Thyrosonics: Learning-based detection and classification of thyroid nodules from ultrasound images

Hari Sreedhar

UNIVERSITÉ CÔTE D'AZUR

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

THÈSE DE DOCTORAT

Thyrosonics

L'apprentissage automatique pour la détection et classification des nodules thyroïdiens dans les images échographiques

Hari SREEDHAR

Équipe-Projet Epione, Centre Inria d'Université Côte d'Azur

Présentée en vue de l'obtention du grade de docteur en Santé numérique **d'**Université Côte d'Azur

Dirigée par : Hervé DELINGETTE, Directeur de recherche, Équipe-Projet Epione, Centre Inria d'Université Côte d'Azur

Co-encadrée par : Charles RAFFAELLI, Praticien hospitalier, Centre Hospitalier Universitaire de Nice

Co-encadrée par : Guillaume LAJOINIE, Assistant Professor, Université de Twente, Enschede, Pays-Bas

Soutenue le : 25 Octobre 2024

Devant le jury, composé de :

Emilie FRANCESCHINI, Directrice de recherche, LMA CNRS/Aix-Marseille Université

Hervé LOMBAERT, Associate Professor, École Polytechnique de Montréal

Adrian BASARAB, Professeur, CREATIS Université Claude Bernard Lyon 1

Pierre-Yves MARCY, Praticien hospitalier, Polyclinique Les Fleurs, Ollioules

Olivier HUMBERT, Professeur des universités-praticien hospitalier, Centre Antoine Lacassagne, Nice

Co-Funded by the European Union

THYROSONICS

L'APPRENTISSAGE AUTOMATIQUE POUR LA DÉTECTION ET CLASSIFICATION DES NODULES THYROÏDIENS DANS LES IMAGES ÉCHOGRAPHIQUES

Thyrosonics

Learning-based detection and classification of thyroid nodules from ultrasound images

Hari SREEDHAR

⋈

Jury :

Président du jury
Emilie FRANCESCHINI, Directrice de recherche, LMA CNRS/Aix-Marseille Université

Rapporteurs
Hervé LOMBAERT, Associate Professor, École Polytechnique de Montréal
Adrian BASARAB, Professeur, CREATIS Université Claude Bernard Lyon 1

Examinateurs
Pierre-Yves MARCY, Praticien hospitalier, Polyclinique Les Fleurs, Ollioules
Olivier HUMBERT, Professeur des universités-praticien hospitalier, Centre Antoine Lacassagne, Nice

Directeur de thèse
Hervé DELINGETTE, Directeur de recherche, Équipe-Projet Epione, Centre Inria d'Université Côte d'Azur

Co-encadrant de thèse
Charles RAFFAELLI, Praticien hospitalier, Centre Hospitalier Universitaire de Nice

Co-encadrant de thèse
Guillaume LAJOINIE, Assistant Professor, Université de Twente, Enschede, Pays-Bas

Université Côte d'Azur

Hari SREEDHAR

Thyrosonics

L'apprentissage automatique pour la détection et classification des nodules thyroïdiens dans les images échographiques

xv+164 p.

Ce document a été préparé avec LaTeX2e et la classe `these-ISSS` version v. 2.10.

Impression : output.tex – 5/11/2024 – 16:02

Révision pour la classe : these-ISSS.cls,v 2.10 2020/06/24 14:16:37 mpelleau

അച്ഛനും അമ്മയ്ക്കും ശിയകുട്ടിക്കും ഒരുപാട് നന്ദി.

Merci à Hervé, Charles, et Guillaume pour les connaissances que vous m’avez transmises avec une patience infinie.

Thank you to everyone at Epione for helping me get through the strange new experience of being a PhD student.

También quiero agradecer el apoyo y la amistad de Jairo.

Salvatore e Lucia, grazie mille per i ravioli e i cannoli.

V’aremercìi touplen lu amic de Nissa la Bella.

Merci à Chris, Guillaume, Loïc, Roma, et Mika pour la bonne bière et la bonne bouffe.

Thanks to the regulars at the Snug for Sunday afternoon books and crosswords.

And thank you, Jung, Clément, and Piano, for all the Magic, cheese, and complicated boardgames.

Thyrosonics

Résumé

L'échographie est une technique indispensable pour l'évaluation du risque de malignité des nodules thyroïdiennes. Malgré son utilité, l'échographie thyroïdienne reste limitée par sa dépendance à l'expérience de l'opérateur, autant pour l'acquisition que pour l'interprétation. C'est pourquoi des algorithmes d'apprentissage automatique, ayant connu de grands succès sur des images naturelles et médicales, ont été proposés aussi pour l'interprétation des images échographiques thyroïdiennes.

L'intérêt suscité dans ce domaine par la promesse de l'IA a mené à un grand nombre de publications proposant des algorithmes pour la détection, segmentation, et classification de nodules, ainsi qu'à la création de plusieurs produits commerciaux pour la pratique clinique. Malgré tous ces outils, l'impact réel sur la pratique des endocrinologues et radiologues français reste faible ; cette limitation correspond dans une large mesure au fait que la majorité de ces algorithmes ne prennent pas en compte le contexte clinique de l'échographie thyroïdienne en France.

L'objet de cette thèse est donc d'explorer les particularités de l'échographie thyroïdienne en France, afin d'identifier les possibles pistes d'amélioration en utilisant les méthodes de l'apprentissage automatique.

Le premier chapitre consiste à examiner la variabilité inter-expert en évaluation de l'échographie thyroïdienne. Une étude multicentrique utilisant des images échographiques acquises au fil de l'eau de la pratique clinique de quatre experts français donnent une indication des points de difficulté pour les médecins. Les résultats permettent d'identifier les caractéristiques échographiques des nodules thyroïdiens dont la description génère des différences significatives entre les praticiens, et entraîne des conséquences sur la prise en charge des patients.

Le deuxième chapitre entre plus dans le détail de l'une des caractéristiques échographiques utilisées par les experts : l'échogénicité. En continuité du chapitre précédent, la possibilité de se servir d'un outil d'apprentissage automatique pour aider les praticiens non-experts à distinguer entre des nodules hyper-/isoéchogènes et nodules hypoéchogènes est explorée. Ensuite, les différences quantitatives entre les images sont étudiees pour évaluer la robustesse de la vérité terrain, et la reproductibilité de l'examen échographique.

Le troisième chapitre s'intéresse à la difficulté d'obtenir des annotations expertes pour l'entraînement et le raffinement d'algorithmes d'apprentissage automatique en échographie thyroïdienne. Á partir des résultats précédents, il est évident que l'obtention d'un consensus sur les étiquettes des experts pour entraîner des algorithmes demanderait un temps considérable. Afin de réduire ce coût pour le développement des algorithmes, des stratégies d'apprentissage actif pour entraîner des réseaux de neurones avec moins d'annotations sont explorées. Ce chapitre présente les limitations de ces stratégies sur des vraies données cliniques, et propose aussi une technique d'apprentissage actif qui mélange des critères de sélection classiques avec la représentativité de l'échantillonnage au hasard.

Le dernier chapitre explore l'échographie quantitative comme piste future pour améliorer l'évaluation des nodules thyroïdiens. En utilisant des simulations numériques de tissus mous et d'une vraie sonde échographique, des réseaux de neurones sont entrainés pour estimer le paramètre non linéaire d'un milieu de propagation à partir du signal brut reçu au niveau de la sonde. La stratégie utilise une combinaison de pulses pour créer un signal plus apte à être traité par le réseau. Les contributions de cette thèse cherchent à mieux contextualiser l'utilisation de l'apprentissage automatique dans l'échographie thyroïdienne, afin de permettre ces techniques d'avancer vers des applications ayant un vrai impact durable sur la pratique clinique.

Mots-clés : Cancer de la thyroïde, Apprentissage automatique, Imagerie médicale.

Thyrosonics

Abstract

Ultrasound imaging is an essential technique for evaluating the risk of malignancy in thyroid nodules. Despite its usefulness, thyroid ultrasound is limited by its operator dependence, both for image acquisition and interpretation. As a result, many machine learning algorithms (which have had great success on natural and medical images) have been proposed to automatically interpret thyroid ultrasound images.

The interest in this area stimulated by the promise of AI has led to an abundance of publications proposing algorithms for the detection, segmentation, and classification of thyroid nodules, as well as to the creation of multiple commercial products marketed to medical practitioners. Despite all of these tools, the actual impact on the daily practice of French endocrinologists and radiologists has been fairly minor; this limitation is largely due to the fact that most of these algorithms do not take into account the clinical context of thyroid ultrasound in France.

The goal of this thesis is therefore to explore the unique aspects of thyroid ultrasound in France, in order to identify potential opportunities for improvement using machine learning.

The first chapter consists of an examination of the inter-expert variability in the evaluation of thyroid ultrasound. A multicentric study using real ultrasound images acquired during the course of the clinical practice of four French experts gives an indication of which aspects of evaluation are difficult for clinicians. The results allow for the identification of ultrasound features of thyroid nodules whose description generates disagreement between practitioners and leads to consequences for the care of patients.

The second chapter goes into more detail about one of the ultrasound features used by experts: echogenicity. Building on the previous chapter, the possibility using a machine learning tool to help non-expert practitioners distinguish between hyper-/isoechoic nodules and hypoechoic nodules is explored. Then, quantitative differences between images are investigated to examine the robustness of expert labels, and the reproducibility of the ultrasound examination.

The third chapter addresses the difficulties of obtaining expert annotations for training and refining machine learning algorithms for thyroid ultrasound. Given the previous results, it is clear that obtaining expert consensus labels to create transparent algorithms is enormously time-consuming. In order to reduce the annotation burden for the development of these algorithms, active learning strategies to train neural networks with fewer labels are explored. This chapter presents the limitations of these strategies on real clinical data, and also proposes an active learning technique that blends classic selection criteria with the representative power of random sampling.

Finally, the last chapter explores quantitative ultrasound as a future means to improve the evaluation of thyroid nodules. By using simulations of soft tissue and of a real ultrasound probe, neural networks are applied to map the nonlinear parameter of a propagation medium based on the raw signal received by the transducer. This strategy uses a combination of pulses to create a signal that is better suited to be analyzed by the network.

The contributions of this thesis seek to better contextualize the use of machine learning for thyroid ultrasound, in order to allow these techniques to advance towards applications with a real, lasting impact on clinical practice.

Keywords: Thyroid Cancer, Machine Learning, Medical Imaging.

Acknowledgements

The authors are grateful to the OPAL infrastructure from Université Côte d'Azur for providing resources and support. We also thank the Association Francophone de Thyroïdologie for their help and support. This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 847581.

Contents

1 Introduction **1**
- 1.1 Background . . . 1
  - 1.1.1 Thyroid Nodule Ultrasound . . . 1
  - 1.1.2 Machine Learning Methods for the Evaluation of Thyroid Nodules . . . 2
- 1.2 Outline of the Thesis . . . 3
- 1.3 Challenges Addressed by this Thesis . . . 4
- 1.4 Publications and Presentations . . . 5

2 Study of French Inter-Expert Variability in Thyroid Nodule Ultrasound **7**
- 2.1 Introduction . . . 9
- 2.2 Background . . . 10
  - 2.2.1 Composition . . . 11
  - 2.2.2 Echogenicity . . . 14
  - 2.2.3 Shape . . . 17
  - 2.2.4 Margin . . . 19
  - 2.2.5 Echogenic Foci . . . 22
  - 2.2.6 Other . . . 24
- 2.3 Methods . . . 24
  - 2.3.1 Study Design . . . 25
- 2.4 Results . . . 27
  - 2.4.1 Images and evaluations . . . 27
  - 2.4.2 EU-TIRADS Results . . . 27
  - 2.4.3 Reproducibility of Sonographic Feature Identification . . . 30
  - 2.4.4 Other Differences in EU-TIRADS Scoring . . . 46
- 2.5 Discussion . . . 58
  - 2.5.1 Limitations . . . 58
  - 2.5.2 Feature Identification . . . 59
  - 2.5.3 Decision Structures . . . 62
- 2.6 Conclusions . . . 64

3 Expert Variability in Thyroid Nodule Echogenicity Evaluation **65**
- 3.1 Introduction . . . 67
- 3.2 Background . . . 67
  - 3.2.1 Analysis Tasks of Machine Learning Methods on Thyroid Ultrasound . . . 67
  - 3.2.2 Thyroid Ultrasound Datasets for Machine Learning . . . 70
- 3.3 Automated Nodule Echogenicity Characterization in the French Context . . . 71
  - 3.3.1 Echogenicity Classification Strategy . . . 72
  - 3.3.2 Echogenicity Classification Results . . . 75
- 3.4 Reproducibility of Expert Echogenicity Classification . . . 77
  - 3.4.1 Intra-Expert Reproducibility Results . . . 78

3.4.2 Implications of Expert Variability in Echogenicity Assessment 79
3.5 Quantitative Echogenicity Analysis 80
3.5.1 Expert Label Agreement with Quantitative Echogenicity Measures 81
3.5.2 Quantitative Analysis Results 83
3.6 Discussion 88
3.6.1 Limitations 89
3.7 Conclusions 90

4 Active Learning Limitations on Clinical Thyroid Ultrasound Data 91
4.1 Introduction 93
4.2 Background 93
4.2.1 Limitations of Active Learning Strategies 93
4.2.2 Active Learning Applied to Thyroid Ultrasound 95
4.3 Methods 96
4.3.1 Image Datasets 96
4.3.2 Rigged Draw Strategy 98
4.3.3 Supervised and Semi-supervised Active Learning Strategies 98
4.4 Results 99
4.4.1 Supervised Learning Results 100
4.4.2 Semi-Supervised Learning Results 103
4.4.3 Initial Set Impact 106
4.5 Discussion 108
4.5.1 Limitations 109
4.6 Conclusion 109

5 A Machine Learning Strategy for Nonlinear Parameter Estimation 111
5.1 Introduction 113
5.1.1 Quantitative Ultrasound 113
5.1.2 Applying Machine Learning for Nonlinear Parameter Estimation 114
5.2 Background 115
5.2.1 Ultrasound Wave Propagation 115
5.2.2 The Attenuation Coefficient α 116
5.2.3 The Nonlinear Parameter $\frac{B}{A}$ 117
5.3 Methods 120
5.3.1 Pulse Division Method for $\frac{B}{A}$ Estimation 120
5.3.2 Simulation 123
5.3.3 Nonlinear Parameter Estimation with a Neural Network 128
5.4 Results 132
5.5 Discussion 134
5.5.1 Limitations 137
5.6 Conclusion 137

6 Conclusion 139
6.1 Contributions 139
6.2 Future Directions 140

Bibliography **143**

List of Figures **151**

List of Tables **159**

Appendices

A Appendix I: Definition of B/A 163

CHAPTER 1

Introduction

1.1 Background

The thyroid gland is a vital endocrine organ located at the base of the neck. Its anatomy was described in 1656 by Thomas Wharton, who assigned to it its name based on an Ancient Greek word for a type of shield with a similar shape (Kelly, 1961). That shape, also referred to as butterfly-like, consists of a left and a right lobe, connected at the middle by a strip known as the isthmus, located anterior to the trachea and inferior to the laryngeal prominence.

The shield-like contour of the thyroid can become distorted, most commonly by goiters, which manifest as large swellings at the base of the neck arising from enlargement of the gland. Indeed, the earliest hints of the organ's function being intertwined with iodine metabolism came from the knowledge in ancient times that seaweed, rich in the element, could be used to prevent goiters (Küpper et al., 2011). Over four millennia later in 1811, the French chemist Bernard Courtois, while experimenting with seaweed as a source of saltpeter for Napoleon's armies discovered violet-hued vapors that would give the element its name (*iode* in French from a Greek word meaning violet). Within the same decade, the newly-discovered element was used for the treatement of goiters (Kelly, 1961).

Further advances in thyroid physiology have served to elucidate the mechanisms of this connection to iodine. Thyroid hormones, which play a central role in many of the body's endocrine cycles and contain iodine in their chemical structures, are synthesized by the follicular cells of the thyroid gland. Indeed the very histologic structure of the thyroid serves this function, with the follicular cells organized into follicles that contain the precursors of thyroid hormones. The synthesis and generation of thyroid hormones are regulated by complex endocrine feedback loops which are necessary for the body's normal physiologic function (Giovanella, 2023).

Therefore, disturbances of the normal structure or functioning of the thyroid gland lead to a variety of endocrine disorders. Of interest in this thesis, however, is when the normal follicular architecture of the gland is disrupted by abnormal nests of cells known as nodules. Some of these lesions ramp up their secretion of thyroid hormones without regard for the normal regulatory mechanisms; these "hot" or hyperfunctioning nodules are rarely malignant (Kant, Davis, & Verma, 2020). Nonfunctional nodules, however, do have a potential for malignancy. Estimates of their global incidence vary as a function of the method used for their detection, but they are responsible for a substantial global disease burden which is increasing in many countries (Uppal, Collins, & James, 2023 ; Sajisevi et al., 2022).

1.1.1 Thyroid Nodule Ultrasound

Thyroid nodules may be detected by palpation, or as incidental findings during imaging procedures in the head and neck region (Kant et al., 2020). The malignant potential of nodules is

relatively low, a fact which when combined with their astonishing prevalence creates a considerable risk of unnecessary intervention (Uppal et al., 2023 ; Sajisevi et al., 2022). The confirmation of malignancy via cytology of fine-needle aspiration (FNA) samples is overly invasive as a first-line diagnostic test (Kant et al., 2020). Therefore, the risk evaluation of thyroid nodules begins with ultrasound imaging.

Ultrasound, by its nature, is uniquely suited as a first-line diagnostic technique for a common soft-tissue lesion. As a non-ionizing, non-invasive method, the risk-benefit profile remains advantageous even with a low true positive rate for malignant nodules. The low cost of ultrasound equipment, while certainly not negligible, renders the modality more accessible than many other medical imaging technologies. This, combined with advances in the portability of ultrasound systems, facilitates the evaluation and risk-stratification of nodules.

Ultrasound evaluation of malignancy risk is limited, however, by its operator-dependence. In order to standardize the evaluation of thyroid nodules, clinicians utilize various TIRADS (Thyroid Reporting Imaging And Data System) systems. These systems, by analogy to well-known BI-RADS framework for mammography evaluation, formalize the reporting of a set of echographic features correlated with malignant nodules. In France, EU-TIRADS is used to allow radiologists and endocrinologists, who conduct the ultrasound evaluation themselves, to quickly sort nodules into different four risk categories that assist with the decision to proceed to FNA (Russ et al., 2017). In the United States, ACR-TIRADS is used help practitioners evaluate static ultrasound images typically acquired by an ultrasound technician, and uses a cumulative point-based system to calculate a risk category (Tessler et al., 2017).

Despite the existence of these frameworks, however, inter-reader reliability remains limited (Grani et al., 2018). Part of this variation may be due to the fact that the identification of the echographic features used by EU-TIRADS and ACR-TIRADS is inherently subjective, and may differ between readers (Solymosi et al., 2023). This subjectivity has attracted interest in the use of machine learning techniques to automate thyroid nodule ultrasound evaluation.

1.1.2 Machine Learning Methods for the Evaluation of Thyroid Nodules

In recent decades, machine learning techniques have been applied to almost all forms of medical imaging to perform tasks such as lesion detection, segmentation of organs, or classification of disease state (Najjar, 2023). The term machine learning refers to techniques allowing algorithms to learn to make predictions from data, in this case medical images, without explicit human intervention to guide model adjustments. This can be accomplished either with supervised learning using ground-truth labels for training, or unsupervised learning which functions without them. Deep learning in turn uses neural network architectures that can independently learn features of the data to be employed for a predictive task.

Thyroid ultrasound has been the target of some of these proposed machine learning algorithms, which perform nodule detection, segmentation, and characterization on the basis of static ultrasound images. There have been online challenges for thyroid nodule segmentation and classification (Grand Challenge, 2020), software tools tested with commercial ultrasound systems (Wei et al., 2020), and large-scale multi-center validation studies (Xu et al., 2023). Nevertheless, the impact of these tools has yet to be felt in clinical practice; this, in part, may be due to the failure of many machine learning techniques to adapt themselves adequately to the nuances of clinical thyroid ultrasound.

This thesis focuses on the applications of machine learning principles to thyroid nodule ultrasound in a way that takes into account the clinical and physical context of this medical imaging technique. Each chapter of this work focuses on a domain-specific aspect of thyroid ultrasound in order to explore clinical limitations and machine learning applications with relevance to the actual care of patients.

1.2 Outline of the Thesis

Chapter 2 of this thesis begins with an exploration of thyroid nodule evaluation on ultrasound. As this is the basis for thyroid nodule characterization in clinical practice, it must be understood in order to contextualize machine learning applications to thyroid ultrasound. While inter-expert variability has been studied previously, there is a gap specific to the French context. To meet this deficiency, we conduct a multi-centric study with real clinical data acquired by four French experts. This study evaluates the degree of inter-expert variability in EU-TIRADS scoring, along with differences in the identification of specific echographic features and how they are associated with score disagreements. This serves to highlight the areas of nodule evaluation in which subjective interpretation limits the reproducibility of EU-TIRADS-based clinical assessment.

Chapter 3 investigates more specifically nodule echogenicity as a feature particularly relevant to thyroid nodule risk stratification. Building on the previous chapter, the possibility of using a neural network based tool to help non-expert practitioners distinguish between hyper-/isoechoic and hypoechoic nodules is explored. Then, quantitative differences in between nodule and reference zones in ultrasound images are investigated to assess the robustness of expert labels and the reproducibility of this feature of the ultrasound examination.

Chapter 4 addresses the difficulties of obtaining expert annotations for training and refining machine learning algorithms for thyroid ultrasound. Given the time required for experts to evaluate images, it is clear that reducing the annotation burden for development of these tools would facilitate clinical implementation. To this end, we explore active learning strategies that attempt to train neural networks using fewer expert labels. This chapter presents the limitations of these strategies on real clinical data and also proposes a new active learning technique that blends classic selection criteria with the representative power of random sampling.

Finally, Chapter 5 explores quantitative ultrasound as a future means to improve the evaluation of thyroid nodules. One promising target for quantitative evaluation of nodules is the nonlinear parameter, an acoustic characteristic of tissue that has been used elsewhere as a surrogate for pathological changes. The *in vivo* estimation of this parameter, however, remains a technical challenge due to the influence of attenuation and of acoustic scattering in tissue, with existing approaches failing to provide accurate measurements. By using simulations of soft tissue and a real ultrasound probe, neural networks are applied to characterize the nonlinear parameter of a propagation medium based on the raw signal received by the transducer. This strategy utilizes a combination of pulses to create a signal that is more easily analyzed by the network, based on the physics of nonlinear propagation. This represents a first step toward a technique for nonlinear parameter characterization in thyroid tissue.

The conclusion reviews the main contributions of the thesis, and explores some future directions for this area of research.

1.3 Challenges Addressed by this Thesis

This thesis seeks to address in particular the following difficulties in thyroid nodule ultrasound and machine learning methods applied to it:

Evaluation of French Thyroid Ultrasound Data
Proper evaluation of machine learning applications to thyroid ultrasound in France requires an understanding of how this technique is actually used. As thyroid ultrasound acquisition in this country is performed by the interpreting radiologist or endocrinologist, and not by a non-physician sonographer, the quality and characteristics of ultrasound depends on the expert. In order to accurately study French thyroid ultrasound, it is necessary to obtain images that were acquired during actual clinical practice. Ideally, these images should also come from multiple French practitioners, in their own clinic or hospital sites, and with the ultrasound systems they use in routine practice. Such a dataset has yet to be assembled an analyzed.

Inter-Expert Variability in French Thyroid Ultrasound
Another important factor impacting thyroid ultrasound practice in France is the degree to which the evaluations of expert practitioners are reproducible. French practitioners use the EU-TIRADS system during their ultrasound acquisitions, and often have been previously trained in the previous systems of thyroid nodule evaluation in France, which may have given them unique biases in practice. The EU-TIRADS framework is meant to standardize the risk-stratification of nodules, but its effectiveness depends on whether users will assign consistent scores. This in turn may depend on experts' ability to identify the specific sonographic features corresponding to composition, echogenicity, shape, margin, and the presence of echogenic foci that form the basis for EU-TIRADS (Russ et al., 2017). Examining the degree of agreement between experts on these measures could identify areas of nodule evaluation for which machine learning methods could have the most impact.

Evaluation of Active Learning Strategies on Real Ultrasound Data
As expert annotations take time to acquire, they represent a substantial cost for training and fine-tuning machine learning algorithms to thyroid ultrasound evaluation. Active learning strategies seek to reduce this cost by guiding the selection of a subset of the most informative images for annotation, and thereby achieving the same training results for an algorithm with fewer ground-truth labels. However, the efficacy of these techniques must be evaluated on real clinical thyroid data to confirm that they actually offer a benefit over random selection. Because the use-case of active learning techniques for thyroid ultrasound is with the expert annotation of only a small number of images, it is also necessary to identify active learning techniques that can consistently outperform the random selection baseline regardless of the effects of the initial annotated subset.

Nonlinear Parameter Measurement in Tissue with Ultrasound
Given that thyroid nodule ultrasound depends on subjective expert evaluation, the skill level of the operator and interpreter may limit the reproducibility of nodule evaluation. An objective standard of evaluation using a quantitative property of the tissue could therefore someday improve nodule risk stratification. One such promising target is the acoustic nonlinear parameter $\frac{B}{A}$ which has been associated with differences in healthy and diseased tissue. However, measurement of this characteristic of tissue *in vivo* is complicated by the effects of acoustic scattering, attenuation, and

the ultrasound probe itself. Steps toward a practical strategy that compensates for these effects could help improve quantitative ultrasound.

1.4 Publications and Presentations

Publications

— **Sreedhar, H.**, Lajoinie, G. P. R., Raffaelli, C., and Delingette, H. Active Learning Strategies on a Real-World Thyroid Ultrasound Dataset. In: Xue, Y., Chen, C., Chen, C., Zuo, L., Liu, Y. (eds) Data Augmentation, Labelling, and Imperfections Workshop. MICCAI 2023. Lecture Notes in Computer Science, vol 14379. Springer, Cham. 127-136. 2024.

— **Sreedhar, H.**, Monpeyssen, H., Ghanassia, E., Marcy, P., Lajoinie, G. P. R., Delingette, H., and Raffaelli, C. Inter-Expert Variability Among French Thyroid Ultrasound Experts. European Thyroid Journal. (In Preparation).

— **Sreedhar, H.**, Raffaelli, C., and Delingette, H., and Lajoinie, G. P. R. Implementation of a Deep Learning Prediction Strategy for the Nonlinear Parameter Value of Simulated Tissue-Like Media. Ultrasonics. (In Preparation).

Presentations

— Ateliers Thyroïde de Sète 2024. Sète, France. May 18, 2024.

— Ateliers Thyroïde de Sète 2023. Sète, France. May 27, 2023.

CHAPTER 2

Study of French Inter-Expert Variability in Thyroid Nodule Ultrasound

The EU-TIRADS framework serves as the basis for standardizing expert evaluation of thyroid nodule ultrasound in France. Despite its utility, this system depends on the subjective identification by experts of different sonographic nodule features, including composition, echogenicity, shape, margin, and the presence of echogenic foci. In order to study inter-expert variability in French thyroid ultrasound evaluation, we assembled a dataset of 303 thyroid nodule images acquired during routine clinical practice by four French experts. These images were then evaluated independently by each of the experts, who assigned to them descriptions of the EU-TIRADS scores as well as of the different sonographic features. Analysis of these results revealed a strong degree of EU-TIRADS score disagreement between the four experts, particularly in association with differences in the identification of certain composition, echogenicity, and shape descriptions. In addition, the four experts did not always consistently assign the same EU-TIRADS score to nodules based on the same combinations of features. These results highlighted the subjective difficulties of thyroid ultrasound evaluation, at least on static images. They also suggest targets of automation that would be most likely to have clinical impact among expert and non-expert practitioners.

2.1 Introduction 9
2.2 Background 10
2.2.1 Composition 11
2.2.2 Echogenicity 14
2.2.3 Shape 17
2.2.4 Margin 19
2.2.5 Echogenic Foci 22
2.2.6 Other 24
2.3 Methods 24
2.3.1 Study Design 25
2.3.1.1 Image Acquisition 25
2.3.1.2 Image Evaluation 25
2.3.1.3 Analysis of Evaluations 26
2.4 Results 27
2.4.1 Images and evaluations 27
2.4.2 EU-TIRADS Results 27
2.4.3 Reproducibility of Sonographic Feature Identification 30
2.4.3.1 Composition 31
2.4.3.2 Echogenicity 34
2.4.3.3 Shape 37
2.4.3.4 Margin 39
2.4.3.5 Echogenic Foci 41
2.4.4 Other Differences in EU-TIRADS Scoring 46
2.4.4.1 Intra-expert variability within sonographic feature combinations 46
2.4.4.2 Inter-expert variability within sonographic feature combinations 49
2.4.4.3 Expert Variation from the EU-TIRADS Guideline 49
2.4.4.4 Modeling Decision Trees 51
2.5 Discussion 58
2.5.1 Limitations 58
2.5.2 Feature Identification 59
2.5.2.1 Composition 59
2.5.2.2 Echogenicity 60
2.5.2.3 Shape 60
2.5.2.4 Margin 61
2.5.2.5 Echogenic Foci 61
2.5.3 Decision Structures 62
2.6 Conclusions 64

2.1 Introduction

As previously discussed, ultrasound is an indispensable tool for the evaluation of thyroid nodules. In France, thyroid ultrasound examinations are conducted and interpreted by medical practitioners such as radiologists and endocrinologists. During the course of the examination, the operator places the ultrasound probe onto the patient's skin and examines the region of the neck around the thyroid, locating and evaluating potential nodules. An example of a thyroid ultrasound image is presented in Figure 2.1.

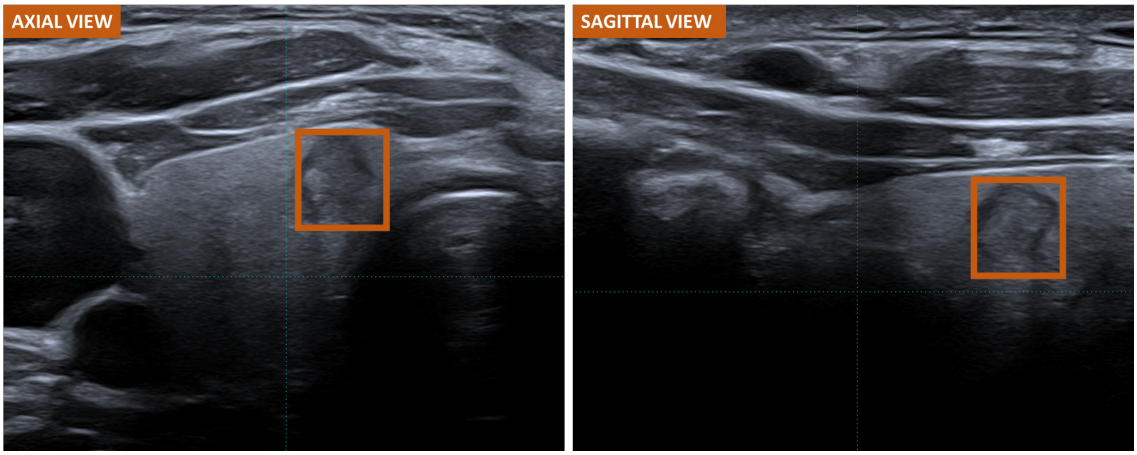

Figure 2.1 – Representative B-mode ultrasound images of a thyroid nodule, indicated by the orange square. (Left) Image acquired in an axial view. (Right) Image acquired in a sagittal view.

Using ultrasound, the practitioner will examine both the right and left thyroid lobes, as well as the isthmus that connects them. Is some patients, an additional pyramidal lobe arises from the isthmus and must be examined as well. By using a knowledge of thyroid anatomy, the operator must examine all of the visible thyroid tissue, as well as its relationships to the adjacent trachea, blood vessels, and muscles, each of which have different appearances on ultrasound, as seen in Figure 2.2. Regions of the thyroid that are deeper may be difficult to visualize due to the depth- and frequency-dependent attenuation of ultrasound waves.

Therefore, thyroid nodule ultrasound depends on the skill of the operator in obtaining views of the region that allow for analysis, both by carefully positioning the patient and the probe, and by adjusting the frequency, time gain compensation, and other acquisition settings of the ultrasound system. If nodules are present in the thyroid, they are examined for features which are correlated with benignity or malignancy. These features have been grouped into standardized reporting systems, referred to as Thyroid Reporting Imaging And Data Systems (TIRADS) by analogy to the systems used for mammography interpretation. The purpose of these systems, which are used in conjunction with a full radiological report, is to facilitate the risk stratification of thyroid nodules so as to guide the decision to proceed to fine-needle aspiration (FNA).

Despite the existence of these standards, thyroid ultrasound is still conducted by human practitioners, who have unique practices and tendencies. In order to understand the nuances of thyroid ultrasound in France, we must therefore examine how practitioners use TIRADS systems and identify the sonographic features on which they are based.

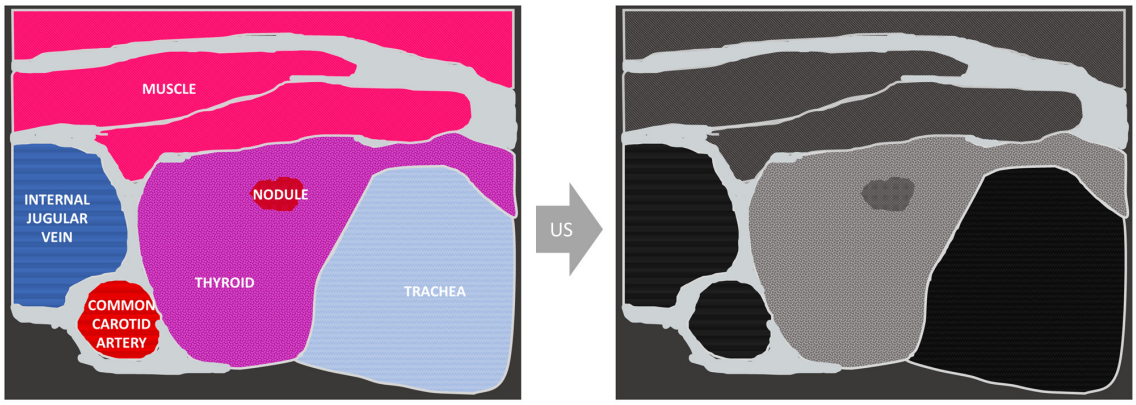

Figure 2.2 – (Left) Simplified illustration of the axial-view anatomy of the region explored by thyroid ultrasound, based on the image in Figure 2.1. (Right) Illustration of the ultrasound view of the image. The fluid-filled vessels and the air-filled trachea appear anechoic, though artifacts may create the appearance of internal structures. The nearby muscles are hypoechoic compared to the thyroid gland. Due to attenuation, deeper structures cannot be seen, especially at higher ultrasound frequencies.

2.2 Background

The TIRADS system used in France is EU-TIRADS, proposed in 2017 (Russ et al., 2017). This system combines sonographic features in a risk stratification algorithm summarized in Figure 2.3.

Within this system, nodules are first assessed for sonographic features of high risk: being markedly or very hypoechoic, having an irregular margin, not having an oval shape, or containing microcalcifications. These features, which will be discussed in subsequent sections, are strongly associated with malignant nodules, so their presence leads to the highest EU-TIRADS score of 5. This score is estimated to encompass only around 4% of nodules. It is important to note that this category leads to risk estimates of 26%-87%, which is a wide range. This means that nodules that fall into this score are not necessarily similar in terms of their sonographic caracteristics, but have the highest risk of malignancy. Therefore, this score is associated with a recommendation to conduct an FNA for nodules with a diameter of at least 10 mm; nodules falling below this size threshold are to be subjected to active surveillance (Russ et al., 2017).

As can be seen from Figure 2.3, nodules that do not have any high-risk features can then be stratified into other scores. If a nodule is entirely spongiform in composition, or anechoic, as would correpond to a fluid-filled cyst, the score of EU-TIRADS 2 is assigned. These lesions (about 5% of nodules) have virtually no risk of malignancy, and therefore no intervention is recommended, unless the size of the nodule is sufficient to generate symptoms by compressing nearby structures such as the trachea (Russ et al., 2017).

If a nodule has neither high-risk features nor strong indicators of being benign, it can then be classified in terms of its echogenicity, as compared to adjacent thyroid parenchyma. From Figure 2.3 we see that a hypoechoic nodule would be scored as EU-TIRADS 4 (about 28% of all nodules), with a risk of malignancy of between 6-17%. This leads to a recommendation for FNA

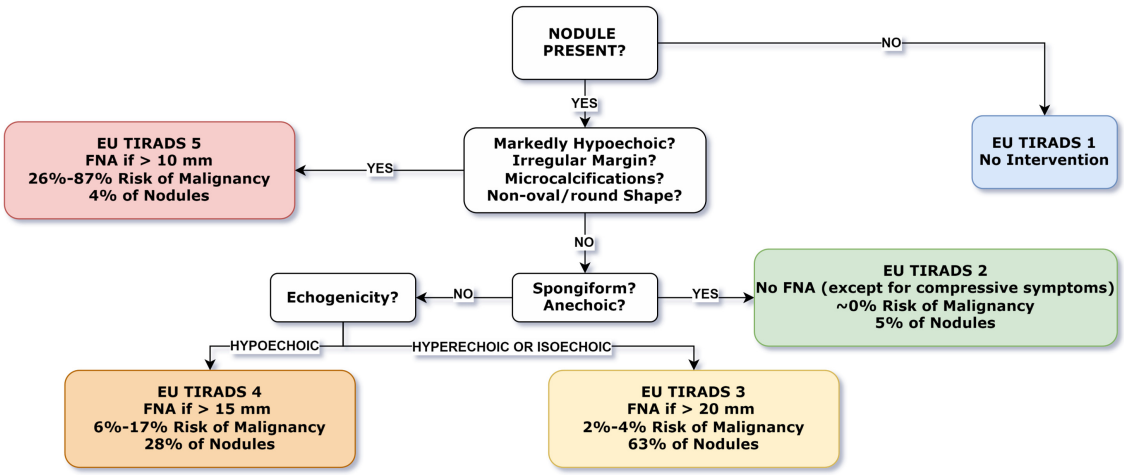

Figure 2.3 – Illustration of the EU-TIRADS algorithm, as proposed in (Russ et al., 2017). The score of EU-TIRADS 1 corresponds to the absence of a nodule; the other scores stratify the risk of malignancy with different indications for FNA.

if the nodule exceeds 15 mm in diameter. If, however, the nodule is hyperechoic or isoechoic as compared to the surrounding thyroid parenchyma, the score of EU-TIRADS 3 is assigned, which corresponds to the majority (63% of nodules). As this score is associated with a risk of malignancy of only 2%-4%, FNA is reserved for cases when these nodules exceed 20 mm in diameter (Russ et al., 2017).

This reporting system is used to standardize the ultrasound evaluation of thyroid nodules in France, though it must be noted that it is used in conjunction with a full description of the examination, as well as other techniques such as Doppler imaging and elastography. Other TIRADS systems also exist, and are used in other countries. For example, the ACR-TIRADS (American College of Radiology TIRADS) system used in the United States examines similar sonographic features, but assigns points on the basis of the features in order to determine a final risk score (Tessler et al., 2017). The recently updated K-TIRADS from South Korea uses a set of ultrasound patterns to describe nodules, that combine low- and high-risk features to stratify the risk of malignancy in a slightly different manner (Ha, Na, & Baek, 2021). For all of these systems, proper implementation and reproducibility of scoring depend on how faithfully practitioners follow the guidelines, as well as on how consistently they can identify the sonographic features used to assign scores.

Therefore, in order to understand thyroid nodule ultrasound in France, we must examine the sonographic features that practitioners are expected to identify. We proceed to do this in the following sections, with a discussion of features used by the EU-TIRADS and ACR-TIRADS systems.

2.2.1 Composition

The composition of a nodule is a description of its apparent tissue structure within the ultrasound image. This aspect has a very logical connection to malignancy, as certain structures are

highly unlikely to be malignant. Some different categories of nodule composition are shown in Figure 2.4.

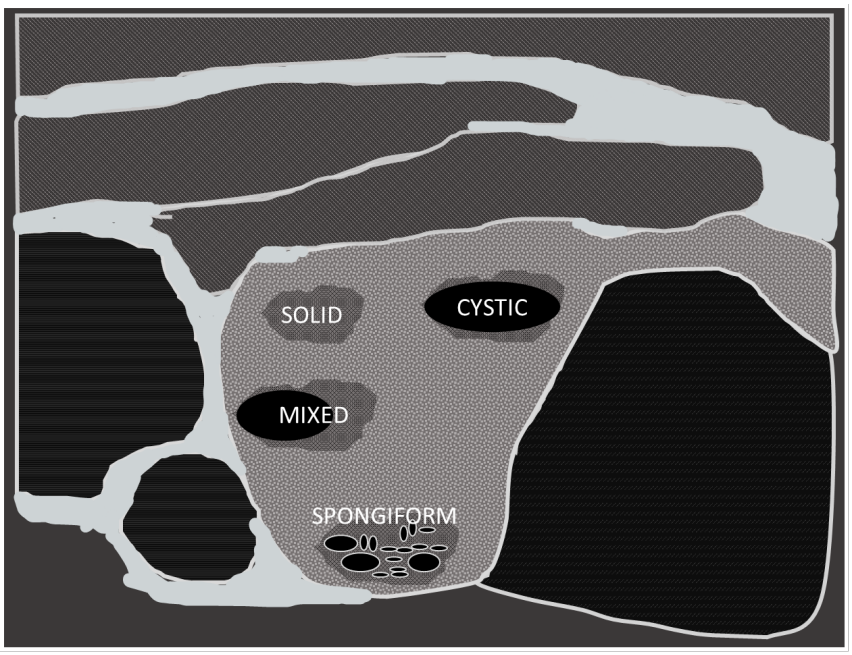

Figure 2.4 – Illustration of different nodule compositions. Solid nodules appear to be composed of solid tissue, while cystic lesions are dominated by large, dark, fluid-filled spaces. Some nodules may have similar proportions of solid and cystic components, and be described as being of mixed composition. Spongiform nodules are unique, in that they are composed of many tiny cystic spaces throughout the entire lesion.

Solid nodules are those which, as the name suggests, appear to be composed of solid tissue, similar to the rest of the thyroid gland (see Figure 2.5). Cystic nodules, by contrast, are primarily composed of large, fluid-filled spaces, which gives their interiors a dark appearance on ultrasound images (see Figure 2.6). Some nodules, however, may not have a purely solid or cystic composition, and may be characterized as being of mixed composition (see Figure 2.7). Finally, spongiform nodules present a special case; they are filled with numerous tiny cystic spaces (see Figure 2.8), which are smaller than for cystic or mixed nodules (Russ et al., 2017 ; Tessler et al., 2017).

