all organs of interest. For each comparison, we use the same network architecture and training protocol, including the hyper parameters, optimizer, learning rate, number of epochs, data augmentation and data preprocessing. In particular, for all experiments a Resnet18 backbone is used, with a fixed learning rate of 0.0001, Adam optimizer, and unlabeled batch size of 16 examinations. For experiments with unsupervised training, fixed number of

epochs (n = 100) is used and model weights of the last epoch are selected for evaluation. For experiments with supervised training, early stopping is performed if no improvement is observed on the validation set for seven epochs to avoid over-fitting. For experiments with semi-supervised training, a fixed amount of iteration steps (5000) is used. Finally for both supervised and semi-supervised training, model weights are selected at the epoch where the model performed best on the validation set. Performance is reported in terms of precision, recall and F1-score for each class and for the weighted average of all classes on the test set. To report these metrics, a class-specific threshold was set to select the probabilities output by the networks; the threshold was defined as the value that maximizes the F1 score of each class during the validation step. Finally, to reflect the performance stability, each experiment was iterated four times, by initializing the random number generator (i.e seed value) differently.

5.3.3 Results

We present the results of the deep clustering method when no labeled data is available during training, and analyze how the added loss term (LSD in (5.6)) impacts the performance. In addition, we present the results of using unlabeled data in the presence of labeled data ranging from 275 to 2742 labeled examples.

Training with Unlabeled Data

Table. 5.1 compares the classification performance of the deep clustering models for all classes on the test set. Multi-labels were assigned to the clusters after training as presented in Algorithm 5.3. Without using any labeled data during training, deep clustering yields reasonable results for all organs. In particular, the proposed constraint on the cluster size distribution (LSD) outperforms on average the PICA baseline for almost all classes with an F1-score Weighted Average of 64.83% and 60.7% respectively with the default hyperparameter $\lambda = 2$ used in the original paper to weight both terms of the objective function. Moreover, LSD seems to be more robust to the change of seeds, in fact the difference between the mean and the best results is almost insignificant in LSD, whereas in PICA, the seed has a greater impact on the performance. We further evaluate the sensitivity of both models to choices of λ by testing different values: $0.5, 1, 2, 5,$ and 10 with four different seeds. Fig. 5.6 displays the box plot of the F1-score values. We can see that PICA is very sensitive to the hyper-parameter λ, while LSD's performance

is more stable, with a higher median.

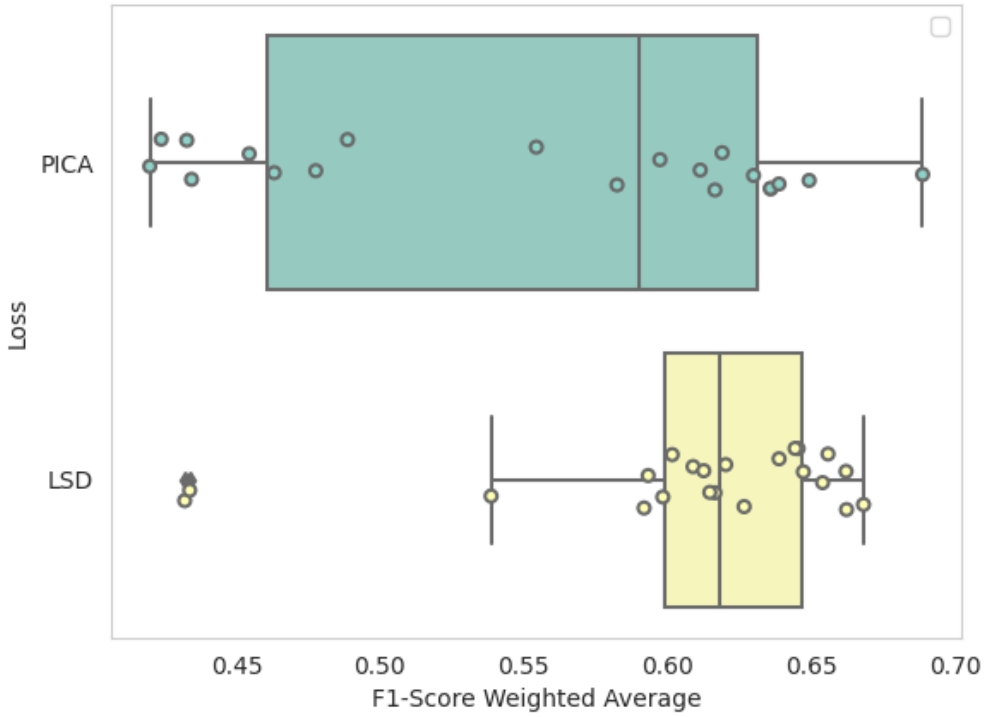


Fig. 5.6.: A box-plot showing F1-score weighted average values for each loss using five different values of λ over four experiments with different seeds.

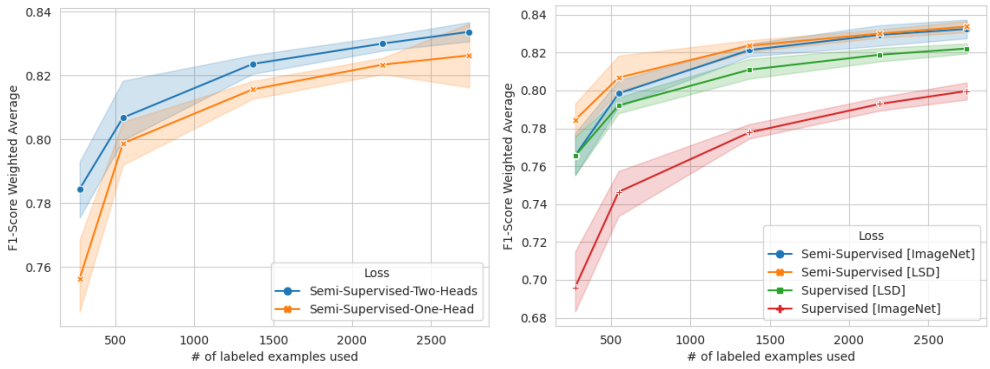


Fig. 5.7.: Mean and 95% confidence intervals of F1-score weighted average using all unlabeled images ($n_u = 84967$) with 10%, 20%, 50% and 100% of labeled images ($n_s = 2742$). On the left: One-Head vs Two Head semi-supervised learning models presented in Fig. 5.4. On the right: Semi-Supervised vs Supervised Learning models with different pretraining methods.

Training with both Labeled and Unlabeled Data

Performance is evaluated for several scenarios presented in Fig. 5.1, namely using unlabeled data simultaneously with labeled data during training in a semi-supervised manner, and/or during pre-training with deep clustering as a

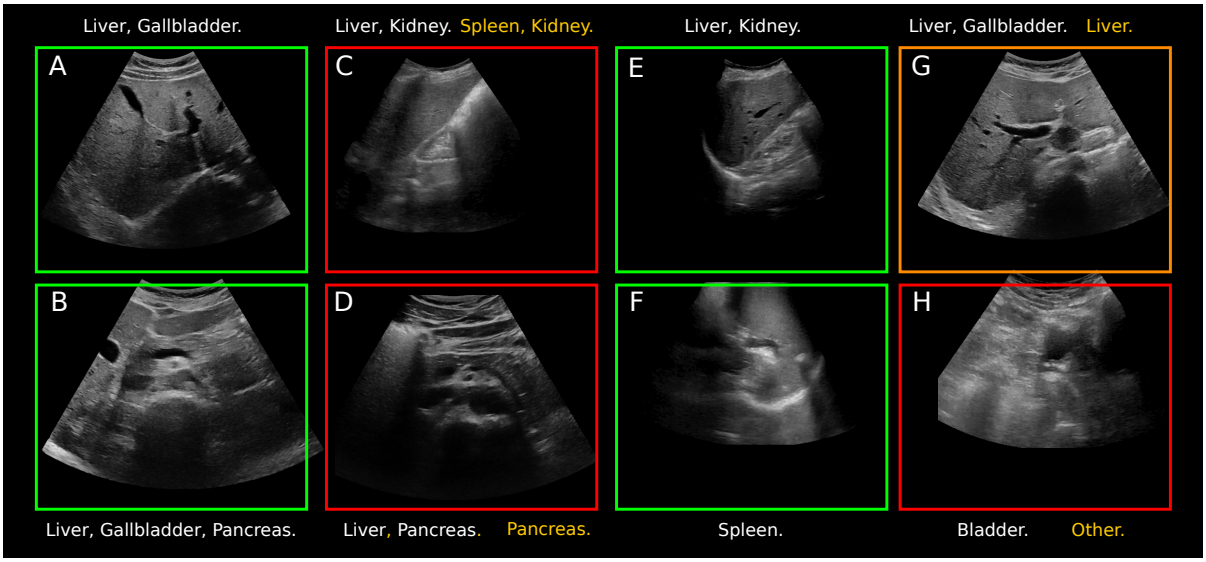


Fig. 5.8.: The images highlighted by a green and a red square represent successful and unsuccessful cases respectively. Labels in white represent the true classes assigned to the image, and yellow labels refer to the predicted classes when the classification is incomplete or (partially) incorrect. The image highlighted by an orange square represents a case where the "true" class *Gallbladder* assigned to the image is likely to be inaccurate.

self-supervised objective function.

Semi-supervised Learning:
Fig. 5.7 compares performances in terms of mean and 95% confidence intervals of F1-score weighted average for both semi-supervised methods presented in Fig. 5.4. We find that using the Two-Heads model, rather than the One-Head model, yields better results regardless of the amount of annotated data used. One possible explanation is that by having only a multi-label objective function, classes with probability below 0.5 are automatically considered negative pseudo-labels, which propagates errors during training.

Deep Clustering as Pre-training Model:
We select the weights of the best performing deep clustering model (Section 5.2.2) to initialize our supervised (Section 5.2.2) and semi-supervised (Section 5.2.3) frameworks and compare the performance with ImageNet weight's initialization. Fig. 5.7 shows these performances in terms of mean and 95% confidence intervals of F1-score weighted average. We observe that deep clustering pre-training outperforms ImageNet pre-training in all settings. An even more interesting result, is the performance difference using only 275 labeled example images (10% of the labeled set) with 69.2% vs 76.9% F1-score weighted average and 76.5% vs 79.3% in the supervised setting and semi-supervised setting respectively.

	Liver	Gallbladder	Other	Kidney	Pancreas	Spleen	Bladder	weighted avg
F1-score	0.89	0.78	0.83	0.87	0.66	0.75	0.89	0.84
Precision	0.87	0.90	0.86	0.82	0.55	0.76	0.98	0.84
Recall	0.90	0.70	0.80	0.93	0.81	0.74	0.81	0.85
n_{labels}	555	115	315	316	72	135	58	1566

Tab. 5.2.: Abdominal organ classification results for the best performing model:A two-head semi-supervised learning model pre-trained with deep clustering and trained using 2742 labelled examples images. Performance is obtained on a test set of 101 abdominal exams ($n_{images} = 1242, n_{labels} = 1566$).

Best performance is obtained with semi-supervised learning and 2742 labeled example images with an F1-score weighted average of 84.1% as shown in Table. 5.2. Fig. 5.8 showcases images classified by the best performing model with four examples of successful cases and four other cases where at least one label was misclassified or not detected. In particular, Picture (A) shows a *liver* and a *gallbladder*, (B) a *liver*, a *gallbladder* and a *pancreas*, (E) a *liver* and a *kidney*, and (F) a *spleen*. The image highlighted by an orange square (G) represents a case where the class *Gallbladder* assigned to the image is likely to be inaccurate after consulting with several experts, and where only the *liver* is visible. Images highlighted by a red square represent cases where at least one label was misclassified or not detected. Picture (D) shows a *liver* and a *pancreas*, (C) shows a *kidney* and *liver*, here the model correctly classified the *kidney* but confused the *liver* with the *spleen*, and finally (H) the model classified the image as *other* instead of *bladder*.

5.4 Discussion

In this work, we analyze the benefits of using unlabeled ultrasound data when the available labeled dataset is limited, in contrast to related studies using transfer learning where the model is retrained, after being initially trained to classify color photographs on ImageNet. We observed that deep clustering can be used as an unsupervised model for the classification of abdominal organs with reasonable performance. We further improved the performance by taking into account the possible imbalance of classes in a given batch, using a symmetric Dirichlet prior, which incidentally makes the method less sensitive to the choice of the hyperparameter λ. In addition, an extensive study was conducted to analyze how large amounts of unlabeled data could be used to improve the performance of a model trained on few labeled data and to determine the extent to which the size of the labeled data set impacts performance.
Of the three questions stated in the introduction, we addressed the first one by using the weights of the deep clustering model as initialization to a supervised

multi-label classification model, and showed that the features learned in this self-supervised manner were more useful for downstream tasks on the same domain compared to features learned on ImageNet, regardless of the amount of labeled data used. Regarding the second question, we showed that the performance obtained by leveraging labeled and unlabeled data simultaneously during training was slightly better than that obtained by using them consecutively as proposed by the self-supervised methods. As for the third question we showed that there is indeed a relevance in combining the two paradigms through deep clustering when the amount of labeled data is extremely limited. In other words, the gap between deep clustering and ImageNet pre-training in the semi-supervised setting decreases as the amount of labeled data increases, suggesting that depending on the amount of labeled data available, ImageNet pre-training in the semi-supervised setting may suffice.