These different descriptions have been associated in various studies with benign or malignant lesions, as confirmed by histopathology. Lesions that are spongiform on ultrasound are notable for being almost always benign; the same is true for entirely cystic lesions (Moon et al., 2008 ; Bonavita et al., 2009). Most thyroid carcinomas have a more solid composition, though this is not to say that most solid nodules are necessarily carcinomas (Henrichsen et al., 2010 ; Moon et al., 2008). These factors have been taken into account by the structures of the EU-TIRADS and ACR-TIRADS guidelines (Russ et al., 2017 ; Tessler et al., 2017).

In the EU-TIRADS system, a spongiform label corresponds to an EU-TIRADS 2 score, as seen in Figure 2.3, which is rarely malignant, and therefore not subjected to FNA. In the ACR-TIRADS system, a spongiform label immediately earns a nodule the lowest possible score, with

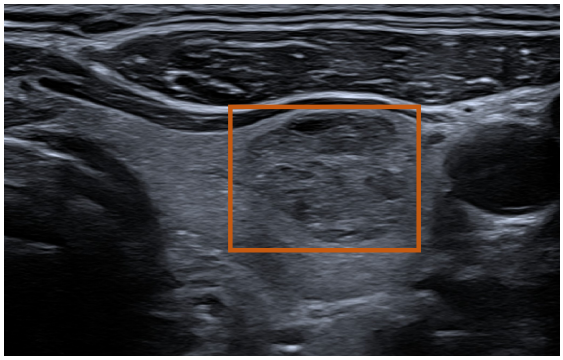

Figure 2.5 – Example of an axial-view ultrasound image containing a nodule judged as solid by four expert practitioners. The solid nodules may contain very small cystic spaces.

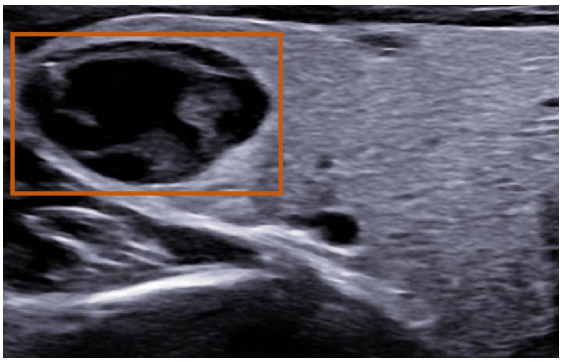

Figure 2.6 – Example of an axial-view ultrasound image containing a thyroid nodule judged as cystic by four expert practitioners. These nodules are primarily composed of large, fluid-filled spaces.

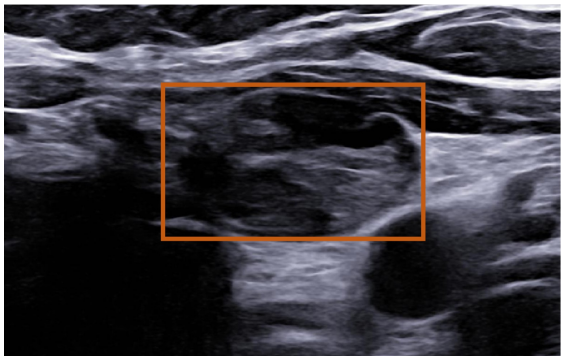

Figure 2.7 – Example of an axial-view ultrasound image containing a thyroid nodule judged as mixed cystic and solid by four expert practitioners.

Figure 2.8 – Example of an axial-view ultrasound image containing a thyroid nodule judged as spongiform by four expert practitioners. These nodules are composed of numerous tiny cystic spaces.

no further consideration of other risk features. Cystic lesions, which are anechoic, can receive an EU-TIRADS score of 2, and in ACR-TIRADS this description does not contribute towards increasing a nodule's risk score. Other compositions qualify for higher scoring on the basis of their other sonographic features in EU-TIRADS, with ACR-TIRADS making a slight distinction between mixed cystic / solid lesions (which receive fewer points) and fully solid lesions (which receive an additional point) (Russ et al., 2017 ; Tessler et al., 2017).

These descriptions call for a degree of subjective judgment on the part of the reader when assessing mixed nodules. Distinguishing between a spongiform nodule composed of small cystic spaces, and a partly solid nodule with a few cystic spaces could be a potential source of ambiguities.

2.2.2 Echogenicity

The echogenicity of a tissue or lesion is a description of the intensity of the incident ultrasound pulse that it reflects back to the probe to generate the signal. On a standard B-mode image, echogenicity is appreciated as the brightness of the pixels within a region, but is affected by numerous factors. The dynamic range and the contrast settings used for visualization impact the perception of echogenicity by the operator, so these must be adjusted carefully. In addition, acquisition settings such as time-gain compensation, and structural effects such as acoustic enhancement or shadowing can alter the perceived brightness of a structure independent of its echogenicity. Therefore, operator skill is critical to evaluating this sonographic caracteristic of tissue.

Because the absolute intensity of pixels depends so much on operator and image settings, echogenicity in thyroid ultrasound is described by comparison to anatomic references. As can be seen from Figure 2.1, the strap muscles are hypoechoic relative to normal thyroid tissue, which means that they reflect less signal and appear darker on the image. The adjacent air-filled trachea and liquid-filled blood vessels appear anechoic, generating no reflected signal and appearing black, though some image artifacts may create the appearance of echogenic structures.

In both the EU-TIRADS and ACR-TIRADS systems, lesions are stratified in terms of risk on the basis of their level of echogenicity, as compared to other tissues (Russ et al., 2017 ; Tessler et al., 2017). As with composition, the difference in echogenicity may be associated with differences in the histologic structure of a nodule, which varies between benign and malignant lesions. Different categories of echogenicity that are used to describe lesions are shown in Figure 2.9.

Hyperechoic nodules are brighter than the surrounding normal thyroid tissue, while isoechoic nodules have a similar brightness to adjacent thyroid tissue (see Figure 2.10). These two categories are usually considered together. If a nodule is darker than the adjacent thyroid tissue, it is considered hypoechoic (see Figure 2.11), and if it is also darker than the nearby muscles, it is labeled as very or markedly hypoechoic (see Figure 2.12). The darkest of all lesions are anechoic fluid-filled cysts, which appear similar in brightness to the blood vessels near the thyroid (see Figure 2.13). Finally, in some cases, it may not be possible to assess the echogenicity of a nodule if its interior is obscured from view by an intervening area of calcification.

While these descriptions may seem straightforward, practitioners may have subjective differences in terms of their perception of different grayscale intensity levels. If the echogenicity of a nodule is not uniform, readers may come to different judgements about which level of echogenicity is dominant. In addition, because of the need to compare a nodule to surrounding thyroid tissue or muscles, echogenicity evaluation depends on having appropriate references visible to the reader. If the echogenicity of the surrounding thyroid tissue is altered due to inflammation (in thyroiditis),

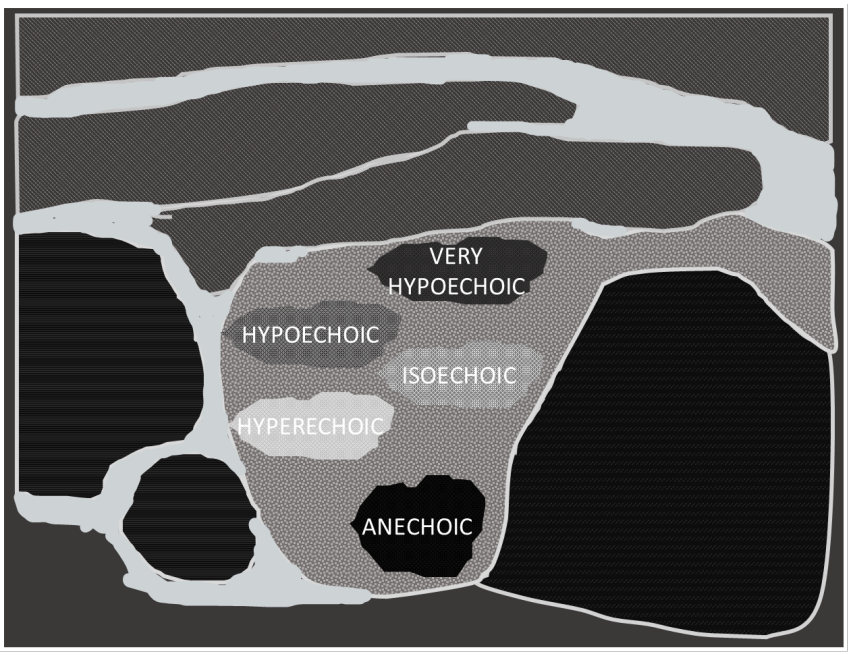

Figure 2.9 – Illustration of different lesion echogenicities. Hyperechoic nodules are brighter than the surrounding normal thyroid tissue, while isoechoic nodules have a similar level of intensity to their surroundings. Hypoechoic nodules are darker than the surrounding thyroid tissue, while very hypoechoic nodules are even darker than the adjacent muscles. Anechoic nodules are cystic, filled with fluid, and appear dark like the blood vessels near the thyroid.

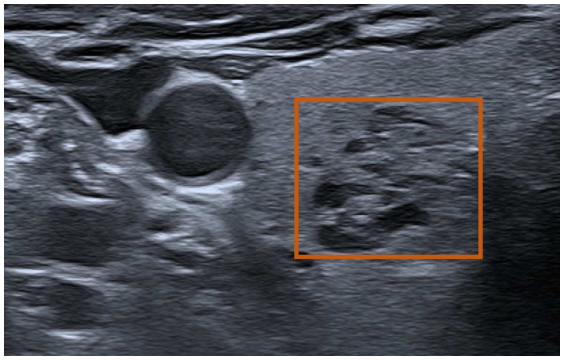

Figure 2.10 – Example of an axial-view ultrasound image containing a thyroid nodule judged as hyperechoic or isoechoic by four expert practitioners. This description is made relative to the echogenicity of nearby thyroid tissue.

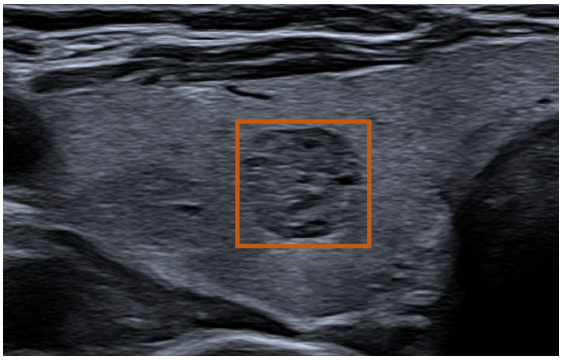

Figure 2.11 – Example of an axial-view ultrasound image containing a thyroid nodule judged as hypoechoic by four expert practitioners. This description is made relative to the echogenicity of nearby thyroid tissue.

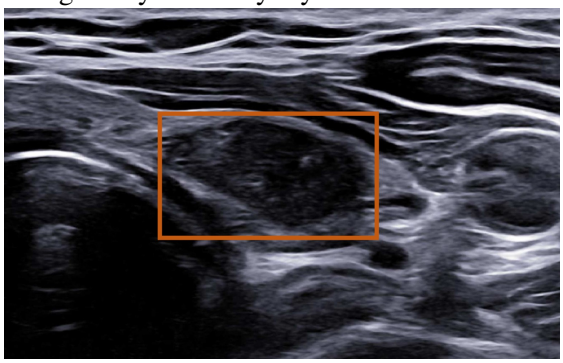

Figure 2.12 – Example of an axial-view ultrasound image containing a thyroid nodule judged as very hypoechoic by four expert practitioners. This description is made relative to the echogenicity of nearby muscles.

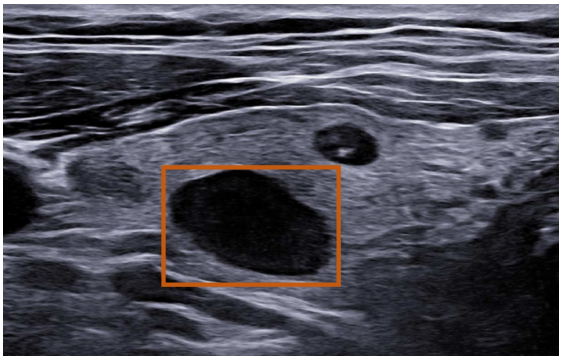

Figure 2.13 – Example of an axial-view ultrasound image containing a thyroid nodule judged as anechoic by four expert practitioners. This description corresponds to fluid-filled lesions.

or if a strap muscle is not visible, it may not be possible to evaluate the echogenicity from a single image.

Despite these difficulties, hyperechogenicity and isoechogenicity have been found to be indicators of benign nodules (Moon et al., 2008 ; Bonavita et al., 2009). Hypoechogenicity compared to adjacent normal thyroid tissue has been found to have an association with malignant nodules, but not as strongly as more marked hypoechogenicity (compared to the muscles) (Moon et al., 2008). These factors make the description of nodule echogenicity extremely valuable for EU-TIRADS scoring.

As can be seen from Figure 2.9, a very/markedly hypoechoic nodule is scored as EU-TIRADS 5, and will be subjected to FNA or surveillance. An anechoic nodule, on the other hand, is likely to receive a score of EU-TIRADS 2 and not be investigated further. Hyperechoic and isoechoic nodules, in the absence of other considerations, are likely to be scored as EU-TIRADS 3, while hypoechoic nodules will be scored as EU-TIRADS 4, with a lower diameter threshold for FNA. A similar stratification of risk is present within the ACR-TIRADS system (Tessler et al., 2017).

2.2.3 Shape

Another aspect of a nodule is its shape, or more specifically its proportions. Nodules are assigned perpendicular axes by readers to in order to generate an estimate of nodule diameter. The relative proportions of these diameters are used as a sign associated with the risk of malignancy. This is presented from an axial or transverse view in Figure 2.14. A taller-than-wide shape is associated with malignancy (Moon et al., 2008). A wider-than-tall, or oval shape, is defined as

having an anteroposterior diameter which is shorter than its tranvserse diameter (see Figure 2.15). In EU-TIRADS, a non-oval shape earns a nodule a score of EU-TIRADS 5, as can be seen in Figure 2.3 (Russ et al., 2017). A taller-than-wide nodule, by contrast, has an anteroposterior diameter which is longer than its transverse diameter (see Figure 2.16). This earns multiple risk score points in ACR-TIRADS (Tessler et al., 2017). When the diameters are estimated to be equal, the nodule is considered to be round and non-oval in EU-TIRADS, and as wider-than-tall in ACR-TIRADS (Russ et al., 2017 ; Tessler et al., 2017).

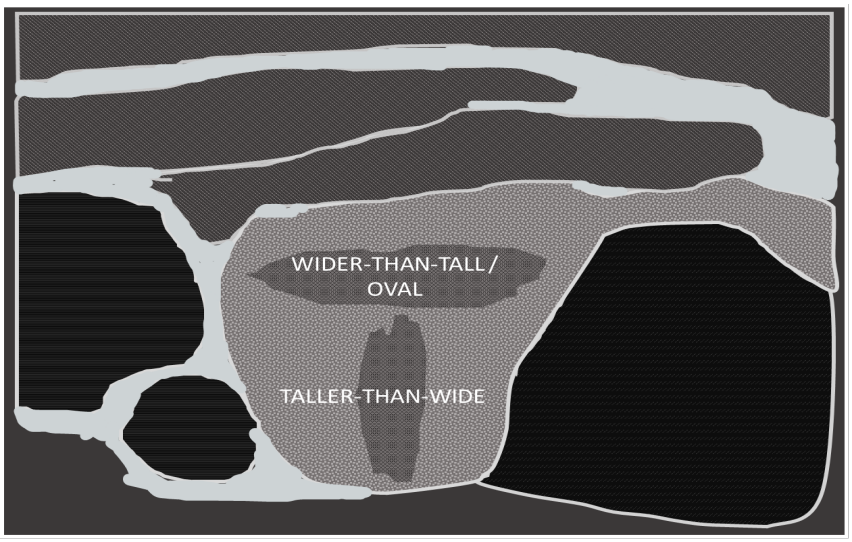

Figure 2.14 – Illustration of different nodule shapes, in an axial or transverse view. A wider-than-tall, or oval, shape is defined as a nodule whose anteroposterior diameter is less than its transverse diameter. A taller-than-wide shape, having the opposite ratio of dimensions, is more associated with malignancy (Russ et al., 2017 ; Tessler et al., 2017).

In clinical practice, variability could potentially arise if experts differ in their identification and estimation of the lengths of these axes.

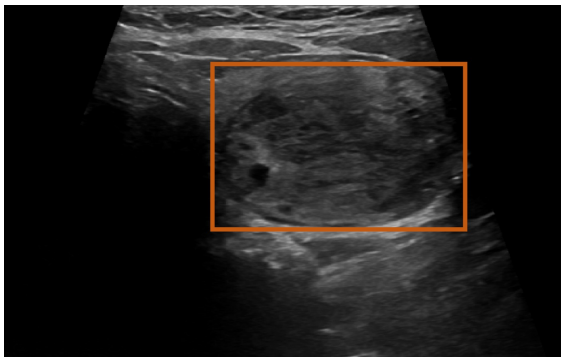

Figure 2.15 – Example of an axial-view ultrasound image containing a thyroid nodule judged as wider than tall by four expert practitioners.

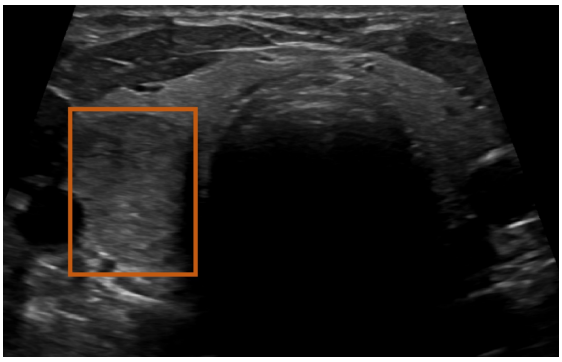

Figure 2.16 – Example of an axial-view ultrasound image containing a thyroid nodule judged as taller than wide by four expert practitioners.

2.2.4 Margin

Moving now outside of the center of the nodule, the margin of a lesion is also an important characteristic related to the risk of malignancy. Different margin descriptors are presented in Figure 2.17.

If the edges of a nodule are clearly demarcated, without irregularities, it may be described as smooth in both EU-TIRADS and ACR-TIRADS, with no associated risk of malignancy (see Figure 2.18). An ill-defined margin is one which, by contrast, is not readily distinguishable from the thyroid parenchyma (see Figure 2.19). In both EU-TIRADS and ACR-TIRADS, this condition does not increase risk scoring (Russ et al., 2017 ; Tessler et al., 2017).

Irregular margins have protrusions that disrupt the smooth curve of a nodule's edge (see Figure 2.20). This category groups together spiculated protrusions, which are more angular, with lobulated protrusions, which are rounded (see Figure 2.17), both of which have been found to be more frequent in malignant nodules (Moon et al., 2008 ; E.-K. Kim et al., 2002). In EU-TIRADS, this finding leads to a score of EU-TIRADS 5, while in ACR-TIRADS it adds a substantial number of points to the risk score (Russ et al., 2017 ; Tessler et al., 2017).

Finally, the label of extra-thyroidal extension suggests that a nodule has invaded the tissues surrounding the thyroid (see Figure 2.21) (Hoang, Lee, Lee, Johnson, & Farrell, 2007 ; Koike et al., 2001). This finding adds multiple risk points in ACR-TIRADS (Tessler et al., 2017). In EU-TIRADS, it is not incorporated into the formal score definition, but is nevertheless a finding that would be reported in addition to the EU-TIRADS score, and lead to further investigation (Russ et al., 2017).

The assessment of the margin depends on the operator's ability to completely examine the edges of a nodule. An individual reader's evaluation of the margin may depend on a subjective decision as to whether or not a slight deviation from a smooth boundary constitutes a lobulation or a spiculation. In addition, it may not be possible to examine some sections of a margin if they are obscured by large calcifications.

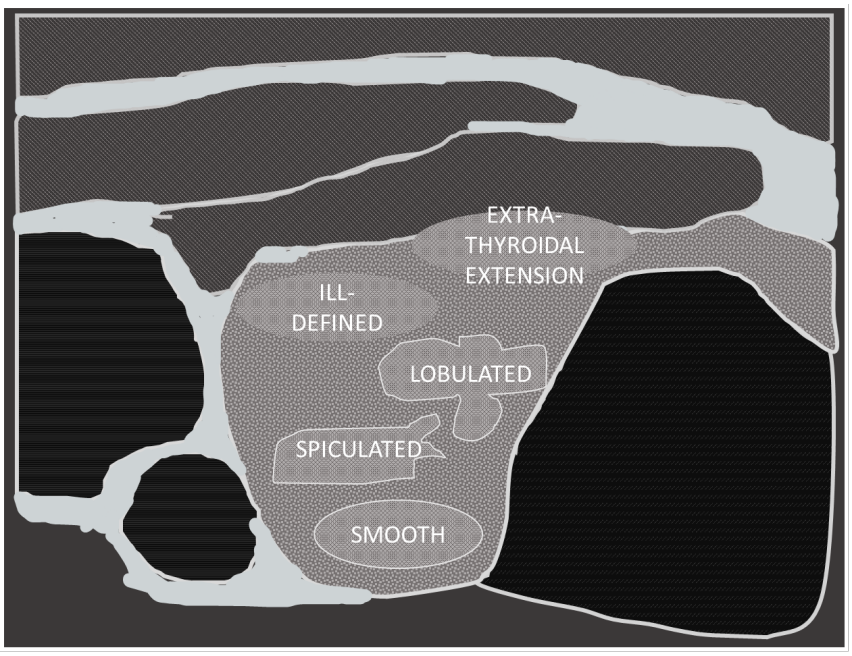

Figure 2.17 – Illustration of different lesion margins. Smooth margins are clearly visible demarcations between the nodule and the surrounding thyroid parenchyma. Ill-defined margins are not readily distinguishable from the thyroid parenchyma. Spiculated and lobulated margins are both considered irregular; the former have sharp, angular protrusions while the latter have smooth, round bumps. Finally, extra-thyroidal extension is an important feature of margins, and describes when a nodule appears to extend beyond the thyroid capsule into adjacent structures.

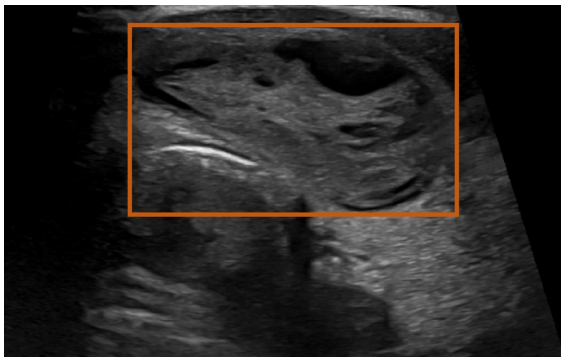

Figure 2.18 – Example of an axial-view ultrasound image containing a thyroid nodule judged to have a smooth margin by four expert practitioners.

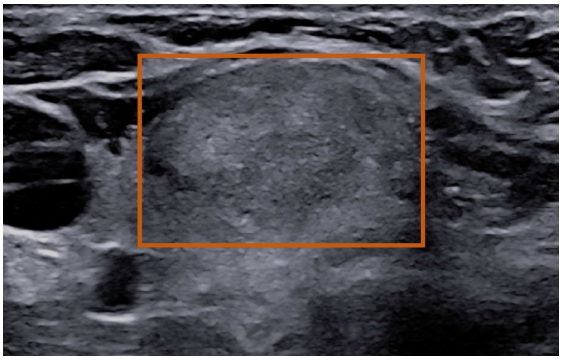

Figure 2.19 – Example of an axial-view ultrasound image containing a thyroid nodule judged to have an ill-defined margin by four expert practitioners.

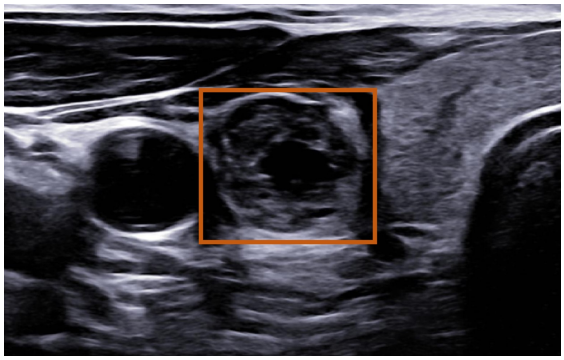

Figure 2.20 – Example of an axial-view ultrasound image containing a thyroid nodule judged to have an irregular margin by four expert practitioners.

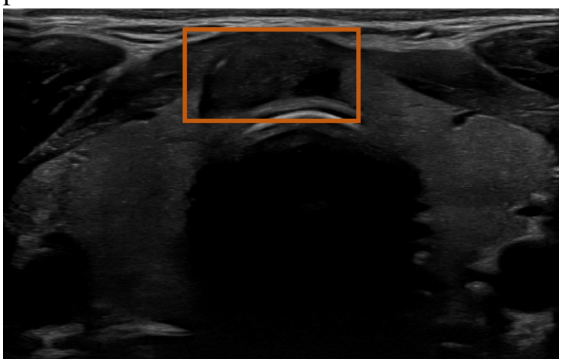

Figure 2.21 – Example of an axial-view ultrasound image containing a thyroid nodule judged to show extra-thyroidal extension by three out of four expert practitioners.

2.2.5 Echogenic Foci

The previous sonographic features are described using mutually-exclusive labels for a particular feature of a nodule's appearance on ultrasound. Echogenic foci represent a collection of different structures which may independently be present or absent within nodules. Broadly speaking, these are small, intensely hyperechoic structures found within the nodule. These are shown in Figure 2.22.

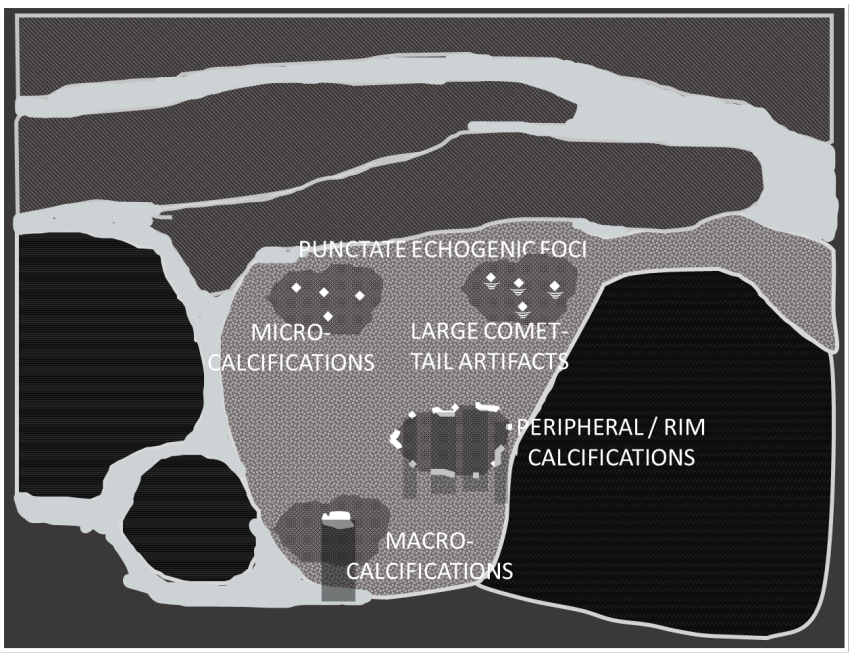

Figure 2.22 – Illustration of different echogenic foci. Punctate echogenic foci are small hyperechoic spots. These may correspond to benign signs such as colloid crystals (which are associated with large comet-tail artifacts) or the back walls of small cysts. However, they may also be associated with microcalcifications, which are associated with malignancy (and do not show large comet-tail artifacts). Macrocalcifications are larger, and generate acoustic shadows behind them. Peripheral or rim or eggshell calcifications are located around the margin of the nodule. (Russ et al., 2017 ; Tessler et al., 2017)

The first category of echogenic foci to consider are punctate echogenic foci, small hyperechoic spots around 1 mm in diameter (see Figure 2.23). This finding can correspond to a number of tissue structures, including microcalcifications, which are highly suggestive of malignancy, being associated with the psammoma bodies often found in papillary thyroid carcinomas (Chammas et al., 2008 ; Malhi et al., 2014 ; Tessler et al., 2017). This finding leads to an EU-TIRADS 5 score (see Figure 2.3), and also adds a substantial number of risk score points in ACR TIRADS. However, the punctate echogenic foci arising from microcalcifications must be distinguished from similar signs that arise from the posterior wall reinforcement of microcystic areas, as well as from colloid crystals. Because colloid crystals typically produce large comet-tail artifacts (see

Figure 2.22), this feature is formally used in ACR-TIRADS to exclude microcalcifications (Malhi et al., 2014 ; Russ et al., 2017 ; Tessler et al., 2017).

Figure 2.23 – Example of an axial-view ultrasound image containing a thyroid nodule judged to have punctate echogenic foci without significant comet-tail artifacts by three out of four expert practitioners.

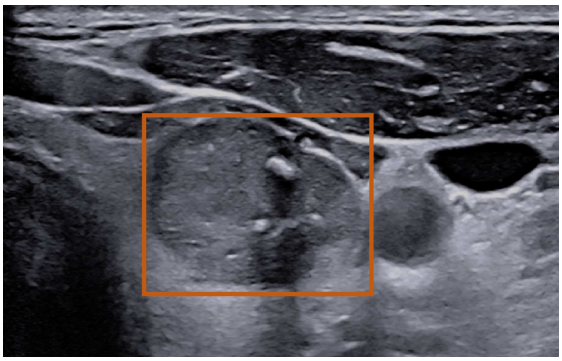

Figure 2.24 – Example of an axial-view ultrasound image containing a thyroid nodule judged to have macrocalcifications by three out of four expert practitioners. These generate posterior acoustic shadows because they reflect much of the ultrasound signal.

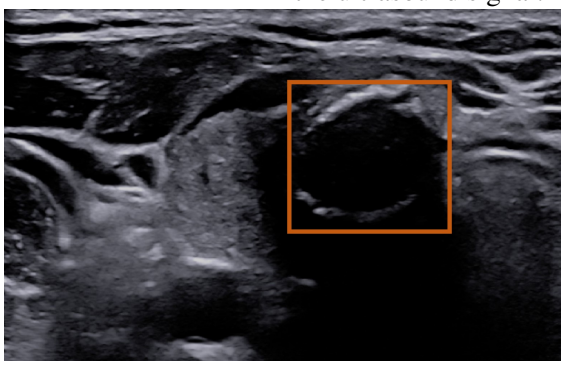

Figure 2.25 – Example of an axial-view ultrasound image containing a thyroid nodule judged to have peripheral calcifications by three out of four expert practitioners. These calcifications can obscure the interior of the nodule.

Macrocalcifications, as the name suggests, are larger than microcalcifications. Their size and intensely hyperechoic nature lead to the generation of acoustic shadows in posterior regions of tissue (see Figure 2.24), because very little ultrasound signal reaches this area and then returns to the probe after traversing the macrocalcification. Their presence may be more frequent in cancerous nodules (Na, Kim, Kim, Ryoo, & Jung, 2016). While macrocalcifications do not figure into the formal EU-TIRADS score, they do add risk score points in ACR-TIRADS (Russ et al., 2017 ; Tessler et al., 2017).

Finally, peripheral, rim, or eggshell calcifications are located in the periphery of the nodule, as seen in Figure 2.25, and may be correlated with malignancy (B. M. Kim et al., 2008). They may not necessarily be continuous, but they can generate acoustic shadows like macrocalcifications,

and thus obscure the interior of the nodule. This can render it impossible to determine the composition or echogenicity of some parts of the nodule. In ACR-TIRADS, peripheral calcifications add risk score points to the nodule evaluation (Tessler et al., 2017).

While the definitions of macrocalcifications and peripheral calcifications are fairly straightforward, the identification of microcalcifications may depend greatly on expert experience. Distinguishing this sign from other causes of punctate echogenic foci is critical in both EU-TIRADS and ACR-TIRADS in order to properly risk-stratify a nodule.

2.2.6 Other

The sonographic features described above are essential to EU-TIRADS and ACR-TIRADS scoring. However, they do not represent the entirety of thyroid ultrasound evaluation. Practitioners are expected to provide a report detailing all important findings, in addition to those used for a TIRADS score. Other signs, such as the presence of a halo, whether the nodule is within a thyroid affected by thyroiditis, or if there are compressive symptoms from the mass of a goiter will all be reported and play an important clinical role in nodule management.

Furthermore, clinical evaluation in France does not consist of interpreting static B-mode images. Practitioners examine nodules from multiple views, repositioning the patient if necessary. They also use additional techniques such as Doppler imaging to examine vascularization within lesions, and elastography to characterize tissue stiffness.

With that said, the features described above are the core of thyroid nodule evaluation. Understanding the utility of machine learning applications to thyroid ultrasound will therefore depend on this foundation.

2.3 Methods

With this review of clinical thyroid ultrasound complete, we are now better-prepared to identify useful machine-learning applications in this domain. However, knowledge of relevant ultrasound features defined by the literature and guidelines is not the same as an appreciation for their role in practice. As with any clinical endeavor, there is variability arising from individual operators, leading to a recent proposal for an international standard lexicon in order to describe nodules more reproducibly (Durante et al., 2023).

Inter-reader variability has been examined in the European context by studies comparing TIRADS scores and relevant feature labels assigned by different European practitioners. Grani et al. performed a comparison between two Italian readers to evaluate the agreement on ultrasound images acquired by a single operator using multiple different TIRADS systems, including ACR-TIRADS, EU-TIRADS, and K-TIRADS (Grani et al., 2018). While the two readers were not always in agreement on nodule descriptions, the TIRADS guidelines allowed them to issue uniform predictions on FNA recommendations, particularly after a series of consensus discussion meetings. This was an encouraging finding, albeit one limited by the fact that both experts were from the same institution (Grani et al., 2018).

A much larger group of European experts, reading from ultrasound clips, was more recently studied by Solymosi et al. (Solymosi et al., 2023). These clips were all acquired at a single clinic and examined by 7 experts from different European countries. Their thorough analysis revealed a high degree of inter-reader disagreement on the identification of a number of ultrasound fea-

tures (Solymosi et al., 2023). This finding highlighted the significant differences between expert operators, though only one French expert participated in the study.

In our case, the variability among French experts in particular must be studied for a complete understanding of thyroid nodule ultrasound in France. This assessment also demands that ultrasound images be acquired from across different centers in the country, to reflect the spectrum of operator practice as well.

To this end, we conducted a multicentric study of the inter-reader reproducibility of thyroid nodule ultrasound interpretation. This involved evaluation by multiple expert readers, representing the standard of clinical practice in this country, of B-mode images acquired in the course of routine clinical practice. The goal was to examine agreement and disagreement in the identification of the sonographic features described above, as well as to study differences between experts' EU-TIRADS scores arising from factors not captured by these well-described features.

2.3.1 Study Design

In order to best capture inter-reader variability, the study was structured to allow clinical experts to independently evaluate axial-view thyroid ultrasound images acquired from different clinical centers. Four experts in thyroid nodule ultrasound, each with over 15 years of experience, acquired images according to their standard clinical practice. All experts then independently labeled all of the acquired images with EU-TIRADS scores in addition to an inventory of sonographic characteristics.

2.3.1.1 Image Acquisition

All images used in this study were anonymized and collected during the course of clinical practice, in accordance with local GDPR regulations. Each of the four participating experts submitted images acquired during routine thyroid nodule ultrasound, with the following exclusion criteria :

— Nodule too small for clinical consideration, as judged by the expert
— Multinodular goiter rendering impossible the identification of a single nodule of interest
— Nodule without adjacent parenchyma or strap muscles as a reference for echogenicity

The images were acquired in axial and sagittal views, with the nodules centered in the field of view. They were saved without annotations, and exported in DICOM format with anonymization of the metadata.

2.3.1.2 Image Evaluation

The anonymized images were then evaluated independently by all four experts. Only the axial views were used, to save time and to avoid adjudication of cases in which a particular nodule feature was present in one view but not the other. Before beginning this evaluation, the experts were familiarized with the evaluation procedure. They were presented with each of the feature definitions that would be used, and given the opportunity to question and refine these criteria in order to synchronize their reading of the static images.

The evaluation process itself consisted of reviewing the axial-view images one by one, during sessions of two to three hours. Because of the potential evolution of experts' tendencies or biases during this process, images were presented in a sequence alternating between sites of origin, so as to distribute the effects of expert adaptation to the task across all four sets of images.

The evaluation process consisted of assigning each image an EU-TIRADS score, in addition to an inventory of other sonographic features: composition, echogenicity, shape, margin, and the presence of echogenic foci. An overview of the process is presented in Figure 2.26. The first step upon initially viewing the image was to quickly provide a subjective score, akin to a Likert scale, to capture the expert's initial impression of the nodule. This was obtained first in order to provide an estimate of the evaluator's subjective judgment, and to prime them to evaluate as they would in a clinical context.

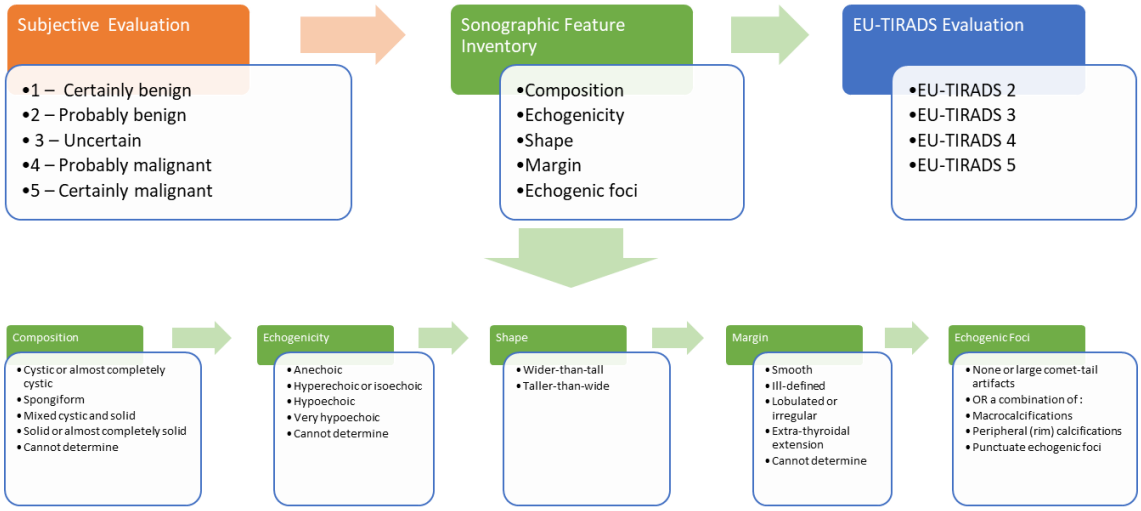

Figure 2.26 – Evaluation inventory used by the experts. Evaluation began with a subjective evaluation based on an initial impression, followed by an inventory of sonographic features, before culminating in an EU-TIRADS score.

Afterwards, the experts proceeded through an inventory akin to the ACR-TIRADS categories, describing different features of the nodule. These categories were used as a system for describing sonographic characteristics, and not as a formal basis for nodule evaluation, as the experts all use EU-TIRADS in their daily practice. No points were added up to imitate ACR-TIRADS scoring, and experts were asked to complete the entire inventory even if, for example, the nodule was judged to be spongiform (unlike in ACR-TIRADS) (Tessler et al., 2017). After this inventory, the experts assigned an EU-TIRADS category according to their personal practice.

This process was then repeated for the next image, in the same order for all experts.

2.3.1.3 Analysis of Evaluations

Following the phase of individual evaluations, the assigned labels were compared between experts in order to assess three different aspects of expert agreement:

1. the inter-reader reproducibility of EU-TIRADS scores and sonographic feature labels
2. the associations of inter-expert EU-TIRADS disagreements with disagreements in specific sonographic feature labels
3. the degree to which EU-TIRADS score disagreements arose despite identification of the same sonographic features

The purpose of this approach was to investigate EU-TIRADS score disagreement, the specific sonographic features most often associated with these disagreements, as well as possible differences within the mental models of EU-TIRADS evaluation of the different experts.

2.4 Results

2.4.1 Images and evaluations

The details of the images acquired are given in Table 2.4.1. As attested to in the table, the four experts each submitted images acquired during the course of their routine clinical practice, using their own ultrasound systems. An important exception was the case of Expert 1 who also submitted images acquired at the site of Expert 3, as a result of his own occasional clinical practice there. The frequency of acquisition used was also guided by the operator's own clinical practice and intuition, and therefore not standardized (though it was limited by the frequency ranges of their ultrasound probes).

System	Site	Operator	Frequency	Number of Images
Esaote MyLab Nine	Site 1	Expert 1	4-15 MHz	55
Canon Aplio 800 Prism	Site 2	Expert 2	10-18 MHz	33
Supersonic Explorer Mach 30 SSI	Site 3	Expert 3	5-18 MHz	149
Supersonic Explorer Mach 30 SSI	Site 3	Expert 1	5-18 MHz	8
GE Healthcare Logic E9	Site 4	Expert 4	2.4-18 MHz	58
Total				303

Images were evaluated during multiple independent sessions by all four experts. The average evaluation time per image was 1 minute and 14 seconds, with a standard deviation of 49 seconds. This represented a substantial time investment on the part of the experts, who scheduled sessions over a ten week period.

2.4.2 EU-TIRADS Results

The results of the annotation process generated labels from all four experts on all 303 images. The experts were not always in agreement on the EU-TIRADS score, as illustrated in Figure 2.27. A consensus of three out of four or four out of four experts, referred to here as a strong consensus, was only obtained for around 68% of the images. In all other cases, there was no consensus, except in about 14% of images for which two experts agreed on a label while the other two each submitted a different label. However, this form of weak consensus most likely signals an ambiguous image, given that four expert readers were able to produce three different labels.

Evaluation of the distribution of different EU-TIRADS scores among the acquired images is difficult in the absence of ground truth labels. However, we can examine both the total number of times each EU-TIRADS score was assigned by each expert, as well as the number of cases for which a strong consensus was obtained in order to determine which labels were over- and underrepresented.

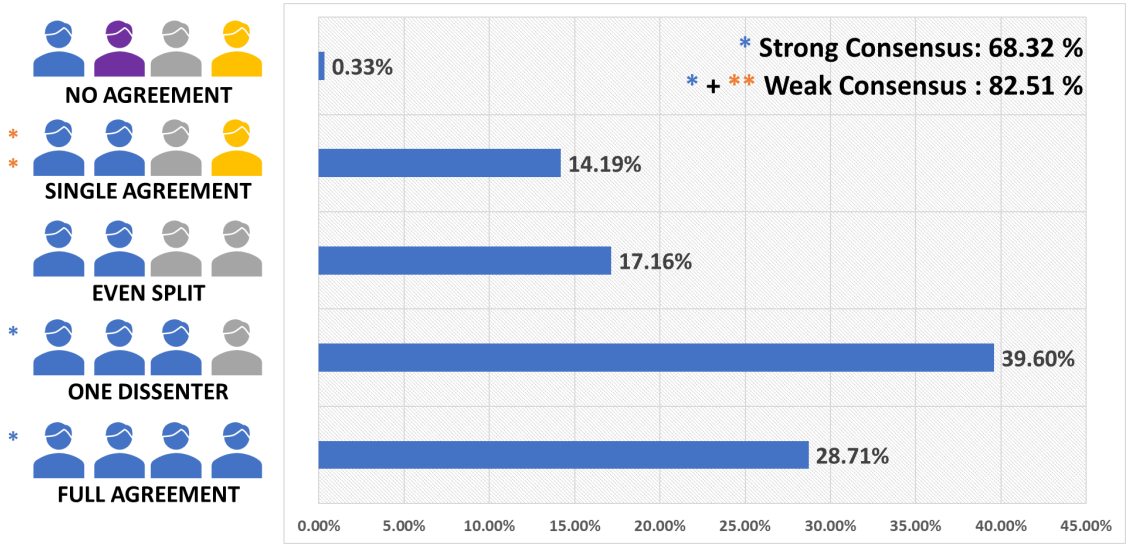

Figure 2.27 – Percentage agreement among the four experts on EU-TIRADS labels for the 303 images.

The total numbers of images assigned each EU-TIRADS score by each expert, as well as the average number of times each scores was assigned across all four experts, are presented in Table 2.2. EU-TIRADS 2, the score corresponding to nodules which are most likely benign, was used uniformly infrequently by all experts, while EU-TIRADS 3 and EU-TIRADS 4 were more common. When compared to the estimated frequencies of label use given by the EU-TIRADS guidelines (see Figure 2.3), it appears that, on average EU-TIRADS 3 scores were assigned somewhat less frequently than expected (about 35% of cases instead of the expected 65%), and EU-TIRADS 5 scores were assigned somewhat more frequently than expected (about 22% of cases instead of the expected 4%).

It is also evident from Table 2.2 that different experts did not have the same proclivity towards all scores; Expert 4, for example, was far more likely to assign the score EU-TIRADS 5, and less likely to use EU-TIRADS 4 than any other expert. Expert 3, by contrast, very rarely used the EU-TIRADS 5 score. Experts 1 and 2 were similar to each other in the frequency of score use, though Expert 1 used EU-TIRADS 3 slightly less and EU-TIRADS 5 slightly more than Expert 2.

Expert	EU-TIRADS 2	EU-TIRADS 3	EU-TIRADS 4	EU-TIRADS 5
Expert 1	25	81	124	73
Expert 2	25	118	115	45
Expert 3	26	116	122	39
Expert 4	24	103	71	105
Average	25	104.5	108	65.5

Table 2.2: Total labels assigned by each expert for each EU-TIRADS category. The mean value across all experts is presented in the final row.

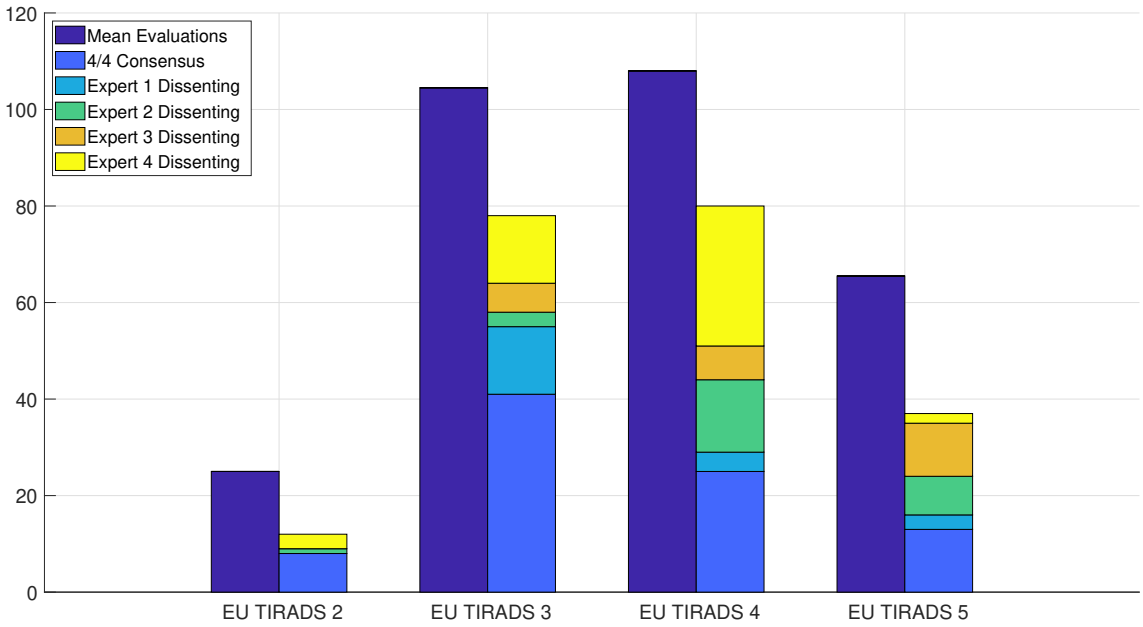

Figure 2.28 – Mean number of overall labels and strong consensus labels assigned by EU-TIRADS score for the 303 images. The bar on the left for each score represents the average number of times the score was assigned across all four experts. On the right, cases of strong consensus for that score are shown, with the bottom bar representing unanimous consensus. The remaining stacked segments of the right-hand bars are cases on which 3 out-of-four experts agreed on the score, sorted by the expert who was the lone dissenter.

These score totals show inter-expert variability in terms of the use of these scores, but not how often these scores were assigned to the same images. To study this, we examine the distribution of scores for which a strong consensus was obtained. These cases, with either four-out-of-four or three-out-of-four experts in agreement, are presented in Figure 2.28, with the cases of three-out-of-four consensus are further broken down by which expert was the lone dissenter against the majority score.

It is clear from Figure 2.28 that, as suggested by the overall consensus rate, the average number of times each score was used far exceeded the number of times the score was assigned with a strong consensus. Strong consensus agreement on EU-TIRADS 2 was limited compared to how frequently the score was used on average. Of the few cases of strong consensus on EU-TIRADS 2, most were unanimous. Among the rare cases of 3-out-of-4 consensus, Expert 4 was the most frequent lone dissenter.

For EU-TIRADS 3, the proportion of cases with strong consensus was more substantial. About half of these consensus cases came from 3-out-of-four agreements, with Experts 1 and 4 being the most frequent lone dissenters. EU-TIRADS 4 had a similar proportion of strong consensus cases, though more of these came from 3-out-of-4 consensus. Experts 2 and 4 were the most common dissenters in these cases, with Expert 4 being the lone dissenter for more 3-out-of-4 cases than there were cases of unanimous agreement. This aligns with the relative underutilization of the EU-TIRADS 4 score by Expert 4 (see Table 2.2).

For EU-TIRADS 5, a strong consensus was obtained only about half as frequently as the average number of uses of this label, with Experts 2 and 3 being the most frequent lone dissenters. From Table 2.2 we see that these were also the least frequent users of this EU-TIRADS score.

It is clear that the four experts did not reach a consensus on a significant proportion of images, and that individual experts had different tendencies in terms of assigning different EU-TIRADS scores. Conceptually, the differences between experts can be divided into two categories: differences in the identification of specific sonographic features and differences in scores assigned for nodules with the same sonographic features, as a result of an expert's own mental framework for EU-TIRADS classification. In the following sections, we examine these two aspects of inter-expert differences.

2.4.3 Reproducibility of Sonographic Feature Identification

We begin by examining differences in the identification of sonographic features. Using the sonographic inventory completed by the experts (listed in Figure 2.26), we can determine how reproducible the detailed characterization of nodules is between experts. The values of Fleiss' kappa for the EU-TIRADS evaluation and for each element of the sonographic inventory are presented in Table 2.3 (Fleiss, 1971). The values of this measure of inter-reader agreement do not have absolute thresholds for interpretation, though values between 0.4-0.6 have been suggested as indicative of moderate agreement (albeit in a two-reader binary case) (Landis & G., 1977). Rather than use these magnitudes as absolute assessments, we can instead compare them to observe that some features were less reproducible than others; most notably the presence or absence of punctate echogenic foci varied greatly among experts. The composition and shape labels have the highest values of Fleiss's kappa, though they remain limited.

Feature	κ	Number of Categories
EU-TIRADS	0.38	4
Composition	0.54	5
Echogenicity	0.39	5
Shape	0.46	2
Margin	0.23	5
Punctate Echogenic Foci	0.004	2
Peripheral Calcifications	0.32	2
Macrocalcifications	0.37	2

Table 2.3: Fleiss' kappa scores among the four experts for each subcategory of the sonographic feature inventory. The number of possible classes for each category from the sonographic feature inventory is also provided.

We will now examine the reproducibility of each of these feature labels among the four experts. In addition to examining the disagreements among these labels, we must also study their associations with disagreements in EU-TIRADS scores. Among the four possible EU-TIRADS

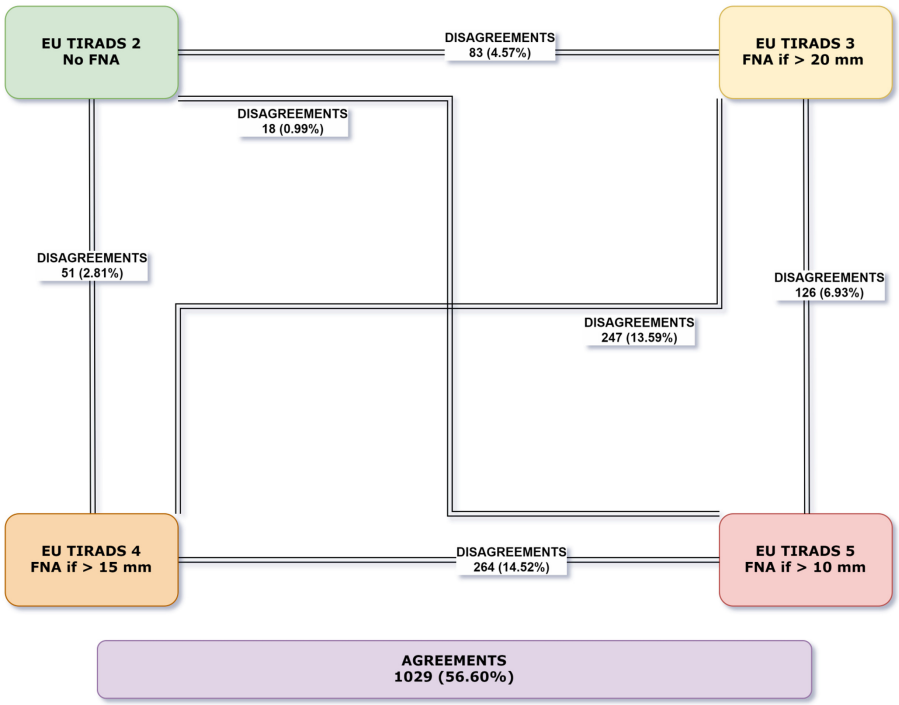

Figure 2.29 – Agreements and disagreements among the four experts on EU-TIRADS labels for the 303 images. The $\binom{4}{2} = 6$ pairwise comparisons for each image yield a total of 1818 agreements or disagreements.

scores, there are $\binom{4}{2} = 6$ possible disagreements for each image, each with a different impact on clinical management in terms of the size threshold for proceeding to FNA (see Figure 2.29).

Between the four experts, there are also $\binom{4}{2} = 6$ pairwise comparisons presenting opportunities for disagreement, the prevalence of which are also shown on Figure 2.29. For each disagreement in EU-TIRADS, we can examine which disagreements among sonographic feature labels are most often associated with it. If cases of disagreement between two particular EU-TIRADS scores are often associated with a specific disagreement in a particular sonographic characteristic, there may be a link worth investigating. While the relationship is not necessarily causal, this approach may allow for the identification of features that are tied to ambiguities in EU-TIRADS scoring.

2.4.3.1 Composition

As seen in previous sections, the composition label is important to nodule evaluation. The percentage consensus among the experts for the composition label is given in Figure 2.30. The overall degree of strong consensus was high compared to the consensus over EU-TIRADS scores (see Figure 2.27), but this was likely affected by an imbalance in the composition labels.

The overall number of evaluations given for each label is presented in Table 2.4, where it is evident that there was an overwhelming predominance of the label "Solid or Almost Completely Solid", while the "Spongiform" label was only rarely used. The "Cystic or Almost Completely Cystic" label was also quite rare, with a more consistent use of the label "Mixed Cystic and Solid".

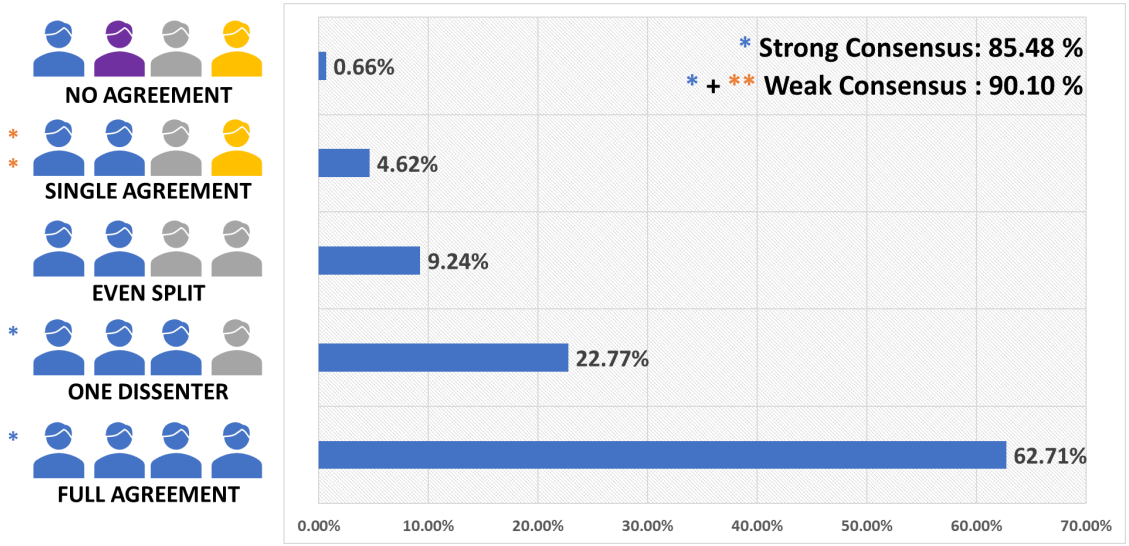

Figure 2.30 – Percentage agreement among the four experts on composition labels for the 303 images.

Only in very few cases did any of the experts signal that the composition could not be determined. The frequency of use of the labels appeared roughly symmetric across the four experts.

Expert	Spongiform	Cystic or Almost Completely Cystic	Mixed Cystic and Solid	Solid or Almost Completely Solid	Cannot Determine
Expert 1	10	20	69	199	5
Expert 2	17	13	45	227	1
Expert 3	13	15	62	211	2
Expert 4	17	37	51	197	1
Average	14.25	21.25	56.75	208.5	2.25

Table 2.4: Total labels assigned by each expert for each composition category. The mean value across all experts is presented in the final row.

In Figure 2.31, we see the cases of strong consensus on composition label. Most of these consensus cases were from nodules judged to be solid, which may have skewed the perception of overall consensus. The strong consensus labels for solid nodules were also overwhelmingly unanimous.

For the other categories, the rate of strong consensus was lower compared to the average number of evaluations received. Cystic and mixed nodule labels had consensus rates of around half of their average number of uses, with no one expert being a noticeably frequent dissenter in cases of 3-out-of-4 consensus. Agreement about the spongiform label, already very infrequently used by all four experts, was almost non-existent. We also examined which composition disagreements

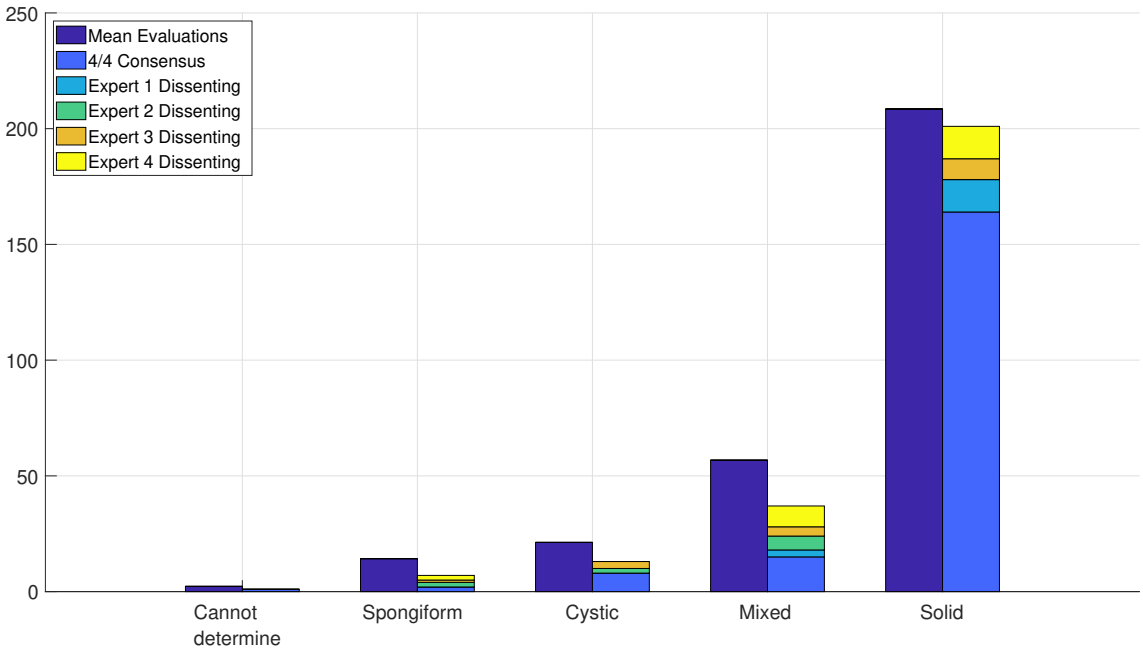

Figure 2.31 – Mean number of labels and strong consensus labels assigned by composition category for the 303 images. The strong consensus labels also include 3/4 consensus cases sorted by the expert who was the lone dissenter.

were most commonly associated with differences in EU-TIRADS score (associated with at least a third of a particular score disagreement) in Figure 2.32. The strongest association was with disagreements between the scores EU-TIRADS 2 and EU-TIRADS 5, of which in 50% of cases the expert who assigned the score EU-TIRADS 2 labeled the composition as cystic, while the expert who assigned the score EU-TIRADS 5 labeled the composition as solid instead. The overall number of EU-TIRADS 2 vs. EU-TIRADS 5 disagreements, however, was low.

The other strong association was between EU-TIRADS 2 and EU-TIRADS 4 disagreements, which were aligned in 47.06% of cases with a disagreement over the spongiform (corresponding to the expert who assigned EU-TIRADS 2) vs. solid (corresponding to the expert who assigned EU-TIRADS 4) composition labels. However, the EU-TIRADS 2 to EU-TIRADS 4 disagreements were also rare.

Overall, the composition labels were dominated by solid nodules, with little agreement among the four experts on spongiform or cystic labels. Disagreements between these latter two categories and solid nodules were the composition disagreements most often associated with disagreements in EU-TIRADS scores.

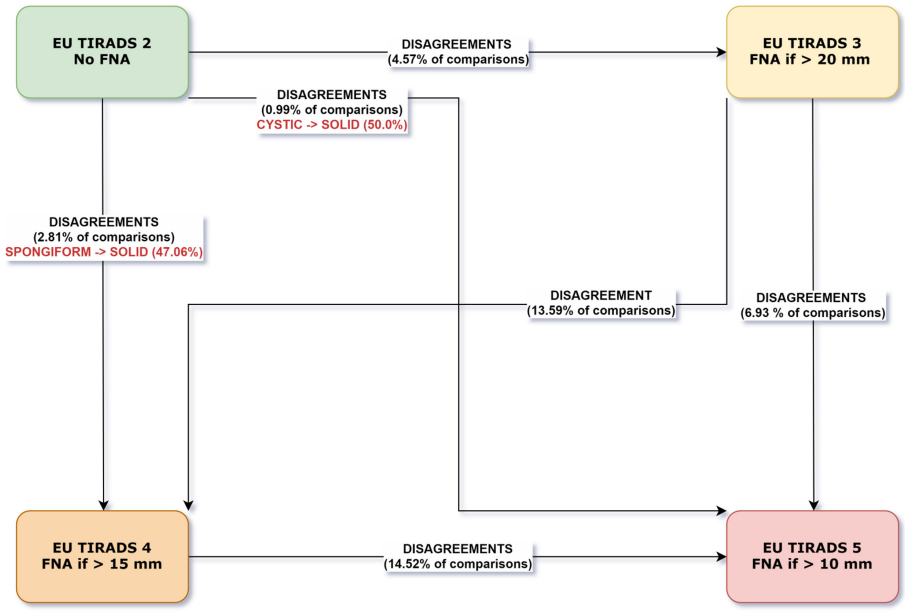

Figure 2.32 – The composition disagreements most commonly associated with disagreements in EU-TIRADS label.

2.4.3.2 Echogenicity

Echogenicity is a sonographic feature which, as discussed, has many potential labels that depend on comparison with other regions of the image. The percentage consensus among the experts for the echogenicity labels assigned to images is given in Figure 2.33. The overall degree of strong consensus is reasonably high, at around 76%.

When examining the overall number of labels assigned by each expert in Table 2.5, we notice some imbalances. There is a marked predominance of the labels "Hyperechoic or Isoechoic" and "Hypoechoic", with very few "Anechoic" labels. In the category of "Very Hypoechoic", we notice an extreme divergence in propensity towards use of this label. Expert 4 judged nodules to be very hypoechoic more frequently than all of the other experts combined, while Expert 3 used the label only 8 times. It was relatively rare for all experts to indicate that they could not determine the echogenicity of a nodule.

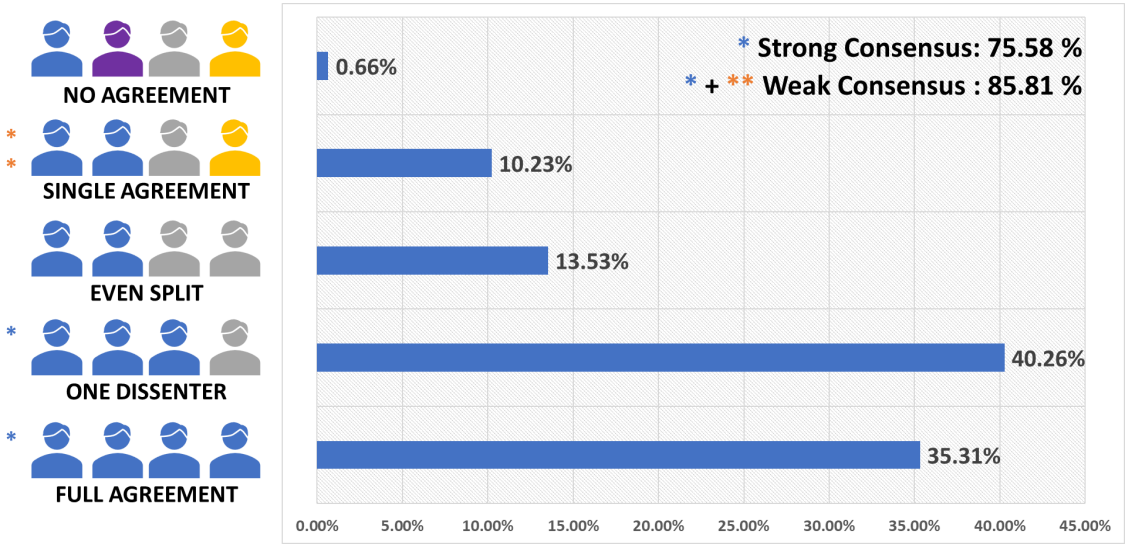

Figure 2.33 – Percentage agreement among the four experts on echogenicity labels for the 303 images.

Expert	Anechoic	Hyperechoic or Isoechoic	Hypoechoic	Very Hypoechoic	Cannot Determine
Expert 1	17	97	144	43	2
Expert 2	10	131	148	12	2
Expert 3	16	131	147	8	1
Expert 4	12	104	110	76	1
Average	13.75	115.75	137.25	34.75	1.5

Table 2.5: Total labels assigned by each expert for each echogenicity category. The mean value across all experts is presented in the final row.

In terms of labels with strong consensus, seen in Figure 2.34, the rate of agreement between three or four of the experts varied with each category. Compared to the overall low rate of use of the anechoic label, a substantial proportion of cases received strong consensus. The same was true for the hyperechoic/isoechoic label, with about half of consensus being unanimous, and Experts 1 and 4 being the most frequent dissenters from a 3-out-of-4 majority. For the hypoechoic label, there was a similar proportion of unanimous consensus, with most cases of 3-out-of-4 agreement being with the dissension of Experts 2 or 4.

The very hypoechoic label had a very low rate of consensus compared to its overall use, with very few unanimous consensus labels. The most frequent 3-out-of-4 dissenter was Expert 3, congruent with that expert's low overall usage of this label in Table 2.5.

Turning to associations with EU-TIRADS disagreements, the strongest by far was between the hyper-/isoechoic to hypoechoic labels and the EU-TIRADS 3 to EU-TIRADS 4 disagreement (see Figure 2.35). In nearly 90% of these disagreements, the expert who assigned the EU-TIRADS 3 score assigned a hyper-/isoechoic label, while the expert who assigned the score of EU-TIRADS 4 assigned a hypoechoic label. The disagreement between hyper-/isoechoic and hypoechoic labels was also associated, albeit less frequently, with EU-TIRADS 2 vs. EU-TIRADS 4 disagreements as well as EU-TIRADS 3 vs. EU-TIRADS 5 disagreements.

Disagreements about very hypoechoic labels were also associated about in 50% of cases with EU-TIRADS 2 (with anechoic) vs. EU-TIRADS 5 disagreements and EU-TIRADS 4 (hypoechoic) to EU-TIRADS 5 disagreements.

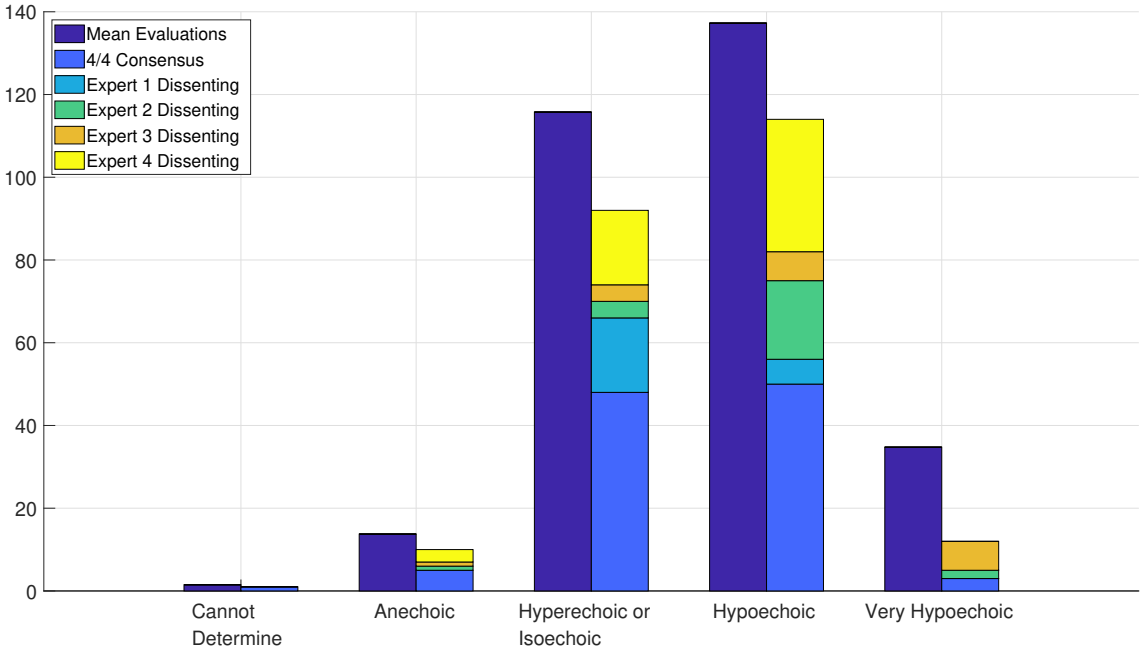

Figure 2.34 – Mean number of labels and strong consensus labels assigned by echogenicity category for the 303 images. The strong consensus labels also include 3/4 consensus cases sorted by the expert who was the lone dissenter.

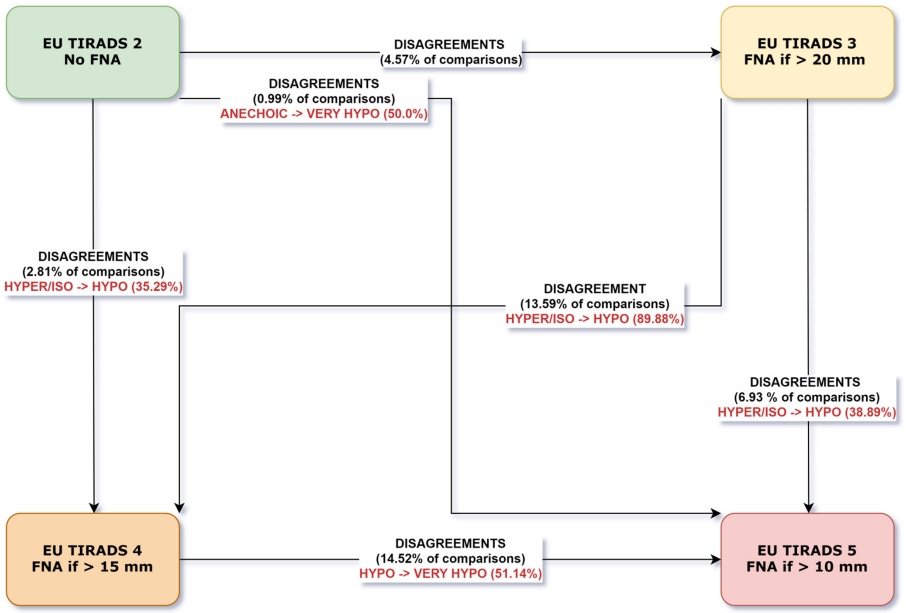

Figure 2.35 – The echogenicity disagreements most commonly associated with disagreements in EU-TIRADS label.

2.4.3.3 Shape

The shape category, having only two possible labels, admits less possibility of disagreement. Such an evaluation is seemingly straightforward, but as the consensus results in Figure 2.36 suggest, there were differences between experts that prevented unanimity in around 23% of cases.

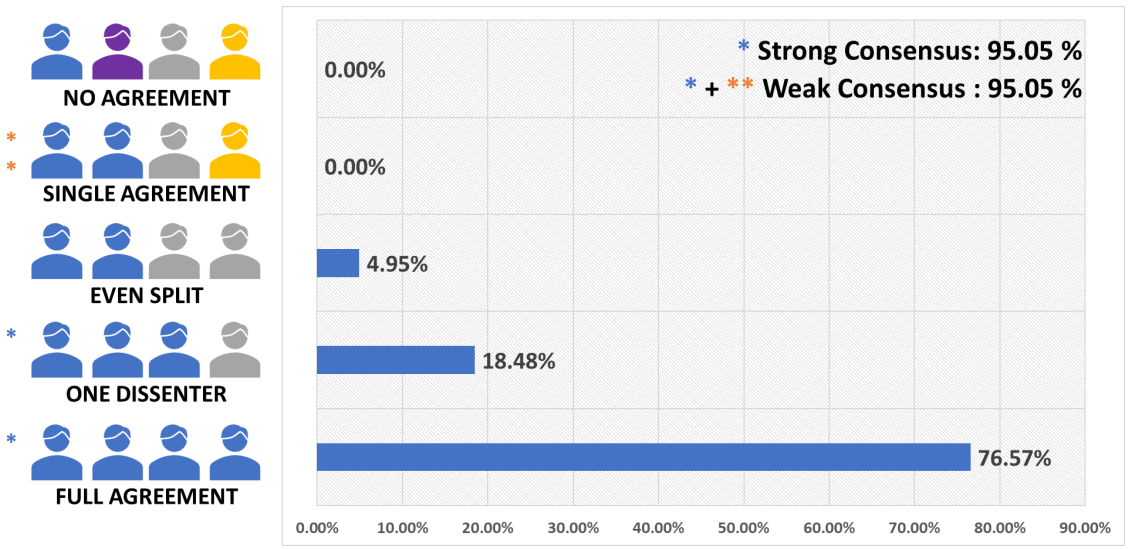

Figure 2.36 – Percentage agreement among the four experts on shape labels for the 303 images.

The total evaluations assigned for the two shape labels are given in Table 2.6. It is apparent that the overwhelming majority of labels were assigned as "Wider-than-Tall", by all experts. Experts 1 and 4 were more likely than the other two to assign the contrary "Taller-than-Wide" label. Virtually all of the images received by necessity a strong consensus.

The strong consensus data in Figure 2.37 reveals that Expert 4 was by far the most frequent lone dissenter against 3-out-of-4 agreements for the wider-than-tall label. The rate of consensus for the taller-than-wide label, however, was only about half the average rate of use of the label. The most frequent dissenter from a 3-out-of-4 taller-than-wide label was Expert 3.

Expert	Wider-than-Tall	Taller-than-Wide
Expert 1	255	48
Expert 2	277	26
Expert 3	281	22
Expert 4	237	66
Average	262.5	40.5

Table 2.6: Total labels assigned by each expert for the two shape categories. The mean value across all experts is presented in the final row.

From Figure 2.38 we see that the only frequent association was between a shape disagreement and a disagreement between EU-TIRADS 3 (wider-than-tall) or EU-TIRADS 5 (taller-than-wide), but this was only observed in around 43% of such disagreement cases.

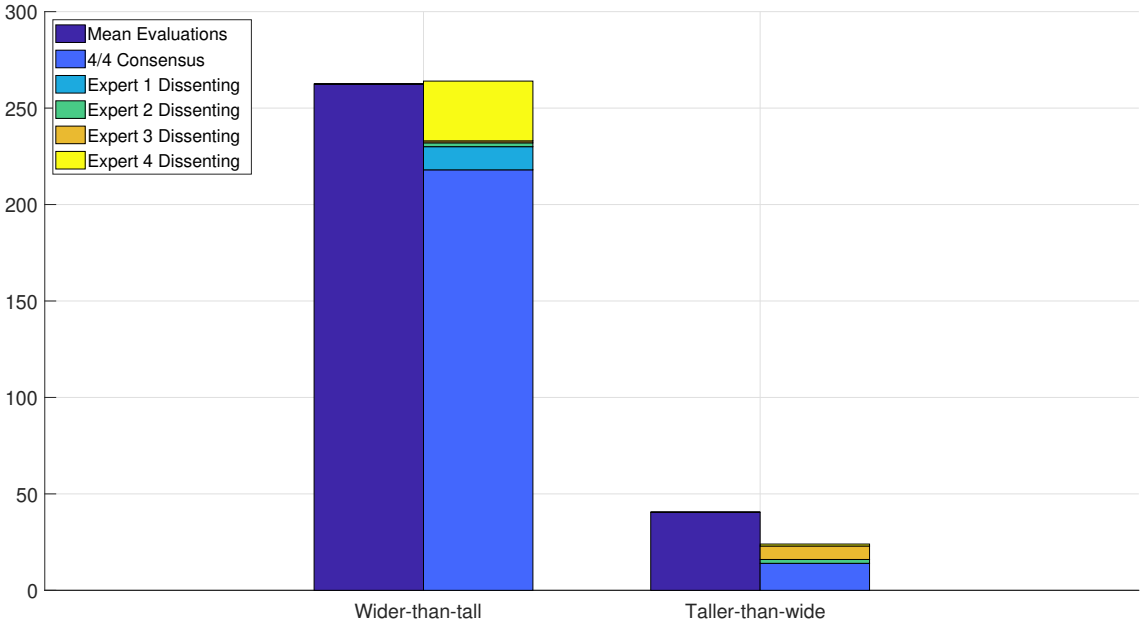

Figure 2.37 – Mean number of labels and strong consensus labels assigned by shape category for the 303 images. The strong consensus labels also include 3/4 consensus cases sorted by the expert who was the lone dissenter.

Overall, the shape category had a high level of consensus due to being a binary label with an overwhelming predominance of the wider-than-tall label. Disagreement over the minority taller-than-wide label was substantial, though there were few associations with this binary disagreement and EU-TIRADS score disagreements.

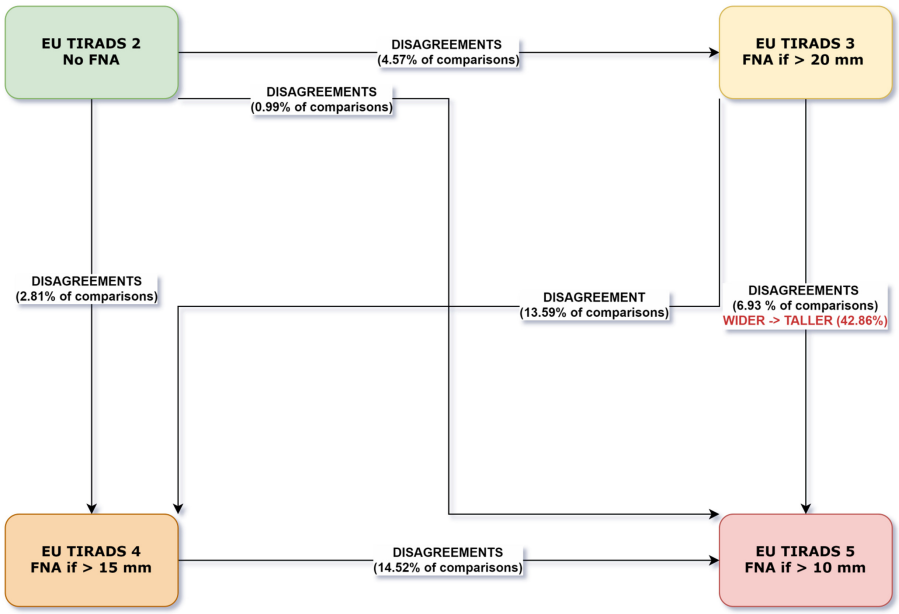

Figure 2.38 – The shape disagreements most commonly associated with disagreements in EU-TIRADS label.

2.4.3.4 Margin

The margin label, describing the edges of the nodule, had an overall rate of strong consensus of around 75%, as shown in Figure 2.39. When we examine the total number of margin labels assigned by category in Table 2.7, we can see that the "Smooth" label was by far the most frequently used. Expert 4 was significantly less likely to use this label than the other experts.

There were fewer labels assigned to "Ill-Defined" and "Lobulated or Irregular", with a great variability in the frequency of use of the latter between experts. Expert 4 used this label more often than the other three experts combined. As for the label of "Extra-Thyroidal Extension", it was quite was rare, with Expert 2 using it most frequently. The experts were almost always able to evaluate the margin.

Expert	Smooth	Ill-Defined	Lobulated or Irregular	Extra-thyroidal Extension	Cannot Determine
Expert 1	228	48	16	9	2
Expert 2	221	45	21	16	0
Expert 3	220	45	29	6	3
Expert 4	142	63	93	4	1
Average	202.75	50.25	39.75	8.75	1.5

Table 2.7: Total labels assigned by each expert for each margin category. The mean value across all experts is presented in the final row.

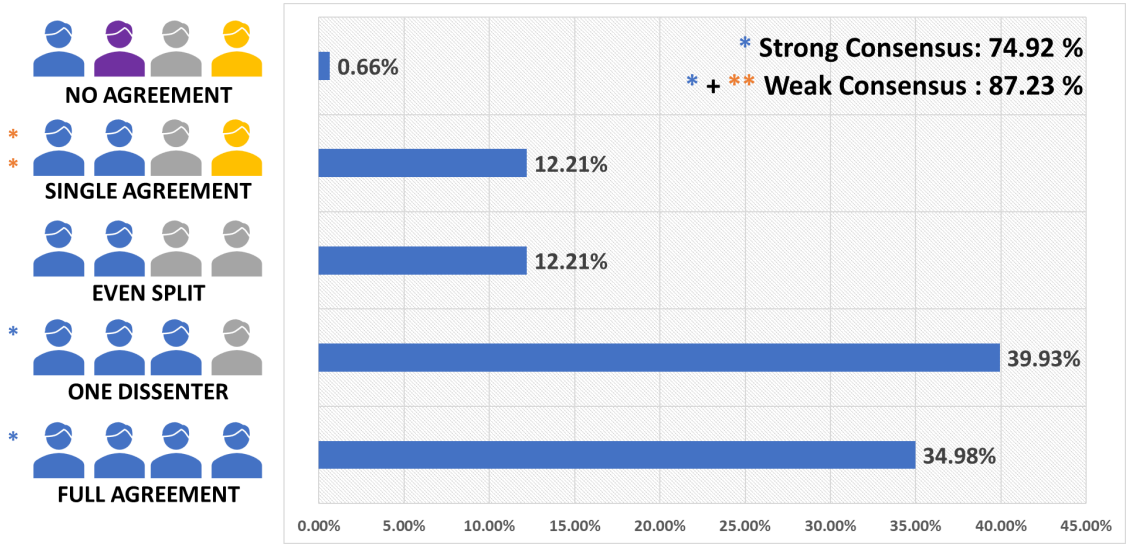

Figure 2.39 – Percentage agreement among the four experts on margin labels for the 303 images.

When examining the labels with a strong consensus in Figure 2.40, it is clear that most of the inter-expert agreement was for the dominant smooth label. The most frequent dissenter from 3-out-of-4 agreements was Expert 4. For the ill-defined, lobulated/irregular, and extra-thyroidal extension margin labels, consensus was rare, even relative to the low average frequency of use of these labels.

Therefore, the margin evaluation was largely dominated by smooth label, with little agreement about any other categories. However, unlike for the previous sonographic features, none of the disagreements in margin label were strongly associated with particular disagreements in EU-TIRADS score.

Overall, the margin category shows a marked predominance of smooth margins, with less agreement over other definitions. No clear associations between these definitions and EU-TIRADS disagreements exist.