Our study has several limitations: first to be able to use the deep clustering method, multi-label classification was transformed to a single-label classification. Doing so, the method cannot benefit from the potential relationships between labels as they are considered independent. Second, by adding a Dirichlet prior on the cluster size distribution, an additional hyperparameter is introduced which in our opinion should remain fixed but further investigation on the values of this hyperparameter is needed. The effect of self-supervised and semi-supervised learning for the task of organ classification in abdominal ultrasound images was evaluated on a single data set using only two methods. Thus exploration of other methods on multiple data sets is warranted. Finally, an in-depth investigation on the effect of class imbalance for semi-supervised learning is needed, the proposed method could benefit from taking into account the imbalance in the pseudo-labels generated during the training.

In summary, we provided a framework for the classification of abdominal organs in a large-scale multi-vendor ultrasound database. Specifically, and in contrast to the aforementioned related studies, we use an unrefined database with a very limited number of labeled examples. Both self-supervised and semi-supervised learning methods were explored to leverage the unlabeled data. In addition, we adapted these methods to a multi-label framework that, to our knowledge, has not been addressed before.

5.5 Conclusion

Classification of abdominal organs in large ultrasound databases is an important step for future work on US based diagnosis. Indeed, several steps are necessary to build and pre-process the database before training a model for such a task

[Dadoun, 2022b]. First the abdominal exams are selected, then amongst the images of the exam, the ones containing the organ(s) of interest are manually selected by a trained operator. All (or part) of the selected images are further annotated by a panel of experts according to the predefined task (e.g., detection of abnormalities in the kidney), and finally used to train a machine learning model. Our study can help to automate some of these processes at minimal cost. In general, US studies have the potential to speed up the democratization of access to medical imaging in developing countries where healthcare providers consider lack of training to be the main limitation to the use of US [Shah, 2015].

6 Conclusion

In this thesis we proposed a framework for the automatic analysis of abdominal US images. Specifically, we detailed preliminary solutions to address the challenges mentioned in the introduction: Lack of a shared database, presence of uncontrolled measurements and annotations inside the image, non-standardized examinations, and lack of expert annotations. In the following, we summarize our contributions and discuss remaining challenges and potential future directions.

6.1 Main Contributions

AbdoUS: A large dataset for abdominal ultrasound image analysis.
In chapter 2 we detailed the construction of a large abdominal ultrasound dataset called Abdo-US. The dataset consisted of 8011 abdominal ultrasound examinations (120 593 images) from 6482 patients (along with the corresponding radiological reports). In total, 10516 images were annotated according to a list of diseases associated with each organ, conforming to medical experts' recommendations. The dataset included labels annotated by expert caregivers, noisy labels produced by language processing models, and finally standard reference evaluation sets labeled by physicians. In addition, we included 6913 unlabeled exams ($n = 110,053$ images) from 5417 patients that could be used for unsupervised learning methods.

A preprocessing tool for fan-beam extraction on ultrasound images The preprocessing of the ultrasound images is addressed in chapter 3. Although essential to avoid a biased training set, there was a lack of tools to automatically isolate the fan-beam and remove annotations on 2D ultrasound images in JPEG/PNG format, without access to the metadata potentially present in a DICOM format. More recently a *ready-to-use* library named ITK Point-of-Care Ultrasound was developed to allow pre-processing and streaming of point-of-care ultrasound[1]. However this framework uses machine-dependent hard-coded rules to crop the region of interest, and only works if the image comes from one of the machines supported by the library. On the other hand, the method developed in chapter

[1] Point-of-care ultrasound refers to scenarios where portable ultrasound machines may be used.

3 allows for the automatic pre-processing of ultrasound image, without using any information on the machine or the transducer. The method relies on a parametric probabilistic approach to generate a training dataset with region of interest (ROI) segmentation masks. This data is then used to train a deep U-Net network to perform the same task in a supervised manner, thus considerably reducing the calculation time of the method, one hundred and sixty times faster. These images are then processed with existing inpainting methods to remove annotations present within the signal area.

Evaluation of label noise in the dataset

The annotation of ultrasound images is both time-consuming and expensive, which explains why only a part of the images in the dataset has been annotated by experts. We have shown in chapter 2 that due to the complexity and the lack of context on the ultrasound images, the annotation was subject to error and provided precise evaluation of inter-expert variability. Next, we studied the use of a hard-coded annotation tool and a natural language processing model for annotating radiological reports. We showed that automatic annotation of medical reports can also result in noisy labels being assigned to images. First, the report may include human errors, or mention anomalies not visible on the ultrasound. Second, the tools developed are not perfectly accurate. Finally, the link between the images and the report is only partial.

Detection, Localization, and Characterization of Focal Liver Lesions in Abdominal US with Deep Learning

In chapter 4, we conducted a study to train and evaluate the performance of a deep learning-based network on a specific task around ultrasound imaging, when given access to a reasonable amount of strongly labeled noise-free data, and compare its performance to that of caregivers with different levels of expertise. The detection, localization, and characterization of focal liver lesions (FLLs) in B-mode ultrasound images were chosen as the setting for this study. Indeed, the accurate detection and assessment of FLLs is a critical public health issue due to the increasing incidence of primary liver malignancies. Furthermore, previous work on this topic has focused on the diagnosis of FFLs, i.e., either the entire liver is classified as benign or malignant or the lesions are first localized by an expert and then classified. Evaluation of methods for lesion localization in the liver and specific characterization of each lesion remained to be explored. Experiments on a test set and comparison with experts showed that the DETR vision transformation network can serve as a visual attention companion for the examiner by helping him or her to focus on areas of interest. This can also provide insights to

non-expert caregivers and facilitate screening of FLLs with a potential to increase early detection of HCC in the future.

Automatic selection and organization of Abdominal US dataset around organs of interest
In chapter 5 we explored how deep clustering can be used to select ultrasound images of interest and cluster them around organs of interest in an unsupervised manner. We showed that this method may be used to reduce the annotation effort in abdominal ultrasound studies, but can also serve in a self-supervised way to pre-train an organ classification network. In particular, we showed that the deep clustering approach is an effective unsupervised model for the classification of abdominal organs with reasonable performance. We further improved the performance by taking into account the possible imbalance of classes in a given batch, using a symmetric Dirichlet prior, which incidentally makes the method less sensitive to the choice of the hyperparameter λ. Next, we adapted two state-of-the-art multi-class methods to the multi-label classification setting: deep clustering with PICA, and semi-supervised learning with FixMatch. We evaluated the use of deep clustering in self-supervised learning, and showed that the learned features transfer better to the classification task, with performance higher than that of ImageNet initialization. In particular, this pre-training method improved classification performance compared to ImageNet pre-training, regardless of the number of annotated images available for training. Finally, we showed that combining deep clustering pre-training with semi-supervised learning yields robust results, even when the number of labelled examples is extremely limited.

6.2 Future Research

The proposed methods provide a starting point for future research on the analysis of abdominal ultrasound images.

Improving the Quality of the Abdo-US Dataset
First of all, the Abdo-US dataset could be complemented by examinations from other centers to allow multi-centric clinical studies. We worked with the *Entrepôt des Données de Santé* to enrich this dataset with abdominal examinations from 25 APHP's hospitals, and the added data are presented in Appendix D. Multi-institutional data helps include relevant patient demographics and disease state, which is of high importance in medical studies to train unbiased models. Also the incorporation of video clips, often present in the examination file, would allow to evaluate the performance of the models in real time and to have access to

background images during training, in the absence of which the trained models might not be applicable in real time.

The noise in the labels of the dataset must be addressed before models can be trained to detect abnormalities in the ultrasound images. To support this claim, we show below the results of a simple classification model. The model was trained using a standard cross-entropy loss on the images of the training dataset presented in Section 2.2.1 and tested on a subset of the test set. The subset was chosen specifically to include images annotated by six raters, three of which form the adjudication panel with the majority vote used to define the ground truth. The performances of the remaining three experts were compared against the model's predictions. Expert 1, Expert 2 and Expert 3 annotated approximately 10%, 15% and 25% of the training dataset as well. One can see that the model's performance is close to that of the expert that annotated the highest number of images, with a macro-average of 67.97% for the model and 69.97% for Expert 3.

		Precision				Recall				F1-score				support
		Expert 1	Expert 2	Expert 3	Model	Expert 1	Expert 2	Expert 3	Model	Expert 1	Expert 2	Expert 3	Model	
Liver	Normal	90.09%	88.37%	80.30%	84.48%	88.50%	67.26%	93.81%	86.73%	89.29%	76.38%	86.53%	85.59%	113
	Abormal	75.00%	40.00%	57.14%	56.00%	60.00%	73.33%	40.00%	46.67%	66.67%	51.76%	47.06%	50.91%	30
Gallbladder	Normal	87.50%	84.21%	73.68%	64.00%	82.35%	94.12%	82.35%	94.12%	84.85%	88.89%	77.78%	76.19%	17
	Abormal	100.00%	92.31%	87.50%	75.00%	85.71%	85.71%	50.00%	42.86%	92.31%	88.89%	63.64%	54.55%	14
Kidneys	Normal	94.23%	92.00%	82.26%	84.48%	89.09%	83.64%	92.73%	89.09%	91.59%	87.62%	87.18%	86.73%	55
	Abormal	95.65%	88.46%	89.47%	73.08%	88.00%	92.00%	68.00%	76.00%	91.67%	90.20%	77.27%	74.51%	25
Spleen	Normal	100.00%	100.00%	78.26%	79.31%	70.00%	66.67%	60.00%	76.67%	82.35%	80.00%	67.92%	77.97%	30
	Abormal	80.00%	35.71%	40.00%	44.44%	66.67%	83.33%	33.33%	66.67%	72.73%	50.00%	36.36%	53.33%	6
micro avg		90.91%	77.74%	78.55%	77.36%	82.76%	75.86%	78.28%	78.97%	86.64%	76.79%	78.41%	78.16%	290
macro avg		90.31%	77.63%	73.58%	70.10%	78.79%	80.76%	65.03%	72.35%	83.93%	76.72%	67.97%	69.97%	290
weighted avg		90.94%	84.12%	77.98%	77.53%	82.76%	75.86%	78.28%	78.97%	86.41%	78.32%	77.19%	77.76%	290
samples avg		90.47%	75.97%	79.30%	78.60%	84.88%	74.88%	78.99%	80.31%	86.73%	74.91%	78.59%	78.50%	290

Tab. 6.1.: Performance on a test set for the task of abnormality detection in the liver, gallbladder, spleen and kidneys.

To overcome this problem several approaches are possible. The first one would be to use images annotated by more than three annotators to estimate the confidence we have on each annotator and to include this prior to weight the examples used during training. The second approach is to consider the inter-expert variability measured for each class to estimate the noise distribution and to apply class-dependent label smoothing during training. Finally, another method called Confident learning (CL) [Northcutt, 2021] relies on label quality to train a model in the presence of a noisy dataset. CL does not assume any prior knowledge about how the labels were retrieved. Instead training is done by characterizing and identifying label errors in the datasets. In particular, they rely on principles of pruning noisy data, counting with probabilistic thresholds to estimate the noise, and ranking of examples.

Joint Representation Learning from Radiological Reports and US Images
Traditionally, the learning process is done in two steps, first a pre-training is done on a large dataset of annotated natural images, with the objective of learning visual representations that allow to better separate the classes. Then we refine this representation on a smaller dataset, such as a dataset of annotated ultrasound images. In the same way that we used deep clustering to learn more relevant visual representations for the organ classification task, we could use textual descriptions associated with images to learn relevant visual representations for the abnormality detection task. To do so, we need to analyze text and image jointly, assuming access to image and text pairs, which is not exactly our case since a report describes a set of images. We have taken as an example the detection of focal liver lesions to illustrate our point. We start by detecting the sentences in the report that deal with the liver, then we detect among the images of the associated examination those that contain the liver. A visual model is trained to provide latent representations close to those of the text model when the image and the text come from the same pair, as presented in [Zhang, 2020]. The visual representations learned from this pre-training are projected in a 2D space and compared to the representations of the same model pre-trained on a dataset of labeled natural images as shown in Figure 6.1. The same process is repeated for the kidneys, and the t-SNE visualizations of encoded kidney images are shown in Figure E.3. One can see that the representations seems to better separate the classes of interest for the model pre-trained with textual descriptions. A detailed analysis can be found in Appendix E [Dadoun, 2023]. Another potential future direction is the use of multiple instance learning [Ilse, 2018] with the global labels given by the report model presented in section 2.4.3.