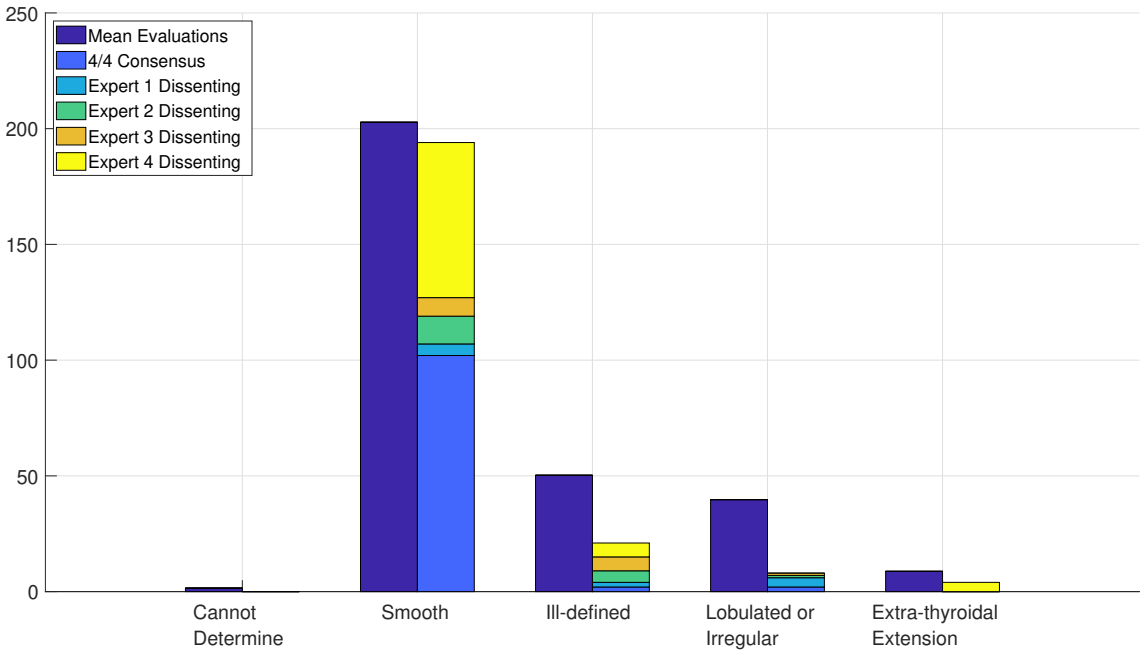

Figure 2.40 – Mean number of labels and strong consensus labels assigned by margin category for the 303 images. The strong consensus labels also include 3/4 consensus cases sorted by the expert who was the lone dissenter.

2.4.3.5 Echogenic Foci

The final category of sonographic feature, echogenic foci, is different, because the labels used to describe it are not mutually exclusive findings. Many nodules could simply be described as not possessing any of these labels. The overall relatively high rate of consensus (84%) for the combinations of echogenic foci labels is given in Figure 2.41.

The breakdown of the labels assigned is presented in Table 2.8, as paired "Present" or "Absent" labels. For all three types of echogenic foci, absence was far more common than presence. For the label "Punctate Echogenic Foci", Expert 4 applied the label for more than a third of all images, while Expert 3 only used it four times. This naturally led to virtually no strong consensus agreement on the presence of punctate echogenic foci corresponding to microcalcifications, as seen in Figure 2.42. The disparities in label use for peripheral calcifications and macrocalcifica-

tions was less significant (see Table 2.8). Figures 2.43 and 2.44 show that consensus for these categories was almost entirely among the dominant absent labels, with unanimous recognition of present labels being virtually nonexistent.

Expert	Punctate Echogenic Foci		Peripheral Calcifications		Macrocalcifications	
	Present	Absent	Present	Absent	Present	Absent
Expert 1	11	292	4	299	10	293
Expert 2	53	250	5	298	25	278
Expert 3	4	299	1	302	2	301
Expert 4	135	168	8	295	26	277
Average	50.75	252.25	4.5	298.5	15.75	287.25

Table 2.8: Total labels assigned by each expert for each echogenic foci category. The mean value across all experts is presented in the final row.

From the associations between echogenic foci label disagreements and EU-TIRADS score disagreements in Figure 2.45, the only strong associations were between disagreements about punctate echogenic foci presence or absence and disagreements about a label of EU-TIRADS 5. From each of the three other EU-TIRADS labels, about 40% of disagreements with EU-TIRADS 5 were associated with a disagreement about punctate echogenic foci, or microcalcifications.

Overall, the scarcity of positive labels in this category makes it difficult to analyze their importance. Among these, the greatest variability seemed to be in the label of punctate echogenic foci being present, which was associated with disagreements about an EU-TIRADS 5 score.

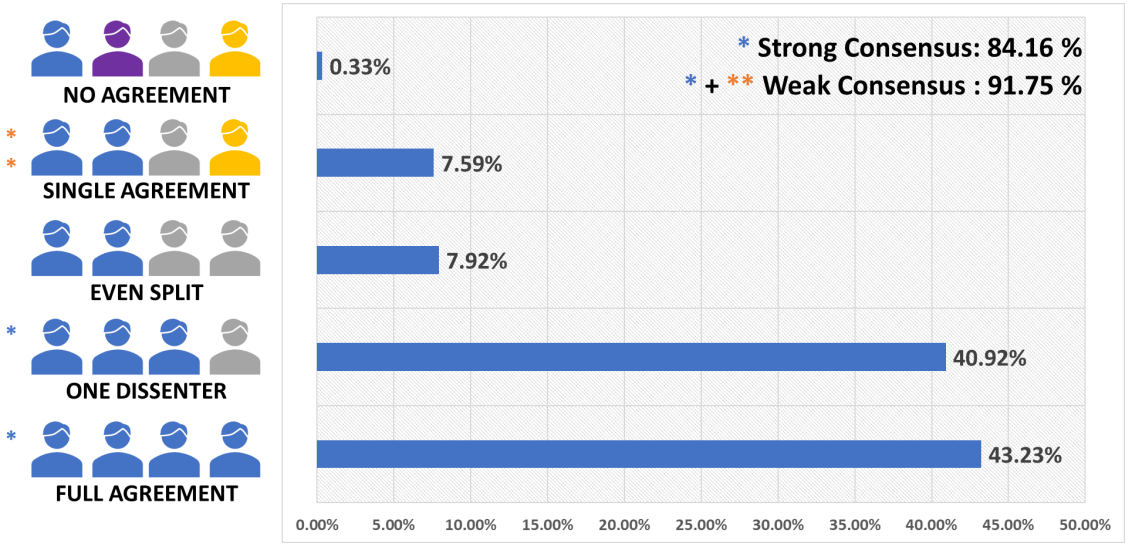

Figure 2.41 – Percentage agreement among the four experts on echogenic foci labels for the 303 images.

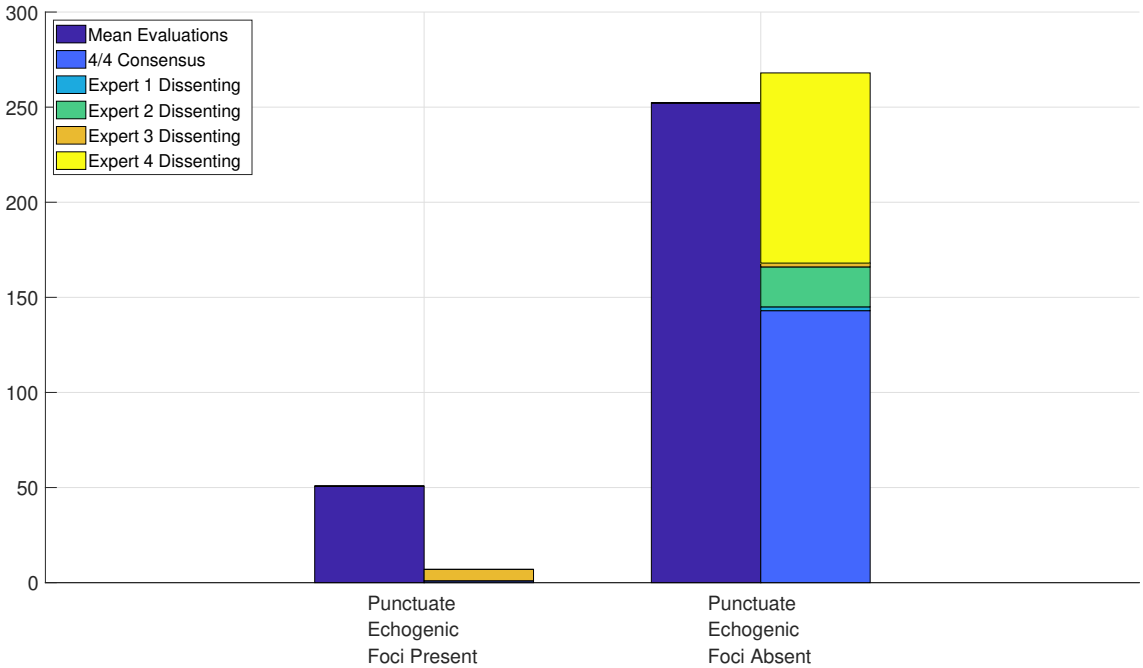

Figure 2.42 – Mean number of labels and strong consensus labels assigned for punctuate echogenic foci for the 303 images. The strong consensus labels also include 3/4 consensus cases sorted by the expert who was the lone dissenter.

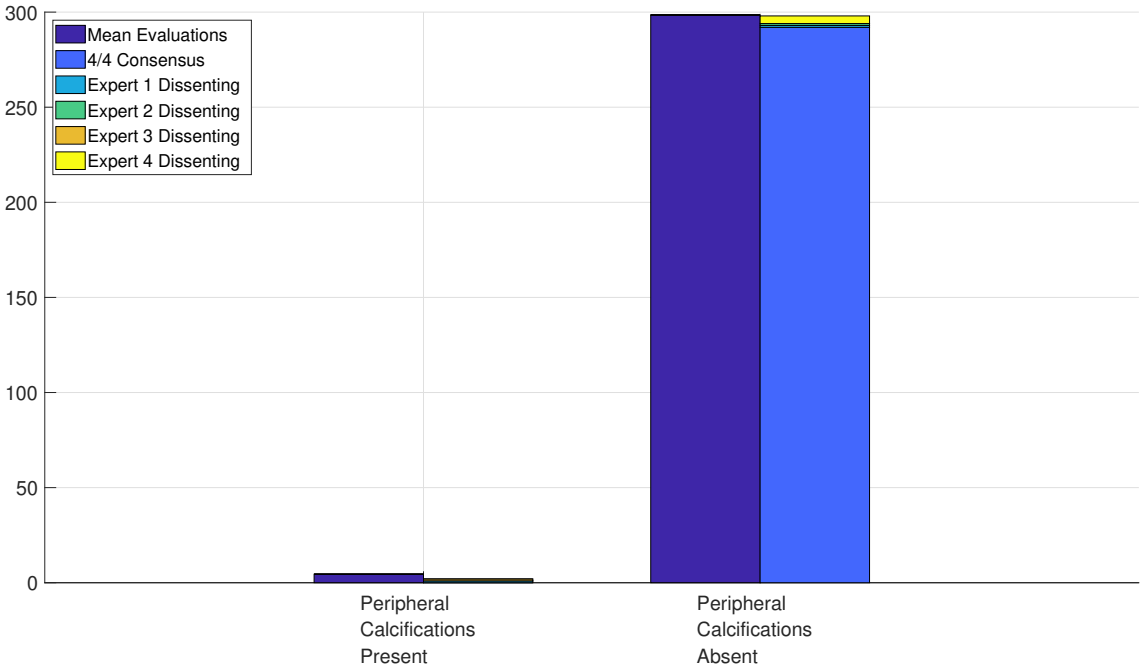

Figure 2.43 – Mean number of labels and strong consensus labels assigned for peripheral calcifications for the 303 images. The strong consensus labels also include 3/4 consensus cases sorted by the expert who was the lone dissenter.

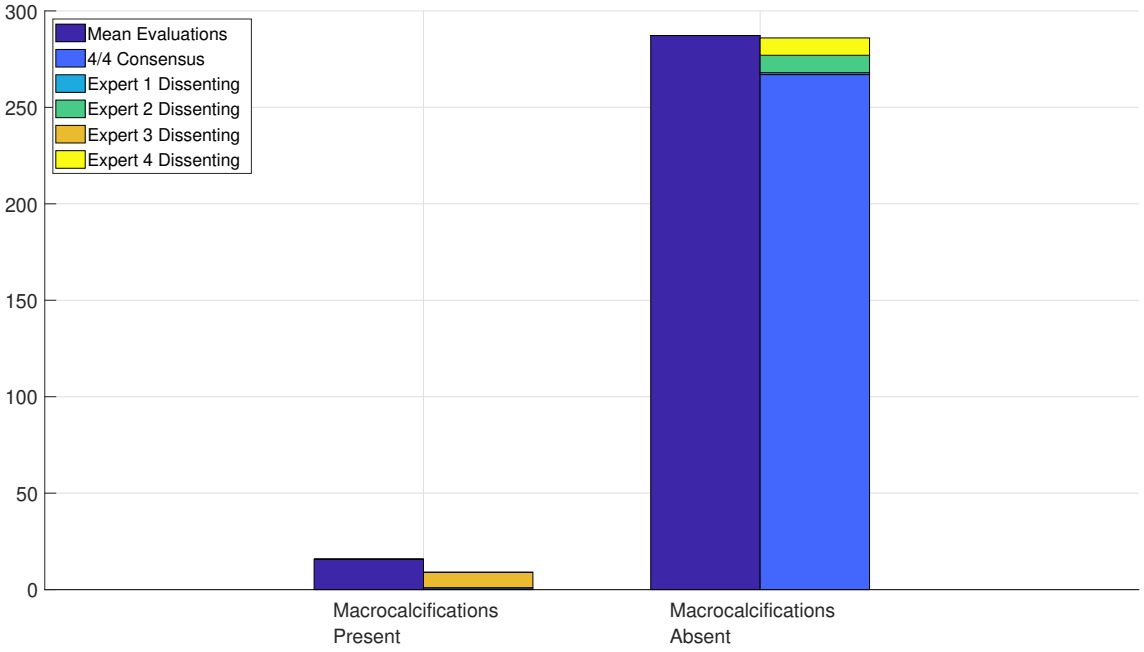

Figure 2.44 – Mean number of labels and strong consensus labels assigned for macrocalcifications for the 303 images. The strong consensus labels also include 3/4 consensus cases sorted by the expert who was the lone dissenter.

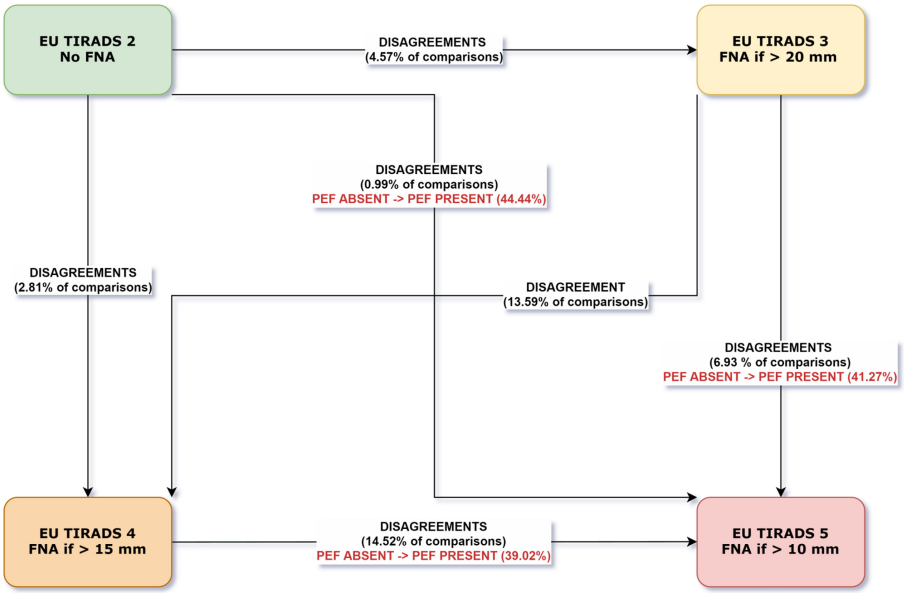

Figure 2.45 – The echogenic foci disagreements most commonly associated with disagreements in EU-TIRADS label. PEF - Punctate echogenic foci.

2.4.4 Other Differences in EU-TIRADS Scoring

In addition to the differences in identification of sonographic features, experts may also assign different EU-TIRADS scores when their sonographic inventory labels are the same. These irregularities arise from differences in the application of the clinical EU-TIRADS algorithm seen in Figure 2.3, and are studied here by examining how experts assign EU-TIRADS scores for nodules having the same combination of all of the sonographic feature labels previously discussed.

These differences can be examined in four ways:

— On an individual expert level, do EU-TIRADS evaluations vary when the same sonographic features have been described?

— Between experts, do EU-TIRADS scores vary when the same combinations of sonographic features have been assigned?

— Do individual expert evaluations differ from a standardized EU-TIRADS guideline?

— Do individual experts weigh sonographic features differently in their EU-TIRADS evaluations?

2.4.4.1 Intra-expert variability within sonographic feature combinations

We begin with an examination of whether individual expert evaluations vary for nodules that received the same sonographic feature labels. For each nodule, the experts assigned a particular combination of these labels as described in Figure 2.26, with multiple nodules possibly receiving the same combination. Other combinations might never be used, or only be used once.

An indication of the differences not captured by the sonographic inventory would therefore be the number of cases in which an expert assigned different EU-TIRADS scores to nodules having the exact same combination of feature labels. The number of unique combinations of feature labels assigned by each expert, as well as the number of combinations that each expert reused for multiple evaluations, is presented in Figure 2.46.

The number of combinations used by each expert varied, with many only having been applied to a single nodule. Expert 4 used the most unique combinations of feature labels, and had the highest number of repeated combinations, with no single combination being used more than 20 times. The other three experts used fewer unique combinations, with a few being applied far more frequently than others. Expert 3 in particular was notable for having used only 40 combinations of feature labels to describe all 303 images, and using one of these to describe upwards of 80 nodules.

The unique combinations of feature labels used by each expert are further explored in Figure 2.47. The combinations that were reused multiple times by each expert were filtered down to those for which the expert did not always apply the same EU-TIRADS score. This was only 6 combinations for Experts 1 and 2, 12 combinations for Expert 3, and 25 for Expert 4. Within these combinations, the corresponding images were then separated into those which received the most commonly-used EU-TIRADS score for that feature combination, and those that did not.

This second group of images that received minority EU-TIRADS labels within their combination of feature labels represents the proportion of nodules for which each expert's EU-TIRADS score differed for reasons independent from the sonographic features that were assigned. For Experts 1 and 2, this represented less than 4% of images, and for Expert 3 this was less than 6%. Expert 4 was the most likely to show variability within a given combination of feature labels, with 41 out of 303 images, or about 13.5% of cases.

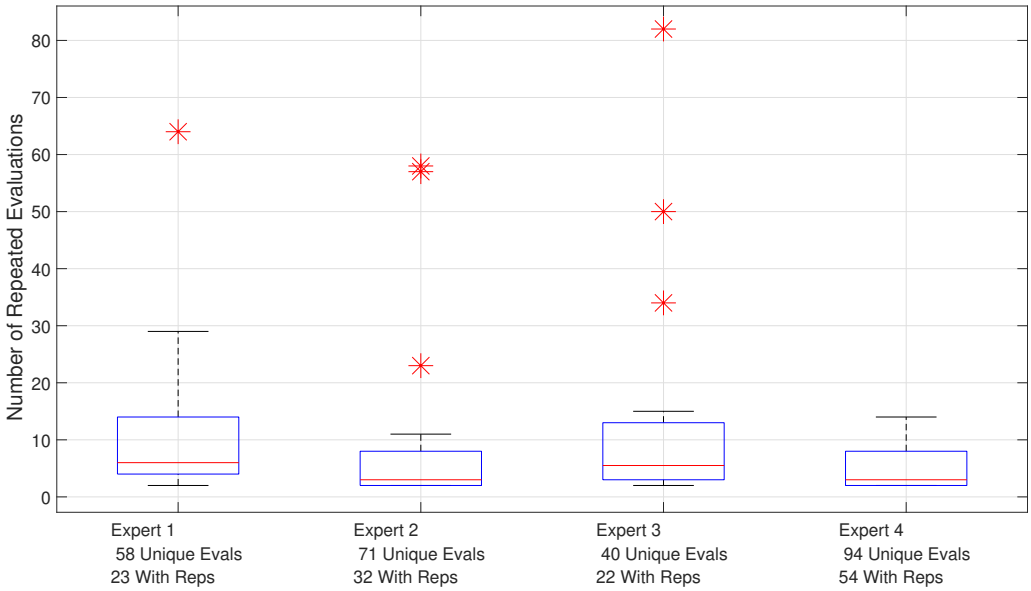

Figure 2.46 – The frequency of use of repeated combinations (used by the same expert on multiple images) of sonographic feature labels by each expert presented as a boxplot. The total number of unique evaluations per expert and the total number of repeated combinations per expert are presented as well.

Overall, the four experts varied in the diversity of feature label combinations that they used. Within combinations, experts rarely varied in their EU-TIRADS score, with the slight exception of Expert 4.

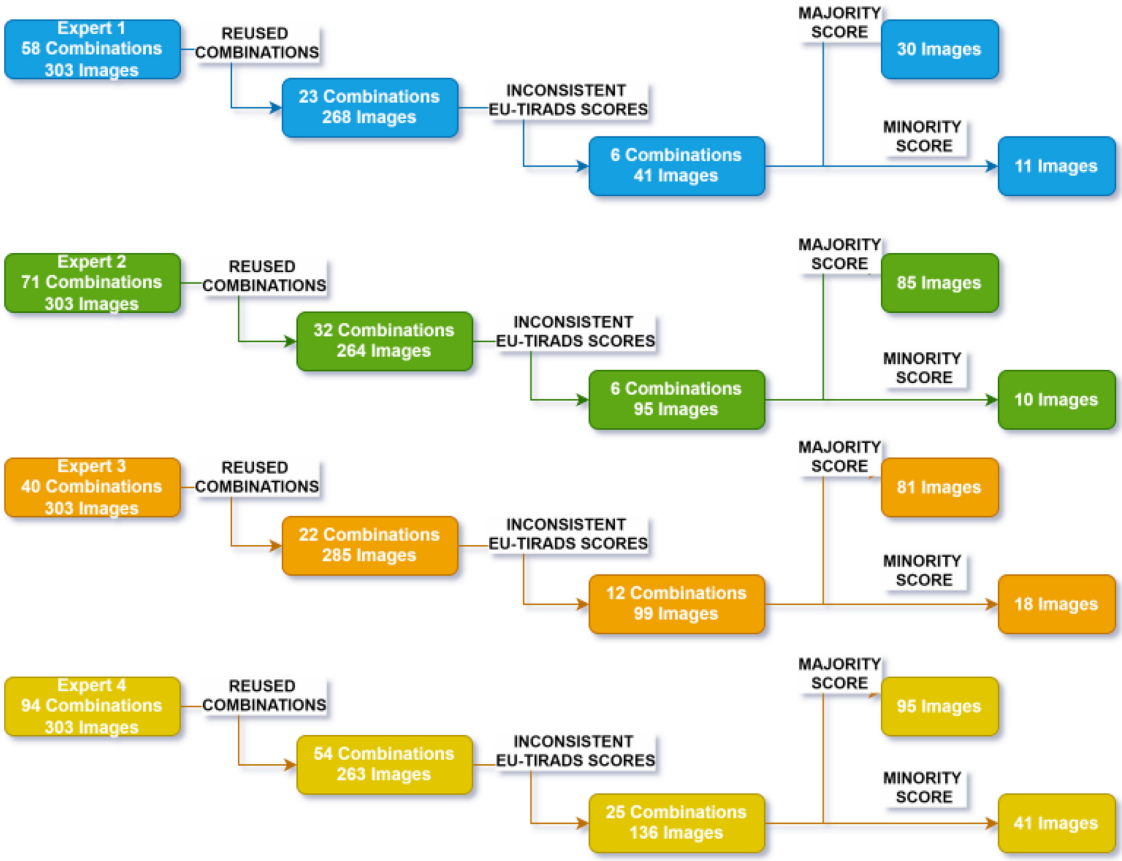

Figure 2.47 – The number of unique feature combinations used by each expert are presented on the left, along with the subset of combinations that were reused for multiple images. These are further filtered down to combinations of feature labels for which the expert did not always apply the same EU-TIRADS score in all cases. Finally, the images corresponding to these categories are separated into those which were labeled in accordance with the majority EU-TIRADS score for their feature combination, and the smaller subset that differed. This latter subset represents potential inconsistencies in EU-TIRADS score attribution on the basis of sonographic feature labels.

2.4.4.2 Inter-expert variability within sonographic feature combinations

After examining the intra-expert variability in scoring for nodules with the same sonographic feature labels, we turn to differences between experts. Given that the four experts frequently disagreed on which feature labels to assign, it is not possible to make this comparison on an image-per-image basis. Therefore, we examined the EU-TIRADS scores assigned by each expert on the basis of combinations of feature labels that all four experts used, even if they were not on the same image. Figure 2.48 shows the combinations of features used by all four experts.

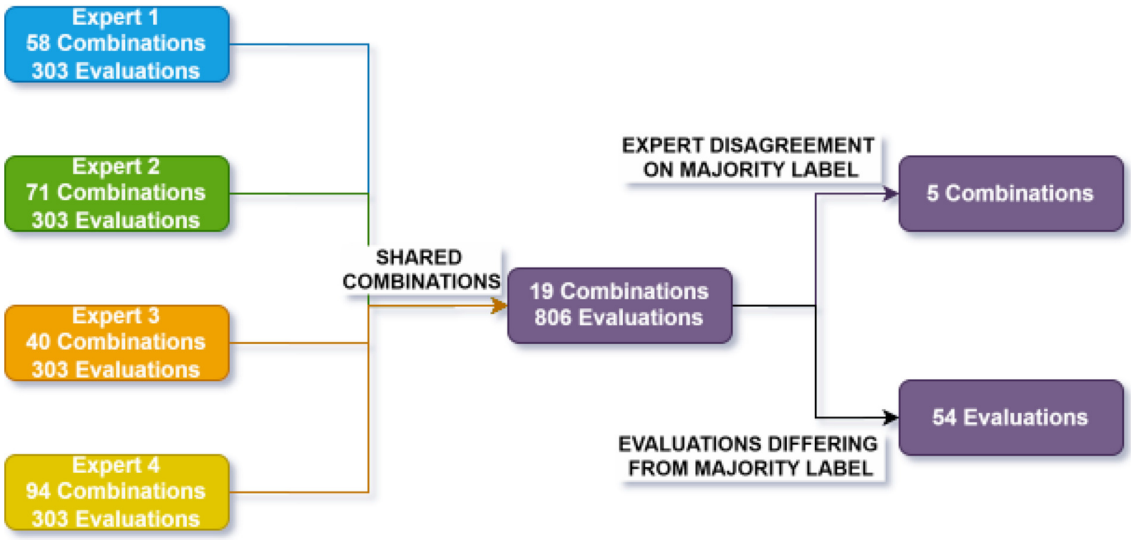

Figure 2.48 – The shared combinations of echographic features used by all four experts. The combinations for which the four experts were not in unanimous agreement about the EU-TIRADS score most often assigned is shown, along with the small number of cases across all shared combinations that disagreed from the majority EU-TIRADS score.

Each expert proposed 303 evaluations, one for each image, using a certain number of unique combinations of sonographic features. Of these combinations, 19 were used by all four experts for a total of 806 evaluations, though not necessarily on the same images. This represented a substantial majority of the 1212 total evaluations. Among these shared combinations, only in 5 did the four experts not agree on the most common EU-TIRADS score. Among the 806 evaluations from 19 shared combinations, only 54 (about 6.7%) differed from the majority EU-TIRADS score.

Overall, the combinations of feature labels used by all four experts accounted for a substantial proportion of evaluations ($\frac{806 \text{ evaluations}}{4 \text{ experts} \times 303 \text{ images}} \approx 67\%$). Comparison on this limited set of shared combinations showed very little variation between experts in terms of assigning EU-TIRADS scores.

2.4.4.3 Expert Variation from the EU-TIRADS Guideline

Another measure of variability within combinations of sonographic feature labels is expert deviation from a standard. We compared expert EU-TIRADS scores to an EU-TIRADS score calculated using the sonographic inventory labels and the algorithm presented in Figure 2.3. The

results of this comparison are presented as confusion matrices in Figure 2.49, and are summarized in Table 2.9.

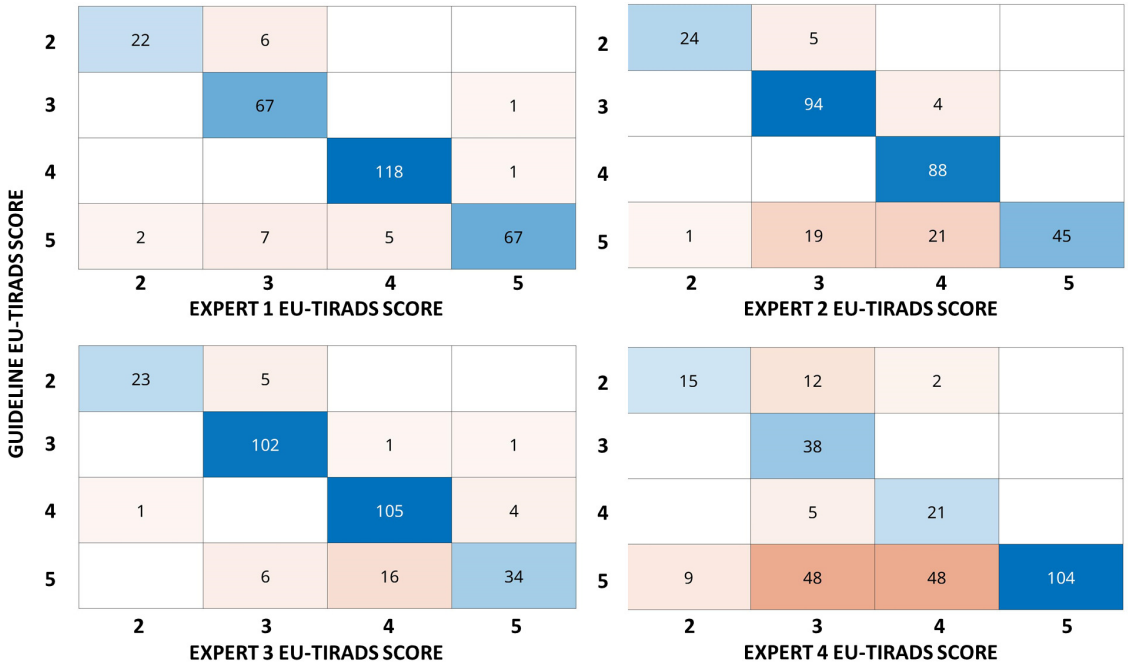

Figure 2.49 – The confusion matrices of the four experts' EU-TIRADS scores, as compared with the guideline-based EU-TIRADs score calculated from the sonographic feature inventory. It was not possible to calculate an EU-TIRADS score for some images with indeterminate composition or echogenicity, thus totals are not the same across all experts.

The degree of agreement by expert with the guidelines varied substantially. Expert 1 did not differ frequently from the guideline score, with most disagreements being with the guidelines-based EU-TIRADS 5 score. Expert 3 had a similar profile, albeit with slightly more frequent disagreements, also mostly around EU-TIRADS 5 scores. Experts 2 and particularly Expert 4 differed far more frequently from the guideline-based EU-TIRADS score.

Expert 2 also had disagreements predominately concentrated in an under-scoring of EU-TIRADS 5 to EU-TIRADS 4 or EU-TIRADS 3. Expert 4 had even more frequent disagreements of this nature, in addition to a substantial number of disagreements from EU-TIRADS 2 to EU-TIRADS 3.

Expert	Images with Guideline EU-TIRADS Score	Images with Different Expert Score	% with Different Score
Expert 1	296	22	7.4%
Expert 2	301	50	16.6%
Expert 3	298	34	11.4%
Expert 4	302	124	41.1%

Table 2.9: Expert differences from guideline-based EU-TIRADS score, calculated on the basis of expert-assigned features. Not all images could be used to calculate a guideline-based EU-TIRADS score, if the composition or echogenicity had been assigned a "Cannot Determine" label.

Overall, the frequency of disagreements from the guideline score seem to indicate significant individual differences for Experts 2 (50 images or 16.6%) and 4 (124 images or 41.1%) for given combinations of feature characteristics. At least as based on the sonographic characteristics identified by each expert, Experts 1 (22 images or 7.4%) and 3 (34 images or 11.4%) seem to follow the EU-TIRADS guideline quite closely.

2.4.4.4 Modeling Decision Trees

If the experts made decisions to assign EU-TIRADS scores in a way that differed from the guideline, their idiosyncrasies would be reflected in the observed score classification algorithm. In order to examine differences in the observed scoring patterns of experts, we considered their EU-TIRADS scores on the basis of identified features compared with a decision tree based on the EU-TIRADS guideline.

The decision tree that was used to imitate the guideline from Figure 2.3 reconstructed the clinical algorithm as a series of binary choices on the basis of identified sonographic features. Multiple potential configurations were possible, but we studied one sequence of label decisions across all experts for the purposes of standardization. Each expert's EU-TIRADS scoring was then studied as it agreed with and departed from the guideline-based decision tree. For the purposes of standardized scoring, expert sonographic inventory labels were converted to EU-TIRADS-relevant descriptors as in the previous section. As the anechoic echogenicity label was meant to correspond only to cystic lesions, these two labels were combined.

Beginning with Expert 1, the guideline decision tree with expert differences is given in Figure 2.50. Expert 1 had relatively few departures from the guideline-based labeling, though most were from EU-TIRADS 5 scores. These differences arose from not assigning an EU-TIRADS 5 score despite the presence of a high-risk feature (see Figure 2.3). Most frequently, a taller-than-wide shape was the high-risk feature that did not lead to a guideline-based EU-TIRADS 5 score. Among these nodules, most but not all appeared to follow the guideline apart from ignoring the taller-than-wide label.

The other common departures for Expert 1 from the guideline were in the scoring of guideline-based EU-TIRADS 2 nodules. This involved assigning EU-TIRADS 3 scores despite the presence

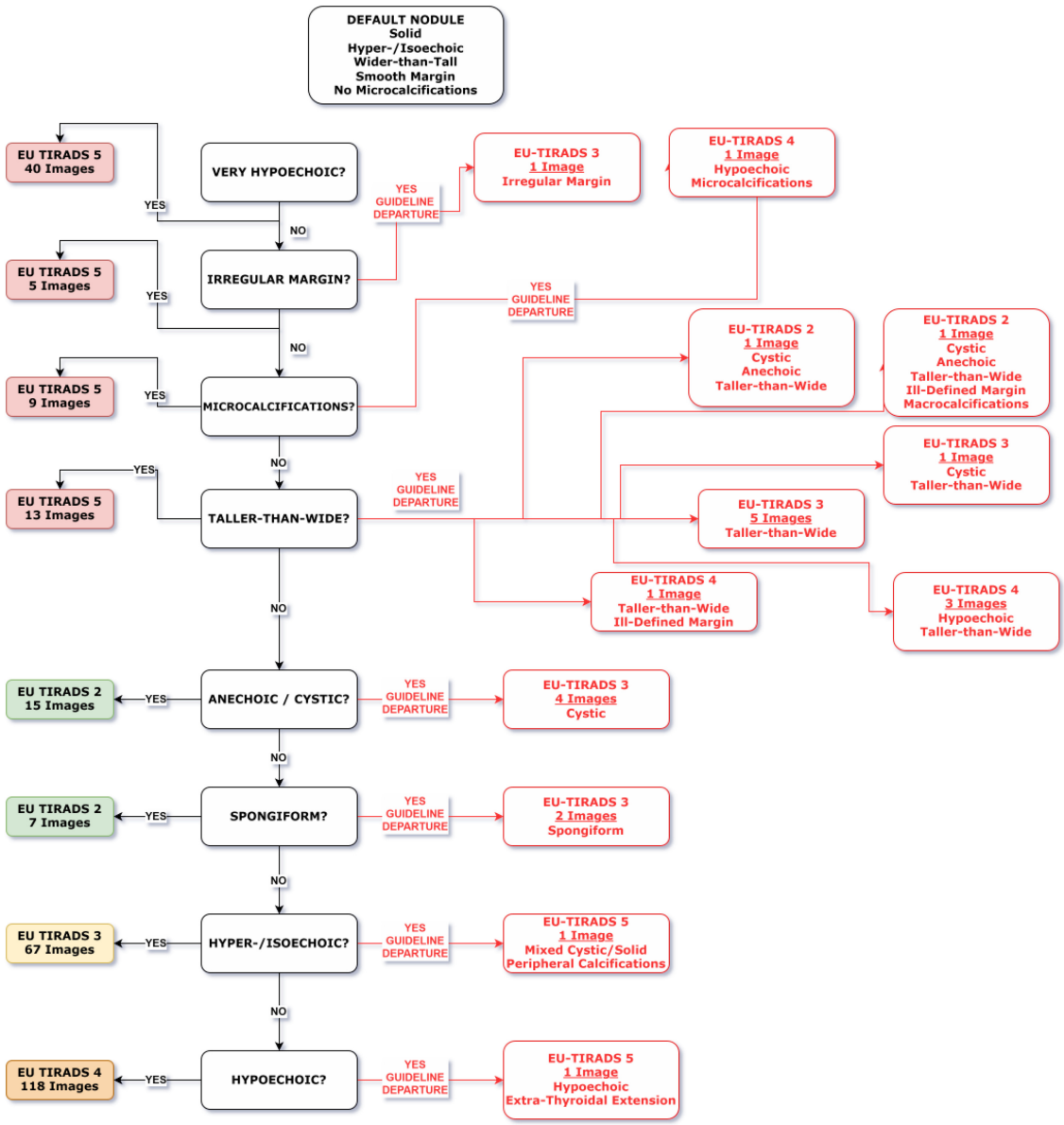

Figure 2.50 – Expert 1's departures from a guideline EU-TIRADS algorithm. The default nodule labels are given at the top, with decisions about labels proceeding downwards. On the left are listed the scores that were assigned in accordance with the guideline, while departures from that guideline are in red on the right.

of spongiform or cystic/anechoic features. These departures were less frequent than those from the taller-than-wide decision.

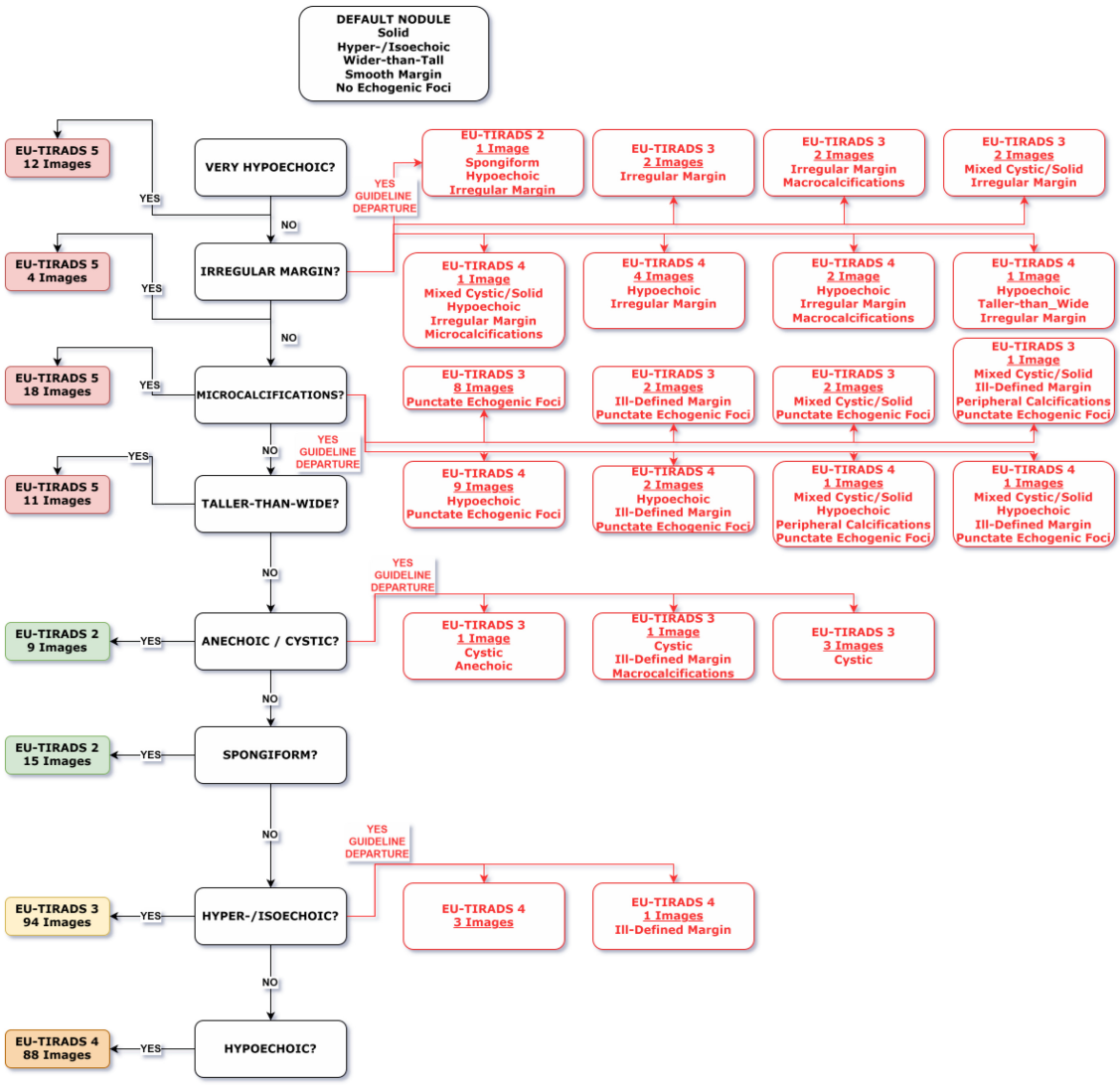

Figure 2.51 – Expert 2's departures from a guideline EU-TIRADS algorithm. The default nodule labels are given at the top, with decisions about labels proceeding downwards. On the left are listed the scores that were assigned in accordance with the guideline, while departures from that guideline are in red on the right.

For Expert 2, the departures from the guideline were different, as shown in Figure 2.51. The most common departure was from assigning a score of EU-TIRADS 5 for the presence of punctate echogenic foci, corresponding to true microcalcifications (see Figure 2.22). In general, the deviations from guideline at this point of the decision tree otherwise followed the EU-TIRADS algorithm when ignoring the impact of punctate echogenic foci, with all nodules being non-spongiform and non-cystic and correspondingly assigned to EU-TIRADS 3 if hyper-/isoechoic or EU-TIRADS 4 if hypoechoic.

The next largest category of departures was from assigning an EU-TIRADS 5 category for an irregular margin. Once again, these nodules were otherwise categorized according to the guideline after ignoring an irregular margin, with the lone spongiform nodule being scored as EU-TIRADS 2, the hyper-/isoechoic nodules scored as EU-TIRADS 3, and the hypoechoic nodules scored as EU-TIRADS 4.

In terms of departures from the EU-TIRADS 2 guideline, Expert 2 had a few images that received a EU-TIRADS 3 label in accordance with hyper-/isoechogenicity after ignoring cystic/anechoic features. There were also a few hyper-/isoechoic solid nodules which were scored as EU-TIRADS 4, though these were very rare compared to the images that were scored according to the guideline.

Moving on to Expert 3, we see a different pattern of departure from the EU-TIRADS guideline in Figure 2.52. By far the most frequent departure for Expert 3 was from an EU-TIRADS 5 score for the presence of an irregular margin. Apart from this omission, the corresponding images were scored as EU-TIRADS 3 and EU-TIRADS 4 according to their echogenicity, as would be expected. The other guideline departures for Expert 3 were scattered, with the next most frequent being from an EU-TIRADS 4 label for hypoechogenicity. A few nodules in this category were scored as EU-TIRADS 5, despite not having guidelines-based high risk factors.

Finally, the most departures from guideline were seen from Expert 4, and are split between Figures 2.53 and 2.54. From the former, we see that many guideline departures arose from departures from EU-TIRADS 5 guideline scores on the basis of high-risk features. The high-risk feature that most often did not lead to an expected EU-TIRADS 5 score was the presence of punctate echogenic foci, corresponding to microcalcifications. Many different combinations of sonographic features including punctate echogenic foci were involved, with their scoring not necessarily following the guidelines even apart from the microcalcification decision.

In addition, we see from Figure 2.53 that many nodules did not receive an EU-TIRADS 5 score despite the presence of an irregular margin or a very hypoechoic label. A few images also differed from the guideline on the basis of a taller-than wide label.

Moving down to the part of the guideline decision tree seen in Figure 2.54, we see fewer differences from the guideline for samples not associated with high-risk features. The most frequent were EU-TIRADS 3 and 4 scores assigned despite the presence of cystic or spongiform labels. In these cases, the hyper-/isoechoic vs. hypoechoic distinction was not always followed after omitting the benign feature. For a few hypoechoic images without any high-risk or low-risk features, a score of EU-TIRADS 3 was also assigned.

Overall, the four experts differed from the guideline EU-TIRADS algorithm in different ways. Expert 1 rarely departed from the guideline, with the most common exception being not assigned EU-TIRADS 5 scores for taller-than-wide nodules. For Expert 2, not assigning an EU-TIRADS 5 score despite the presence of punctate echogenic foci and irregular margins was the most frequent departure. Expert 3 also most often had differences from EU-TIRADS 5 scores on the basis of an irregular margin, but otherwise did not have consistent patterns.

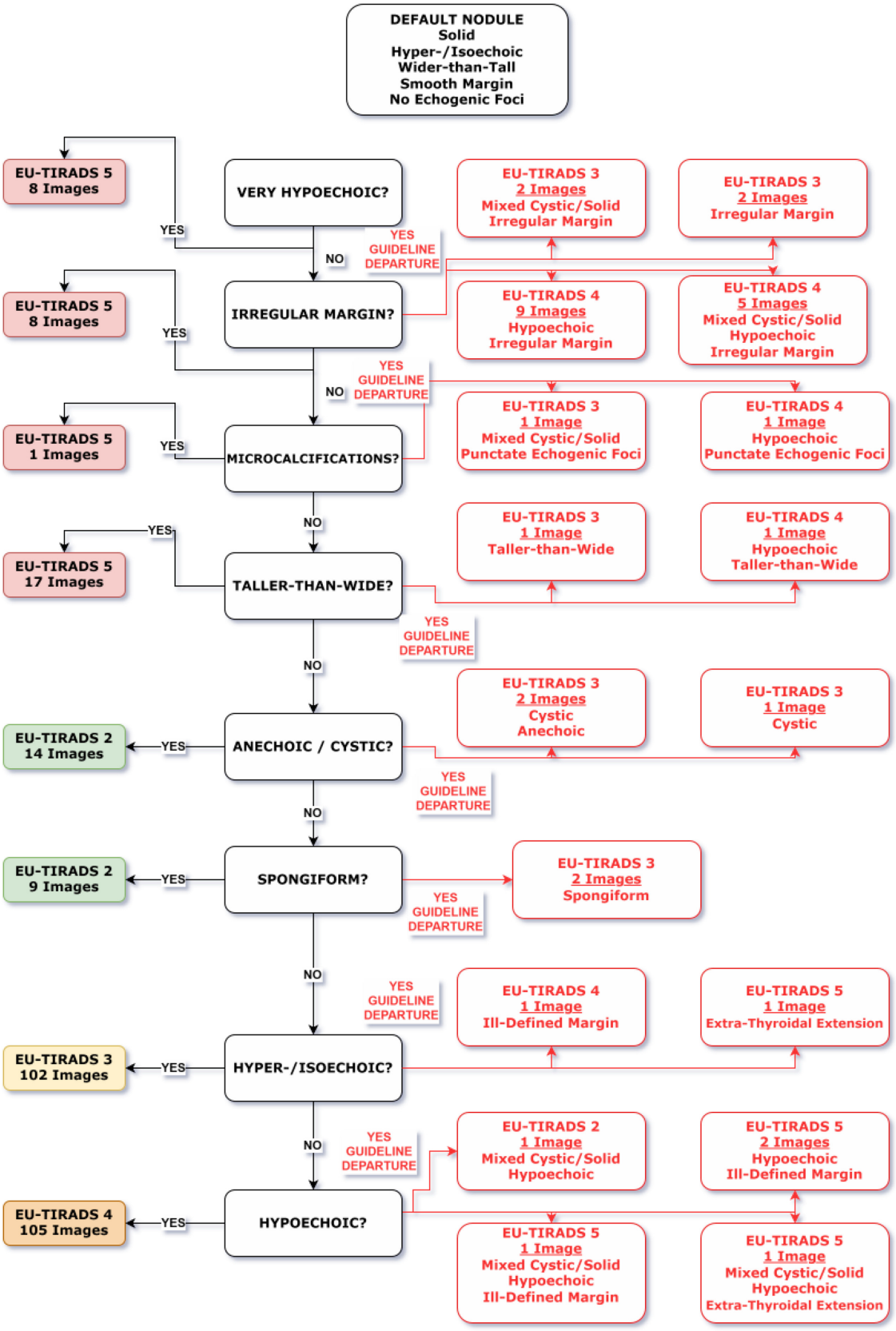

Figure 2.52 – Expert 3's departures from a guideline EU-TIRADS algorithm. The default nodule labels are given at the top, with decisions about labels proceeding downwards. On the left are listed the scores that were assigned in accordance with the guideline, while departures from that guideline are in red on the right.

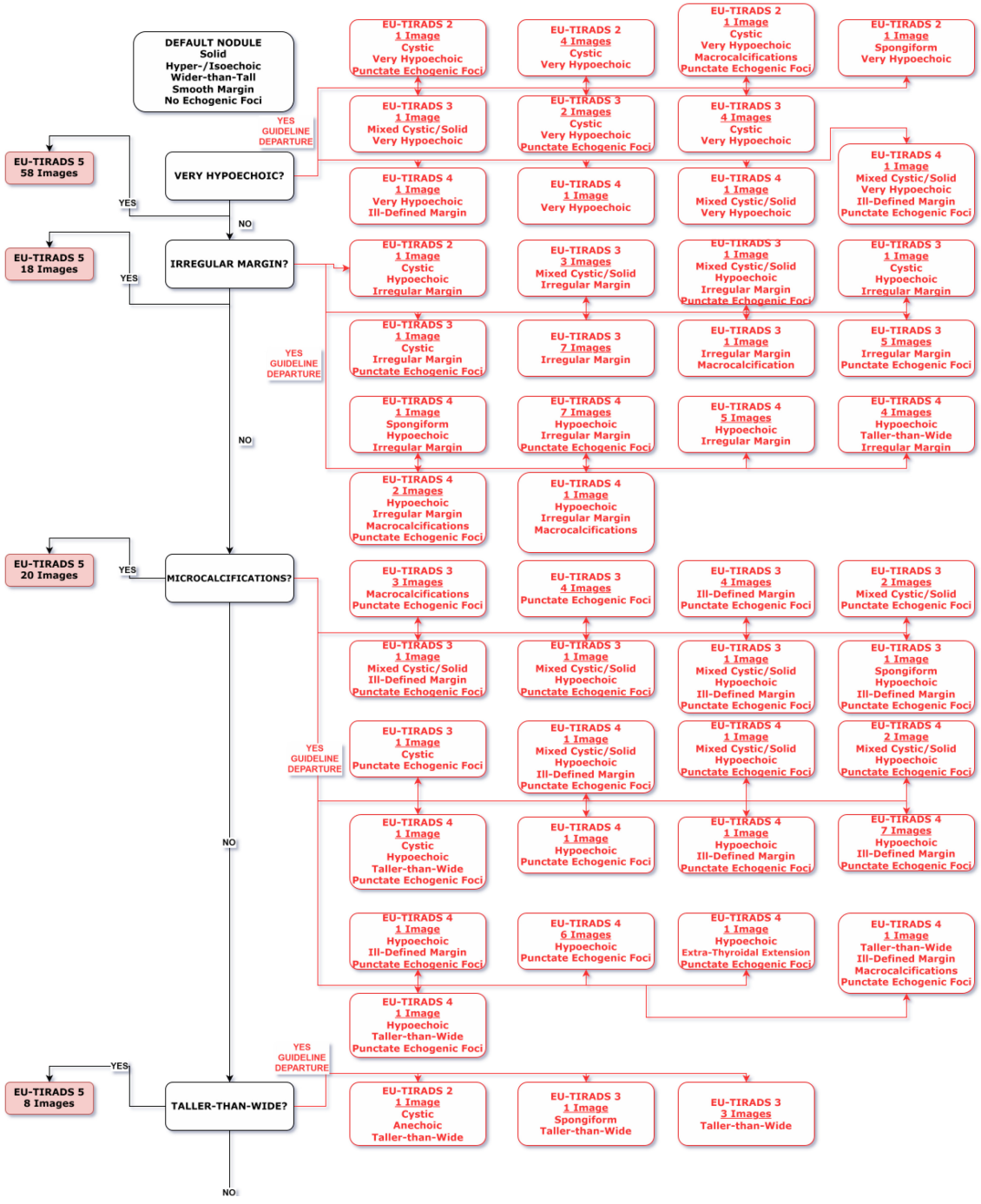

Figure 2.53 – The top half of Expert 4's departures from a guideline EU-TIRADS algorithm. The default nodule labels are given at the top, with decisions about labels proceeding downwards. On the left are listed the scores that were assigned in accordance with the guideline, while departures from that guideline are in red on the right. Continues in Figure 2.54.

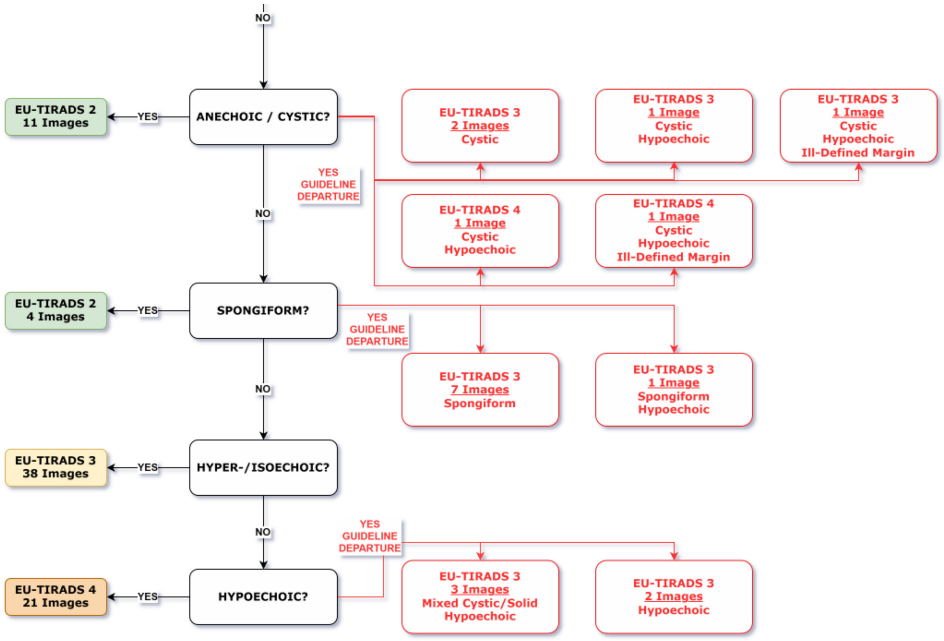

Figure 2.54 – The bottom half of Expert 4's departures from a guideline EU-TIRADS algorithm. The default nodule labels are given at the top, with decisions about labels proceeding downwards. On the left are listed the scores that were assigned in accordance with the guideline, while departures from that guideline are in red on the right. Continues from a negative Taller-than-Wide determination in Figure 2.53.

Expert 4 showed the greatest differences, with the vast majority having to do with not assigning EU-TIRADS 5 labels for multiple high-risk features, particularly for punctate echogenic foci, which this expert frequently assigned. This expert also differed from the guideline on the identification of EU-TIRADS 2 scores.

2.5 Discussion

The results of the inter-reader study uncovered multiple sources of variability. In order to understand them, we must first address the aspects of the evaluation which differed from clinical practice, which limit how representative these results are of real ultrasound. We may then explore in detail the differences between experts both in terms of feature identification as well as in how they assign TIRADS scores. This is particularly significant given that our four experts each have at least 15 years of experience in thyroid ultrasound; average French practitioners are likely to have an even lower inter-reader reliability.

These findings may have implications for how thyroid nodule ultrasound practice can be improved, so as to avoid unnecessary or missed interventions.

2.5.1 Limitations

To begin with, evaluation on a fixed, axial-view image of a nodule in isolation did not resemble the clinical reality of thyroid nodule evaluation in France. French practitioners acquire their own thyroid ultrasound images and evaluate them during the examination, rather than on captured still images. Being unable to dynamically adjust settings such as frequency and transducer focus could understandably introduce ambiguities into image interpretation, especially if the image had been optimized for the preferences of another operator.

The nature of French practice also means that the four experts may be used to categorizing nodules with a gestalt approach. The process of systematically assigning labels to different sonographic features in isolation is different from providing a typical thyroid ultrasound report. The features described during the evaluation process are also unlikely to all be independent of one another; certain combinations are far more likely than others, which may have affected the experts' identifications and evaluations. This represents an important limitation of this analysis.

In addition, clinicians would typically also examine images from other imaging modes to glean further information. Doppler imaging in particular, by allowing for an evaluation of the vascularization of a nodule, would be regularly used to interrogate most nodules. Some experts might also rely on elastography to further evaluate a lesion. Furthermore, a full clinical report consists of more than merely an EU-TIRADS score; practicitioners would be expected to include additional signs and findings that contribute to an overall evaluation.

Finally, the proportions of nodules receiving each EU-TIRADS classification must also be put into context. The experts each contributed images acquired during their routine practice, though one site predominated with more than half of all images (see Table 2.4.1). This imbalance may introduce a geographical bias to the sampled population. Furthermore, since experts contributed images that they acquired during routine practice (without sending images from the same patient twice), a proportion of the patients may have been those who were scheduled for follow-up exams because of the higher-risk features of their nodules, creating a bias toward higher TIRADS scores.

In order to understand, therefore, where machine learning tools and automation can improve clinical thyroid ultrasound, we must take these limitations into account when examining the inter-expert variability in sonographic feature identification and EU-TIRADS scoring.

2.5.2 Feature Identification

Identification of sonographic features forms the basis of the EU-TIRADS system. The differences in the use of sonographic feature labels between experts could be due to a failure to standardize definitions for all readers. Though the features of the sonographic inventory were explained and discussed prior to beginning the evaluation process, an expert's years of experience might lead to in-built variations from the guideline. This might also vary over the course of long evaluation sessions with cumulative fatigue.

With this in mind, we consider the findings for each sonographic feature category in light of the clinical context of thyroid nodule ultrasound.

2.5.2.1 Composition

Understanding inter-reader variabiltiy in terms of composition label is difficult due to the overwhelming predominance of solid nodules, for which there was also a high rate of consensus. Identification of the minority categories posed the most problems (see Figure 2.31).

Beginning with the difficulty of identifying spongiform labels, we find a disagreement that was often associated with a change in score from EU-TIRADS 2 to EU-TIRADS 4 (see Figure 2.32). Even if this was a relatively rare disagreement, its consequences are important, triggering a difference in decision to proceed to FNA or not. When asked about their experience applying composition labels, experts agreed that the definition of a spongiform nodule was one of the most difficult to apply precisely. As seen in Figure 2.4, a spongiform nodule is composed of tiny cystic spaces; judging the size of these spaces and the proportion of the nodule that they occupy leads to borderline cases. The experts cited distinguishing spongiform from mixed cystic and solid nodules as a difficult judgement; this latter category also does not exist in EU-TIRADS.

The other disagreement often associated with a difference in EU-TIRADS scoring was between cystic and solid nodules, though this was once again a rare score disagreement. From the description in Figure 2.4, it would seem that these two categories are quite distinct. In fact, in true clinical practice, these two types of composition are readily distinguishable. While a single, static B-mode image may create confusion between a dark, anechoic fluid-filled cavity and a dark, hypoechoic but solid nodule, the same ambiguity does not exist in real practice. With the ability to adjust gain and examine a cyst from multiple views, the distinction would be simplified. With the addition of Doppler imaging, which would show vascularization within the weakly echoic regions of a solid nodule but not from a fluid-filled cyst, the distinction becomes trivial. Therefore, disagreement about this distinction is likely an artifact of the examination of static images.

In light of the low rate of consensus and the disagreements about the EU-TIRADS 2 score arising from the spongiform label, it would seem that this aspect of composition deserves particular attention. The definition is difficult for experts to agree upon, and leads to a significant divergence in potential intervention. Identification or ruling out of this definition is an important target for machine learning algorithms seeking to improve clinical practice.

2.5.2.2 Echogenicity

Moving onto echogenicity, we again faced a category with a paucity of images receiving consensus for minority categories. Most nodules were identified by consensus as being either hyper-/isoechoic or hypoechoic (see Figure 2.34), with very little agreement on assigning the very hypoechoic label. In particular, the frequency of use of the very hypoechoic label varied quite by almost a factor of ten between Experts 3 and Expert 4 (see Table 2.5).

Such a high degree of variation suggests different thresholds for judging a nodule to be hypoechoic or very hypoechoic. The definitions of these terms, as seen in Figure 2.9, depend on comparison to the echogenicity of adjacent healthy thyroid parenchyma (for hypoechoic nodules) or to adjacent muscles (for very hypoechoic nodules). The inclusion criteria specified that appropriate reference tissue should be visible within the image, but even with a reference, comparison was not always straightforward.

A difficulty commonly cited by experts in conjunction with this feature was when the echogenicity of a nodule or of the surrounding reference tissue was not homogeneous. If it was difficult to decide which echogenicity label predominately defined a nodule relative to the nearby parenchyma or muscle, then the overall label was ambiguous. In addition, experts noted that it was not always easy to find adquate reference zones of normal tissue to serve as a basis for the hyper-/isoechoic vs. hypoechoic comparison.

An additional disagreement that was associated with EU-TIRADS 2 to EU-TIRADS 5 score differences was between anechoic and very hypoechoic nodules. As discussed previously, this corresponds to the artificial difficulty of identifying a fluid-filled cyst on a static B-mode image, and not to clinical reality.

Overall, identification of nodule echogenicity varied significantly between experts. This variability is of great clinical importance given the central role of echogenicity in stratification between multiple EU-TIRADS scores (see Figure 2.3). Distinguishing between hyper-/isoechoic, hypoechoic, and very hypoechoic nodules with an automated algorithm could be useful to increase reproducibility between the four experts. When considering non-expert practitioners with less experience adjusting acquisition parameters to be able to distinguish between the greyscale intensities of different tissues, it is clear that there is an opportunity for machine learning to standardize EU-TIRADS evaluation.

2.5.2.3 Shape

Shape differed from the previous features in being a binary label, with a straightforward definition. The lack of consensus on the minority category of taller-than-wide (see Figure 2.37) was therefore surprising. Given the association of this disagreement with a disagreement about the score EU-TIRADS 5 (see Figure 2.38), its impact is also important.

The question as to whether or not this is a clinically realistic ambiguity depends on whether readers are better able to assess the proportions of a nodule while operating an ultrasound system, perhaps using calipers on the image. However, experts agreed that they had an inbuilt idea of proportions based on the familiar anatomy of structures such as the trachea, and ought to be able to make the judgment visually.

Discussion with the experts after the evaluations were completed did present a difference in practice: the four experts did not agree on how to orient the axes of measurement. Depending on the orientation of the nodule's long axis, experts who considered measurements relative to the plane of the image rather than relative to the nodule itself could generate different assessments.

This was confirmed by experts looking at the same image together and indicating their preferred axes for measurement.

Therefore, it is possible that clinical reproducibility could be improved by simply reinforcing a standardized definition for practitioners. The added value of machine learning solutions for this determination appears limited, though a standardized segmentation or bounding box could facilitate this aspect of nodule evaluation provided the overall orientation of the probe during acquisition was made clear.

2.5.2.4 Margin

The margin category was also affected by a class imbalance, as well as by stark differences between experts' utilization of the label lobulated or irregular margins (see Table 2.7). In terms of consensus, Expert 4 disagreed often with the smooth label, and very little agreement was present about the minority labels.

Notably, some labels in this category do not have an impact on the EU-TIRADS score, as only the presence of an irregular margin would qualify as a high-risk sign indicating EU-TIRADS 5 (see Figure 2.3). A finding of extra-thyroidal extension is extremely important and would always be highlighted in the report, but not directly impact the score (Russ et al., 2017).

Perhaps given the scarcity of lobulated or irregular labels, no strong associations were found between margin label disagreements and EU-TIRADS score disagreements. During consensus review meetings, it was discussed that one expert inspected the margin more closely than the others. The definitions of lobulations or irregularities such as spiculations (see Figure 2.17) depend on a reader's sensitivity to small perturbations in the margin.

It is unclear whether an automated algorithm for detection of irregular margins would improve EU-TIRADS reproducibility in the clinic. At any rate, the sheer difference in the frequency of this label's use between experts suggests that a standardized definition would be useful.

2.5.2.5 Echogenic Foci

The final sonographic feature category, echogenic foci, was dominated by an absence of positive labels (see Table 2.8) for macrocalcifications, peripheral calcifications, and punctate echogenic foci. The latter category did have over a hundred positive labels from Expert 4, albeit with only 4 from Expert 3, leading to virtually no consensus.

A lack of labels renders difficult the analysis of this category. The identification of punctate echogenic foci was associated with disagreements about the EU-TIRADS 5 category (see Figure 2.45). Because this label was intended to distinguish true microcalcifications (see Figure 2.22), it contributes to decisions about EU-TIRADS 5 scores.

However, it appears that the label was interpreted differently by Expert 3 and Expert 4, who agreed only once on its presence. The echogenic foci portion of the inventory began, akin to ACR-TIRADS, with a question asking whether there were no echogenic foci, or only echogenic foci with large comet-tail artifacts, which would be more likely to correspond to colloid crystals (see Figure 2.22). If not, then the second stage of the assessment was to indicate whether macrocalcifications, peripheral calcifications, or punctate echogenic foci were present.

Therefore, the distinction between large comet tail artifacts and small ones could create an ambiguity in the judgment about punctate echogenic foci. The definition of this category would need to be much more explicit to be useful for clinical practice. The detection of microcalcifications

would make an attractive target for an automated method, though obtaining sufficient reference samples to create and validate an algorithm for this purpose might be difficult.

2.5.3 Decision Structures

With these differences in the experts' identification of sonographic features considered, we turn to an examination of differences in their assigning of EU-TIRADS scores on the basis of the identified features. To begin, this investigation is limited by the limited representation of different combinations of feature labels among experts.

From Figure 2.46, it is clear that certain experts were more limited than others in their use of combinations of feature labels; Experts 3 and 4 represent two extremes. As seen previously, Expert 3 assigned very few labels for certain features, such as the presence of punctate echogenic foci or very hypoechoic nodules. As a consequence, this expert had very few unique combinations of nodule features, with the most common feature combination being applied to over a quarter of all nodules. By contrast, Expert 4 had a great diversity of combinations of nodule features, with no single combination being used more than twenty times.

This suggests a difference in the evaluation styles among the experts. Expert 3, and to a lesser extent Experts 1 and 2, described nodules as often falling into the same few combinations of features, while Expert 4 described them more unique combinations. This difference might be interpreted as assigning feature labels more or less independently of a gestalt perception of a nodule, though this is merely a speculative observation.

Looking to the combinations that were repeatedly used by experts, we can form an idea of how consistently the combination of features assigned to a nodule was associated with a single EU-TIRADS score. From Figure 2.47, we see that for Experts 1 through 3, very few feature combinations had inconsistent EU-TIRADS scores, and less than 6% of all images had an EU-TIRADS score that differed from the most commonly used score for their combination of feature labels. These numbers were slightly higher for Expert 4, with around 14% of images differing from the most common EU-TIRADS score for their label, but overall it seems that for each expert, a particular combination of features led often to a consistent EU-TIRADS score. Comparing the inter-expert consistency of EU-TIRADS scoring on the basis of specific feature combinations was not feasible due to the limited number of shared feature combinations (see Figure 2.48). The slight variability that was noted within experts could be due to the fact that the sonographic inventory that was collected may not have sufficiently described all nodule features consciously used by the experts for their analysis. In this case, these differences could be related to differences in nodule characteristics that might be cited in radiology reports, such as a halo sign, or the position of a nodule such that its form is distorted by being pressed next to the trachea.

Of course, these differences in terms of undescribed features may also be intertwined with subjective differences in the mental EU-TIRADS algorithms of each expert. On the basis of their own experience with benign and malignant nodules, experts may learn to assign different weights to nodule characteristics when assigning en EU-TIRADS score. Understanding these inter-expert differences which are not captured by the sonographic feature inventory is equally important to understanding thyroid ultrasound evaluation in France.

The comparison with a guideline-based EU-TIRADS score based on sonographic inventory features provided a means of investigating these differences. From Figure 2.49, it is evident that the most common deviation from the guideline for all experts was to assign a score of EU-TIRADS 3 or 4 to a combination of features that would receive a score of EU-TIRADS 5 based on the

guideline. The next most common deviation, albeit far less frequent, was to assign a score of EU-TIRADS 3 to a combination of features that would receive a score of EU-TIRADS 2 according to the guideline.

Looking to deviations from the EU-TIRADS 5 guideline score, we must first observe that the nature of the guideline algorithm means that this score is assigned if any high-risk features are present (see Figure 2.3). Therefore, the expert departures from guideline-based EU-TIRADS 5 scores corresponded to cases for which an expert assigned a high-risk feature label, but did not assign the EU-TIRADS 5 score. For Expert 1, this was most often due to taller-than-wide labels. For Experts 2, 3, and 4, irregular margins were a frequent high-risk feature that did not lead to an EU-TIRADS 5 score as expected. Experts 2 and 4 shared frequent deviations from the guideline on the basis of punctate echogenic foci, which correspond to true microcalcifications. Expert 4 also had deviations from EU-TIRADS 5 on the basis of a very hypoechoic label.

These differences from the guideline are important, given that an EU-TIRADS 5 score has the lowest threshold for FNA, and demands monitoring for smaller nodules (Russ et al., 2017). Some of these differences may be due to an imperfect equivalence between the sonographic inventory categories and the constructed guideline. For example, Expert 4 stated the description of an irregular or lobulated margin was considered during the annotation task to be positive even for a single spiculation or lobulation (see Figure 2.17). However, in clinical practice the same expert said he would more carefully examine the entire nodule, as compared to a single image, before coming to this determination.

For the undervaluation of punctate echogenic foci, it may not have been clear to experts that the definition of this feature corresponded to true microcalcifications, as discussed previously. As for taller-than-wide labels, Expert 1 stated that in some positions, a nodule might appear taller-than-wide due to its position next to structures such as the trachea, though the dimensions did not reflect this. In addition, given that experts had slightly different ways of orienting the axes for a taller-than-wide determination, a strict taller-than-wide definition might lead to cases in which a slightly taller-than-wide nodule earned that feature label, but was not extreme enough to surpass the expert's threshold for suspicion during EU-TIRADS scoring.

The other most common departure from the guideline was with feature combinations that would received EU-TIRADS 2 scores that were instead scored as EU-TIRADS 3, though these were far rarer than the previously discussed departures. Within the framework of EU-TIRADS (see Figure 2.3), this score is assigned for nodules without high-risk features that are either spongiform or anechoic. The anechoic label properly corresponds to a cystic composition, though this was not always the case in expert evaluations. All four experts had evaluations including the cystic label that did not receive the score EU-TIRADS 2. Experts 1, 3, and 4 also had spongiform nodules that did not receive the score EU-TIRADS 2.

One explanation for this may be an imprecise correlation between the sonographic feature descriptions and the EU-TIRADS clinical guidelines. In EU-TIRADS, the definitions used for EU-TIRADS 2 depend on a nodule being "purely" cystic or "entirely" spongiform (Russ et al., 2017). The four experts may have assigned labels for an predominately spongiform or cystic composition, but may not have considered these equivalent to the EU-TIRADS 2 criteria.

Overall, given how rare the EU-TIRADS 2 guideline disagreements were, it is uncertain whether they have clinical impact. For all of the experts, however, it seems that the high-risk features were not always taken into consideration for EU-TIRADS 5 scores. Particularly for taking into consideration irregular margins and punctate echogenic foci, a standardization could improve inter-expert reproducibility.

2.6 Conclusions

The purpose of this study was both to understand French thyroid ultrasound practice and where automated nodule characterization algorithms could meaningfully contribute to reproducibility. Many of the differences in feature identification and application of the EU-TIRADS guidelines are likely related to differences between the evaluation of fixed images and actual clinical practice; however, others are important targets for machine learning tools.

Distinguishing hyper-/isoechoic, hypoechoic, and very hypoechoic nodules could be a useful function of an automated tool to standardize EU-TIRADS evaluation, with an impact on standardizing scores. Being able to automatically classify spongiform nodules might also be helpful due to its impact on EU-TIRADS 2 scores that do not require FNA.

Automating the detection of true microcalcifications is also important. However, training a robust algorithm for this purpose would be difficult in terms of collecting sufficient samples of these rare findings in order to distinguish them from colloid crystals with large comet-tail artifacts or the acoustic enhancement of back walls of cysts.

Otherwise, the other aspects of variability in feature identification and application of EU-TIRADS guidelines could be addressed with reinforcement of standards among French practitioners. In addition, overcoming the limitations of evaluation on static images would give a clearer image of the difficulties in real practice. This could involve using additional view of each nodule, acquiring video clips of ultrasound sweeps, or even having patients directly examined by multiple practitioners.

A final conclusion is that inter-expert variability in nodule characterization has implications for the implementation of machine learning algorithms. Obtaining expert annotations as ground-truth references is time consuming, especially when considering the need for multiple readers to obtain consensus. When training machine learning models based on these labels, the ambiguities that exist on the interpretation of static B-mode images must also be taken into account.

CHAPTER 3

Expert Variability in Thyroid Nodule Echogenicity Evaluation

Nodule echogenicity evaluation, particularly distinguishing between hypoechoic and hyperechoic or isoechoic nodules, is an important part of expert interpretation of thyroid ultrasound. This determination has an impact on the EU-TIRADS score used in France to guide follow-up and biopsy decisions, but is difficult because of its subjective nature. In this chapter, we explore recent applications of machine learning methods to automate thyroid nodule analysis, and apply one to attempt to reproduce expert echogenicity labels. We also examine inter-expert and intra-expert variability in distinguishing between hypoechoic and hyperechoic or isoechoic nodules, with an analysis of the factors that the experts identify as contributing to uncertainty in their labels. The results of these analyses suggest that expert echogenicity labels, at least when applied on static images, can vary significantly between readers; in some cases, they may even vary upon repeat examination by the same reader. Heterogeneity of echogenicity levels within the nodule appears to be associated with expert disagreement. The results of this analysis highlight the limitations of expert-label based evaluation systems, suggesting that quantitative or automated ultrasound analysis approaches could help better standardize thyroid nodule evaluation.

3.1 Introduction 67
3.2 Background 67
3.2.1 Analysis Tasks of Machine Learning Methods on Thyroid Ultrasound 67
3.2.1.1 Detection and Segmentation 68
3.2.1.2 Characterization 69
3.2.2 Thyroid Ultrasound Datasets for Machine Learning 70
3.3 Automated Nodule Echogenicity Characterization in the French Context 71
3.3.1 Echogenicity Classification Strategy 72
3.3.2 Echogenicity Classification Results 75
3.4 Reproducibility of Expert Echogenicity Classification 77
3.4.1 Intra-Expert Reproducibility Results 78
3.4.2 Implications of Expert Variability in Echogenicity Assessment 79
3.5 Quantitative Echogenicity Analysis 80
3.5.1 Expert Label Agreement with Quantitative Echogenicity Measures 81
3.5.2 Quantitative Analysis Results 83
3.6 Discussion 88
3.6.1 Limitations 89
3.7 Conclusions 90

3.1 Introduction

The results of the previous chapter have demonstrated that even among French experts, there exists a great deal of inter-reader variability in the evaluation of thyroid ultrasound images, at least when those images are static axial views taken in isolation. It is not surprising therefore that in response to these limitations, many groups have proposed machine learning algorithms to automate the task of thyroid nodule ultrasound evaluation; for example, experts from the ACR-TIRADS committee participated in the development of a deep learning algorithm for nodule evaluation shortly after the publication of that system (Buda et al., 2019). Since then, there has been a public challenge for nodule segmentation and benign-malignant classification (Grand Challenge, 2020) with hundreds of participants, as well as multiple algorithms marketed to clinicians by private companies (Szczepanek-Parulska et al., 2020 ; Reverter, Vázquez, & Puig-Domingo, 2019).

In this chapter, we begin with an overview of these existing machine learning tools for thyroid nodule ultrasound assessment. We then examine whether a specific aspect of the expert evaluation, that of echogenicity, can be adequately reproduced using a machine learning algorithm on our French dataset. This could prove useful for helping non-expert practitioners in France, as the results of the previous chapter showed that disagreements on echogenicity labels also led to disagreements on EU-TIRADS scores.

In addition, we further examine the variability among experts in terms of echogenicity labels. The intra-reader reproducibility of expert echogenicity labels has important implications for both current clinical practice and the value of these labels as targets for machine learning strategies. This is examined through another labeling study with the four experts, keeping in mind their observations about which features of axial-view thyroid images make nodule echogenicity more or less clear. Finally, quantitative measures related to nodule echogenicity are examined for associations with image features in order to identify potential sources of expert variability.

This investigation into echogenicity labels specifically should allow for insights into the challenges and limitations of working with thyroid ultrasound data in both a clinical and machine learning context.

3.2 Background

To begin with, we must understand the work that has been conducted in machine learning applications to thyroid nodule ultrasound characterization. To sort through the abundance of proposed algorithms and the studies seeking to validate them, we can segregate them on the basis of the analysis tasks that they seek to perform, as well as the input data they use for this purpose.

3.2.1 Analysis Tasks of Machine Learning Methods on Thyroid Ultrasound

The first aspect to assess is the analysis task that an algorithm seeks to perform. The function of an algorithm is naturally relevant to its clinical applicability, in terms of whether or not it adequately addresses a real problem faced by practitioners. In addition, the specific task also influences the nature of the ground truth labels necessary to train and validate the algorithm. In particular, uncertainties or ambiguities within these reference labels will affect the reliability and generalizability of these algorithms to other users, ultrasound machines, and healthcare systems.

With this in mind, analysis tasks on thyroid nodule ultrasound can generally be divided into three categories: detection, segmentation, and characterization.

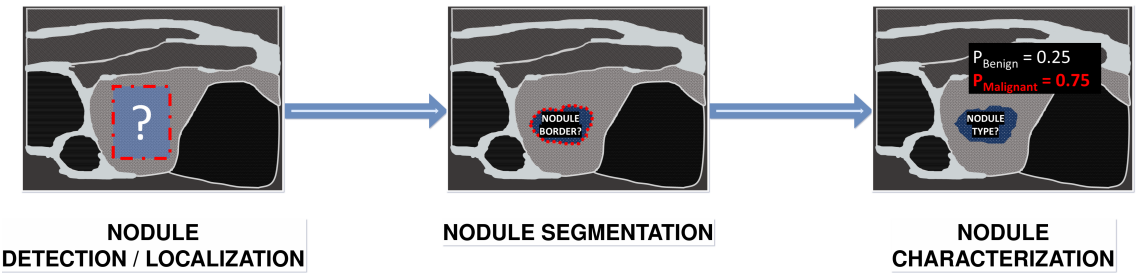

Figure 3.1 – Summary of different automated evaluation tasks for thyroid nodule ultrasound. Detection and localization of nodules within the thyroid gland go together. This is often a prerequisite for the task of nodule segmentation, identifying the border of the nodule within the image; this can be useful for characterizing nodule size and margin properties. Finally, nodule characterization can be a prediction of the risk of malignancy, or a classification according to various sonographic features.

3.2.1.1 Detection and Segmentation

The first of these tasks is the detection of a nodule or nodules within an ultrasound image (see Figure 3.1). The utility of this task depends greatly on the input data and the clinical context. Given that thyroid ultrasound is often used as a tool to risk-stratify incidentally-discovered thyroid lesions, mere detection of nodule presence within a saved static image is not clinically useful (Kant et al., 2020). For an inexperienced operator, however, automated detection during live ultrasound sweeps could signal the presence of a nodule meriting further examination. It would also be a useful component of other algorithms that search for lesions within pre-recorded video clips of ultrasound sweeps, in order to perform further analysis on any detected nodules.

The analysis of detected nodules often includes the task of segmentation (see Figure 3.1). Signaling a nodule's contour is inherently useful to assessing its margin, a feature employed in EU-TIRADS and ACR-TIRADS (Russ et al., 2017 ; Tessler et al., 2017). Automatic segmentation is also useful for the estimation of nodule diameter and volume, which are important to determining guideline-based FNA recommendations. These are also important to consider for surveillance, even though current guidelines do not find reliable evidence that nodule growth rate can predict cancer (Russ et al., 2017 ; Tessler et al., 2017). Therefore, segmentation has long been a target of machine learning algorithms for thyroid ultrasound.

The task of automated nodule detection and segmentation from static B-mode images has been addressed by too many publications to review here. For example, a public challenge for thyroid nodule ultrasound evaluation in 2020 included the target of nodule segmentation, and received hundreds of algorithm submissions (Grand Challenge, 2020). The top three submissions had similar performances using different architectures, with nearly identical IoU scores of 0.83, 0.82, and 0.82; in fact, the top 55 submissions all had IoU scores greater than 0.81 (Grand Challenge, 2020)(M. Wang et al., 2021 ; Chen et al., 2021 ; Tang & Ma, 2021). Clearly, many different neural network architectures can be applied to segment nodules with very similar results.

Among the abundance of proposed algorithms, there are a few publications that distinguish themselves by their use of training strategies specifically adapted to the clinical context of thyroid ultrasound. One such example is that of Gong et al., who considered the problem of thyroid nodules being falsely detected and segmented outside of the thyroid gland (Gong et al., 2021, 2023). With the exception of cases of extra-thyroidal extension, the anatomical considerations dictate that

nodules should only be found within the part of an image corresponding to the thyroid; therefore, this group proposed a two-stage strategy of segmenting the entire gland, and then segmenting lesions only within this area (Gong et al., 2021, 2023).

Another example, taking into account the variability between images acquired by different operators and ultrasound systems, is the publication by Xu et al. investigating the generalizability of deep learning models on tasks including nodule detection and segmentation. They applied a You-Only-Look-Once system for detection and a UNet for segmentation, using a dataset of thyroid ultrasound images from over ten thousand patients from 208 hospitals, with ultrasound equipment from twelve different vendors (Xu et al., 2023). This scale of evaluation is unique in its clinical relevance, as it genuinely reflects the variability between different practitioners, ultrasound machines, and healthcare institutions.

Another interesting consideration for clinical relevance is the fact that, as we have seen previously, different practitioners may not assess nodule features such as diameter consistently when relying on static B-mode images. For French practitioners, thyroid ultrasound consists of the evaluation of the region in real time, as the probe is being manipulated by the operator. 3D data acquired with specialized ultrasound systems, or even simple video clips acquired by an operator during a sweep of the probe, can allow for more spatially complete representation of a nodule. Taking advantage of this fact, a study of nodule volumetry showed reduced inter-reader variability using a commercial 3D ultrasound tool with automated segmentation when compared to standard clinical assessment based on 2D cross-sections (Krönke et al., 2022).

Another useful application of automated detection and segmentation on 3D data would be to standardize the evaluation of multinodular cases. When multiple nodules are located in close proximity, it can be difficult and time-consuming for a practitioner to identify all of them, leading the EU-TIRADS guidelines to suggest evaluating at least the three most significant nodules if there are numerous lesions (Russ et al., 2017). For their part, the ACR-TIRADS guidelines observe that a confluence of many similar nodules with few suspicious features may reasonably be surveilled either without FNA or with FNA of only the largest nodules (Tessler et al., 2017). In such cases, automated segmentation could potentially reduce ambiguity in nodule identification for monitoring.

A few groups have proposed algorithms for automated detection and segmentation on 3D data, including the commercial tool previously mentioned (Krönke et al., 2022). The most notable is the work by Liu et al. using over one thousand clips from thyroid ultrasound examinations (D. Liu, Yang, Zhang, Xiao, & Zhao, 2024). They applied an analysis strategy including automated detection with a ResNet34 to find nodules within video clips (D. Liu et al., 2024). This kind of application could be useful to allow even unskilled operators to detect nodules by simply performing adequate sweeps of thyroid lobes and the isthmus.

3.2.1.2 Characterization

Once a nodule is identified within an image, the final and most important aspect of the evaluation is the characterization of its risk of malignancy. Algorithms can be trained to make a simple benign-malignant prediction using training data with a ground-truth reference taken from either a cytologic or histopathologic diagnosis (see Figure 3.1). It is not possible to obtain histopathologic confirmation unless biopsy or thyroidectomy is performed to obtain a tissue sample; cytological confirmation is more likely to be available because it requires only FNA. It bears reflection, therefore, that ultrasound images of nodules that have cytologic or especially histopathologic con-

firmation likely represent higher-risk cases, which would be disproportionately subjected to these invasive procedures.

Many articles have been published on benign/malignant classification using static B-mode images. We can look to the same public challenge from 2020, which also included a binary benign-malignant classification component with ground-truth based on biopsy results (Grand Challenge, 2020). Once again, the top three results showed similar performances, with F1 scores of 0.86, 0.85, and 0.85 (Zhang, Lai, & Yang, 2021 ; J. Lu, Ouyang, Liu, & Shen, 2021 ; Shen, Ouyang, Liu, & Shen, 2021). Xu et al., using a large dataset across more than two hundred Chinese hospitals, applied a DenseNet classifier to their images, using a ground-truth determination from surgical pathology results (Xu et al., 2023). And finally, Liu et al. used a ResNet18 to classify nodules detected within video clips as being either benign or malignant, again based on surgical pathology results (D. Liu et al., 2024).