Learning an Ensemble of Organ-specific and/or Disease-specific Models
As discussed throughout this manuscript, the variability of abdominal ultrasound images, i.e, various organs, shapes and sizes, and overall variable ultrasound patterns makes it difficult to learn a single model for diagnostic guidance. Instead, we believe that a first step would be to learn organ-specific models on top of a global organ classification model. An interesting direction is the focus on sub-type-aware models, as proposed by [Liu, 2021]. In this framework, the model has access to class labels (e.g. normal liver or abnormal liver), and assumes that each class consists of n sub-types to be identified and taken into account in learning. The authors argue that unlabeled sub-types of a class can be very diverse and form an underlying local distribution. This also implies that two sub-types may share patterns that are not exclusive to a class, which in turn complicates class-level discrimination. In particular, these observations suggest that, when available, fine-grained labels may be useful to learn organ-specific models. For

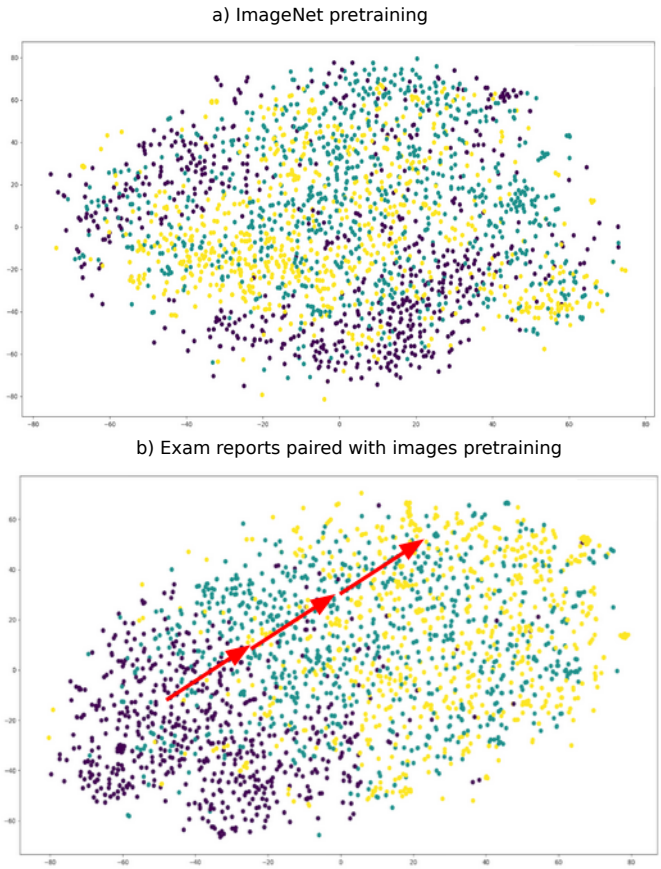


Fig. 6.1.: t-SNE visualizations of encoded image from different pre-training methods. Purple points correspond to images of a homogeneous liver, green points to images of liver with benign lesion, and yellow to images of liver with malignant lesion.

abdominal organs, the fine-grained labels may be defined based on the defined list of diseases associated with each organ, presented in Table 2.2. Alternatively unsupervised deep clustering can be used for sub-type exploration [He, 2022].

In summary, the constructed dataset and the developed methods provide sound baselines for exploring the utility of using learning methods for diagnostic assistance in abdominal ultrasound. Many challenges remain unexplored and future work is needed to develop diagnostic guidance tools. However we found that some tasks such as organ detection or classification of standard views were simpler to learn than tasks such as anomaly detection that are highly variable and challenging. Therefore, we believe that two areas would be interesting to explore for future research on this topic. First, the use of deep learning during the examination to focus the examiner's attention on areas of interest, e.g., segmenting organs in the image to reassure the person performing the ultrasound, and second, the use of deep learning as a training tool. In this sense, the study on the detection and localization of focal lesions as well as the study on organ

classification in ultrasound images for instance may serve as a training ground for caregivers interested in receiving ultrasound training.

A Appendix: Patient Re-Identification Risk Analysis

In this section we present the detailed procedure to minimize patient's re-identifiation risk. First data is transformed in order to no longer refer to a real person, and second it is generalized such that related attributes are no longer specific to a specific person but a group of people. This work is based on the recommendations of the EDPS (European Data Protection Committee)-formerly the G29 Working Party-, and on the book "Santé et Intelligence Artificielle" (Health and Artificial Intelligence) written under the direction of Bernard Nordlinger and Cédric Villani [Villani, 2018].

A.1 Data Transformation

An ultrasound examination consists of a text report and 10 to 20 ultrasound images. We transform each of these elements first by removing all identifying or almost-identifying data and metadata (last name, first name, sex, date of birth, patient identification number, examination number, date and time of examination, name of the hospital, name and first name of the physician who took care of the patient) both in the report and in the ultrasound image.

- Transformation of the images:

    1. Removing the banner identifying the image: The images are saved by default in DICOM format (.dcm), we use the metadata of each file to retrieve the coordinates of the band containing the identifying data on the image, these coordinates are used to crop the image (in order to remove the identifying band) which is then exported in JPEG and PNG formats. No metadata is kept after this operation. A random number is assigned to each exam, the images of the same examination are then named from 1 to the total number of images in the examination.

2. Removal of machine features and other annotations outside of the ultrasound area through precise segmentation of the ultrasound area. These operations are presented in Chapter 3.

3. Filling in annotations inside the ultrasound area using inpaiting methods, also presented in Chapter 3.

- Transformation of reports: Only key words, whose interest is validated by expert physicians, are kept to summarize the text of the report. The Unified Medical Language System® (UMLS®) is used to perform a named entity recognition and extract relevant key words, a detailed presentation of the tool can be found in 2.4.1. The use of keywords is intended to orient the patient according to the general presentation of the pathology (urgent or not) and to help the caregivers during screening.

A.2 Generalization of the Data

The aim of data generalization is to prevent inference attacks based on low level features. Instead data is transformed by replacing relatively low level values with higher level concepts ($\mathbf{K}$-anonymization/$\mathbf{L}$-diversity). In other words, instead of grouping images by specific pathologies (described in the sub-category column of Table 2.2), we group images by group of pathologies (described in the label column of Table 2.2). Thus we have $\mathbf{K}$-images per label and $\mathbf{L}$-labels.

We believe that the risk of correlation is very low with reasonable means for two main reasons:

1. A correlation attack would require access to the only two existing datasets containing these images. An attack by someone who already has access to the source dataset would not provide any additional information.

2. The source dataset is stored on a secure server within the Assistance Publique Hôpitaux de Paris (APHP). Access to this server, managed by the APHP's information systems department, is only possible with a unique identifier and a secure password (SSI measures). The storage is General Data Protection Regulation (GDPR) compliant, and the various actions and requests on the health data of the APHP are traced.

B

Appendix: Estimating Inter-Rater Reliability in Ill-Structured Measurement Designs (ISMDs)

Here we report the outputs for Fitting a Random Effects Model with Ratees and Raters Treated as Crossed Random Factors for each of the four organs of interest : liver, kidneys, gallbladder and spleen.

Listing B.1.: Liver: R Output for Fitting a Random Effects Model with Ratees and Raters Treated as Crossed Random Factors

```
# Linear mixed model fit by REML ['lmerMod']
# Formula: Rating ~ 1 + (1 | Ratee_ID) + (1 | Rater_ID)
# Data: tab_Foie

# REML criterion at convergence: 2524.4

# Scaled residuals:
#    Min      1Q  Median      3Q     Max
# -2.7548 -0.4472 -0.1101  0.4107  2.7051

# Random effects:
#   Groups    Name        Variance Std.Dev.
# Ratee_ID (Intercept) 0.11988  0.34623
# Rater_ID (Intercept) 0.00481  0.06936
# Residual             0.09718  0.31174
# Number of obs: 2586, groups:  Ratee_ID, 737; Rater_ID, 8

# Fixed effects:
#   Estimate Std. Error t value
# (Intercept)  0.34472    0.02863   12.04
```

Listing B.2.: Kidneys: R Output for Fitting a Random Effects Model with Ratees and Raters Treated as Crossed Random Factors

```
# Linear mixed model fit by REML ['lmerMod']
# Formula: Rating ~ 1 + (1 | Ratee_ID) + (1 | Rater_ID)
# Data: tab_Reins

# REML criterion at convergence: 667.7

# Scaled residuals:
#    Min      1Q  Median      3Q     Max
# -3.4735 -0.1603 -0.0753  0.0652  3.8275

# Random effects:
#   Groups   Name        Variance Std.Dev.
# Ratee_ID (Intercept) 0.127285 0.3568
# Rater_ID (Intercept) 0.000876 0.0296
# Residual             0.048973 0.2213
# Number of obs: 1418, groups:  Ratee_ID, 405; Rater_ID, 8

# Fixed effects:
#   Estimate Std. Error t value
# (Intercept)  0.23905    0.02177   10.98
```

Listing B.3.: Gallbladder: R Output for Fitting a Random Effects Model with Ratees and Raters Treated as Crossed Random Factors

```
# Linear mixed model fit by REML ['lmerMod']
# Formula: Rating ~ 1 + (1 | Ratee_ID) + (1 | Rater_ID)
# Data: tab_VB

# REML criterion at convergence: 271.5

# Scaled residuals:
#  Min      1Q  Median      3Q     Max
# -3.8329 -0.1127 -0.0700  0.1800  3.9254

# Random effects:
#   Groups   Name        Variance Std.Dev.
# Ratee_ID (Intercept) 0.15990  0.3999
# Rater_ID (Intercept) 0.00000  0.0000
# Residual             0.04453  0.2110
# Number of obs: 545, groups:  Ratee_ID, 175; Rater_ID, 8

# Fixed effects:
#   Estimate Std. Error t value
# (Intercept)  0.28001    0.03196    8.76
# optimizer (nloptwrap) convergence code: 0 (OK)
# boundary (singular) fit: see ?isSingular
```

Listing B.4.: Spleen: R Output for Fitting a Random Effects Model with Ratees and Raters Treated as Crossed Random Factors

```
# Linear mixed model fit by REML ['lmerMod']
# Formula: Rating ~ 1 + (1 | Ratee_ID) + (1 | Rater_ID)
# Data: tab_Rate

# REML criterion at convergence: 396.4

# Scaled residuals:
#    Min       1Q   Median       3Q      Max
# -2.70889 -0.32013 -0.09974  0.07318  3.13976

# Random effects:
#   Groups   Name        Variance Std.Dev.
# Ratee_ID (Intercept) 0.089737 0.29956
# Rater_ID (Intercept) 0.004484 0.06696
# Residual             0.071660 0.26769
# Number of obs: 550, groups:  Ratee_ID, 183; Rater_ID, 8

# Fixed effects:
#   Estimate Std. Error t value
# (Intercept)  0.21646    0.03527   6.137
```

C

Appendix: Automatic Title-Based Filtering of Abdominal Examinations

In this section we present the set of titles detected in the electronic radiological reports and detail which ones were kept in the final database.

Original Title	Translated Title	Keep Examination
Échographie rénale, vésicale et prostatique,	Renal, bladder and prostate ultrasound,	no
Echographie réno-vésico-prostatique,	Renal-vesico-Prostatic ultrasound,	no
Écho-Doppler veineux des membres inférieurs ,	Venous Doppler ultrasound of the lower limbs,	no
Echographie réno-vésicale,	Renal-vesical ultrasound,	yes
Échographie rénale et vésicale ,	Renal and bladder ultrasound ,	yes
Echographie abdominale ,	Abdominal ultrasound ,	yes
Échographie abdominale et du greffon rénal ,	Abdominal and renal graft ultrasound ,	no
Echographie rénale,	Renal ultrasound,	yes
Biopsie écho-guidée ,	Ultrasound-guided biopsy ,	no
Écho-Doppler de greffon rénal,	Ultrasound Doppler of renal graft,	no
Echo-doppler hépatique,	Hepatic Doppler ultrasound,	yes
ECHOGRAPHIE ABDOMINALE ET DU GREFFON RENAL,	ABDOMINAL AND RENAL GRAFT ULTRASOUND,	no
Échographie abdomino-pelvienne ,	Abdominal and pelvic ultrasound ,	no
ÉCHO-DOPPLER DE L'AORTE et DE SES BRANCHES,	Ultrasound Doppler of the Aorta and its branches,	no
Échographie hépato-biliaire,	Hepato-biliary ultrasound,	yes
Échographie pelvienne ,	Pelvic ultrasound ,	no
Ponction biopsie hépatique échoguidée,	Ultrasound-guided liver biopsy,	no
Échographie du système urinaire ,	Ultrasound of the urinary system ,	yes
Echographie hépatobiliaire,	Hepatobiliary ultrasound,	yes
Microbiopsie sous échographie du greffon rénal,	Microbiopsy under ultrasound of the renal graft,	no
Échographie des aires ganglionnaires axillaires ,	Ultrasound of the axillary lymph nodes,	no
ECHO-DOPPLER ABDOMINAL ,	ABDOMINAL ECHO-DOPPLER ,	yes
Repérage pour ponction pleurale,	Spotting for pleural puncture,	no
Repérage avant ponction d'ascite,	Spotting before ascites puncture,	yes
Échographie de repérage avant ponction ,	Ultrasound of location before puncture,	yes
Écho-Doppler des artères rénales ,	Ultrasound Doppler of the renal arteries ,	yes
ECHOGRAPHIE RENALE, VESICALE ET ECHO-DOPPLER SCROTAL,	RENAL AND BLADDER ULTRASOUND AND SCROTAL ECHO-DOPPLER,	no
ECHOGRAPHIE TESTICULAIRE,	TESTICULAR ULTRASOUND,	no
Examen réalisé par voies sus pubienne et endovaginale ,	Examination performed by suprapubic and endovaginal routes,	no
PONCTION D'UNE COLLECTION SOUS CUTANEE,	PUNCTURE OF A SUBCUTANEOUS COLLECTION,	no
ECHOGRAPHIE RENALE BILATERALE ET ECHOGRAPHIE DE CONTRASTE',	BILATERAL RENAL ULTRASOUND AND CONTRAST ULTRASOUND,	yes
ECHO DOPPLER AORTE ABDO,	DOPPLER ECHO OF THE ABDOMEN,	yes
ECHO-DOPPLER DE L'AORTE ABDOMINALE,	ECHO-DOPPLER OF THE ABDOMINAL AORTA,	yes
ECHO DOPPLER ABDOMINAL,	ABDOMINAL DOPPLER ECHO,	yes
Echographie abdomino-pelvien,	Abdominal and pelvic ultrasound,	no
ECHOGRAPHIE HEPATIQUE ET FIBROSCANNER,	HEPATIC ULTRASOUND AND FIBROSCANNER,	yes
Echographie de paroi abdominale ,	Ultrasound of the abdominal wall,	no
ECHOGRAPHIE-DOPPLER DES ARTERES RENALES ,	DOPPLER ULTRASOUND OF RENAL ARTERIES,	yes
BIOPSIE DU GREFFON RENAL,	RENAL GRAFT BIOPSY,	no
ECHO DOPPLER DU GREFFON RENAL,	DOPPLER ECHO OF THE RENAL GRAFT,	no
Biopsie systematique d'un greffon renal,	Systematic biopsy of a renal graft,	no
BIOPSIE HEPATIQUE,	HEPATIC BIOPSY,	no
Doppler du greffon renal,	Doppler of the renal graft,	no
Echographie des aires ganglionnaires cervicales,	Ultrasound of the cervical lymph nodes,	no
ECHO-DOPPLER DES ARTERES DES MEMBRES SUPERIEURS,	ECHO-DOPPLER OF THE ARTERIES OF THE UPPER LIMBS,	no
ECHO DOPPLER ARTERIEL DES MEMBRES SUPERIEURS,	ARTERIAL DOPPLER ULTRASOUND OF THE UPPER LIMBS,	no
ECHOGRAPHIE AXILLAIRE DROITE, INGUINALE ET POPLITEE BILATERALE,	ULTRASOUND OF THE RIGHT AXILLA, INGUINAL AND POPLITEAL AREAS,	no
Reperage d'ascite,	ascites detection,	yes
voies sus-pubienne et endo-vaginale,	suprapubic and endo-vaginal tracts,	no
ECHOGRAPHIE HEPATIQUE,	HEPATIC ULTRASOUND,	yes
ECHOGRAPHIE DES PARTIES MOLLES DU COUDE DROIT,	ULTRASOUND OF THE SOFT PARTS OF THE RIGHT ELBOW,	no
DRAINAGE ECHOGUIDE,	ECHOGUIDE DRAINAGE,	no
Doppler hepatique,	Hepatic Doppler,	yes
Echographie des voies urinaires,	Ultrasound of the urinary tract,	yes
ECHO DOPPLER RENAL,	RENAL DOPPLER ULTRASOUND,	yes
Echographie perineale,	Perineal ultrasound,	no
Echographie pelvienne,	Pelvic ultrasound,	no
Echographie vesicale,	Bladder ultrasound,	yes
Echographie de contraste,	Contrast ultrasound,	yes
ECHOGRAPHIE ABODMINALE,	ABDOMINAL ULTRASOUND,	yes
ECHOGRAPHIE ABDOMINAL,	ABDOMINAL ULTRASOUND,	yes
ECHOGRAPHIE DOPPLER ABDOMINALE,	ABDOMINAL DOPPLER ULTRASOUND,	yes
Ponction biopsie splenique echoguidee ,	Echoguided splenic biopsy,	no
ECHO-DOPPLER ARTERIEL CERVICO-ENCEPHALIQUE,	CERVICO-ENCEPHALIC ARTERIAL ECHO-DOPPLER,	no
DOPPLER PULSE ET ECHOGRAPHIE DES TRONCS SUPRA-AORTIQUES,	PULSE DOPPLER AND SUPRA-AORTIC TRUNK ULTRASOUND,	no
ECHOGRAPHIE POST BIOPSIE GREFFON RENAL,	POST RENAL GRAFT BIOPSY ULTRASOUND,	no
ECHOGRAPHIE DE PROSTATE (par voie endo-rectale),	ECHOGRAPHY OF PROSTATE (by endo-rectal way),	no
Examen realise par voies sus-pubienne et endo-vaginale,	Examination performed by suprapubic and endo-vaginal routes,	no
Echo-doppler veineux du membre inferieur gauche,	Venous Doppler ultrasound of the left lower limb,	no
ECHO-DOPPLER ARTERIEL CERVICO-ENCEPHALIQUE,	CERVICO-ENCEPHALIC ARTERIAL ECHO-DOPPLER,	no
ECHO-DOPPLER DES GREFFONS RENAL ET PANCREATIQUE,	ECHO-DOPPLER OF RENAL AND PANCREATIC GRAFTS,	no
ECHOGRAPHIE DE PROSTATE et RENO-VESICALE,	ECHOGRAPHY OF PROSTATE and RENO-VESICAL,	no
ECHOGRAPHIE INGUINALE ,	INGUINAL ULTRASOUND ,	no
Echographie des bourses,	Ultrasound of the bursa,	no
Reperage pour ponction d'un epanchement intra-abdominal,	Location for puncture of intra-abdominal effusion,	yes
Echographie prostatique,	Prostate ultrasound,	no
Echographie Endovaginale et sus pubienne,	Endovaginal and suprapubic ultrasound,	no
Echo-doppler de cuisse,	Thigh Doppler ultrasound,	no
ECHOGRAPHIE VESICO-PROSTATIQUE,	VESICO-PROSTATIC ULTRASOUND,	no
ECHOGRAPHIE RENAL, DU GREFFON ET DE LA VESSIE,	RENAL, GRAFT AND BLADDER ULTRASOUND,	no
Echodoppler cervical,	Cervical Doppler ultrasound,	no
Echographie des aires ganglionnaires axillaire et inguinale,	Ultrasound of the axillary and inguinal lymph nodes,	no

Tab. C.1.: Set of titles detected in the electronic radiological reports

D

Appendix: Multi-Centric Dataset from the *Entrepôt des Données de Santé*

In this appendix, we present the specifics related to an additional multi-center dataset built in collaboration with the *Entrepôt des Données de Santé* and the *Health Data Hub*.

D.1 Data Extraction

We retrieved all abdominal ultrasound examinations performed in 25 hospitals of Paris between 9/15/2014 to 11/12/2018 provided that the description of the examination refers to an abdominal ultrasound. In total, 23,065 examinations ($n_{images} = 290,707$) from 18,843 patients were extracted. The examinations consisted of freeze frames paired with their medical report, with a median of 13 images per examination. Of the 23,065 examinations retrieved, 98.13% corresponded to abdominal, abdominal-pelvic, or renal ultrasound, and the remainder corresponded to ultrasound biopsy, drainage, or abdominal puncture as shown in Table D.1.

D.2 Data Distribution

A detailed presentation of the distribution of examinations per year, hospital, patient's age, and gender can be found in Figure D.1. One can see that the majority

Examination Type	**# Examinations**
Abdominal, abdomino-pelvic, renal ultrasound	13521
Abdominal biopsy under ultrasound	116
Abdominal drainage under ultrasound	75
Abdominal puncture under ultrasound	66

Tab. D.1.: Distribution by type of examination

U/S Manufacturer	U/S Model	Count
Unknown	Unknown	6122
GE	LOGIQ E9	2665
	VOLUSON	288
SUPERSONIC	AIXPLORER	773
TOSHIBA	APLIO	3121
	APLIO 500	8646
	APLIO XG	568
	APLIO XV	790
	XARIO	81

Tab. D.2.: Distribution of examinations per U/S machine

of the examinations were performed in 2017 and 2018, and that examinations are unevenly distributed among hospitals, with four out of 25 hospitals accounting for 81% of examinations. Patients had a median age of 57 years old. Patient's gender is distributed almost evenly, with 46% male and 54% female.

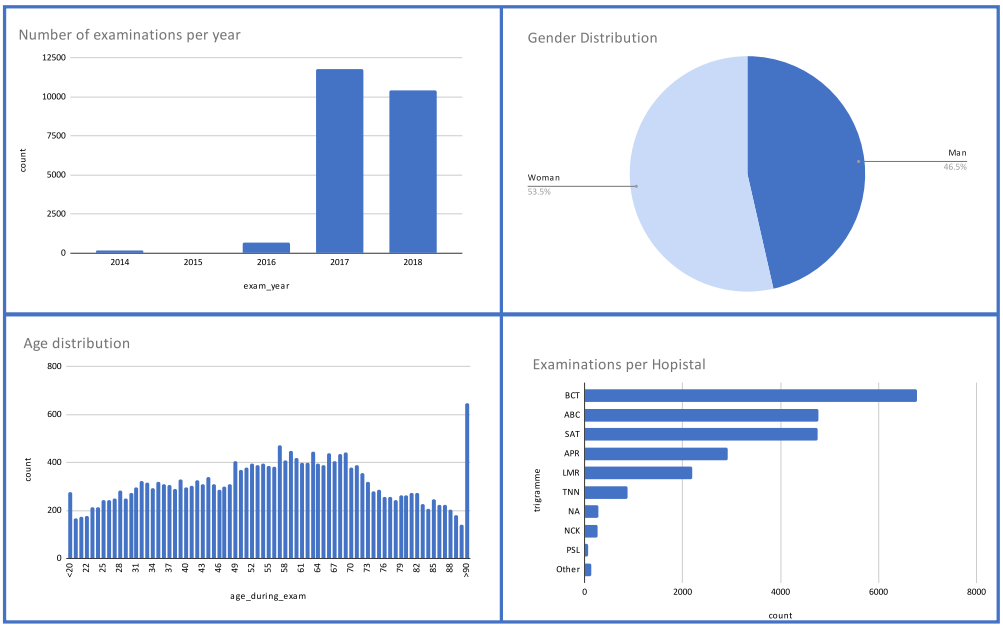


Fig. D.1.: Distribution of examinations per year, hospital, patient's age, and gender.

At least four different manufacturers are represented in the imaging data, including nine different ultrasound machines, detailed in Table D.2.

E

Appendix: Joint Representation Learning from Radiological Reports and US images

Abstract In this study, we explore the value of using a recently proposed multimodal learning method as an initialization for anomaly detection in abdominal ultrasound images. The method efficiently learns visual concepts from radiological reports using natural language supervision and constrastive learning. The underlying requirement of the method is simply the availability of image and textual descriptions pairs. However, in abdominal ultrasound examinations, radiological reports are associated with several images and describe all organs observed during the examination. To address this shortcoming, we automatically construct image and text pairs using 1) deep clustering for abdominal organ classification on ultrasound images and 2) natural language processing tools to extract the corresponding description on the report. We show that pre-training the model with these constructed pairs yields representations that better separate normal classes from abnormal ones on ultrasound images for the kidneys, compared to ImageNet-based representations, with a 10% improvement in macro-average accuracy. This work will be submitted to a conference [Dadoun, 2023].