Of course, not all nodules undergo FNA or biopsy, particularly if they are assessed as being likely benign. This means that the patient population examined by the typical thyroid ultrasound practitioner may differ from the population examined by studies that require FNA or surgical biopsy for inclusion. Such considerations may limit the clinical applicability of many proposed algorithms.

In addition, the legal and ethical concerns intrinsic to medical practice make the use of black-box approaches to prediction difficult to justify; without human-interpretable methods, practitioners may feel hesitant to rely on a machine learning algorithm's prediction. By contrast, prediction of features relevant to TIRADS classification could standardize the identification of these features, provide greater transparency, and assist less experienced practitioners with thyroid ultrasound evaluation.

Therefore, some groups propose algorithms for nodule characterization using expert labels of TIRADS-relevant features as a ground-truth reference. The work of Buda et al., who used an ensemble of models to predict the features for the ACR-TIRADS score is a prime example, especially because it was developed with the participation of members of the committee that defined ACR-TIRADS. Some of these tools are also available commercially, sometimes integrated into ultrasound systems. Multiple private companies have developed such algorithms, and have published studies testing their applicability in a clinical context (Reverter et al., 2019 ; Chambara, Liu, Lo, & Ying, 2021 ; H. L. Kim, Ha, & Han, 2019 ; Wei et al., 2020 ; Szczepanek-Parulska et al., 2020 ; Barczyński, Stopa-Barczyńska, Wojtczak, Czarniecka, & Konturek, 2020 ; Ye et al., 2021 ; Y. Lu, Shi, Zhao, Song, & Li, 2019 ; Li et al., 2020).

These tools have a powerful advantage in the fact that they can be easily integrated into a practitioner's TIRADS-based clinical workflow. For example, Wei et al. showed that the use of one such commercial system to inform benign-malignant classification of nodules increased the accuracy of two relatively inexperienced radiologists (with 1 and 4 years of experience) to the level of that of an experienced practitioner (with 9 years of experience). A TIRADS-based prediction tool could therefore be of great utility to new practitioners, or those without much experience in thyroid nodule ultrasound.

3.2.2 Thyroid Ultrasound Datasets for Machine Learning

As discussed in the previous chapter, many different types of data are used by thyroid ultrasound practitioners, and may be used as input for machine learning tools. We have seen that most algorithms work with static B-mode images saved during ultrasound examination. In this case,

saved images typically center on the nodule cross-section with the greatest diameter, or the most relevant ultrasound features. There may or may not be indications of nodule position or measurements of nodule diameter superimposed upon this image. This form of static input is different from the input perceived by the practitioner, who can take into account all of what is visible with ultrasound over the course of the examination.

One means of increasing the input available to an algorithm is to use multiple complementary images of a nodule, such as with orthogonal axial and sagittal views. This could provide more information by showing additional regions of the nodule and of the thyroid parenchyma, as well as other adjacent anatomic structures. However, an even more complete set of input data could be acquired by recording a video clip during a sweep of the ultrasound probe through a region of the thyroid containing a nodule. This could also take the form of a 3D acquisition system that combines spatial information with the image data. Multiple publications have highlighted the advantages of using these forms of more thorough imaging data (Krönke et al., 2022 ; D. Liu et al., 2024).

In addition to these variants of B-mode images, other types of ultrasound data may also be used as input. This can include color Doppler images or spectral Doppler traces to assess blood flow to a nodule, or elastography data to analyze the stiffness of lesions. Supplemental information, such as the contents of the report prepared by a practitioner after interpreting an ultrasound examination, or the metadata in a DICOM file has been used in algorithm training strategies (Hu et al., 2020). Even non-ultrasound clinical data, such as the results of laboratory tests, could also be utilized as input data.

Yet another source of input data, rarely explored, is the raw signal, referred to as the radiofrequency (RF) signal received by the ultrasound probe. This signal is typically modified according to an assumed speed of sound and time-gain compensation to generate an ultrasound image. However, this signal may also be utilized as input data for a classification algorithm, as in the example of Liu et al. who used RF data in addition to B-mode images to train a network to perform benign-malignant classification (Z. Liu et al., 2021).

Despite these options, static B-mode images remain the most common form of input data. They are easily exported from ultrasound machines, and compatible with many existing image-processing algorithms. Therefore, their practical applicability as an input for algorithms that assist clinical practitioners is unrivaled.

3.3 Automated Nodule Echogenicity Characterization in the French Context

As we have seen, many groups have already created machine learning algorithms for nodule detection, segmentation, and characterization on the basis on static B-mode images and 3D data. Our objective here, with the limited French dataset that we have assembled, is not to attempt to match the performance of carefully-crafted networks trained and validated on far more extensive datasets, often with histological confirmation of benign or malignant status.

Rather, our aim is to investigate machine learning applications to relevant aspects of French thyroid nodule ultrasound practice. From the previous chapter, we have seen that even expert practitioners are not always in agreement on many ultrasound features of nodules when working on static B-mode images. Some of this disagreements were often associated with disagreements about EU-TIRADS scores. Of these disagreements, the only one for which our dataset provides

adequate samples with an expert consensus is for the distinction between hyper-/isoechoic nodules and hypoechoic nodules.

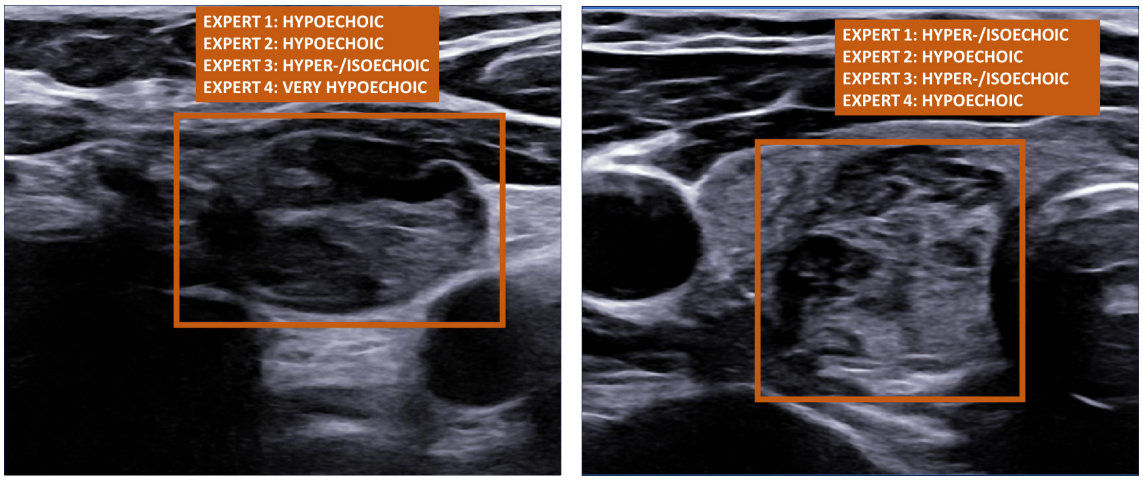

Figure 3.2 – Example images for which the experts were not in agreement on the echogenicity description.

By training a neural network to make this distinction, we can assess the feasibility of a pedagogical tool that would be instructive for less experienced French practitioners, who might have even greater difficulty in distinguishing between hyper-/isoechoic and hypoechoic nodules. This discrimination is useful for deciding between scores of EU-TIRADS 3 and EU-TIRADS 4, which have different thresholds for FNA (Russ et al., 2017). Such an algorithm could reproduce the consensus of French experts to serve as a guide for new practitioners.

3.3.1 Echogenicity Classification Strategy

Therefore, we were interested in training a neural network on the images from the previous chapter to distinguish between hyper-/isoechoic nodules and hypoechoic nodules. The images used were those for which the experts reached at least a weak consensus on either label for the nodule's echogenicity.

Of the 303 images labeled by the four experts, 105 images had at least a weak consensus for hyper-/isoechogenicity, and 122 had at least a weak consensus for hypoechogenicity. Consensus about the other echogenicity categories was too infrequent to be used.

Because the definition of this echogenicity distinction depends, as discussed in Chapter 2, upon a difference between the perceived echogenicity of the nodule and normal thyroid parenchyma, axial-view images were annotated by a non-expert to provide masks for these two regions. The area of the nodule was manually drawn onto the image, and the area of the ispilateral lobe, excluding the nodule and other lesions, was also drawn as a separate mask. The non-lesion tissue of the ipsilateral lobe was chosen because it avoided having the non-expert select a specific comparison region, and also because tissue in the isthmus or seen in the contralateral lobe would be at the edges of the ultrasound probe and therefore might not have comparable intensity to central regions of the image.

Threshold filters were then applied to these masks to exclude areas corresponding to essentially anechoic cystic fluid-filled spaces as well as hyperechoic foci such as calcifications. This was

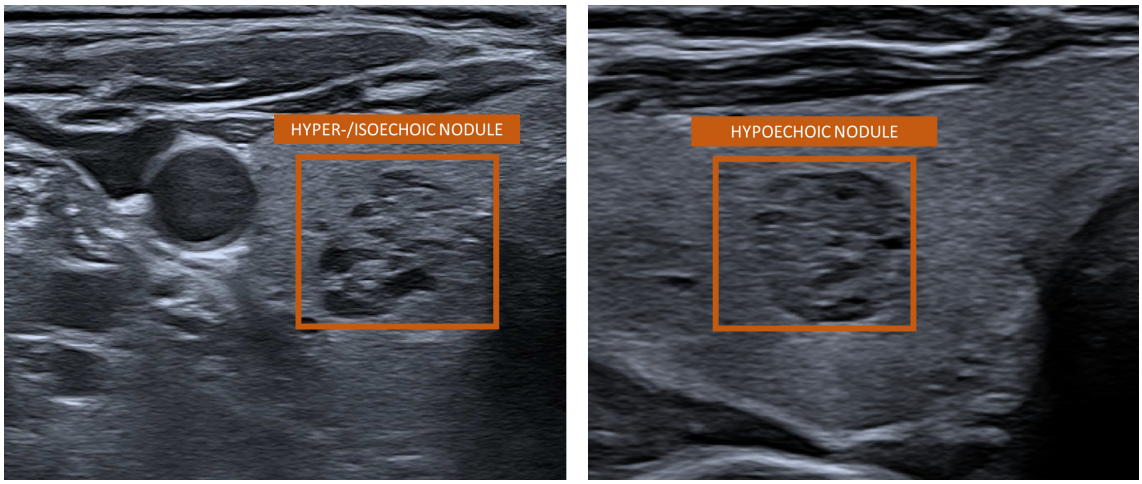

Figure 3.3 – (Left) Example of an axial-view image with a nodule labeled as hyper-/isoechoic by all four experts. (Right) Example of an axial-view image with a nodule labeled as hypoechoic by all four experts.

necessary so that only the echogenicity of soft tissue would be compared between the nodule and reference regions. A lower bound was established based on the intensity in the regions of the trachea and blood vessels in the images, and an upper bound was established based on the intensity of calcifications or the highly echogenic areas in front of the trachea or carotid. These were established via visual inspection of the masks created on the images.

The 227 images were split into a test set of approximately 20% and a set containing the rest of the images for cross-validation training, using stratified random sampling to maintain class balance. The cross-validation set was used for training three networks under 3-fold cross validation to be used for an ensemble prediction on the test set to reduce the effects of overfitting (Mohammed & Kora, 2023). Each sample consisted of a greyscale ultrasound image along with the two binary masks, and was subjected to data augmentation in the form of random cropping, rotation, rescaling, addition of Gaussian noise, and Gaussian blur.

The 3-fold cross-validation set was used to train a ResNet50 model pretrained on the ImageNet dataset (He, Zhang, Ren, & Sun, 2015). The choice of this model was guided by the fact that it had been recently used to perform benign/malignant classification of thyroid ultrasound images (Alghanimi, Aljobouri, & Al-shimmari, 2024). This result suggested that the architecture was sufficient to learn a significant binary difference from B-mode data. In addition, the fact that it was recent allowed for a more up-to-date comparison in terms of the resolution and contrast of modern thyroid ultrasound images.

The initial input to the network was a three channel image extracted from the DICOM, similar to the format used by the ResNet50 model (He et al., 2015). This was initially used with the pretrained ResNet with frozen weights on the initial layers, but the predicted area under the curve of the binary classification ROC for the test set was only 0.59. To provide more relevant information, the entire network's weights were unfrozen, and retrained on image data including the aforementioned masks for the nodule and reference regions. If the network were to learn to imitate the experts' method of evaluation, this information would be necessary.

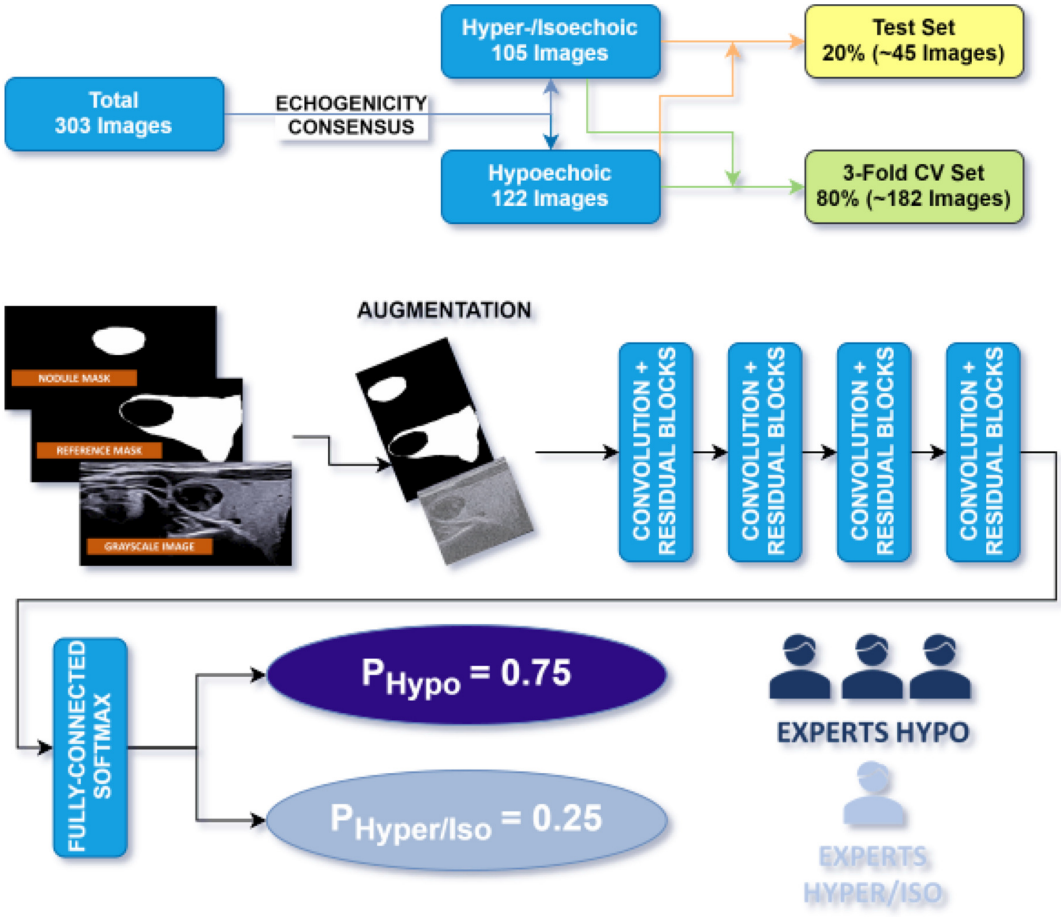

Figure 3.4 – (Top) Sample partition for cross-validation training and testing of ResNet50 for binary classification between hyper-/isoechoic and hypoechoic nodules. Training partitions were randomized with proportionate representations of both classes. (Bottom) Images for training within 3-fold cross validation were treated with data augmentation techniques, and used with class labels based on expert consensus with a ResNet50 architecture.

Therefore, the final input format used with the network was that of a three-channel image, consisting of the two masks and the greyscale image data. This provided, in theory, the spatial and echogenicity information necessary to make the distinction. Formats with the non-image channel containing only the pixels of the regions corresponding to the masks were also tested, but showed no improvement in results. The output from the network, passing through a softmax layer, was used as the probability estimate associated with either class, summing to one.

The networks were trained with supervised learning using labels from the expert evaluation of the echogenicity. Initially, binary ground-truth labels were generated based on the majority evaluations of experts. However, given that there was not always perfect agreement on the consensus, the target label was modified to be the proportion of experts who favored one label over the other, normalized to sum to one across both classes. This generated soft labels to use as prediction targets, e.g. with a value of 1 if all experts agreed or 0.75 if three out of four agreed (see Figure 3.4).

Binary cross entropy loss for the predicted probability of the hyper-/isoechoic label compared to the expert label (with a value between 0 and 1) was used for training:

$$\mathcal{L} = -\sum_{n}^{N} \left[y_{n,hyper} \cdot \log(p_{n,hyper}) + (1 - y_{n,hyper}) \cdot \log(1 - p_{n,hyper}) \right], \tag{3.1}$$

where $p_{n,hyper}$ is the predicted probablility of the hyper-isoechoic label for the nth sample, $y_{n,hyper}$ is the target label, and the loss is calculated across all N samples.

Each model was trained until overfitting occurred, and the weights for evaluation were determined by the epoch having the lowest binary cross-entropy loss on the validation set for each fold. The networks trained on each fold were then used to generate predictions of hypoechoic class probability on the reserved test set, which were averaged.

Multiple random partitions of test sets and cross validation splits were tested, to account for variability as a function of which images were in the cross-validation set.

3.3.2 Echogenicity Classification Results

The results of the predictions on the test sets for different random partitions are given in Table 3.1. Some examples on specific images are shown in Figure 3.5.

Random Partition Number	Hyper-/Isoechoic Test Samples	Hypoechoic Test Samples	Classification AUROC
1	22	26	0.797
2	21	24	0.669
3	22	25	0.707
4	21	24	0.659
Average	21.5	24.75	0.708

Table 3.1: AUROC and test set composition for the binary classification network across different sample partitions.

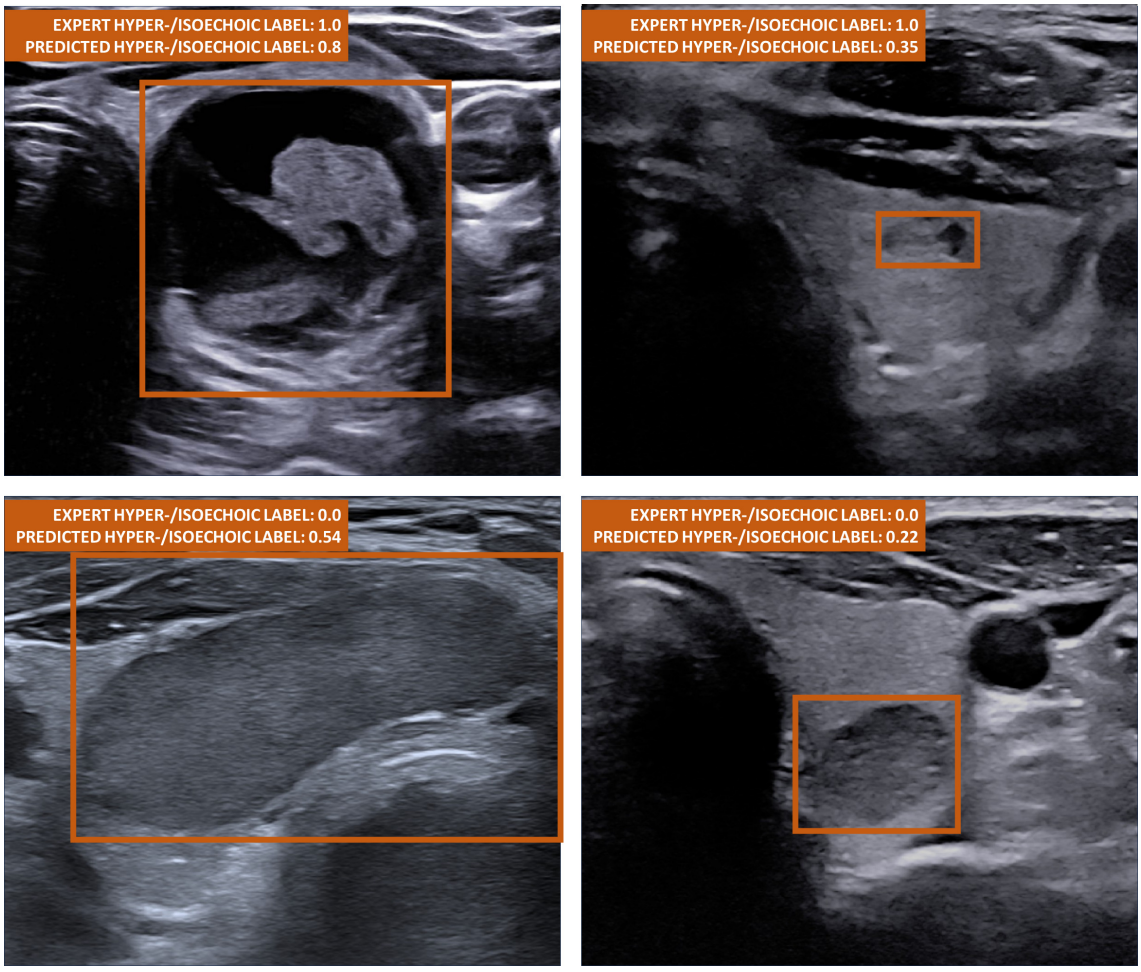

Figure 3.5 – Examples of predictions on images, with the reference and predicted values for the hyper-/isoechoic labels listed.

For a binary classification, the AUROC scores on the test set were not particularly impressive. This was a slight improvement over the AUROC of 0.59 obtained on the test set from training with a three-channel image and no masks; no difference was noted when the other channels of the input image were restricted to pixels of the grayscale image in the regions corresponding to the masks, with non-included pixels set to zero.

Given that a similar network architecture was successful in learning a binary classification of benign/malignant nodules, there may be an issue with the proposed task that makes it difficult to learn (Alghanimi et al., 2024).

3.4 Reproducibility of Expert Echogenicity Classification

Given that, as we have seen in the previous chapter, experts can vary significantly in their description of ultrasound features, it could be that the lackluster echogenicity classification results were due to the substantial label noise. If the expert labels lacked a reproducible connection to the images, it would be difficult for a network to learn to distinguish classes in a fashion that was generalizable to a withheld test set. This could be investigated by using expert re-evaluations of the same images, to examine the intra-expert variability.

Following the individual expert evaluations, a series of consensus meetings were organized to evaluate the images for which the EU-TIRADS score was not agreed upon by at least three out of the four experts. This had initially been intended to generate consensus on all such images; however, discussion among the four experts could take over 30 minutes to reach a conclusion on a single image. Therefore, only ten images were subjected to this re-evaluation. These images were ordered randomly, and re-evaluated using the consensus process detailed in Figure 3.6.

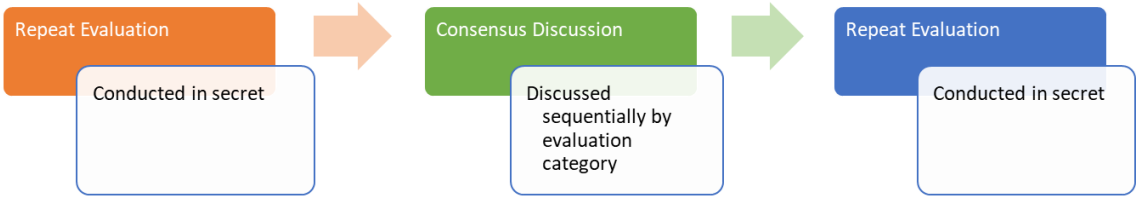

Figure 3.6 – Process followed for the consensus meetings. All participants secretly re-evaluated the image. A discussion among all four experts then proceeded sequentially along areas of the evaluation for which they did not reach a perfect consensus: composition, echogenicity, shape, margin, echogenic foci, and EU-TIRADS score. Finally, the images were re-evaluated again, secretly.

The intention was to generate consensus through discussion if possible, without forcing an agreement to which a majority was not convinced. The first step was for each expert to individually repeat the evaluation described in the previous chapter; this was conducted completely independently, with no discussion allowed until all four readers had finished. This served as a starting point for consensus discussions, and provided a measure of intra-reader reproducibility when compared to the initial evaluations. It should be noted, however, that the independent re-evaluation performed by each expert would be affected by the experience having gone through the process of consensus discussion with the other experts on prior images.

After the initial re-evaluation, a consensus discussion was led by a moderator. Each aspect of the evaluation discussed in the previous chapter was sequentially addressed by the moderator, who indicated whether or not there was unanimous agreement on each feature, including echogenicity. If there was a disagreement, discussion ensued over the label, before proceeding to the next item. Finally, the experts individually repeated the evaluation according to their best judgment. The purpose of this final individual re-evaluation was to allow for continued disagreement rather than to create an artificial consensus if some experts were not convinced. This would allow for preservation of genuine ambiguity in the expert assessment of EU-TIRADS and sonographic feature labels.

3.4.1 Intra-Expert Reproducibility Results

For the ten images that underwent re-evaluation, the evaluations provided by each of the four experts on the first and second rounds of independent evaluation are shown in Table 3.2.

Expert 1 First	Expert 1 Second	Expert 2 First	Expert 2 Second	Expert 3 First	Expert 3 Second	Expert 4 First	Expert 4 Second
Hyper/Iso	Same	Hyper/Iso	Same	Anechoic	Hyper/Iso	Very Hypo	Anechoic
Anechoic	Same	Hyper/Iso	Same	Hyper/Iso	Same	Very Hypo	Hyper/Iso
Hypo	Same	Hyper/Iso	Hypo	Hyper/Iso	Same	Hypo	Same
Hyper/Iso	Hypo	Hypo	Hyper/Iso	Hyper/Iso	Same	Hyper/Iso	Same
Hyper/Iso	Hypo	Hypo	Hyper/Iso	Hyper/Iso	Hypo	Hypo	Same
Hyper/Iso	Hypo	Hypo	Same	Hyper/Iso	Hypo	Hyper/Iso	Hypo
Hyper/Iso	Hypo	Hyper/Iso	Same	Hypo	Same	Hypo	Same
Anechoic	Same	Hyper/Iso	Same	Hyper/Iso	Same	Hypo	Anechoic
Hypo	Very Hypo	Hyper/Iso	Hypo	Hypo	Same	Hyper/Iso	Hypo
Very Hypo	Same	Hypo	Very Hypo	Hypo	Very Hypo	Very Hypo	Same

Table 3.2: Intra-expert variability in echogenicity label. The first and second labels assigned individually by experts are listed, with cases of altered labels during the re-evaluation being highlighted in orange.

For each of the re-reviewed cases, one or more of the experts changed their echogenicity label during independent evaluation. An example case is shown in Figure 3.7. Overall, almost half of all expert evaluations changed between the first and second rounds. It must be highlighted that these images had no consensus on EU-TIRADS score during the first round of annotation, and therefore might be considered as being intrinsically difficult to evaluate on the static axial images. However, the intra-expert variability still suggests that each individual's method of evaluation may not produce reproducible results for echogenicity. It is possible that these results would improve when not selecting for potentially difficult nodules, and when using more complete image information to allow for thorough echogenicity comparison.

During the discussion process, the moderator took note of which characteristics of nodules experts used to justify their opinions on echogenicity to each other in case of disagreement. Following the re-annotation, experts were invited to give their observations about which aspects of

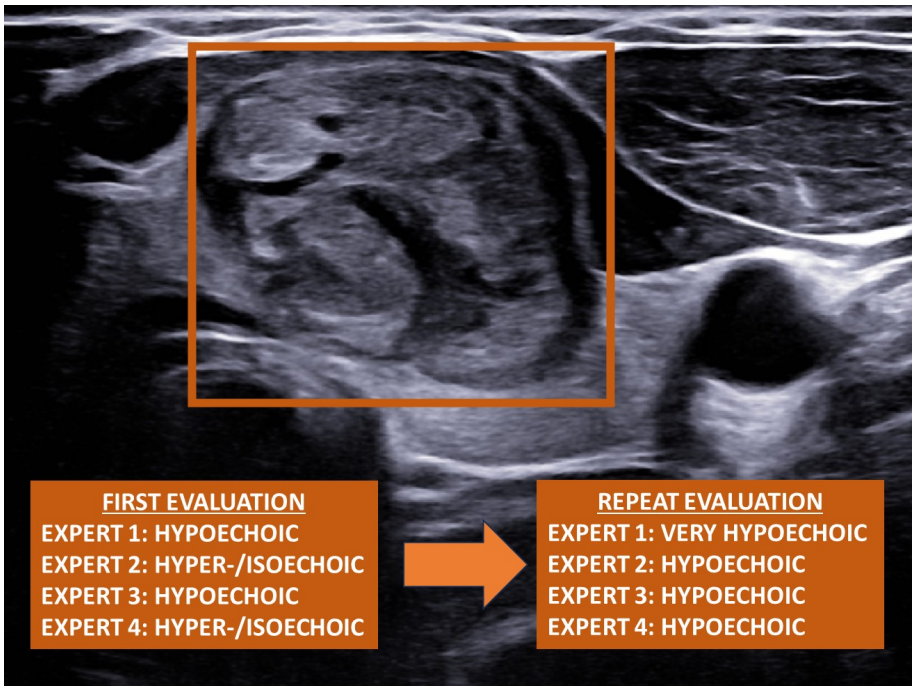

Figure 3.7 – Example image for which three out of four experts changed their assigned echogenicity label.

images they felt made characterization of echogenicity more difficult. The most common source of discussion appeared to be when experts would signal different regions of the nodule, in case of heterogeneous echogenicity, as a basis for judging the entire lesion. The experts agreed that it was difficult to judge in some cases whether a nodule was predominately hyper-/isoechoic or hypoechoic if the nodule was not uniform.

The experts also frequently observed that not finding an adequate reference zone within the thyroid tissue in the image to make the comparison was a common difficulty. This was seen during discussions, as individual experts would draw attention to different regions of the thyroid parenchyma in case of heterogeneity in order to convince their colleagues. All of the experts agreed that having more reference tissue, such as being able to look for this on the sagittal view, could have helped them more confidently assess the echogenicity. In particular, one expert cited a difficulty in evaluating cases with large nodules that, in an axial view, dominated most of the thyroid cross-section and left little in the way of reference tissue.

3.4.2 Implications of Expert Variability in Echogenicity Assessment

Evidently, the four experts showed a great deal of intra-reader variability, in addition to the inter-reader variability seen in the last chapter. Some of this could be due to the limitations of evaluation on a static axial-view image. Looking to the literature, we find a recent study of inter- and intra-reader variability using video clips by Solymosi et al. They reported a mean intra-rater Cohen's kappa value of 0.67 among 7 raters for the hyper-/isoechoic vs. hypoechoic distinction among 74 nodules (Solymosi et al., 2023). There may therefore still be variability with improved reference data such as video-clips.

While the intra-expert variability was measured on only ten images, the inter-expert variability across all images is worth considering. The reasons cited by the experts and observed as sources of difficulty for subjective evaluation merit further study. It could be that expert evaluation is uncertain only when the difference in overall nodule and reference zone intensity are not strongly pronounced; this issue could be addressed with better image contrast. The observations about nodule and background heterogeneity should also be investigated, as they may required more precision in label definitions. Finally, the relative size of the nodule and of the reference zone, as commented upon by one of the experts, could be a useful metric for quality control of diagnostic images.

These sources of error merit examination with more quantifiable metrics of echogenicity.

3.5 Quantitative Echogenicity Analysis

While expert echogenicity evaluation may be limited in terms of reproducibility, the B-mode image data can also be analyzed quantitatively to assess differences in echogenicity. As previously explained, the distinction between hyper-/isoechoic nodules and hypoechoic nodules is made by comparing the tissue within the nodule to normal thyroid parenchyma (see Figure 3.8). Quantitatively, the intensity of pixels within the nodule region of a static image can therefore be compared to the intensity of a region of thyroid parenchyma.

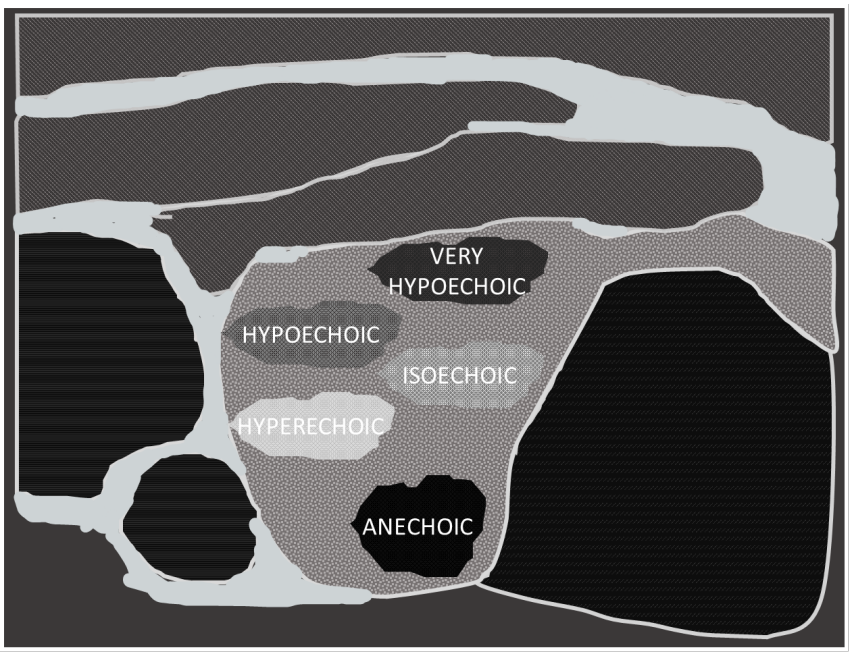

Figure 3.8 – Illustration of nodule echogenicity labels, which are determined by comparing the brightness or intensity of a nodule to that of a nearby reference area. Hyperechoic nodules are brighter than the surrounding normal thyroid parenchyma, while isoechoic nodules have a similar level of intensity to their surroundings. Hypoechoic nodules, by contrast, are darker than the surrounding thyroid parenchyma.

An important consideration to be made is that the images acquired by different practitioners on different ultrasound systems result in different settings with regards to image contrast and dynamic range. The intensity of pixels on an image, typically a log-scale compression of the signal intensity received by the ultrasound probe, is affected by the strength of the pulse sent in, as well as the gain applied to the returning signal. In clinical ultrasound, time-gain compensation is applied to progressively amplify the signal so as to counteract the depth-dependent effects of attenuation. This can be adjusted by the operator in order to better visualize different regions of the tissue being examined.

These settings mean that different images, even those acquired by the same operator on the same machine during different sessions, are not standardized in terms of the distribution of pixel intensity. However, all experts were instructed to acquire images with suitable settings to allow for EU-TIRADS analysis, including description of echogenicity. This would typically involve ensuring that a noticeable intensity difference existed between the normal thyroid parenchyma and the strap muscles. Therefore, the images acquired in this study ought to have the required information used by human experts to make echogenicity determinations.

The quantitative approach has been previously studied by Wu et al., who compared inter-expert reproducibility of echogenicity labels with automated analysis using a commercial software (Wu et al., 2016). Their quantitative measure consisted of the difference between the mean pixel intensities of nodule and reference (thyroid tissue or muscle) areas of B-mode images. When compared between the populations of benign and malignant nodules (as confirmed by surgical biopsy), a statistically significant difference was found (Wu et al., 2016). In addition, these values were more accurate than expert hypoechogenicity label as a predictor of malignancy (Wu et al., 2016).

3.5.1 Expert Label Agreement with Quantitative Echogenicity Measures

Among our images, we examined the quantitative differences in echogenicity. Having no ground-truth cytological or histopathological confirmation, we focused on comparing these differences to inter-expert variability, in order to investigate potential connections between expert labels and pixel intensity distributions.

As discussed previously, for each image with at least a weak consensus for either the hyper-/isoechoic or the hypoechoic label, a non-expert annotator manually segmented the nodule and all non-lesion areas of the ipsilateral thyroid lobe judged to not be in a pronounced acoustic shadow (see Figure 3.9). The pixels included in these masks were then filtered with two thresholds: a lower bound judged visually to exclude nearly anechoic cystic zones (which would be of a similar intensity to the trachea or blood vessels), and an upper bound to exclude echogenic foci (brighter than normal thyroid tissue). This operation was conducted in order to utilize only the solid soft tissue components of the nodule and the rest of the thyroid lobe for comparison.

With these filtered zones, pixel intensities (stored in the DICOM as values from 0 to 255) were normalized to a 0-1 range. The difference between the distributions of pixels in the nodule region and the reference region were then compared for images having different degrees of expert agreement (see Figure 3.10).

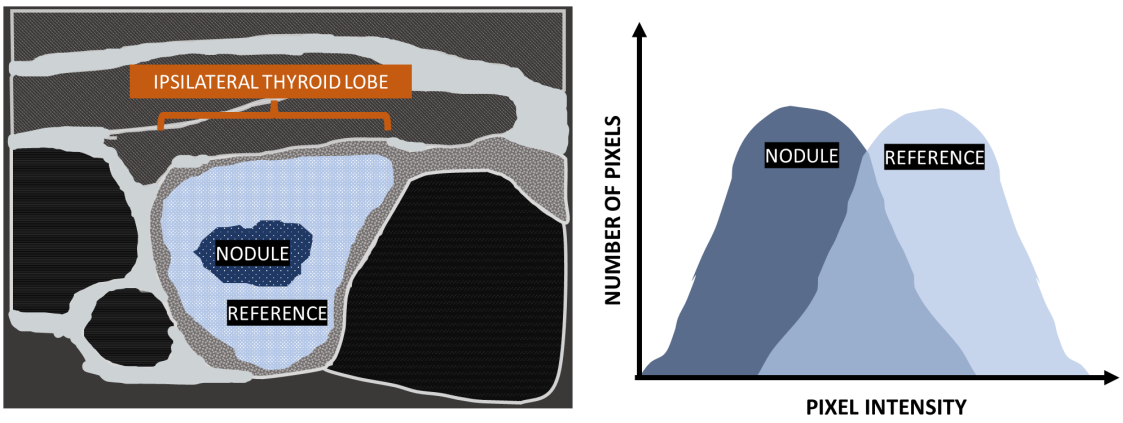

Figure 3.9 – (Left) Illustration of ultrasound image with masks selecting the nodule region and available healthy thyroid parenchyma in the ipsilateral lobe as a reference area. These zones were filtered with pixel intensity thresholds to exclude cystic areas and echogenic foci such as microcalcifications. (Right) Illustration of pixel intensity plots being considered; for each image, the distributions of nodule pixel intensity were compared between the nodule and reference area to look for differences associated with the expert echogenicity labels.

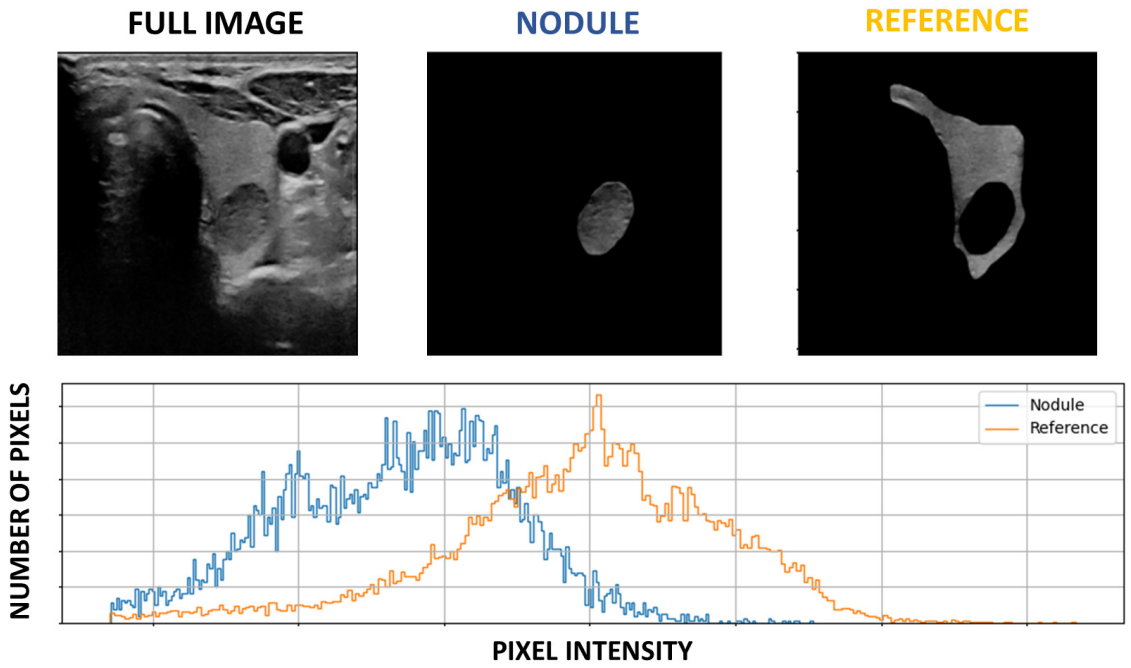

Figure 3.10 – (Top) The entire image and the zones selected as nodule and reference regions. (Bottom) The pixel intensity distributions for the nodule and reference regions. In this case, the nodule zone's distribution seems to have lower pixel intensities than the reference zone, corresponding with the experts' unanimous hypoechoic label.

3.5.2 Quantitative Analysis Results

The first consideration in analyzing inter-expert variability on echogenicity labels was whether the experts were more often in agreement on images in which the overall differences between the nodule and reference zones were strong. This was measured by using mean pixel intensities of the two regions. The difference between the mean pixel intensity in the reference zone $\mu_{\text{reference}}$ and the mean pixel intensity in the nodule zone μ_{nodule} was compared for different degrees of expert agreement on the hyper/isoechoic vs. hypoechoic label. Examples of images with high and low values of this metric are presented in Figure 3.11, with a lower value corresponding in theory to nodules identified by most experts as hypoechoic, and a higher value to nodules identified by most experts as hyper-/isoechoic. The distribution of these metrics as a function of the overall expert agreement on a label is presented in Figure 3.12.