E.1 Introduction

In this study, we focus on abnormality detection in abdominal ultrasound images by considering a binary classification task (i.e. normal vs. abnormal organ) with access to limited labeled data. Diseases associated to a given organ may alter its shape, size, contour, position, or textural appearance, resulting in highly variable differences in ultrasound patterns, all grouped into a single "abnormal" category. For this reason, transferring the model weights from ImageNet[Deng, 2009] pre-training can result in poor performance, as the features learned on natural images are not suited to capture the fine-grained visual features necessary to separate the normal class from the abnormal class. Alternatively, in the absence of large

number of annotated datasets, the model can be pre-trained on an unlabeled set of ultrasound images using self-supervised learning methods. These image-based self-supervised methods were proven to enhance performance in specific settings [Jiao, 2020; Chen, 2019; Bai, 2019]. In the case of abdominal ultrasound, we observed that for tasks with high inter-class variability, these methods can be useful. However, on a task such as pathology detection, in which there is greater similarity between classes, and more variability within classes, these methods often fail to achieve better results. More recently, a study [Zhang, 2020] proposed the use of multi-modal pre-training to learn fine-grained representations required by medical imaging tasks. They argue that medical reports, as opposed to image labels are often produced by medical experts in their routine workflow and are therefore easily accessible. The approach takes advantage of the medical reports associated to medical images, to learn better latent representations. The underlying assumption of the method is simply the presence of pairs of images and text describing the image. This method was evaluated on four different medical image classification tasks covering 2 different specialties with encouraging results. Yet, the considered hypothesis in which image pairs and a textual description are always available is not consistent with the abdominal ultrasound setting in which a textual radiology report describes a set of ultrasound images. In this work, we present a method to automatically build pairs of text descriptions and ultrasound images using deep clustering for images and named entity recognition for text. We evaluate the pre-trained image encoder on two criteria: its ability to extract discriminative features for the anomaly classification task, and its performance when fine tuning the model on a labeled set. We provide results for the normal/abnormal kidney detection task in an abdominal ultrasound examination.

E.2 Image and Text pairs generation

E.2.1 Data

During an ultrasound examination, the sonographer performs a complete scan of the area of interest and takes captures, also known as freeze frames, of the standard scanning plane views and potential visible abnormalities. The freeze-frames along with a textual documentation of the examination form the ultrasound examination report. Our dataset consisted of 8011 abdominal ultrasound examinations ($n_{images} = 120,593$) from 6482 patients with an average of 12.5 images per examination. The images are not restricted to kidneys, and can contain other abdominal organs. We show in the following, how images

containing the kidneys are selected among all images. Likewise we show how the sentences in the report that describe the kidneys are detected.

E.2.2 Data Partition

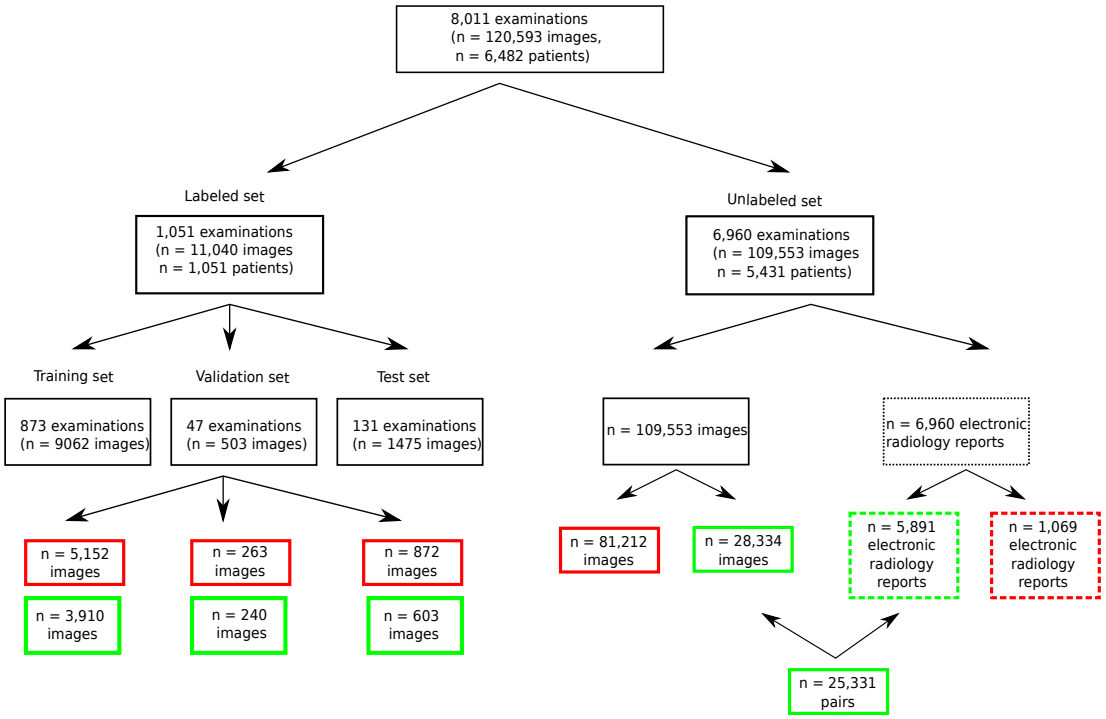


Fig. E.1.: Data partition. Green boxes are used to indicate that the images or text refer to the kidneys, red boxes are used for all other organs.

A subset of 1,051 ultrasound examinations was randomly selected to constitute the labeled sets of all freeze frames linked to the examinations. For each image the annotators had to assign a label, either normal kidneys, abnormal kidneys or "absent" if the image did not contain any kidneys. 873 of those examinations ($n_{images} = 3,910$ images of kidneys) were assigned to the training set, 47 examinations ($n_{images} = 240$ images of kidneys) to the validation set and 131 examinations ($n_{images} = 603$ images of kidneys) to the test set. The remaining 6,960 unlabeled examinations were used to construct the pre-training set. The image and text data of the pre-training set were processed in three steps which are summarized in Figure E.2. First images containing the kidneys are selected, then sentences in the medical report describing the kidneys are selected. Finally, pairs of image and text are constructed for the kidneys.

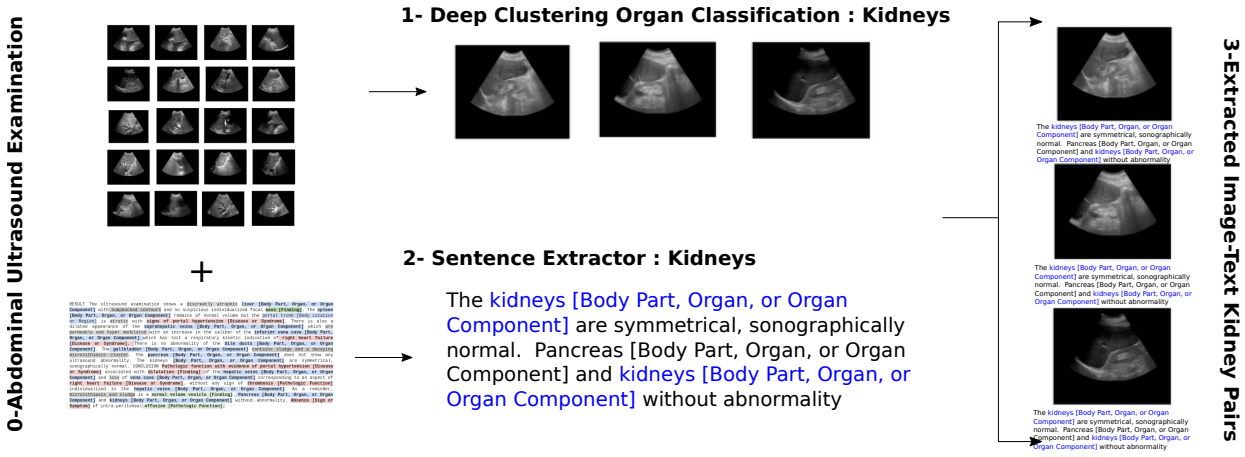


Fig. E.2.: Generation Pipeline of Image-Text Kidney Pairs for the Pre-training Set.

E.2.3 Clustering U/S Images

Since we are only interested in kidney images in this study, we use deep clustering to identify a set of clusters linked to this organ. Specifically, we use the method presented in [Dadoun, 2022a] where a framework for abdominal organ clustering using unlabeled ultrasound images is presented. Two different augmentation schemes $(\tilde{\cdot})$ and $(\hat{\cdot})$ are applied to all input images before passing through the CNN. Then a loss term is used to encourage both augmented versions of the image to be assigned to the same class with high probability. Finally, to avoid all images being assigned to the same cluster, an additional loss term is introduced to constrain the distribution of the cluster size to follow a symmetric Dirichlet distribution. This method achieves reasonable performance with an F1-score weighted average of 66.75% and an F1-score of 71.5% for the kidney's class. All images ($n = 87,696$) of our pre-training set are processed by the deep clustering method, and only images assigned to "Kidney" clusters are kept which amounts to a total 28,334 images from 6,531 examinations.

E.2.4 Text Data

To select the sentences mentioning kidneys in the medical report, we use a tagging tool based on the Unified Medical Language System (UMLS). UMLS is a meta-system that unifies concepts from several dozen terminologies in the biomedical domain. Each UMLS concept is assigned a unique concept identifier (CUI), a set of terms (or synonyms), possibly in multiple languages, and a semantic type. The labeling tool (QuickUMLS) [Soldaini, 2016], is a fast, unsupervised biomedical concept extraction tool from medical texts that works for multiple languages, including French. Given an unstructured textual medical report, we

first divide it into sentences. For each sentence, we search for a word related to the semantic type "Body part, organ or organ component" and check whether its unique concept identifier refers to kidneys. If so, the sentence is added to the list of sentences assigned to the organ. A total of 5891 examinations, out of the original 6960 examinations, mention kidneys in the report.

E.2.5 Image-Text Pairing

For each image where the abdominal organ clustering predicted the presence of a kidney, we look if the associated report mentions the kidney. If so, the sentences mentioning the kidney in the report are all associated with the image. This means that multiple images from the same examination will have the same description. Of the 28,334 images, 3003 could not be linked to any textual description (i.e. the associated examination did not mention the kidneys), which leaves us with a final dataset of 25,331 pairs of text and images from 5,513 examinations.

E.3 Joint Representation Learning

Here we detail the model architecture for the renal anomaly detection task. First, the text and image encoders are trained jointly to project the data into the same dimensional space while ensuring coherence between text and image representations. The image encoder is then fine-tuned on a labeled dataset of renal ultrasound images.

E.3.1 Architecture

We base the pre-training approach on a model (ConVIRT) developed by Zhang *et al* [Zhang, 2020]. The model is composed of a text encoder and an image encoder. For the image encoder, we use the ResNet50 architecture. The input images are randomly augmented with different data augmentations: cropping, horizontal flipping, affine transformation, color jittering and Gaussian blur before passing to the encoder. For the text encoder, we use the *CamemBERT* [Martin, 2020] model which is a state-of-the-art language model pre-trained on a French corpus *OSCAR*, based on the *RoBERTa* [Liu, 2019b] architecture. Since the text encoder was pre-trained on generic text, it is essential to consider words that are specific to the domain on which we want to refine the model (abdominal ultrasound radiological reports). To do so, we re-train a word-piece Tokenizer to find the set of words that minimize the number of tokens needed to reconstruct the reports in

our training set. Each input image and text are then converted into d-dimensional vector representation using a Linear Layer as shown in Figure **??**.

E.3.2 Pre-training Objective Function

The model is trained to predict which image goes with which description and conversely. This is achieved using two InfoNCE [Oord, 2018] losses based on the cosine similarity between the transformed image and text vectors. Let I, T be the d-dimensional vectors of image and text respectively, and $\langle I, T \rangle$ be the cosine similarity between the two, the objective function introduced in [Zhang, 2020] is as follows:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} (\lambda \cdot L_i^{I \to T} + (1 - \lambda) \cdot L_i^{T \to I})$$

where $L_i^{I \to T}$ is a text-to-image contrastive loss, whose goal is to predict I_i, T_i as the true pair among all possible descriptions and $L_i^{T \to I}$ is an image-to-text contrastive loss, whose goal is to predict I_i, T_i as the true pair among all possible images.

$$L_i^{I \leftarrow T} = -\log \frac{exp(\langle I_i, T_i \rangle / \tau)}{\sum_{j=1}^{n} exp(\langle I_i, T_j \rangle)} \quad \text{and} \quad L_i^{T \leftarrow I} = -\log \frac{exp(\langle I_i, T_i \rangle / \tau)}{\sum_{j=1}^{n} exp(\langle I_j, T_i \rangle)}$$

	Precision		Recall		F1-Score		support
	Baseline	**ConVIRT**	**Baseline**	**ConVIRT**	**Baseline**	**ConVIRT**	
Normal Kidneys	0.91	**0.92**	0.63	**0.77**	0.74	**0.84**	427
Abnormal Kidneys	0.49	**0.60**	**0.85**	0.84	0.62	**0.70**	176
accuracy					0.69	**0.79**	603
macro-average	0.70	**0.76**	0.74	**0.80**	0.68	**0.77**	603
weighted average	0.79	**0.83**	0.69	**0.79**	0.71	**0.80**	603

Tab. E.1.: Fine-tuning performance on the 603 images of the test set. The baseline initializes the model weights with those of the ImageNet pre-training, whereas ConVIRT initializes them with those of the image-text pre-training.

E.3.3 Fine-tuning Objective Function

We evaluate our pretrained image encoder on the binary classification task of abnormality detection (normal vs. abnormal kidneys). Both the CNN weights and the linear head are fine-tuned on a labeled training set of 3,910 images. We

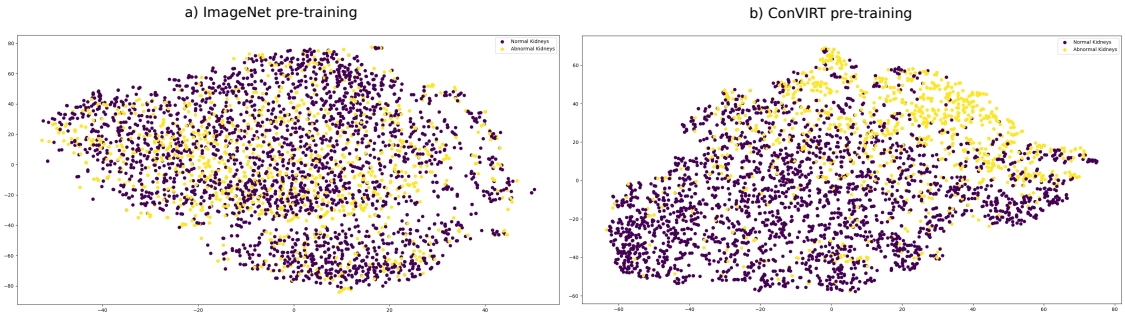


Fig. E.3.: t-SNE visualizations of encoded image from different pre-training methods. Purple points correspond to images of healthy kidneys, and yellow to images of abnormal kidneys.

use the generalized cross entropy loss introduced in [Zhang, 2018] to train deep neural networks with noisy labels.

E.4 Results

E.4.1 T-SNE Visualization of Extracted Features during Pre-training

First we evaluate how joint representation pre-training impacts the resulting image features, in comparison to the same encoder trained on ImageNet, using t-SNE visualization [Van der Maaten, 2008]. t-SNE is a stochastic method for visualizing high-dimensional data. We use the scikit-learn [Pedregosa, 2011] implementation with the default parameters. We can see in Figure E.3 that joint representation pre-training helps separate normal classes from abnormal ones in its encoding low-dimensional space.

E.4.2 Classification Results after Fine-tuning

In order to quantitatively assess the detection performance of this approach after fine-tuning, we evaluated the model on the test cohort presented in Section E.2.2. We measured the algorithm's performance using the precision and recall rates as well as the F1-score. One can see in Table E.1 that the Image-Text joint pre-traning using ConVIRT yields a 10% improvement in macro-average accuracy compared to the baseline with ImageNet pre-training. Specifically, we found that for ImageNet and ConVIRT pre-training, the negative predictive value (0.91 vs. 0.92) and sensitivity (0.85 vs. 0.84) respectively, were similar. On the other

hand, the positive predictive value (0.49 vs. 0.60) and specificity (0.63 vs. 0.77) respectively, were both higher for ConVIRT compared to ImageNet pre-training.

E.5 Conclusion

In this study we explored the value of using unstructured radiological reports to pre-train a model to better separate normal and abnormal kidney ultrasound images. Although a direct link between images and their descriptions is not provided in abdominal ultrasound examinations, we were able to build pairs of images and text using different unsupervised methods. Finally, we showed that this matching strategy, combined with conVIRT pre-training, provided a 10% increase in accuracy during fine-tuning compared to ImageNet pre-training.

Bibliography

[Almajalid, 2018] Rania Almajalid, Juan Shan, Yaodong Du, and Ming Zhang. "Development of a deep-learning-based method for breast ultrasound image segmentation". In: *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE. 2018, pp. 1103–1108 (cit. on p. 4).

[Alsharid, 2022] Mohammad Alsharid, Harshita Sharma, Lior Drukker, Aris T Papageorgiou, and J Alison Noble. "Weakly Supervised Captioning of Ultrasound Images". In: *Annual Conference on Medical Image Understanding and Analysis*. Springer. 2022, pp. 187–198 (cit. on p. 3).

[Andrews, 2000] Matthew W Andrews. "Ultrasound of the spleen". In: *World journal of surgery* 24.2 (2000), pp. 183–187 (cit. on p. 2).

[Bai, 2019] Wenjia Bai, Chen Chen, Giacomo Tarroni, Jinming Duan, Florian Guitton, Steffen E Petersen, Yike Guo, Paul M Matthews, and Daniel Rueckert. "Self-supervised learning for cardiac mr image segmentation by anatomical position prediction". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 541–549 (cit. on p. 104).

[Bajaj, 2021] Retesh Bajaj, Xingru Huang, Yakup Kilic, Ajay Jain, Anantharaman Ramasamy, Ryo Torii, James Moon, Tat Koh, Tom Crake, Maurizio K Parker, et al. "A deep learning methodology for the automated detection of end-diastolic frames in intravascular ultrasound images". In: *The international journal of cardiovascular imaging* 37.6 (2021), pp. 1825–1837 (cit. on p. 4).

[Baloescu, 2020] Cristiana Baloescu, Grzegorz Toporek, Seungsoo Kim, Katelyn McNamara, Rachel Liu, Melissa M Shaw, Robert L McNamara, Balasundar I Raju, and Christopher L Moore. "Automated lung ultrasound B-line assessment using a deep learning algorithm". In: *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* 67.11 (2020), pp. 2312–2320 (cit. on p. 4).

[Bardes, 2021] Adrien Bardes, Jean Ponce, and Yann LeCun. "Vicreg: Variance-invariance-covariance regularization for self-supervised learning". In: *arXiv preprint arXiv:2105.04906* (2021) (cit. on p. 66).

[Baumgartner, 2017] Christian F Baumgartner, Konstantinos Kamnitsas, Jacqueline Matthew, Tara P Fletcher, Sandra Smith, Lisa M Koch, Bernhard Kainz, and Daniel Rueckert. "SonoNet: real-time detection and localisation of fetal standard scan planes in freehand ultrasound". In: *IEEE transactions on medical imaging* 36.11 (2017), pp. 2204–2215 (cit. on pp. 4, 37).

[Bejnordi, 2017] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al. "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer". In: *Jama* 318.22 (2017), pp. 2199–2210 (cit. on p. 10).

[Bimbraw, 2020] Keshav Bimbraw, Xihan Ma, Ziming Zhang, and Haichong Zhang. "Augmented Reality-Based Lung Ultrasound Scanning Guidance". In: *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis*. Springer, 2020, pp. 106–115 (cit. on p. 3).

[Bonmati, 2021] E Bonmati, Y Hu, A Grimwood, GJ Johnson, G Goodchild, MG Keane, K Gurusamy, B Davidson, MJ Clarkson, SP Pereira, et al. "Voice-assisted Image Labelling for Endoscopic Ultrasound Classification using Neural Networks." In: *IEEE Transactions on Medical Imaging* (2021) (cit. on p. 66).

[Bray, 2018] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries". In: *CA: a cancer journal for clinicians* 68.6 (2018), pp. 394–424 (cit. on p. 46).

[Buades, 2005] Antoni Buades, Bartomeu Coll, and J-M Morel. "A non-local algorithm for image denoising". In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Vol. 2. IEEE. 2005, pp. 60–65 (cit. on p. 43).

[Byrd, 1995] Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. "A limited memory algorithm for bound constrained optimization". In: *SIAM Journal on scientific computing* 16.5 (1995), pp. 1190–1208 (cit. on p. 40).

[Cadier, 2017] Benjamin Cadier, Julie Bulsei, Pierre Nahon, Olivier Seror, Alexis Laurent, Isabelle Rosa, Richard Layese, Charlotte Costentin, Carole Cagnot, Isabelle Durand-Zaleski, et al. "Early detection and curative treatment of hepatocellular carcinoma: a cost-effectiveness analysis in France and in the United States". In: *Hepatology* 65.4 (2017), pp. 1237–1248 (cit. on p. 48).

[Carion, 2020] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. "End-to-end object detection with transformers". In: *European Conference on Computer Vision*. Springer. 2020, pp. 213–229 (cit. on p. 51).

[Caron, 2018] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. "Deep clustering for unsupervised learning of visual features". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 132–149 (cit. on pp. 67, 68).

[Chapelle, 2009] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. "Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]". In: *IEEE Transactions on Neural Networks* 20.3 (2009), pp. 542–542 (cit. on p. 67).

[Chen, 2019] Liang Chen, Paul Bentley, Kensaku Mori, Kazunari Misawa, Michitaka Fujiwara, and Daniel Rueckert. "Self-supervised learning for medical image analysis using image context restoration". In: *Medical image analysis* 58 (2019), p. 101539 (cit. on p. 104).

[Cheng, 2017] Phillip M Cheng and Harshawn S Malhi. "Transfer learning with convolutional neural networks for classification of abdominal ultrasound images". In: *Journal of digital imaging* 30.2 (2017), pp. 234–243 (cit. on p. 64).

[Choi, 2011] Brian G Choi, Monica Mukherjee, Praveen Dala, Heather A Young, Cynthia M Tracy, Richard J Katz, and Jannet F Lewis. "Interpretation of remotely downloaded pocket-size cardiac ultrasound images on a web-enabled smartphone: validation against workstation evaluation". In: *Journal of the American Society of Echocardiography* 24.12 (2011), pp. 1325–1330 (cit. on p. 36).

[Chou, 2021] Tsung-Hsien Chou, Hsing-Jung Yeh, Chun-Chao Chang, Jui-Hsiang Tang, Wei-Yu Kao, I-Chia Su, Chien-Hung Li, Wei-Hao Chang, Chun-Kai Huang, Herdiantri Sufriyana, et al. "Deep learning for abdominal ultrasound: A computer-aided diagnostic system for the severity of fatty liver". In: *Journal of the Chinese Medical Association* 84.9 (2021), pp. 842–850 (cit. on p. 4).

[Craig, 2020] Amanda J Craig, Johann Von Felden, Teresa Garcia-Lezana, Samantha Sarcognato, and Augusto Villanueva. "Tumour evolution in hepatocellular carcinoma". In: *Nature reviews Gastroenterology & hepatology* 17.3 (2020), pp. 139–152 (cit. on p. 47).

[Dadoun, 2021] Hind Dadoun, Hervé Delingette, Anne-Laure Rousseau, Eric de Kerviler, and Nicholas Ayache. "Combining Bayesian and Deep Learning Methods for the Delineation of the Fan in Ultrasound Images". In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2021, pp. 743–747 (cit. on pp. 6, 7, 36, 50).

[Dadoun, 2022a] Hind Dadoun, Hervé Delingette, Anne-Laure Rousseau, Eric de Kerviler, and Nicholas Ayache. "Deep Clustering for Abdominal Organ Classification in US imaging". preprint submitted to a journal. 2022 (cit. on pp. 6, 7, 64, 106).

[Dadoun, 2022b] Hind Dadoun, Anne-Laure Rousseau, Eric de Kerviler, Jean-Michel Correas, Anne-Marie Tissier, Fanny Joujou, Sylvain Bodard, Kemel Khezzane, Constance de Margerie-Mellon, Hervé Delingette, et al. "Detection, Localization, and Characterization of Focal Liver Lesions in Abdominal US with Deep Learning". In: *Radiology: Artificial Intelligence* (2022) (cit. on pp. 6, 7, 46, 83).

[Dadoun, 2023] Hind Dadoun, Hervé Delingette, Anne-Laure Rousseau, Eric de Kerviler, and Nicholas Ayache. "Joint Representation Learning from Radiological Reports and Ultrasound images". preprint submitted to a conference. 2023 (cit. on pp. 7, 89, 103).

[Dahdouh, 2015] Sonia Dahdouh, Elsa D Angelini, Gilles Grangé, and Isabelle Bloch. "Segmentation of embryonic and fetal 3D ultrasound images based on pixel intensity distributions and shape priors". In: *Medical image analysis* 24.1 (2015), pp. 255–268 (cit. on p. 4).

[Dehaene, 2020] Olivier Dehaene, Axel Camara, Olivier Moindrot, Axel de Lavergne, and Pierre Courtiol. "Self-Supervision Closes the Gap Between Weak and Strong Supervision in Histology". In: *arXiv preprint arXiv:2012.03583* (2020) (cit. on p. 61).

[Deng, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255 (cit. on p. 103).

[Dosovitskiy, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020) (cit. on p. 51).

[Dove, 2018] Edward S Dove. "The EU general data protection regulation: implications for international scientific research in the digital era". In: *Journal of Law, Medicine & Ethics* 46.4 (2018), pp. 1013–1030 (cit. on p. 10).

[Droste, 2020] Richard Droste, Lior Drukker, Aris T Papageorghiou, and J Alison Noble. "Automatic Probe Movement Guidance for Freehand Obstetric Ultrasound". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2020, pp. 583–592 (cit. on p. 37).