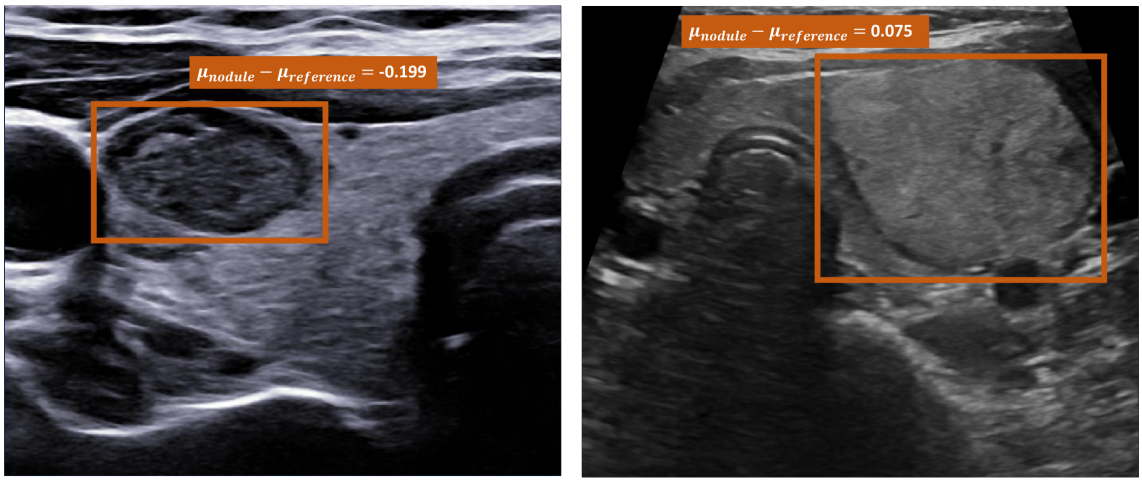

Figure 3.11 – (Left) Example of an image in which the difference between the mean pixel intensities in the nodule and reference regions is on the lower end of the observed range, denoting a more hypoechoic nodule. (Right) Example of an image in which the difference between the mean pixel intensities between the same regions is at the higher end of the observed range, denoting a hyper-/isoechoic nodule.

The value $\mu_{\text{nodule}} - \mu_{\text{reference}}$ was in fact previously compared between regions of nodule and manually selected regions of reference tissue by Wu et al., to compare the association of quantifiable echogenicity differences with malignancy. In that publication, this metric was correlated with malignancy; in our case, we examine the relationship of this value with expert consensus (Wu et al., 2016). If the four experts formed an opinion of hyper-/isoechogenicity or hypoechogenicity on the basis of the global echogenicity difference between the nodule and the reference tissue as measured by the mean, one would expect to see in Figure 3.12 higher values of $\mu_{\text{nodule}} - \mu_{\text{reference}}$ for cases with 4/4 or 3/4 expert consensus on a hyper-/isoechoic label, with values close to (for isoechoic nodules) or exceeding zero (for hyperechoic nodules). By contrast, lower (and consistently negative) values would be expected for cases with 3/4 or 4/4 consensus on a hypoechoic label. While the median values of the extreme consensus groups in that figure somewhat reflect that difference, the distributions of values overlap significantly across all expert consensus groups. Correspondingly, a one-sided Mann-Whitney U test with the alternative hypothesis being that the

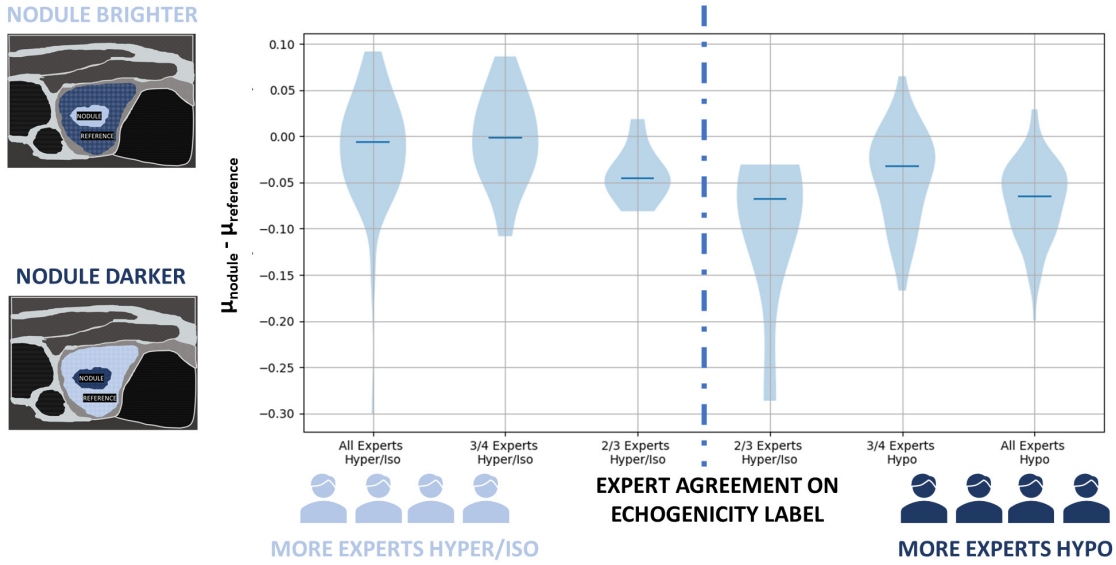

Figure 3.12 – Violin plot of mean pixel intensity distribution differences between the nodule and thyroid parenchyma, separated by the proportion of experts who assigned a hyper-/isoechoic or hypoechoic label. Median values of the distributions are indicated. In some cases, one expert applied a different echogenicity label, so the experts are counted out of three rather than four.

$\mu_{\text{nodule}} - \mu_{\text{reference}}$ would be lower cases with most experts assigning a hypoechoic label than for cases with most experts assigning a hyper-/isoechoic label had a p value of over 0.999.

This may suggest that when the experts assigned echogenicity values on the static axial-view images, they did not use a global reckoning of the pixel intensity in those regions, either because they looked to specific regions in order to make a determination, or because the human eye did not appreciate the difference in pixel intensity in the same way as quantitative analysis. If it is difficult for experts to appreciate global heterogeneity, their labels might be expected to be less consistent for cases in which either the nodule pixel intensities or the reference tissue pixel intensities were more heterogeneous. To examine this, we can look to the standard deviation of the pixel intensity distributions of nodules (in Figure 3.13 and the standard deviation of the pixel intensity distributions of reference zones (in Figure 3.15).

Figure 3.14 shows the distributions of the metric for nodules σ_{nodule} by sorted expert label agreement, and Figure 3.16 shows the same for the reference tissue $\sigma_{\text{reference}}$.

If heterogeneity in the pixel intensities of either region makes it difficult to obtain inter-expert consensus on the hyper-/isoechoic or hypoechoic labels, one would expect lower values of σ_{nodule} or $\sigma_{\text{reference}}$ in both Figures 3.14 and 3.16 for the cases with 4/4 consensus on either the hyper-/isoechoic or hypoechoic labels, and higher standard deviations for cases with less firm consensus. For both cases, the distributions appeared to visibly overlapping across all groups and have similar median values. When grouping cases of 4/4 expert agreement to compare with all less certain cases, a one-sided Mann-Whitney U test showed a p-value of approximately 0.0387 for lower values of σ_{nodule} in the 4/4 group. The p-value for the equivalent test on $\sigma_{\text{reference}}$ was approximately 0.778.

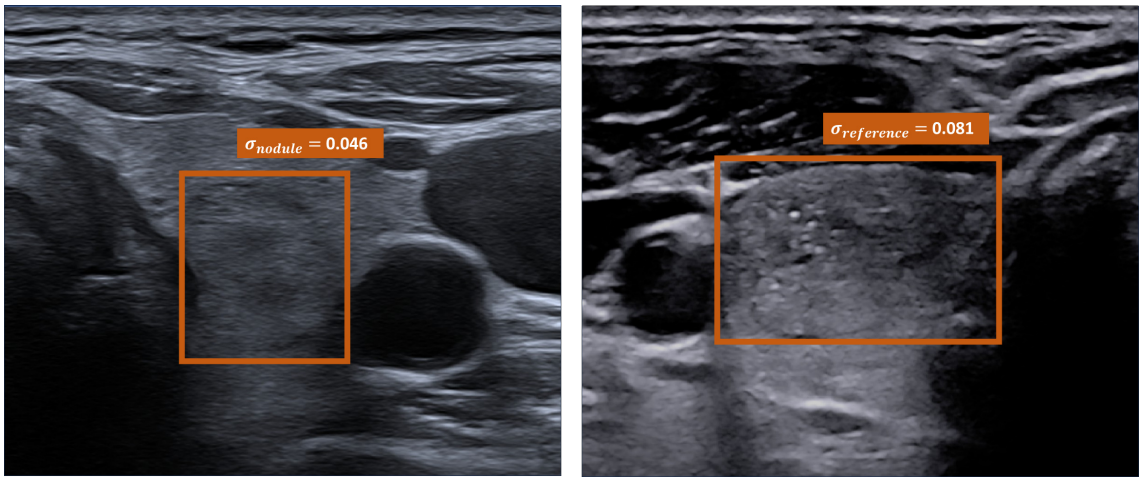

Figure 3.13 – (Left) Example of an image in which the standard deviation of pixel intensities in the nodule is on the lower end of the observed range, denoting a more homogeneous nodule in terms of echogenicity. (Right) Example of an image in which the standard deviation of pixel intensities in the nodule is greater, denoting a more heterogenous nodule in terms of echogenicity.

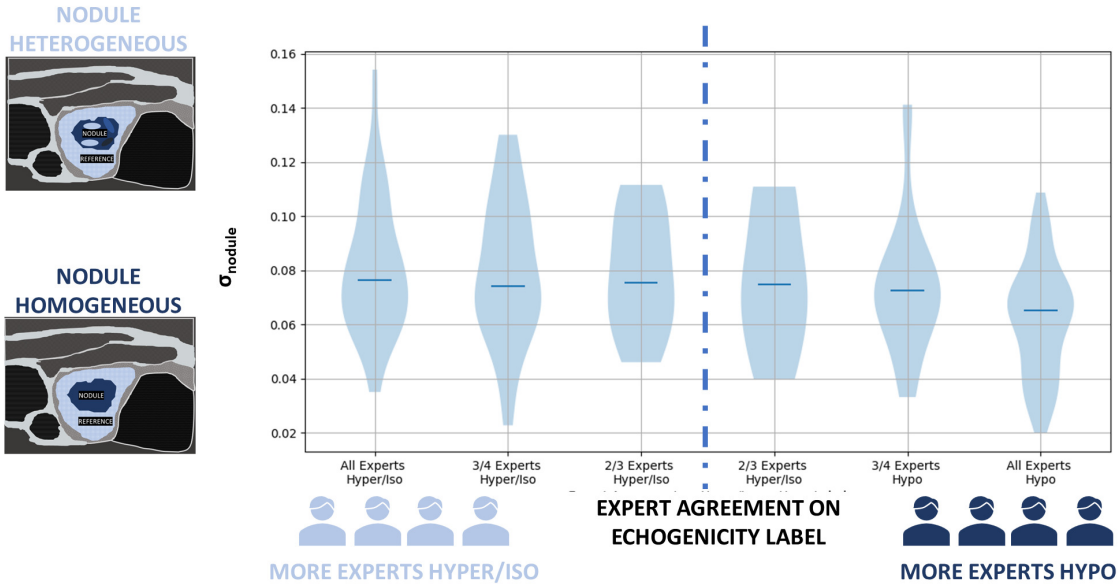

Figure 3.14 – Violin plot of the standard deviations of pixel intensity distributions of nodule regions within ultrasound images, separated by the proportion of experts who assigned a hyper-/isoechoic or hypoechoic label. Median values of the distributions are indicated. In some cases, one expert applied a different echogenicity label, so the experts are counted out of three rather than four.

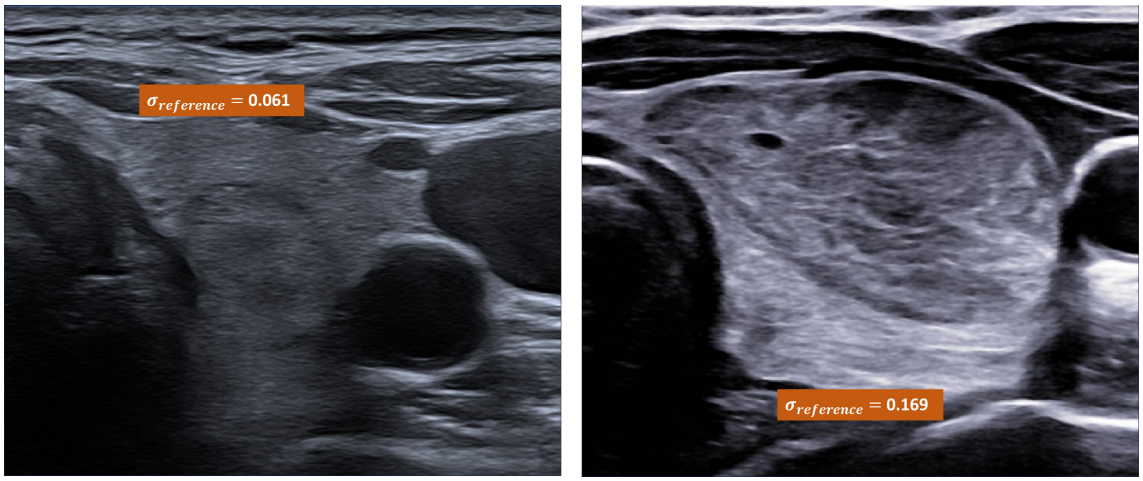

Figure 3.15 – (Left) Example of an image in which the standard deviation of pixel intensities in the reference zone is on the lower end of the observed range, denoting a more homogeneous echogenicity. (Right) Example of an image in which the standard deviation of pixel intensities in the reference is greater, denoting a more heterogeneous echogenicity.

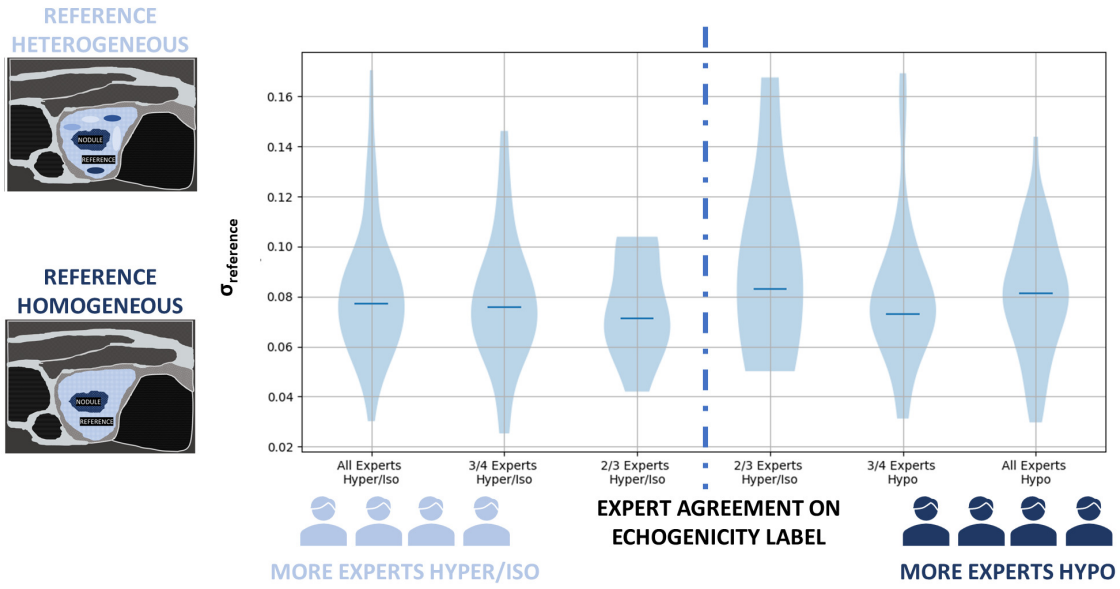

Figure 3.16 – Violin plot of the standard deviations of pixel intensity distributions of reference tissue regions within ultrasound images, separated by the proportion of experts who assigned a hyper-/isoechoic or hypoechoic label. Median values of the distributions are indicated. In some cases, one expert applied a different echogenicity label, so the experts are counted out of three rather than four.

This comparison may suggest that nodule heterogeneity impacted experts' evaluations; this was congruent with their observations after the consensus process. As there was no evident association found with reference zone heterogeneity as measured by $\sigma_{\text{reference}}$ and expert agreement, it could be that experts search for small patches of the reference region that they judge to be sufficient to make an echogenicity distinction. When considering echogenicity, practitioners must be selective in order to avoid areas of acoustic shadow or inflammation in the thyroid lobe (Russ et al., 2017 ; Tessler et al., 2017).

Indeed, experts commented on the difficulty of making an evaluation from a static image if they felt that the static image did not present sufficient areas of normal parenchymal tissue to allow for a clear comparison, this could contribute to expert uncertainty. At they same time, particularly large areas of suitable reference zones led in the discussion to disagreement over where to look. There could therefore be a connection between expert agreement on labels and the size of the reference region. Since this would depend on the acquisition settings of each operator, the ratio between the size of this region and that of the nodule region could be compared.

Figures 3.17 and 3.18 show the ratio of the number of normal parenchymal reference tissue pixels $N_{\text{reference}}$ to the number of nodule tissue pixels N_{nodule} for the different images, separated by the expert label consensus. Visually, no trend separating the extremes of 4/4 expert agreement from less certain cases is evident from this figure. The results of a two-sided Mann-Whitney U test between the 4/4 agreement cases and all other cases gave a p-value of about 0.659.

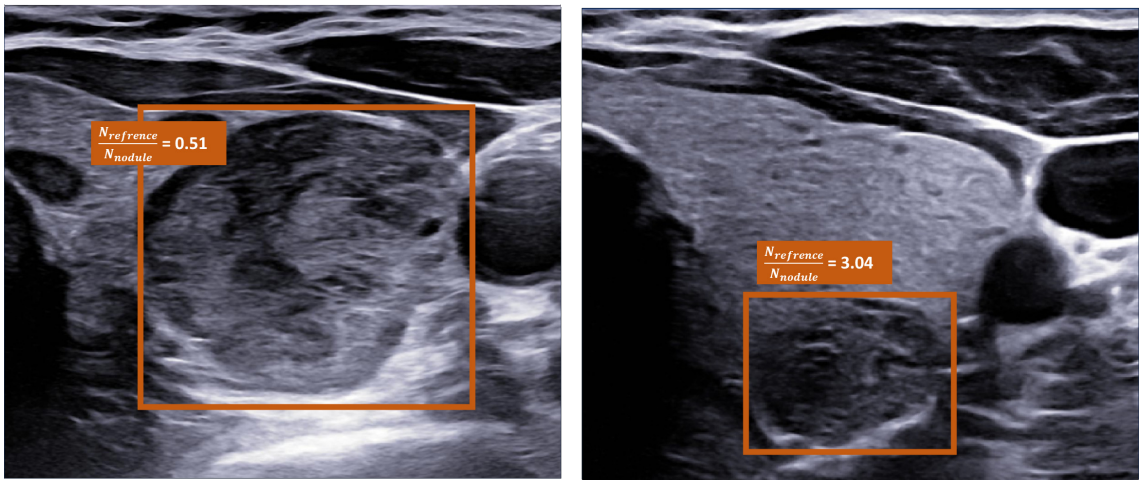

Figure 3.17 – (Left) Example of an image in which the ratio of reference pixels to nodule pixels is small. Note that other nodules present are not included in the area of the reference zone. (Right) Example of an image in which the ratio is much larger.

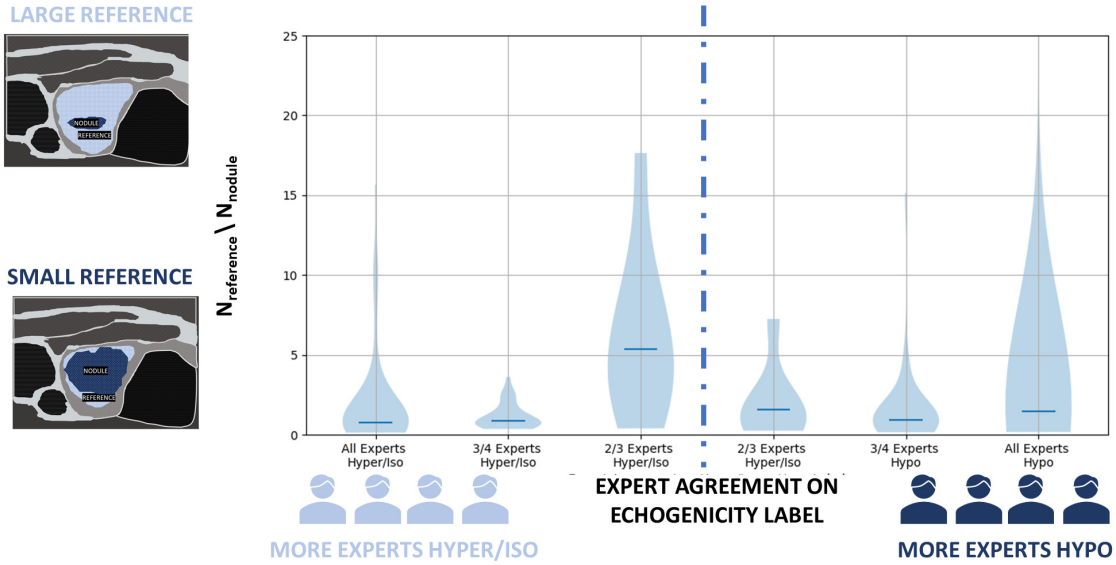

Figure 3.18 – Violin plot of the ratio of number of pixels within the reference region to the number within the nodule region, separated by the proportion of experts who assigned a hyper-/isoechoic or hypoechoic label. Median values of the distributions are indicated. In some cases, one expert applied a different echogenicity label, so the experts are counted out of three rather than four. An outlier is excluded from the all experts hypoechoic group.

3.6 Discussion

The results of the Chapter 2 showed a marked degree of inter-reader variability in the description of the ultrasound features used for EU-TIRADS classification. The difficulty of this task is sufficient motivation for the creation of tools to assist less experienced French practitioners evaluate thyroid nodule images.

In our case, this took the form of utilizing a neural network to predict the difference between hyper-/isoechoic and hypoechoic nodules. This same architecture had worked to perform benign-malignant classification on the basis of biopsy results did not learn from expert consensus labels in a manner that was generalizable to a test set (see Table 3.1) (Alghanimi et al., 2024). This was true even when utilizing position masks to provide necessary information for nodule echogenicity comparison, i.e. the regions of the nodule and of healthy thyroid parenchyma (Russ et al., 2017).

This suggests that the task of learning to predict the expert labels of echogenicity from the axial-view B-mode images with relevant position information was a more difficult task to learn reproducibly. In addition to the limited size of the dataset, this draws attention to two aspects for the learning task: the expert labels used as a target, and the input data used to make the predictions.

It was already apparent from the inter-expert variability seen in the previous chapter that the labels provided by French experts when analyzing static images are noisy. As these practitioners are used to performing evaluation with dynamic control over imaging settings, as well as with more views than a single static axial image, some degree of variability is to be expected. However, analysis of the intra-expert variations (see Table 3.2) suggests that the evaluations made by individual experts on the same still image are not necessarily reproducible.

Evidently, it is not possible to reproducibly learn an ill-defined task with unclear labels. Many studies do have something closer to a ground truth label in the form of cytology or histopathology results from FNA or biopsy; however, relying on these may skew populations by excluding nodules that would not be subjected to any further procedure (Piticchio et al., 2024). Unsupervised learning strategies could be useful in this context, especially with a much larger set of unlabeled images. However, our interest, specific to the French context, drives us to more closely examine the reasons for this label uncertainty.

In order to understand why French experts can vary when applying an echogenicity label, the method by which they apply this label must be examined with respect to the pixel-level information present in the image. If, as had been previously explored by Wu et al., the difference in echogenicity depended on a global assessment of pixel intensity, a difference in mean intensity value between distributions could separate expert labels (Wu et al., 2016). However, expert consensus between hyper-/isoechoic and hypoechoic labels was not associated with perceptible differences between the mean values of nodule and reference areas (see Figure 3.12).

This suggests that the mean value of the pixel intensities in the image may not be perceived by the experts as a basis for assigning the echogenicity label. It is possible that heterogeneity in nodule or reference zone pixel intensity complicates the assessment. This appears to have been the case for heterogeneity within a nodule, which was greater in cases with more expert disagreement.

As for the variability of the reference zone, we did not find evidence of this being related to inter-expert agreement. The ratio of the number of pixels in the reference area to the number of pixels in the reference area also failed to show any advantage in providing more reference pixels. This aligns with findings from others' tests of inter-expert and intra-expert variability in a broader European context using video clips; and the results suggest that more images alone are not a complete solution (Solymosi et al., 2023).

For the moment, therefore, we have few direct insights into how expert echogenicity labels applied to static axial images connect to the pixel intensities in the nodular and parenchymal regions. However, it seems that heterogeneous nodules present a particular challenge for human experts; automated measurements here could be particularly useful to improve reproducibility. The information obtained in this chapter provides useful perspectives for future work in improving thyroid ultrasound.

3.6.1 Limitations

The exploration conducted in this chapter was necessarily limited by the small size of the dataset, though four different French practitioners were represented. This limited the capacity of a pre-trained network to attune to this task.

In addition to the limitation of using only axial-mode images, the sheer fact of limiting practitioners to examining static DICOMS rather than live patients is an important caveat. Expert evaluations could be more standardized if each pracitioner could see the patient under normal clinical circumstances. In addition, the variability of different quantitative measures during live ultrasound examinations could also be explored.

It is also worth noting that this investigation highlighted echogenicity as a category seen in the previous chapter to have a great deal of inter-expert variability with consequences for EU-TIRADS scoring. However, the interactions of this label with other features such as composition would also likely prove enlightening. This would necessitate a far more complete dataset in order to adequately represent rare feature combinations.

3.7 Conclusions

The purpose of this chapter was to explore how deep learning tools could help inexperienced French thyroid ultrasound practitioners evaluate nodule echogenicity, which had been defined as an area of inter-expert difficulty. However, the results that were uncovered threw into question the robustness of expert echogenicity labels when applied to static axial view images.

On the one hand, more rigorous definitions of echogenicity and other labels for EU-TIRADS could help improve French practice. Consensus labels arising from discussion are also preferable to individual, independent labels alone for the development of machine learning tools to help French practitioners. In light of intra-expert variability, multiple rounds of annotation may also be necessary. Though these measures will substantially increase the time and burden of annotation, they are indispensable for true reproducibility.

In addition, looking to more complete datasets as a basis for evaluation, such as with two views or video clips could be necessary, but likely not sufficient. Another means, however, to allow for more reproducible evaluation would be to apply quantitative ultrasound methods in future studies to examine associations with malignancy. This would depend on ground-truth validation while avoiding the sampling bias of nodules not subjected to FNA or biopsy, such as with a cadaver study.

CHAPTER 4

Active Learning Limitations on Clinical Thyroid Ultrasound Data

Machine learning applications in ultrasound imaging are limited by access to ground-truth expert annotations, especially in specialized applications such as thyroid nodule evaluation. Active learning strategies seek to alleviate this concern by making more effective use of expert annotations; however, many proposed techniques do not adapt well to small-scale medical image datasets. In this chapter, we test active learning strategies including an uncertainty-weighted selection approach with supervised and semi-supervised learning to evaluate the effectiveness of these tools for the prediction of nodule presence on a real clinical ultrasound dataset of over one thousand images. Binary classification performance on two other medical image datasets is also assessed, using many repetitions with different random seeds. The results suggest that most active learning strategies struggle to consistently outperform random selection of images for annotation, on ultrasound as well as on other forms of medical imaging. Combining semi-supervised strategies with a degree of random selection can slightly improve performance, but even then, active learning may have limited clinical significance in terms of reducing thyroid ultrasound annotation burden.

This chapter was published as a conference paper (Sreedhar, Lajoinie, Raffaelli, & Delingette, 2024).

4.1 Introduction 93
4.2 Background 93
4.2.1 Limitations of Active Learning Strategies 93
4.2.2 Active Learning Applied to Thyroid Ultrasound 95
4.3 Methods 96
4.3.1 Image Datasets 96
4.3.1.1 External Datasets 98
4.3.2 Rigged Draw Strategy 98
4.3.3 Supervised and Semi-supervised Active Learning Strategies 98
4.4 Results 99
4.4.1 Supervised Learning Results 100
4.4.2 Semi-Supervised Learning Results 103
4.4.3 Initial Set Impact 106
4.5 Discussion 108
4.5.1 Limitations 109
4.6 Conclusion 109

4.1 Introduction

In Chapter 2, we saw many of the limitations inherent in the subjective evaluation of thyroid nodule ultrasound. In Chapter 3, we discussed some of the machine learning algorithms that have been proposed for the analysis of B-mode thyroid images. Whatever the predictive strategy of an algorithm may be, it must be trained and validated on samples that represent the target population.

One limitation for training or fine-tuning these algorithms is the difficulty of obtaining expert annotations of nodule location, margin, or characterization. In our Chapter 2 study, TIRADS assessment alone took an average of 1 minute and 14 seconds per expert per image. During a single session, practitioners could often evaluate only 50 images before requiring a break, and the frequency of sessions depended on clinical work schedules. Furthermore, given the inter-expert variability seen in Chapter 2, and the intra-expert variability in Chapter 3, consensus meetings and/or repeat evaluations might also be in order.

The need for time-consuming expert annotations cannot be obviated by relying exclusively on cytological or histopathological confirmation. The diagnostic criteria based on cytological analysis is evolving with the incorporation of more molecular markers, and even histopathological criteria have recently been redefined (Lebrun & Salmon, 2024). Furthermore, the available data for nodules that have undergone fine needle aspiration or biopsy would be skewed due to the exclusion of nodules judged to be most likely benign on ultrasound.

Therefore, high-quality annotations created by practitioners specifically experienced in thyroid ultrasound are essential to the development or fine-tuning of machine learning tools, and also represent an expensive and time-consuming bottleneck. Practical clinical implementation will therefore depend on training strategies that make intelligent use of the annotations as ground-truth labels.

This is where active learning holds promise, as a means of efficiently utilizing expert annotations. This approach to training machine learning algorithms is based on the premise that, for a large set of unlabeled data, there may exist a smaller subset of observations which would be as effective for supervised learning as the entire set. In terms of medical image analysis, this means starting with a collection of unlabeled images, and having expert annotate only a small subset at random. This initial subset of labeled images is used for supervised learning, though the unlabeled images may be used for semi-supervised learning (Huang, Huang, Wang, Xu, & Liu, 2022 ; Shui, Zhou, Gagné, & Wang, 2020). Additional images are then selected for annotation, with the goal being to choose only those which would help the algorithm improve its performance. In this way, the learning task is accomplished while demanding fewer expert labels.

The intuition behind active learning is appealing, particularly in the context of requiring many expert labels on thyroid ultrasound images. However, the actual utility of these strategies to thyroid nodule analysis must be confirmed, because the use case of active learning is on real clinical data, without the chance to fine-tune. Therefore, a robust evaluation of active learning strategies for training machine learning algorithms for thyroid ultrasound image analysis is necessary.

4.2 Background

4.2.1 Limitations of Active Learning Strategies

In the context of medical images, most active learning approaches use pool-based sampling. An initial group of unlabeled images is gathered, from which an initial set is chosen at random.

After a prediction algorithm is trained on this initial labeled set, additional images are selected for annotation. Once additional images are selected, the algorithm is retrained, and the cycle is repeated (see Fig. 4.1) (Budd, Robinson, & Kainz, 2021).

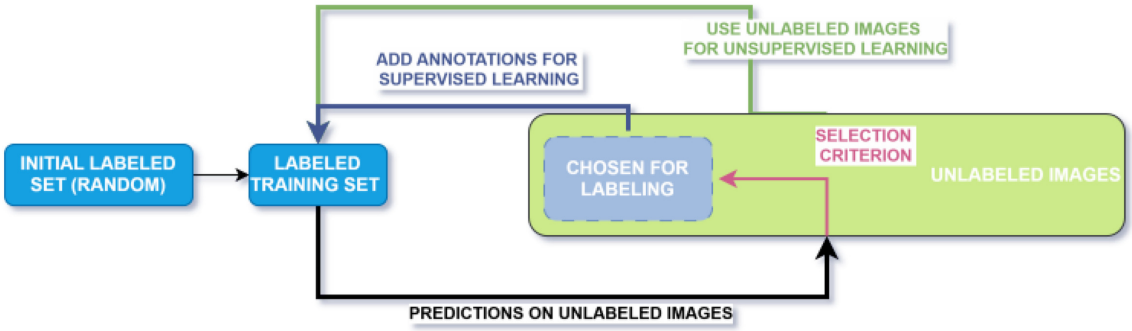

Figure 4.1 – The basic cycle of pool-based active learning: an initial set of images is randomly chosen for annotation, and used for training. In subsequent iterations, further images are chosen for annotation from the unlabeled image pool to retrain the algorithm. The unlabeled images can also be used for semi-supervised strategies.

The criteria for selecting images for annotation vary between strategies. The most commonly considered criterion is uncertainty, i.e. selecting cases in which the algorithm's predictions are no certain in order to improve its performance (Budd et al., 2021 ; Settles, 2009). Relying solely on this measure, however, risks overrepresenting a subset of cases, rather than the entire distribution of images. Therefore, diversity strategies seek to include images dissimilar to each other or to already-labeled images, to prioritize the "representativeness" of the selected instances (see Fig. 4.2) (Yang, Zhang, Chen, Zhang, & Chen, 2017 ; Smailagic et al., 2018).

Whichever specific strategy is chosen, active learning translates logically to the analysis of ultrasound images, because of the cost of manual annotation by expert radiologists. Zhou et al. demonstrated this by combining active learning with transfer learning to fine-tune a convolutional neural network for carotid intima-media thickness interpretation (Zhou et al., 2019). More recently, Huang et al. proposed a framework for segmentation of breast and knee cartilage ultrasound that combined active learning criteria with semi-supervised learning to better adapt to different ultrasound datasets, along with an uncertainty selection strategy modified to avoid redundant image selection (Huang et al., 2022).

Despite these advances, many active learning strategies struggle to outperform the baseline of randomly selecting images for annotation. Munjal et al. observed the inconsistencies in reported active learning performance in the literature, and through testing on non-medical images found that many strategies offered no consistent improvement over random annotations (Munjal, Hayat, Hayat, Sourati, & Khan, 2022). In terms of medical image data (MRI images) Gaillochet et al. demonstrated that active learning failed to consistently outperform random selection (Gaillochet, Desrosiers, & Lombaert, 2023). They addressed this problem by proposing a novel stochastic batch selection strategy to harness the power of random sampling on small-scale datasets (Gaillochet et al., 2023). These examples call into question the feasibility of practical implementation of active learning strategies in a clinical context.

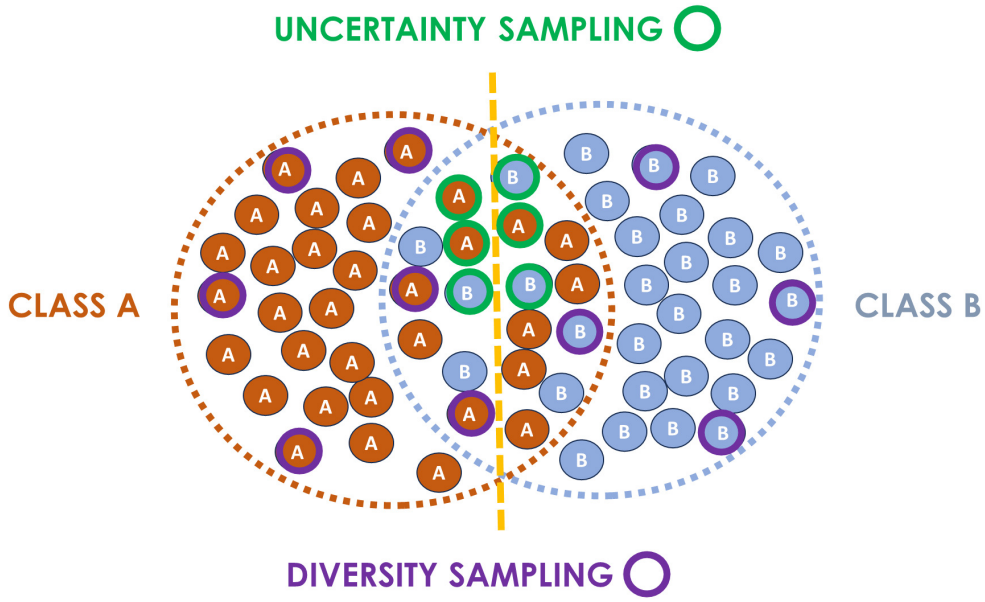

Figure 4.2 – The two main categories of active learning criteria: uncertainty and diversity. Uncertainty sampling chooses cases for which classification is difficult, and may select a subset of images, such as those that are similar between two classes. Diversity sampling attempts to represent more varied samples.

4.2.2 Active Learning Applied to Thyroid Ultrasound

The appeal of active learning techniques for thyroid ultrasound machine learning algorithms was made clear by the time required for expert evaluation of nodules in Chapter 2. From the evaluation sessions in that chapter, it appeared that 50 images would be a reasonable number of annotations to request from an expert at each given step. As far as selecting images for annotation, it could be important to adequately representing different nodule features (e.g. composition, echogenicity, shape, margin, echogenic foci, or nodule hereogeneity) in order to have an adequate training set for a neural network. With this context in mind, we applied active learning on a clinical dataset of thyroid ultrasound images. As with the dataset used in the previous chapters, these images were accumulated during the course of routine clinical practice, and were not acquired according to a standardized protocol. In order to include more images and labels, this separate, larger dataset was composed at a single hospital site.

We conducted a test of active learning strategies of binary classification of the presence or absence of thyroid nodules in these images. This was a real-world implementation adapted to the difficulties of learning on an actual clinical ultrasound data, including using semi-supervised feature extraction to facilitate active learning strategies. The results are assessed with a higher number of repetitions than is typically tested (Gaillochet et al., 2023 ; Shui et al., 2020 ; Zhan et al., 2022) to ensure statistical relevance. In addition, a novel and simple weighted selection active learning strategy to respect the representative power of random selection with small annotation budgets.

4.3 Methods

4.3.1 Image Datasets

A new single-center dataset was created from the stored images of thyroid examinations conducted in the course of routine clinical practice by radiologists at the Centre Hospitalier Universitaire de Nice from August 2021 to June 2022. All scans had been acquired on a Siemens S3000 ultrasound system (Siemens Healthineers, Erlangen, Germany) in accordance with standard practice for the institution.

All images from ultrasound examinations of the thyroid were exported in DICOM format and de-identified. A total of 4,490 images from 300 patients were exported in this fashion. Only the first exam corresponding to a patient was retained, so that the same patient would not be represented twice. As these images had been collected for clinical practice and not research, normal B-mode images were saved along with panoramic images, color Doppler images, and elastography images. The DICOM metadata from these images was therefore used to automatically filter out all panoramic, Doppler, and elastography images, leaving only plain B-mode images. Finally, images were manually sorted by a non-expert reader to only include those in axial views, which would be easier to interpret. These images had to include recognizable anatomical landmarks of the trachea or the carotid vessels to confirm their orientation. This processes yielded a total of 1048 images from 269 patients.

These images were then annotated by a non-expert reader. Initially, the annotations took the form of indicating whether any nodules were present, segmenting each nodule, and assigning a full ACR-TIRADS description with the categories of composition, echogenicity, shape, margin, and echogenic foci as described in Chapter 2. Examples of these annotations can be seen in Figure 4.3. However, following some initial tests of inter-expert variability with results similar to what was seen in Chapter 2, the non-expert evaluations were judged unreliable on the grounds of the reader's inexperience. If experts could disagree on the characterization and even the identification of nodules in multi-nodular cases, a non-experts descriptions and segmentation would be of limited value. Therefore, the labels were converted into a binary label of nodule presence (602 images with nodules of solid, cystic, or mixed composition) or absence (446 images).

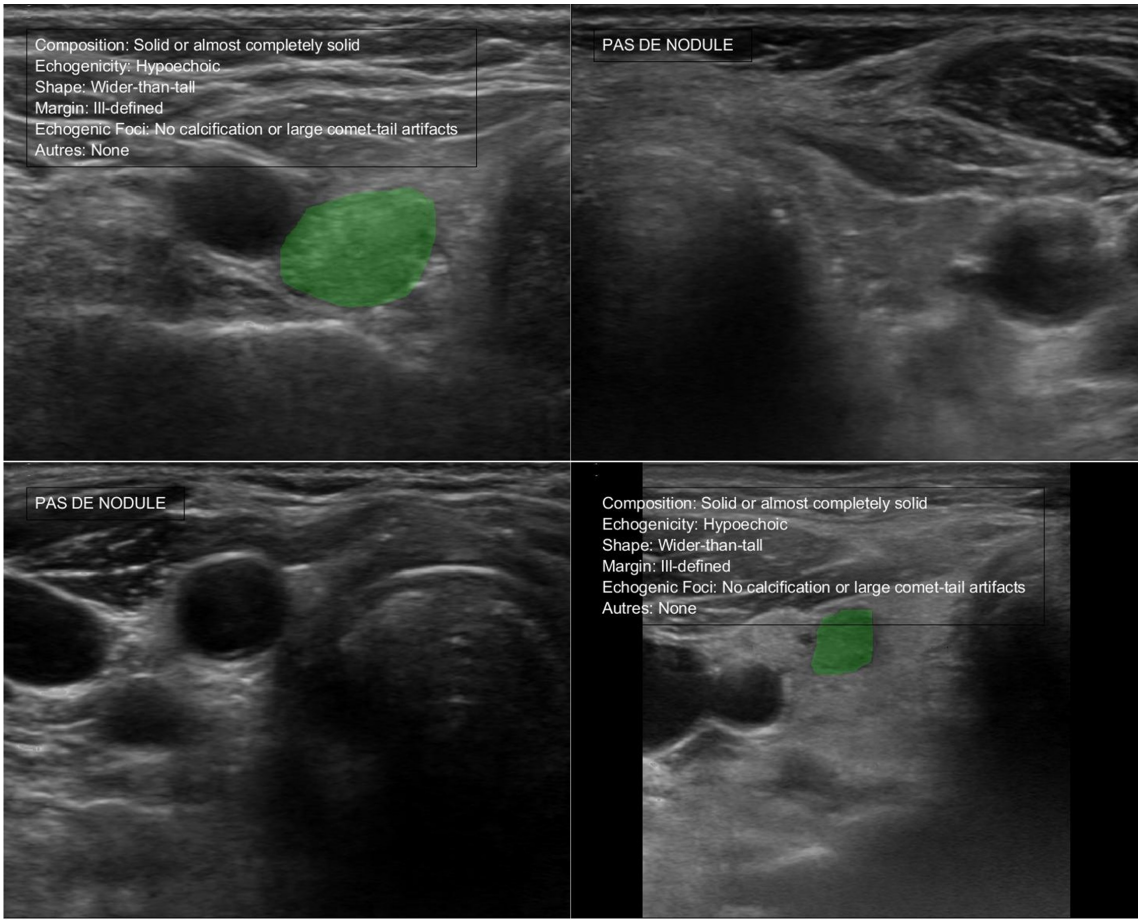

Figure 4.3 – Example ultrasound images from the dataset, with and without nodules, annotated by the non-expert reader.

4.3.1.1 External Datasets

Given the limitations of using non-expert annotations on our dataset, we also conducted equivalent tests of active learning strategies on two public medical imaging datasets randomly downsampled to an equivalent size. The PneumoniaMNIST dataset contains pediatric chest X-ray images with labels for pneumonia vs normal binary classification (Kermany et al., 2018). The BreaKHis dataset contains histopathological images in the context of breast cancer, with labels for benign and malignant diagnoses (Spanhol, Oliveira, Petitjean, & Heutte, 2016). These two sets had also been previously used to evaluate multiple active learning strategies (Zhan et al., 2022).