[Esteva, 2017] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. "Dermatologist-level classification of skin cancer with deep neural networks". In: *nature* 542.7639 (2017), pp. 115–118 (cit. on p. 10).

[Fleiss, 1971] Joseph L Fleiss. "Measuring nominal scale agreement among many raters." In: *Psychological bulletin* 76.5 (1971), p. 378 (cit. on p. 24).

[George, 2022] Mino George and HB Anita. "Analysis of kidney ultrasound images using deep learning and machine learning techniques: A review". In: *Pervasive Computing and Social Networking* (2022), pp. 183–199 (cit. on p. 4).

[Goyal, 2017] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. "Accurate, large minibatch sgd: Training imagenet in 1 hour". In: *arXiv preprint arXiv:1706.02677* (2017) (cit. on p. 51).

[Hanna, 2016] Robert F Hanna, Vesselin Z Miloushev, An Tang, Lee A Finklestone, Sidney Z Brejt, Ranjit S Sandhu, Cynthia S Santillan, Tanya Wolfson, Anthony Gamst, and Claude B Sirlin. "Comparative 13-year meta-analysis of the sensitivity and positive predictive value of ultrasound, CT, and MRI for detecting hepatocellular carcinoma". In: *Abdominal radiology* 41.1 (2016), pp. 71–90 (cit. on p. 60).

[Harvey, 2001] Christopher J Harvey and Thomas Albrecht. "Ultrasound of focal liver lesions". In: *European radiology* 11.9 (2001), pp. 1578–1593 (cit. on p. 61).

[He, 2022] Ju He, Jie-Neng Chen, Shuai Liu, Adam Kortylewski, Cheng Yang, Yutong Bai, and Changhu Wang. "Transfg: A transformer architecture for fine-grained recognition". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 1. 2022, pp. 852–860 (cit. on p. 90).

[Heuvel, 2019] Thomas LA van den Heuvel, Hezkiel Petros, Stefano Santini, Chris L de Korte, and Bram van Ginneken. "Automated fetal head detection and circumference estimation from free-hand ultrasound sweeps using deep learning in resource-limited countries". In: *Ultrasound in medicine & biology* 45.3 (2019), pp. 773–785 (cit. on pp. 3, 37).

[Huang, 2020] Jiabo Huang, Shaogang Gong, and Xiatian Zhu. "Deep semantic clustering by partition confidence maximisation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 8849–8858 (cit. on pp. 68, 69).

[Ilse, 2018] Maximilian Ilse, Jakub Tomczak, and Max Welling. "Attention-based deep multiple instance learning". In: *International conference on machine learning*. PMLR. 2018, pp. 2127–2136 (cit. on p. 89).

[Irvin, 2019] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01. 2019, pp. 590–597 (cit. on p. 10).

[Iwasa, 2021] Yuhei Iwasa, Takuji Iwashita, Yuji Takeuchi, Hironao Ichikawa, Naoki Mita, Shinya Uemura, Masahito Shimizu, Yu-Ting Kuo, Hsiu-Po Wang, and Takeshi Hara. "Automatic segmentation of pancreatic tumors using deep learning on a video image of contrast-enhanced endoscopic ultrasound". In: *Journal of clinical medicine* 10.16 (2021), p. 3589 (cit. on p. 4).

[Izranov, 2019] V Izranov, U Palvanova, V Gordova, S Perepelitsa, and S Morozov. "Ultrasound criteria of splenomegaly". In: *The Radiologist* 1.1002 (2019), pp. 3–6 (cit. on p. 2).

[Jang, 2014] Timothy Jang, Vijai Chauhan, Christopher Cundiff, and Amy H Kaji. "Assessment of emergency physician–performed ultrasound in evaluating nonspecific abdominal pain". In: *The American journal of emergency medicine* 32.5 (2014), pp. 457–460 (cit. on p. 2).

[Jang, 2021] Sung Ill Jang, Young Jae Kim, Eui Joo Kim, Huapyong Kang, Seung Jin Shon, Yu Jin Seol, Dong Ki Lee, Kwang Gi Kim, and Jae Hee Cho. "Diagnostic performance of endoscopic ultrasound-artificial intelligence using deep learning analysis of gallbladder polypoid lesions". In: *Journal of Gastroenterology and Hepatology* 36.12 (2021), pp. 3548–3555 (cit. on p. 4).

[Jiao, 2020] Jianbo Jiao, Richard Droste, Lior Drukker, Aris T Papageorghiou, and J Alison Noble. "Self-supervised representation learning for ultrasound video". In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2020, pp. 1847–1850 (cit. on pp. 66, 67, 104).

[Jing, 2020] Longlong Jing and Yingli Tian. "Self-supervised visual feature learning with deep neural networks: A survey". In: *IEEE transactions on pattern analysis and machine intelligence* 43.11 (2020), pp. 4037–4058 (cit. on p. 66).

[Kameda, 2018] Toru Kameda, Kumiko Uebayashi, Kazuko Wagai, Fukiko Kawai, and Nobuyuki Taniguchi. "Assessment of the renal collecting system using a pocket-sized ultrasound device". In: *Journal of Medical Ultrasonics* 45.4 (2018), pp. 577–581 (cit. on p. 2).

[Kart, 2021] Turkay Kart, Wenjia Bai, Ben Glocker, and Daniel Rueckert. "DeepMCAT: Large-Scale Deep Clustering for Medical Image Categorization". In: *Deep Generative Models, and Data Augmentation, Labelling, and Imperfections*. Springer, 2021, pp. 259–267 (cit. on p. 68).

[Kingma, 2014] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014) (cit. on p. 41).

[Korbar, 2018] Bruno Korbar, Du Tran, and Lorenzo Torresani. "Cooperative learning of audio and video models from self-supervised synchronization". In: *Advances in Neural Information Processing Systems* 31 (2018) (cit. on p. 66).

[Koundal, 2018] Deepika Koundal, Savita Gupta, and Sukhwinder Singh. "Computer aided thyroid nodule detection system using medical ultrasound images". In: *Biomedical Signal Processing and Control* 40 (2018), pp. 117–130 (cit. on p. 4).

[Laine, 2016] Samuli Laine and Timo Aila. "Temporal ensembling for semi-supervised learning". In: *arXiv preprint arXiv:1610.02242* (2016) (cit. on p. 67).

[Lee, 2013] Dong-Hyun Lee et al. "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks". In: *Workshop on challenges in representation learning, ICML*. Vol. 3. 2. 2013, p. 896 (cit. on p. 67).

[Lee, 2020] Lok Hin Lee, Elizabeth Bradburn, Aris T Papageorghiou, and J Alison Noble. “Calibrated bayesian neural networks to estimate gestational age and its uncertainty on fetal brain ultrasound images”. In: *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis*. Springer, 2020, pp. 13–22 (cit. on p. 3).

[Li, 2021] Keyu Li, Yangxin Xu, Ziqi Zhao, and Max Q-H Meng. “Automatic Recognition of Abdominal Organs in Ultrasound Images based on Deep Neural Networks and K-Nearest-Neighbor Classification”. In: *2021 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE. 2021, pp. 1980–1985 (cit. on p. 65).

[Liebo, 2011] Max J Liebo, Rachel L Israel, Elizabeth O Lillie, Michael R Smith, David S Rubenson, and Eric J Topol. “Is pocket mobile echocardiography the next-generation stethoscope? A cross-sectional comparison of rapidly acquired images with standard transthoracic echocardiography”. In: *Annals of internal medicine* 155.1 (2011), pp. 33–38 (cit. on p. 36).

[Lindelius, 2008] A Lindelius, S Törngren, A Sondén, H Pettersson, and J Adami. “Impact of surgeon-performed ultrasound on diagnosis of abdominal pain”. In: *Emergency Medicine Journal* 25.8 (2008), pp. 486–491 (cit. on p. 2).

[Lindelius, 2009] A Lindelius, S Törngren, H Pettersson, and J Adami. “Role of surgeon-performed ultrasound on further management of patients with acute abdominal pain: a randomised controlled clinical trial”. In: *Emergency Medicine Journal* 26.8 (2009), pp. 561–566 (cit. on p. 2).

[Liu, 2019a] Shengfeng Liu, Yi Wang, Xin Yang, Baiying Lei, Li Liu, Shawn Xiang Li, Dong Ni, and Tianfu Wang. “Deep learning in medical ultrasound analysis: a review”. In: *Engineering* 5.2 (2019), pp. 261–275 (cit. on p. 10).

[Liu, 2019b] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. “Roberta: A robustly optimized bert pretraining approach”. In: *arXiv preprint arXiv:1907.11692* (2019) (cit. on pp. 30, 31, 107).

[Liu, 2021] Xiaofeng Liu, Xiongchang Liu, Bo Hu, Wenxuan Ji, Fangxu Xing, Jun Lu, Jane You, C-C Jay Kuo, Georges El Fakhri, and Jonghye Woo. “Subtype-aware unsupervised domain adaptation for medical diagnosis”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 3. 2021, pp. 2189–2197 (cit. on p. 89).

[Liu, 2022] Bin Liu, Konstantinos Blekas, and Grigorios Tsoumakas. “Multi-label sampling based on local label imbalance”. In: *Pattern Recognition* 122 (2022), p. 108294 (cit. on p. 22).

[Luijten, 2019] Ben Luijten, Regev Cohen, Frederik J de Bruijn, Harold AW Schmeitz, Massimo Mischi, Yonina C Eldar, and Ruud JG van Sloun. “Deep learning for fast adaptive beamforming”. In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 1333–1337 (cit. on p. 3).

[Marrero, 2014] Jorge A Marrero, Joseph Ahn, Rajender K Reddy, Practice Parameters Committee of the American College of Gastroenterology, et al. “ACG clinical guideline: the diagnosis and management of focal liver lesions”. In: *Official journal of the American College of Gastroenterology| ACG* 109.9 (2014), pp. 1328–1347 (cit. on p. 47).

[Martin Bland, 1995] J Martin Bland and G Douglas Altman. “Statistics notes: Multiple significance tests: the Bonferronni method”. In: *Br Med J* (1995) (cit. on p. 53).

[Martin, 2019] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, and Benoıt Sagot. “CamemBERT: a tasty French language model”. In: *arXiv preprint arXiv:1911.03894* (2019) (cit. on p. 30).

[Martin, 2020] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoıt Sagot. “CamemBERT: a Tasty French Language Model”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 7203–7219 (cit. on p. 107).

[Matthew, 2022] Jacqueline Matthew, Emily Skelton, Thomas G Day, Veronika A Zimmer, Alberto Gomez, Gavin Wheeler, Nicolas Toussaint, Tianrui Liu, Samuel Budd, Karen Lloyd, et al. “Exploring a new paradigm for the fetal anomaly ultrasound scan: Artificial intelligence in real time”. In: *Prenatal Diagnosis* 42.1 (2022), pp. 49–59 (cit. on p. 3).

[McLachlan, 1975] Geoffrey J McLachlan. “Iterative reclassification procedure for constructing an asymptotically optimal rule of allocation in discriminant analysis”. In: *Journal of the American Statistical Association* 70.350 (1975), pp. 365–369 (cit. on p. 67).

[Meng, 2019] Qingjie Meng, Matthew Sinclair, Veronika Zimmer, Benjamin Hou, Martin Rajchl, Nicolas Toussaint, Ozan Oktay, Jo Schlemper, Alberto Gomez, James Housden, et al. “Weakly supervised estimation of shadow confidence maps in fetal ultrasound imaging”. In: *IEEE transactions on medical imaging* 38.12 (2019), pp. 2755–2767 (cit. on p. 3).

[Menze, 2014] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. “The multimodal brain tumor image segmentation benchmark (BRATS)”. In: *IEEE transactions on medical imaging* 34.10 (2014), pp. 1993–2024 (cit. on p. 10).

[Mjolstad, 2012] Ole Christian Mjolstad, Havard Dalen, Torbjorn Graven, Jens Olaf Kleinau, Oyvind Salvesen, and Bjorn Olav Haugen. “Routinely adding ultrasound examinations by pocket-sized ultrasound devices improves inpatient diagnostics in a medical department”. In: *European Journal of Internal Medicine* 23.2 (2012), pp. 185–191 (cit. on p. 2).

[Moore, 2011] Christopher L Moore and Joshua A Copel. “Point-of-care ultrasonography”. In: *New England Journal of Medicine* 364.8 (2011), pp. 749–757 (cit. on pp. 3, 36).

[Morgan, 2018] Tara A Morgan, Katherine E Maturen, Nirvikar Dahiya, Maryellen RM Sun, and Aya Kamaya. "US LI-RADS: ultrasound liver imaging reporting and data system for screening and surveillance of hepatocellular carcinoma". In: *Abdominal Radiology* 43.1 (2018), pp. 41–55 (cit. on p. 2).

[Northcutt, 2021] Curtis Northcutt, Lu Jiang, and Isaac Chuang. "Confident learning: Estimating uncertainty in dataset labels". In: *Journal of Artificial Intelligence Research* 70 (2021), pp. 1373–1411 (cit. on p. 88).

[Oksuz, 2018] Kemal Oksuz, Baris Can Cam, Emre Akbas, and Sinan Kalkan. "Localization recall precision (LRP): A new performance metric for object detection". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 504–519 (cit. on p. 52).

[Oksuz, 2020] Kemal Oksuz, Baris Can Cam, Emre Akbas, and Sinan Kalkan. "A ranking-based, balanced loss function unifying classification and localisation in object detection". In: *arXiv preprint arXiv:2009.13592* (2020) (cit. on p. 52).

[Oord, 2018] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding". In: *arXiv preprint arXiv:1807.03748* (2018) (cit. on p. 108).

[Organization, 2011] World Health Organization et al. *First WHO global forum on medical devices: context, outcomes, and future actions*. Tech. rep. World Health Organization, 2011 (cit. on p. 1).

[Park, 2013] Hana Park, Jun Yong Park, Sang Hoon Ahn Do Young Kim, Chae Yoon Chon, Kwang-Hyub Han, and Seung Up Kim. "Characterization of focal liver masses using acoustic radiation force impulse elastography". In: *World journal of gastroenterology: WJG* 19.2 (2013), p. 219 (cit. on p. 60).

[Pathak, 2016] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. "Context encoders: Feature learning by inpainting". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2536–2544 (cit. on p. 66).

[Pedregosa, 2011] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. "Scikit-learn: Machine learning in Python". In: *the Journal of Machine Learning Research* 12 (2011), pp. 2825–2830 (cit. on p. 109).

[Putka, 2008] Dan J Putka, Huy Le, Rodney A McCloy, and Tirso Diaz. "Ill-structured measurement designs in organizational research: Implications for estimating interrater reliability." In: *Journal of Applied Psychology* 93.5 (2008), p. 959 (cit. on pp. 22, 23).

[Radford, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. "Learning transferable visual models from natural language supervision". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 8748–8763 (cit. on p. 33).

[Radiology ESR, 2020] European Society of Radiology (ESR). “Position statement and best practice recommendations on the imaging use of ultrasound from the European Society of Radiology ultrasound subcommittee”. In: *Insights into Imaging* 11.1 (2020), p. 115 (cit. on p. 3).

[Randen, 2008] Adrienne van Randen, Shandra Bipat, Aeilko H Zwinderman, Dirk T Ubbink, Jaap Stoker, and Marja A Boermeester. “Acute appendicitis: meta-analysis of diagnostic performance of CT and graded compression US related to prevalence of disease”. In: *Database of Abstracts of Reviews of Effects (DARE): Quality-assessed Reviews [Internet]*. Centre for Reviews and Dissemination (UK), 2008 (cit. on p. 2).

[Ren, 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. “Faster r-cnn: Towards real-time object detection with region proposal networks”. In: *Advances in neural information processing systems* 28 (2015), pp. 91–99 (cit. on p. 51).

[Ren, 2018] Zhongzheng Ren and Yong Jae Lee. “Cross-domain self-supervised multi-task feature learning using synthetic imagery”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 762–771 (cit. on p. 66).

[Rezatofighi, 2019] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. “Generalized intersection over union: A metric and a loss for bounding box regression”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 658–666 (cit. on p. 52).

[Rizve, 2021] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. “In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning”. In: *arXiv preprint arXiv:2101.06329* (2021) (cit. on p. 74).

[Sajjadi, 2016] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. “Regularization with stochastic transformations and perturbations for deep semi-supervised learning”. In: *Advances in neural information processing systems* 29 (2016) (cit. on p. 67).

[Salvadeo, 2014] Denis HP Salvadeo, Isabelle Bloch, Florence Tupin, Nelson DA Mascarenhas, Alexandre LM Levada, Charles-Alban Deledalle, and Sonia Dahdouh. “Denoising based on non local means for ultrasound images with simultaneous multiple noise distributions”. In: *2014 IEEE International Conference on Image Processing (ICIP)*. Ieee. 2014, pp. 2699–2703 (cit. on p. 3).

[Schmauch, 2019] B Schmauch, P Herent, P Jehanno, O Dehaene, C Saillard, Christophe Aubé, Alain Luciani, N Lassau, and S Jégou. “Diagnosis of focal liver lesions from ultrasound using deep learning”. In: *Diagnostic and interventional imaging* 100.4 (2019), pp. 227–233 (cit. on pp. 48, 60).

[Schnittke, 2019] Nikolai Schnittke and Sara Damewood. “Identifying and overcoming barriers to resident use of point-of-care ultrasound”. In: *Western Journal of Emergency Medicine* 20.6 (2019), p. 918 (cit. on p. 3).

[Shah, 2015] Sachita Shah, Blaise A Bellows, Adeyinka A Adedipe, Jodie E Totten, Brandon H Backlund, and Dana Sajed. "Perceived barriers in the use of ultrasound in developing countries". In: *Critical ultrasound journal* 7.1 (2015), pp. 1–5 (cit. on pp. 3, 47, 83).

[Sidhu, 2022] Paul S Sidhu and Vasileios Rafailidis. *Incidentally Detected Gallbladder Polyps at US: Myths and Truths*. 2022 (cit. on p. 2).

[Sloun, 2019] Ruud JG van Sloun, Regev Cohen, and Yonina C Eldar. "Deep learning in ultrasound imaging". In: *Proceedings of the IEEE* 108.1 (2019), pp. 11–29 (cit. on p. 37).

[Sohn, 2020] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. "Fixmatch: Simplifying semi-supervised learning with consistency and confidence". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 596–608 (cit. on pp. 68, 74).

[Soldaini, 2016] Luca Soldaini and Nazli Goharian. "Quickumls: a fast, unsupervised approach for medical concept extraction". In: *MedIR workshop, sigir*. 2016, pp. 1–4 (cit. on pp. 26, 106).

[Strauss, 2007] Simon Strauss, Ella Gavish, Paul Gottlieb, and Ludmila Katsnelson. "Interobserver and intraobserver variability in the sonographic assessment of fatty liver". In: *American Journal of Roentgenology* 189.6 (2007), W320–W323 (cit. on p. 11).

[Su, 2021] Jong-Chyi Su, Zezhou Cheng, and Subhransu Maji. "A realistic evaluation of semi-supervised learning for fine-grained classification". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 12966–12975 (cit. on p. 67).

[Sutskever, 2013] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. "On the importance of initialization and momentum in deep learning". In: *International conference on machine learning*. PMLR. 2013, pp. 1139–1147 (cit. on p. 51).

[Ta, 2018] Casey N Ta, Yuko Kono, Mohammad Eghtedari, Young Taik Oh, Michelle L Robbin, Richard G Barr, Andrew C Kummel, and Robert F Mattrey. "Focal liver lesions: computer-aided diagnosis by using contrast-enhanced US cine recordings". In: *Radiology* 286.3 (2018), pp. 1062–1071 (cit. on pp. 48, 60).

[Tan, 2019] Jeremy Tan, Anselm Au, Qingjie Meng, and Bernhard Kainz. "Semi-supervised learning of fetal anatomy from ultrasound". In: *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*. Springer, 2019, pp. 157–164 (cit. on p. 67).

[Telea, 2004] Alexandru Telea. "An image inpainting technique based on the fast marching method". In: *Journal of graphics tools* 9.1 (2004), pp. 23–34 (cit. on pp. 37, 43).

[Tempkin, 1999] Betty Bates Tempkin, Kristin Dykstra-Downey, and Felicia M Terry. *Ultrasound scanning: principles and protocols*. WB Saunders Company, 1999 (cit. on p. 11).

[Trinchet, 2009] JC Trinchet. “Hepatocellular carcinoma: increasing incidence and optimized management”. In: *Gastroenterologie clinique et biologique* 33.8-9 (2009), pp. 830–839 (cit. on p. 48).

[Van der Maaten, 2008] Laurens Van der Maaten and Geoffrey Hinton. “Visualizing data using t-SNE.” In: *Journal of Machine Learning Research* 9.11 (2008) (cit. on p. 109).

[Van Gansbeke, 2020] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. “Scan: Learning to classify images without labels”. In: *European Conference on Computer Vision*. Springer. 2020, pp. 268–285 (cit. on p. 68).

[Varoquaux, 2022] Gaël Varoquaux and Veronika Cheplygina. “Machine learning for medical imaging: methodological failures and recommendations for the future”. In: *NPJ digital medicine* 5.1 (2022), pp. 1–8 (cit. on p. 10).

[Villani, 2018] Cedric Villani and Bernard Nordlinger. *Santé et intelligence artificielle*. Cnrs, 2018 (cit. on p. 93).

[Xu, 2018] Zhoubing Xu, Yuankai Huo, JinHyeong Park, Bennett Landman, Andy Milkowski, Sasa Grbic, and Shaohua Zhou. “Less is more: Simultaneous view classification and landmark detection for abdominal ultrasound images”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 711–719 (cit. on pp. 4, 64, 65).

[Yang, 2020] Qi Yang, Jingwei Wei, Xiaohan Hao, Dexing Kong, Xiaoling Yu, Tianan Jiang, Junqing Xi, Wenjia Cai, Yanchun Luo, Xiang Jing, et al. “Improving B-mode ultrasound diagnostic performance for focal liver lesions using deep learning: A multicentre study”. In: *EBioMedicine* 56 (2020), p. 102777 (cit. on pp. 48, 60).

[Yao, 2018] Zhao Yao, Yi Dong, Guoqing Wu, Qi Zhang, Daohui Yang, Jin-Hua Yu, and Wen-Ping Wang. “Preoperative diagnosis and prediction of hepatocellular carcinoma: Radiomics analysis based on multi-modal ultrasound images”. In: *BMC cancer* 18.1 (2018), pp. 1–11 (cit. on pp. 48, 60).

[Yap, 2017] Moi Hoon Yap, Gerard Pons, Joan Martí, Sergi Ganau, Melcior Sentı ´s, Reyer Zwiggelaar, Adrian K Davison, and Robert Martí. “Automated breast ultrasound lesions detection using convolutional neural networks”. In: *IEEE journal of biomedical and health informatics* 22.4 (2017), pp. 1218–1226 (cit. on p. 37).

[Yasaka, 2018] Koichiro Yasaka, Hiroyuki Akai, Osamu Abe, and Shigeru Kiryu. “Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced CT: a preliminary study”. In: *Radiology* 286.3 (2018), pp. 887–896 (cit. on p. 60).

[Zenobii, 2016] Maria Francesca Zenobii, Esterita Accogli, Andrea Domanico, and Vincenzo Arienti. “Update on bedside ultrasound (US) diagnosis of acute cholecystitis (AC)”. In: *Internal and emergency medicine* 11.2 (2016), pp. 261–264 (cit. on p. 2).

[Zhang, 2017] Jeffrey Zhang, Sravani Gajjala, Pulkit Agrawal, Geoffrey H Tison, Laura A Hallock, Lauren Beussink-Nelson, Eugene Fan, Mandar A Aras, ChaRandle Jordan, Kirsten E Fleischmann, et al. "A computer vision pipeline for automated determination of cardiac structure and function and detection of disease by two-dimensional echocardiography". In: *arXiv preprint arXiv:1706.07342* (2017) (cit. on p. 37).

[Zhang, 2018] Zhilu Zhang and Mert Sabuncu. "Generalized cross entropy loss for training deep neural networks with noisy labels". In: *Advances in neural information processing systems* 31 (2018) (cit. on p. 109).

[Zhang, 2020] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. "Contrastive learning of medical visual representations from paired images and text". In: *arXiv preprint arXiv:2010.00747* (2020) (cit. on pp. 33, 89, 104, 107, 108).

[Zhu, 1997] Ciyou Zhu, Richard H Byrd, Peihuang Lu, and Jorge Nocedal. "Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization". In: *ACM Transactions on Mathematical Software (TOMS)* 23.4 (1997), pp. 550–560 (cit. on p. 40).

[Zimmer, 2020] Veronika A Zimmer, Alberto Gomez, Emily Skelton, Nooshin Ghavami, Robert Wright, Lei Li, Jacqueline Matthew, Joseph V Hajnal, and Julia A Schnabel. "A multi-task approach using positional information for ultrasound placenta segmentation". In: *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis*. Springer, 2020, pp. 264–273 (cit. on p. 4).

[Zoph, 2020] Barret Zoph, Ekin D Cubuk, Golnaz Ghiasi, Tsung-Yi Lin, Jonathon Shlens, and Quoc V Le. "Learning data augmentation strategies for object detection". In: *European Conference on Computer Vision*. Springer. 2020, pp. 566–583 (cit. on p. 52).

List of Figures

List of Tables