4.3.2 Rigged Draw Strategy

Similarly to the stochastic batch selection strategy of Gaillochet et al., we sought to combine the power of random selection with active learning in order to better represent the variability in a thyroid nodule dataset (Gaillochet et al., 2023). To do this while controlling the relative contribution of the uncertainty criterion, we proposed a weighted selection strategy called rigged draw. In this strategy, the relative weight w_n for selecting any sample in an active learning round is:

$$w_n(\alpha) = 1 + \alpha \frac{c_n}{c_{90}}, \tag{4.1}$$

where c_n is the value of the uncertainty-based criterion for the n$^{\text{th}}$ sample, c_{90} is the 90$^{\text{th}}$ percentile value of the criterion across all unlabeled images, and α is a factor weighting the importance of the uncertainty criterion relative to random selection. The choice to normalize relative to the 90$^{\text{th}}$ percentile was to avoid the effects of outlier maximum values with certain selection strategies.

4.3.3 Supervised and Semi-supervised Active Learning Strategies

We tested supervised learning using only labeled images with a ResNet18 pretrained on natural images. We compared random selection, LeastConfidence (an uncertainty strategy), and KMeans (a diversity strategy) as implemented in Zhan et al. (Zhan et al., 2022 ; D. Wang & Shang, 2014 ; Pedregosa et al., 2011). We also tested rigged draw sampling, defining the uncertainty criterion c_n as the positive entropy contribution of sample n:

$$c_n(p_n) = -p_n \log_2(p_n), \tag{4.2}$$

where p_n is the probability of nodule presence as predicted by the network (between 0 and 1). With this choice, we would preferentially weight images with a predicted probability close to 0.5.

As suggested by Huang et al., learning from ultrasound data may be difficult for active learning strategies that begin with few labeled images (Huang et al., 2022). We therefore also tested semi-supervised learning using the network architecture proposed by Shui et al. for their two-stage WAAL active learning strategy (Shui et al., 2020). This strategy depends on a network which conducts classification upon a feature representation which is in turn trained with a loss function seeking to reduce the distance between labeled and unlabeled images.

Our motivation for using this network was to imitate its approach to learning a useful feature representation from the images that would increase the effectiveness of active learning strategies. In addition to testing the entire WAAL strategy, this network structure was also used separately to test the previously mentioned active learning strategies.

4.4 Results

The active learning strategies were tested with both the supervised and semi-supervised strategies using the DeepAL+ toolkit from Zhan et al. (Zhan et al., 2022). For each test, a base set of 50 images was taken from a training set of 850 images and used to train the network for a fixed number of epochs (60), with subsequent batches of 50 being selected from among the unlabeled images, up to the maximum size of 750 images. The choice of a step size of 50 images was based on the fact that this was the approximate number of images we had noted that an expert could annotate in a single session before requiring a break. A balanced test set on our dataset was established using 199 images from patients not represented in the training set (102 with nodules, 97 without); on the other two datasets the test sets were slightly larger (624 for PneumoniaMNIST and 364 for BreaKHis, as noted in Table 4.1).

Dataset	Training Size	Test Size	Step Size	Max Size
US Dataset	849	199	50	750
Pneumonia MNIST	850	624	50	750
BreaKHis	850	364	50	750

Table 4.1: Active learning test sizes for each dataset.

In order mitigate the effects of different starting sets and the stochastic nature of certain selection strategies, approximately 20 repetitions were used (see Table 4.2). The rigged draw strategy was tested using weights of $\alpha = 5$, $\alpha = 25$, and $\alpha = 50$ to give different importance to the uncertainty criterion during selection.

Dataset	Strategy	Repetitions (Supervised)	Repetitions (Semisupervised)
US Dataset	Random	20	20
	RiggedDraw ($\alpha = 5$)	20	19
	RiggedDraw ($\alpha = 25$)	20	20
	RiggedDraw ($\alpha = 50$)	19	19
	LeastConfidence	19	19
	KMeans	20	20
	WAAL	N/A	19
Pneumonia MNIST	Random	19	20
	RiggedDraw ($\alpha = 5$)	19	19
	RiggedDraw ($\alpha = 25$)	20	19
	RiggedDraw ($\alpha = 50$)	19	20
	LeastConfidence	20	20
	KMeans	20	19
	WAAL	N/A	19
BreaKHis	Random	19	17
	RiggedDraw ($\alpha = 5$)	19	18
	RiggedDraw ($\alpha = 25$)	19	21
	RiggedDraw ($\alpha = 50$)	19	19
	LeastConfidence	20	18
	KMeans	18	17
	WAAL	N/A	16

Table 4.2: Active learning test repetitions with a different initial set for each strategy and dataset.

4.4.1 Supervised Learning Results

We used AUC under the ROC as a measure of classification performance independent of decision threshold. The efficacy of each active learning strategy under supervised learning was measured using the AUC as a function of the cumulative budget of labeled images. The median AUC values attained at each budget size for each active learning strategy including rigged draw with $\alpha = 25$ are plotted for our ultrasound dataset in Figure 4.4, for the PneumoniaMNIST dataset in Figure 4.5, and for the BreakHis dataset in Figure 4.6. The AUCs achieved for the ultrasound

dataset were substantially lower at all budget sizes than those achieved on the PneumoniaMNIST and BreaKHis datasets.

For an overall measure of the efficacy of active learning stratgies, we used the area under the budget curve (AUBC) values, calculated as the area under the curve of classification AUC value vs. the normalized cumulative budget (from 0 to 1) (Zhan et al., 2022). A summary of these AUBC values for the supervised strategies is given in Table 4.3. When the AUBC values from the repeated trials with learning strategies were compared to random selection, no statistically significant difference was found with the two-sample Kolmogorov-Smirnov test.

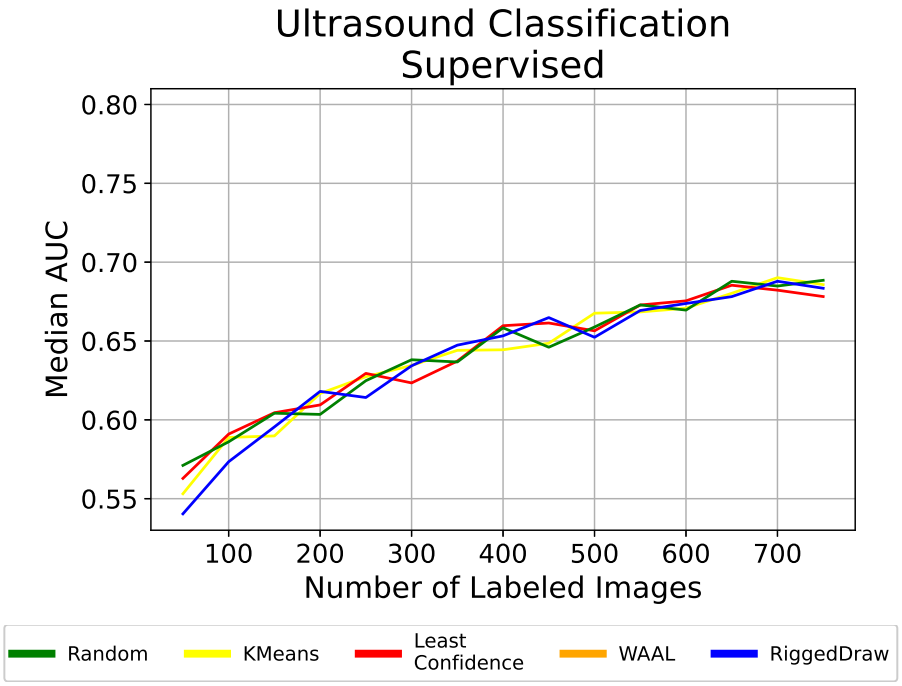

Figure 4.4 – Median AUC values for different active learning strategies with supervised learning on the ultrasound dataset.

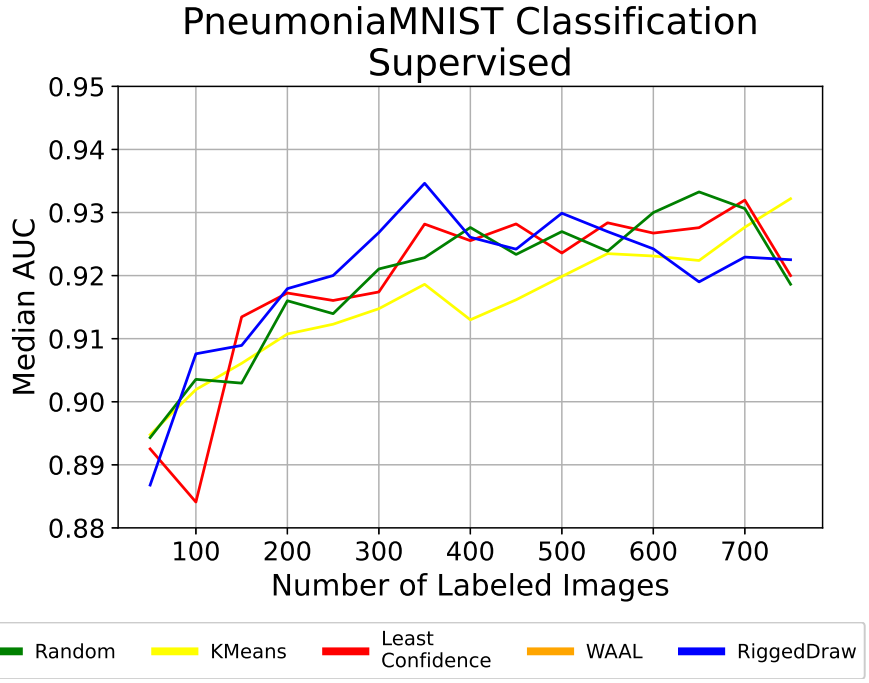

Figure 4.5 – Median AUC values for different active learning strategies with supervised learning on the PneumoniaMNIST dataset.

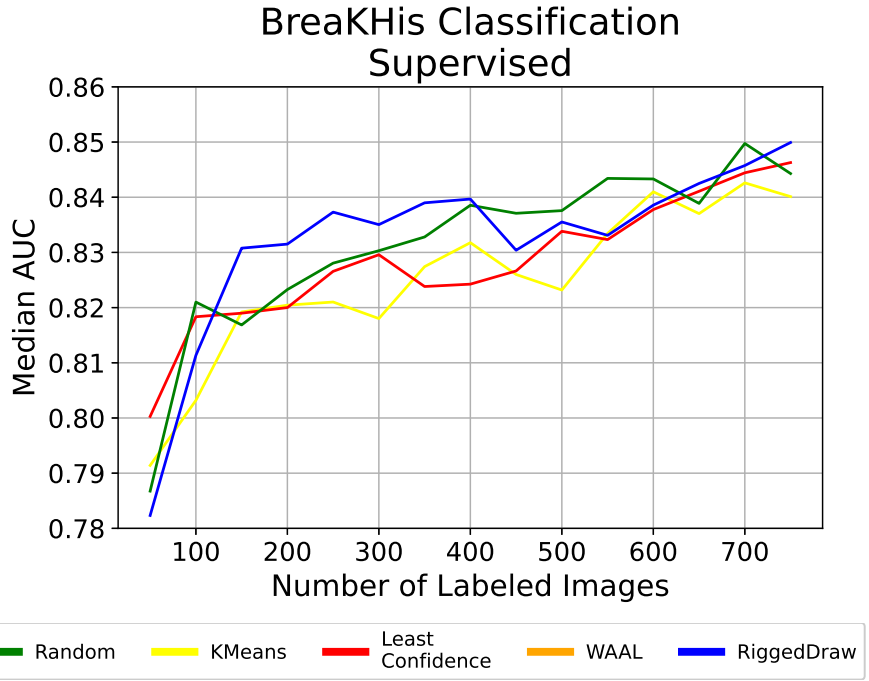

Figure 4.6 – Median AUC values for different active learning strategies with supervised learning on the BreaKHis dataset.

Dataset	Test Set Size	Measure	Rand	LC	KM	RD
US Dataset	199	Mean	0.643	0.642	0.641	0.639
		Median	0.642	0.646	0.641	0.641
		STD	0.010	0.011	0.009	0.012
Pneumonia MNIST	624	Mean	0.918	0.917	0.914	0.919
		Median	0.917	0.917	0.916	0.920
		STD	0.006	0.005	0.005	0.004
BreaKHis	364	Mean	0.832	0.828	0.826	0.832
		Median	0.831	0.829	0.823	0.836
		STD	0.015	0.020	0.017	0.022

Table 4.3: Supervised learning AUBC values. Values closer to 1 indicate a more effective strategy. Rand = Random. LC = Least Certain. KM = KMeans. RD = Rigged Draw (ours) with $\alpha = 25$.

To compare the rigged draw strategy with different values of the weight parameter α, AUBC values for supervised learning with are given in Table 4.4. When the AUBC values from the repeated trials with the rigged draw strategy were compared to random selection, no significant differences were found with the two-sample Kolmogorov-Smirnov test.

Dataset	Measure	$\alpha = 5$	$\alpha = 25$	$\alpha = 50$
US Dataset	Mean	0.639	0.639	0.642
	Median	0.639	0.641	0.640
	STD	0.012	0.012	0.008
Pneumonia MNIST	Mean	0.916	0.919	0.917
	Median	0.915	0.920	0.918
	STD	0.007	0.004	0.005
BreakHis	Mean	0.837	0.832	0.829
	Median	0.842	0.836	0.830
	STD	0.020	0.022	0.026

Table 4.4: Supervised learning AUBC values for different Rigged Draw weights, with * indicating p-values < 0.05 when compared to random selection.

4.4.2 Semi-Supervised Learning Results

For the semi-supervised strategies using the feature representation learned from all images, median AUC values attained at each budget size for different active learning strategies including $\alpha = 25$ are plotted for our ultrasound dataset in Figure 4.7, for the PneumoniaMNIST dataset in Figure 4.8, and for the BreakHis dataset in Figure 4.9. The AUBC values are reported in Table 4.5. When the AUBC values from the repeated trials with the rigged draw strategy were compared to random selection, a p-value of 0.0082 was found via the Kolmogorov-Smirnov test.

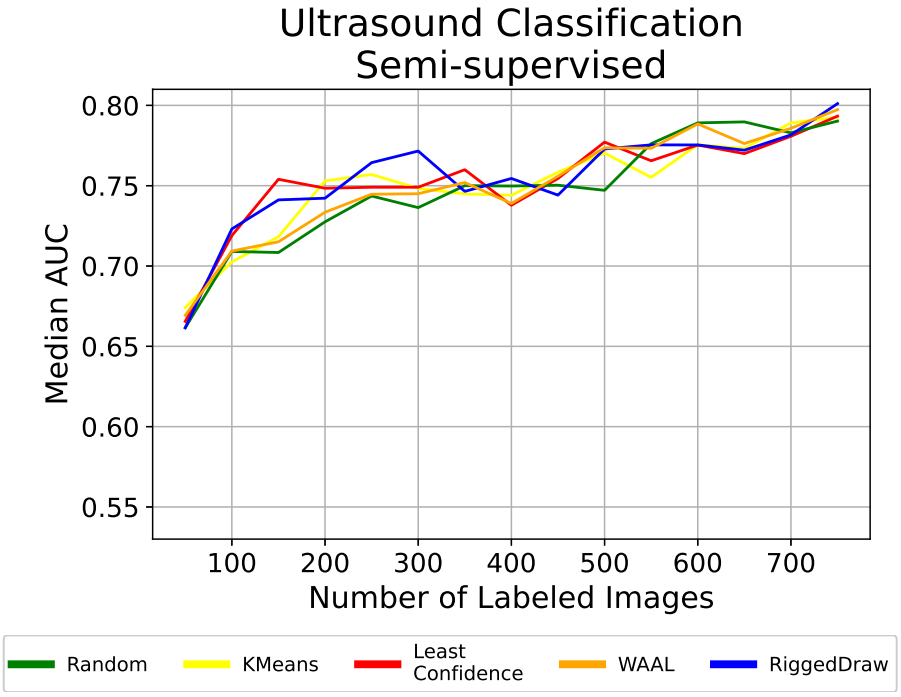

Figure 4.7 – Median AUC values for different active learning strategies with semisupervised learning on the ultrasound dataset.

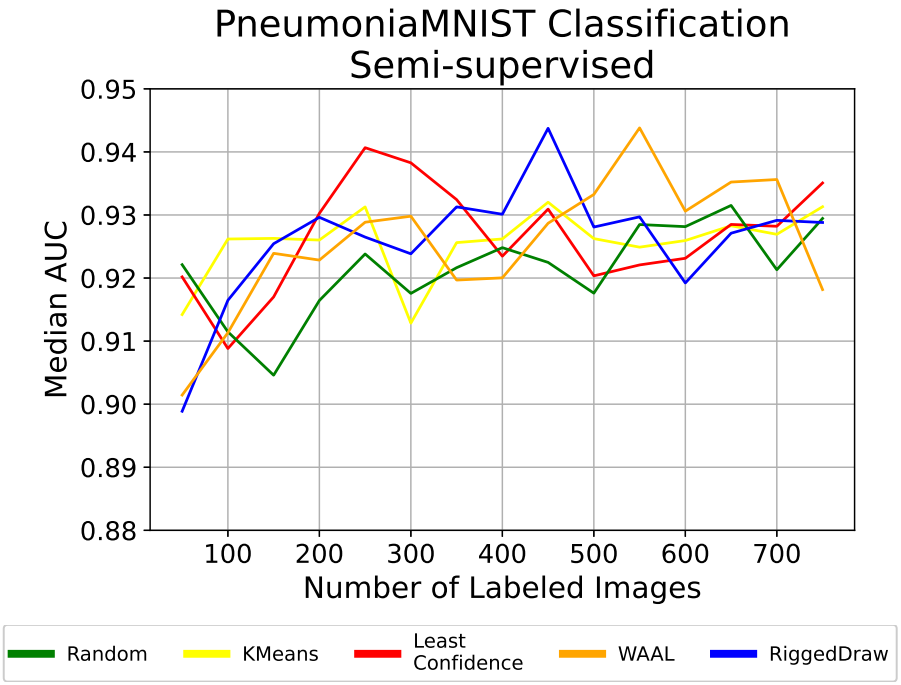

Figure 4.8 – Median AUC values for different active learning strategies with semisupervised learning on the PneumoniaMNIST dataset.

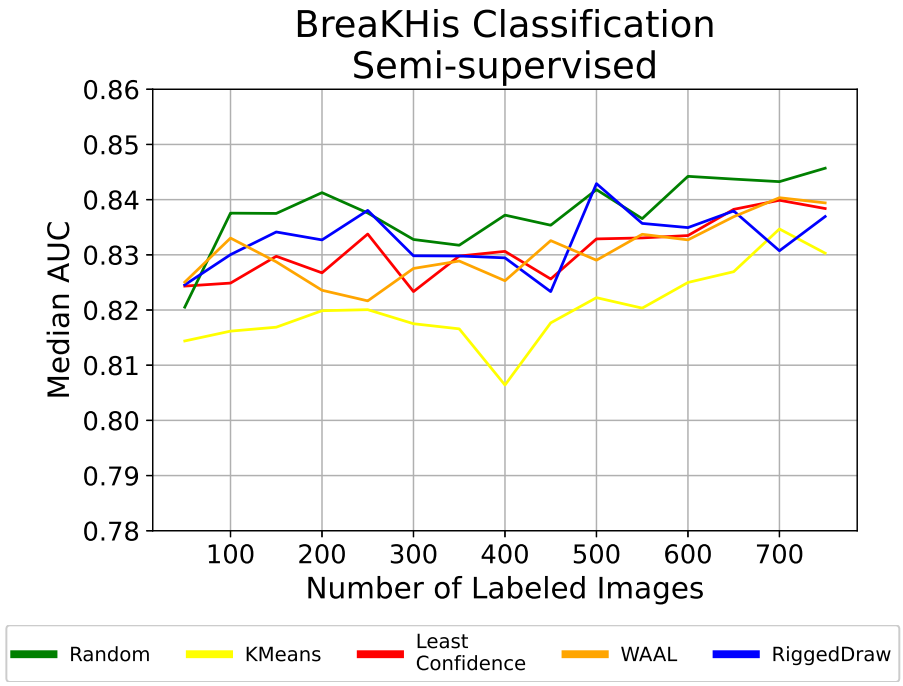

Figure 4.9 – Median AUC values for different active learning strategies with semisupervised learning on the BreaKHis dataset.

Performance for the rigged draw strategy improved substantially for the ultrasound images using the semi-supervised approach, but were not substantially different in terms of magnitude from random selection (see Table 4.5). In addition, for the PneumoniaMNIST and BreaKHis datasets, high AUC values were reached with very few images, and thus no meaningful differences could be observed between strategies (see Figures 4.8 and 4.9).

Dataset	Measure	Rand	LC	KM	WAAL	RD
US Dataset	Mean	0.747	0.751	0.749	0.751	0.754*
	Median	0.748	0.752	0.750	0.751	0.755
	STD	0.009	0.008	0.009	0.008	0.007
Pneumonia MNIST	Mean	0.918	0.923*	0.923	0.923	0.924*
	Median	0.919	0.925	0.922	0.924	0.923
	STD	0.008	0.004	0.008	0.006	0.006
BreaKHis	Mean	0.836	0.828	0.823	0.831	0.833
	Median	0.841	0.830	0.820	0.830	0.834
	STD	0.017	0.026	0.022	0.017	0.018

Table 4.5: Semi-supervised learning AUBC values. Values closer to 1 indicate a more effective strategy, with * indicating p-values < 0.05 when compared to random selection. Rand = Random. LC = Least Certain. KM = KMeans. RD = Rigged Draw (ours) with $\alpha = 25$.

To compare the rigged draw strategy with different values of the weight parameter α, AUBC values for semi-supervised learning are given in Table 4.6. When the AUBC values from the repeated trials with the rigged draw strategy were compared to random selection, significant differences were found for $\alpha = 5$, $\alpha = 25$, and $\alpha = 50$ for the ultrasound dataset, and for $\alpha = 25$ on the PneumoniaMNIST dataset.

Dataset	**Measure**	$\alpha = 5$	$\alpha = 25$	$\alpha = 50$
US Dataset	Mean	0.754*	0.754*	0.751*
	Median	0.755	0.755	0.752
	STD	0.008	0.007	0.012
Pneumonia MNIST	Mean	0.923	0.924*	0.924
	Median	0.923	0.923	0.924
	STD	0.009	0.006	0.006
BreaKHis	Mean	0.823	0.833	0.834
	Median	0.820	0.835	0.830
	STD	0.022	0.018	0.017

Table 4.6: Semi-supervised learning AUBC values for different Rigged Draw weights, with * indicating p-values < 0.05 when compared to random selection.

4.4.3 Initial Set Impact

For both the supervised and semi-supervised tests, there was considerable variation between repeated trials that used different initial sets. To examine this variation at different budget sizes, violin plots of the distribution of AUC values for both supervised and semi-supervised tests are shown for the ultrasound dataset in Figure 4.10, for the PneumoniaMNIST dataset in Figure 4.11, and for the BreakHis dataset in Figure 4.12.

Beginning with the starting random set of 50 images, it is evident that the AUC values varied substantially, particularly for the ultrasound and BreaKHis datasets. In the case of the ultrasound images in Figure 4.10, the mere fact of using the WAAL network architecture for the semi-supervised approach vs. the purely supervised ResNet created a substantial difference even before the effects of active learning could be applied; this difference was perpetuated throughout the active learning steps. There was an overall improvement in the distributions of AUC values for both supervised and semi-supervised results as the sample budget increased for the ultrasound dataset.

For the BreaKHis dataset, a similar difference between supervised and semi-supervised AUC values was seen at the initial random set of 50 images, but then faded away as additional labeled samples were added (see Figure 4.12). For both the PneumoniaMNIST and BreaKHis datasets, there was not much improvement in AUC after 300 images were labeled.

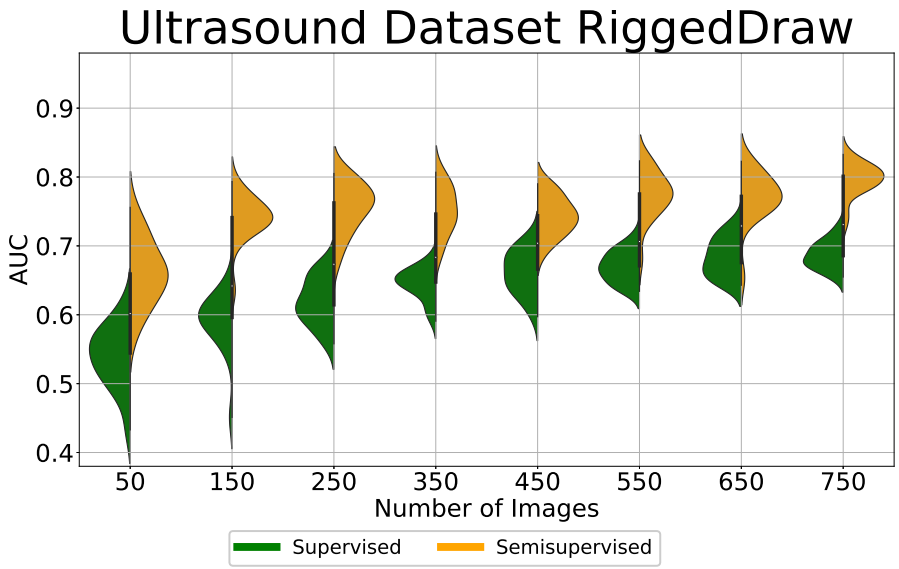

Figure 4.10 – Violin plots of classification AUC values on the at different label budgets with the rigged draw strategy at $\alpha = 25$ on the ultrasound dataset.

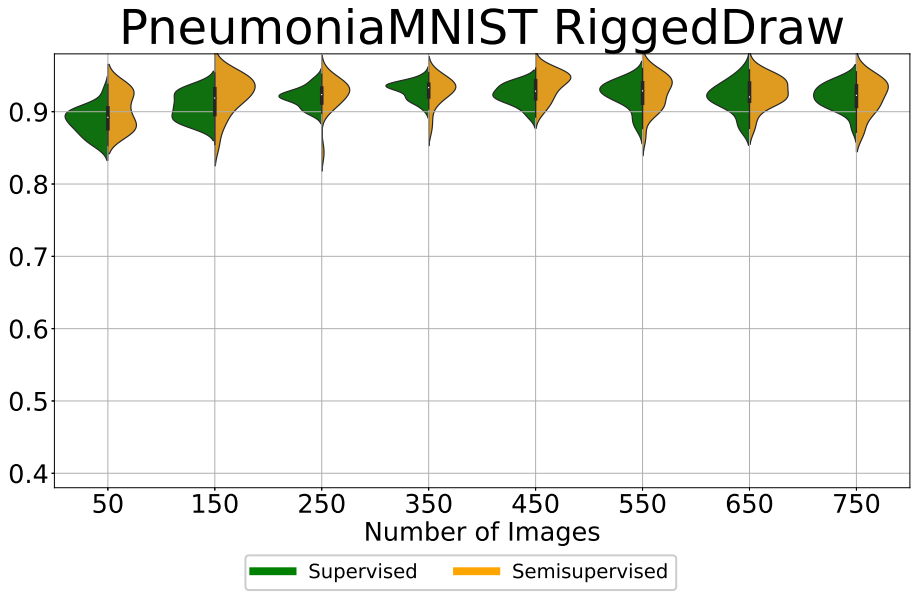

Figure 4.11 – Violin plots of classification AUC values on the at different label budgets with the rigged draw strategy at $\alpha = 25$ on the PneumoniaMNIST dataset.

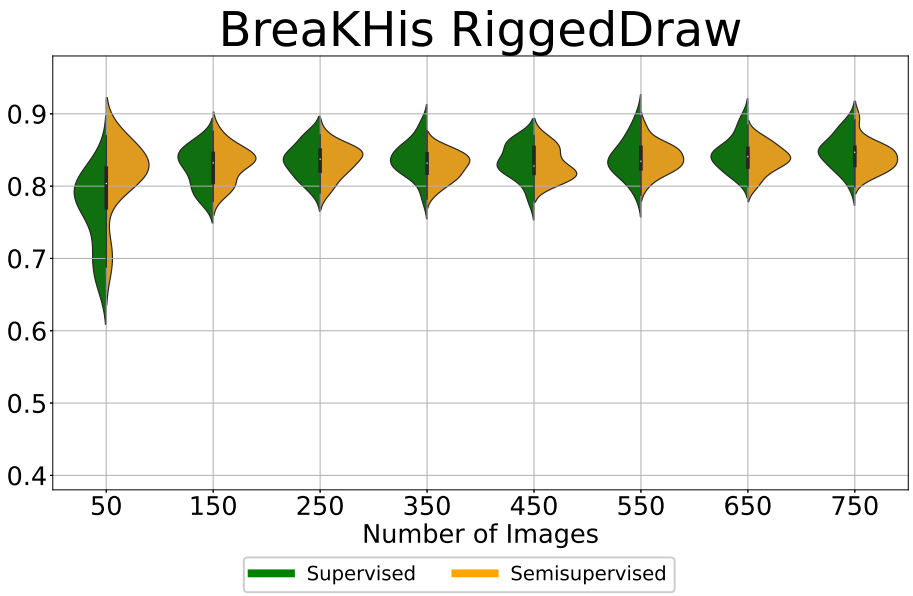

Figure 4.12 – Violin plots of classification AUC values on the at different label budgets with the rigged draw strategy at $\alpha = 25$ on the BreaKHis dataset.

4.5 Discussion

Overall, the results using supervised learning did not show a significant advantage for any active learning strategy compared to random selection on any of the datasets. In addition, classification performance on the ultrasound dataset was poorer than for the others; AUC improvement on the external datasets began to reach a plateau with budgets of only around 300 out of the total 750 images. This difference could be due to limitations inherent to the non expert annotations or the complexity of the classification task. It could also be related to the differences between our clinical ultrasound images and the public dataset images from different imaging modalities.

Performance on the ultrasound dataset was greatly improved, however, by a semi-supervised approach to learn a feature representation to reduce the distance between labeled and unlabeled images. Better results than were possible with the supervised network were attained with only 150 out of the total 750 images. This suggests that some degree of semi-supervised learning is preferable for training on image sets like ours; in an active learning scenario it makes prudent use of unlabeled data for which annotations are expensive.

The semi-supervised approach also showed a statistically significant advantage for the rigged draw strategy over random selection. This was not true of any of the other strategies tested on ultrasound data. However, these observations must be tempered by the fact that differences in AUC were present even from the initial random set of 50 images. In addition, the magnitude of the differences in classification AUC remained minimal, especially in light of the variability within each strategy. This is particularly important as we did test many repetitions of each strategy to compensate for the effects of different starting sets, unlike other comparisons which have used as few as 3 or 5 repetitions (Zhan et al., 2022 ; Gaillochet et al., 2023 ; Shui et al., 2020). In light of the standard deviation of AUBC values as well as the range of AUC values at individual budget

sizes, the impact of active learning on ultrasound data at this scale is unlikely to be clinically relevant.

4.5.1 Limitations

It should be acknowledged that using non-expert annotations could have contributed to poor performance on our dataset. The performance on ultrasound data was worse than for the other two image datasets. More specialized networks or pre-training on ultrasound images could also improve overall performance; however, this would not necessarily increase the relative advantage of active learning strategies.

It is also possible that rigorous optimization of the rigged draw strategy (such as the weight or the percentile for normalization) and of the annotation budget per round could also have improved active learning results specifically; however, the need to fine-tune strategies to this extent further suggests that they would not be suitable for real clinical thyroid ultrasound applications.

Furthermore, the more clinically-relevant tasks of nodule segmentation and characterization were not tested. This was necessary due to the limitations of the non-expert reader. The dataset with evaluations from the four experts from earlier chapters was too small to allow for many active learning steps and retain a useful test set. A more thorough evaluation would need to test strategies on these tasks as well.

4.6 Conclusion

The goal of this chapter was to explore the utility of active learning techniques to train machine learning algorithms for thyroid nodule ultrasound evaluation. The time required for annotation, as well as the inter-expert variability seen in Chapter 2 which reinforces the need for multi-reader consensus, make these approaches conceptually intriguing. We explored only a single task, nodule detection, using non-expert annotations using active learning cycles suitable to thyroid ultrasound evaluation. However, the results on this dataset suggest limitations for the clinical applicability of this technique. This was confirmed by lackluster results for active learning on two other medical image datasets.

In the clinical context of thyroid ultrasound, fine-tuning of an algorithm with active learning would happen once. In order to be sure that active learning was more useful than random selection, the advantage observed would have to be robust over many random initial sets, not merely a subtle difference in the mean AUC values over multiple repetitions. Therefore, at the scale of a thyroid ultrasound dataset from one hospital, the benefits of existing active learning strategies appear to be limited. Semi-supervised approaches, and strategies like rigged draw that harness the power of random selection increase effectiveness; however, further refinement will be necessary to meaningfully reduce annotation burden. Future practical implementation will only be possible with more robust versions of these active learning tools that work consistently in a real hospital setting.

CHAPTER 5

A Machine Learning Strategy for Nonlinear Parameter Estimation

Quantitative ultrasound techniques hold promise as methods for standardizing the evaluation of lesions such as thyroid nodules. One promising target of these methods is the nonlinear parameter $\frac{B}{A}$, which could be used to detect structural changes in tissue. However, the measurement of $\frac{B}{A}$ values in vivo is complicated by its interdependence with the attenuation characteristics of the tissue, the effects of the unknown scatterer distribution, and the diffraction effects while using an ultrasound probe. Compensating for these effects analytically is difficult, and while machine learning methods have been applied to quantitative ultrasound, they work best for narrowly-defined tasks. Therefore, we present a preliminary strategy for nonlinear parameter estimation in simulated tissue-like media by combining a pulse division method for radiofrequency (RF) signal acquisition and processing with a neural network trained to account for these effects when using a specific probe and pulse sequence. The results indicate that this preliminary strategy could be a stepping stone toward more practical estimation of the nonlinear parameter in vivo.

5.1 Introduction 113
5.1.1 Quantitative Ultrasound 113
5.1.2 Applying Machine Learning for Nonlinear Parameter Estimation 114
5.2 Background 115
5.2.1 Ultrasound Wave Propagation 115
5.2.1.1 RF Data Acquisition *In Vivo* 116
5.2.2 The Attenuation Coefficient α 116
5.2.2.1 Effects of Attenuation on Estimation of the Nonlinear Parameter 117
5.2.3 The Nonlinear Parameter $\frac{B}{A}$ 117
5.2.3.1 Generation of Harmonic Waves 118
5.2.3.2 Biological Values of $\frac{B}{A}$ 118
5.2.3.3 Characterization of $\frac{B}{A}$ 119
5.3 Methods 120
5.3.1 Pulse Division Method for $\frac{B}{A}$ Estimation 120
5.3.1.1 Second Harmonic Signal 121
5.3.1.2 Pulse Division 122
5.3.1.3 Limitations of the Pulse Division Strategy 123
5.3.2 Simulation 123
5.3.2.1 Virtual Probe 124
5.3.2.2 Probe Calibration 124
5.3.2.3 Pulse Definition 125
5.3.2.4 Simulated Tissue-Like Media 127
5.3.3 Nonlinear Parameter Estimation with a Neural Network . . 128
5.4 Results 132
5.5 Discussion 134
5.5.1 Limitations 137
5.6 Conclusion 137

5.1 Introduction

In Chapters 2 and 3, we have seen the limitations of thyroid nodule analysis on static B-mode images. The EU-TIRADS, ACR-TIRADS, and other TIRADS, while no doubt useful for standardizing clinical practice, rely on subjective labels (Tessler et al., 2017 ; Russ et al., 2017), leading to inconsistencies as identification of these features differs greatly between readers (Solymosi et al., 2023 ; Grani et al., 2018). This was evident in Chapter 2, as the four experts participating in the study differed substantially in their descriptions of nodule composition, echogenicity, shape, and the presence of echogenic foci, with important consequences for their EU-TIRADS scoring of nodules.

This variability can be attributed to the fact that the acquisition and interpretation of B-mode images are inherently operator dependent. As seen in Chapter 3, even comparative measures such as the relative echogenicity of a nodule versus nearby thyroid parenchyma or muscle do not always provide an objective standard for human experts. Unsurprisingly, this limitation has lead to an interest in direct measurement of the histological differences that the TIRADS feature labels indirectly describe.

For example, a malignant nodule that substantially deviates from the normal follicular architecture of the nodule would also have different material properties, such as density, elasticity, or viscosity. Solid nodules might be denser than spongiform lesions composed of microcystic spaces. The stiffness of a region of thyroid parenchyma might also vary between clusters of malignant cells and normal colloid-filled follicles. Some of these material properties of tissue can and have been measured with quantitative ultrasound techniques as a means to detect pathologic tissue changes.

5.1.1 Quantitative Ultrasound

Among the available quantitative ultrasound techniques, one of the most familiar to thyroid ultrasound practitioners is elastography, which assesses the stiffness of tissues by measuring their response to stress applied by the operator or the acoustic radiation force impulse from an ultrasound pulse (Mena et al., 2023). Multiple elastography techniques are currently available on commercial ultrasound systems, and studies have attempted to show their potential to distinguish between benign and malignant nodules (Shingare, Maldar, Chauhan, & Wadhwani, 2023 ; Mena et al., 2023 ; Ma et al., 2023). However, at the time of publication of ACR-TIRADS and EU-TIRADS, the guideline committees determined that the diagnostic efficacy of these techniques was mixed, and therefore tissue stiffness measurements were not suitable for inclusion as a formal basis for nodule evaluation (Tessler et al., 2017 ; Russ et al., 2017).

Other quantifiable tissue properties, however, could prove more useful for the standardization of thyroid nodule analysis. For example, the speed of sound in a medium depends on its stiffness and density, and can be estimated from the raw radiofrequency (RF) data acquired from an ultrasound probe. Mapping subtle variations in speed of sound arising from local variation in tissue structure has been accomplished, e.g. through beam focusing and spatial coherence methods (Yamaguchi, 2021). Speed of sound mapping has also been tested as a means for the quantification of liver fibrosis (Boozari et al., 2010), but has yet to have a significant clinical impact.

Another category of quantitative ultrasound techniques seeks to characterize the sources of scattering in a tissue. The presence of acoustic scatterers, smaller than the incident wavelength, has an important impact on ultrasound imaging, and generates much of the signal that returns to the probe (Zhou et al., 2024). Since the properties of these scatterers are related to the structure of

tissue at a fine level, the effective medium theory combined with the polydisperse structure factor model has been used for applications such as the analysis of erythrocyte aggregation (de Monchy et al., 2018).

In addition to being scattered, acoustic waves also interact with tissues via absorption, resulting in the diminution of their amplitude. This phenomenon of attenuation is often modeled using power law relationships with frequency that use two tissue-specific parameters: a prefactor attenuation coefficient and the exponent determining the order of the power law (Brandner, Cai, Foiret, Ferrara, & Zagar, 2021 ; Treeby & Cox, 2010). Attenuation is also important for thyroid ultrasound imaging because it must be corrected for with time-gain compensation in order to visualize deeper regions of the image (Abu-Zidan, Hefny, & Corr, 2011). It is also the target of quantitative ultrasound techniques such as spectral difference methods, where it has been used to evaluate hepatic steatosis (Jeon, Lee, & Joo, 2021).

Another tissue property that can be quantified with ultrasound is the nonlinear parameter $\frac{B}{A}$, which determines the amount of nonlinear propagation within a medium. This value has been characterized in multiple biological fluids and tissues as a parameter of potential medical relevance (Bjørnø, 1986 ; Panfilova, van Sloun, Wijkstra, Sapozhnikov, & Mischi, 2021). In most of these cases, the techniques used for this purpose have relied on *ex vivo* laboratory measurements of ultrasound signals transmitted through a sample and measured on the other side (Panfilova et al., 2021). The values of this parameter have been shown to be different between pathologic and healthy liver (Sehgal, Brown, Bahn, & Greenleaf, 1986); this presents a motivation for the use of the nonlinear parameter to detect disease. The incompatibility of current experimental techniques with *in vivo* analysis are therefore unfortunate, because viable methods for the characterization of $\frac{B}{A}$ could provide another quantitative ultrasound modality to someday contribute to more reliable nodule analysis.

Estimating nonlinearity in tissue is therefore an important target for quantitative ultrasound, albeit one that has been relatively unexplored. One of the reasons for this neglect is that most methods of assessing nonlinearity depend on the detection of harmonic signals generated by nonlinear propagation; these signals are weak and strongly affected by attenuation. When adding to this complexity the random distribution of scatterers in the tissue and the diffraction effects from the probe, the difficulties involved in quantifying nonlinear propagation become apparent. Addressing these challenges requires analysis techniques that can learn relevant relationships from intricate signals while accounting for the physical processes affecting ultrasound wave propagation.

5.1.2 Applying Machine Learning for Nonlinear Parameter Estimation

Machine learning methods have been successfully applied to learn from raw RF signals so as to create new beamforming methods, improve B-mode image quality, and perform ultrasound localization microscopy without the need for mathematical models of propagation (Luijten, Chennakeshava, Eldar, Mischi, & van Sloun, 2023). In fact, analysis of RF data in conjunction with thyroid ultrasound images has already been attempted as a means of effecting benign-malignant thyroid classification (Z. Liu et al., 2021). Machine learning combined with physics-based principles could hold promise to estimate tissue properties like $\frac{B}{A}$ which are currently out of reach with classical approaches.

There are many ways to incorporate domain knowledge from physics into machine learning strategies (Karniadakis et al., 2021). Some such strategies use physical laws as loss functions

to guide the training of the algorithm. The samples used for training can also be specifically generated in order to represent the important physical features in the system. Finally, the input data can be pre-processed to make relevant relationships easier to learn, i.e. by compensating for other variables known to interfere with the quantity of interest.

Here, we present a strategy for the estimation of the mean nonlinear parameter of tissue-like media by combining machine learning with a signal acquisition and processing strategy that uses insights from ultrasound physics to highlight the effects of the nonlinear parameter along a single RF data line, reduce the impact of random scatterer distributions, correct for diffraction effects from the probe, and compensate for the influence of attenuation on harmonic generation. This represents a preliminary step towards the implementation of practical techniques for the *in vivo* characterization of these parameters in thyroid nodules.

5.2 Background

To begin, we must examine the physics of ultrasound wave propagation in tissue. This involves first understanding the different physical models used to describe acoustic wave propagation, and the phenomena of attenuation, nonlinear propagation, and diffraction that they take into account. We then discuss the influence of scatterers and the generation of RF data., and explore the parameters used to describe attenuation and nonlinear propagation. Finally, we examine the means of measuring $\frac{B}{A}$ in tissue.

5.2.1 Ultrasound Wave Propagation

The most well-known description of acoustic wave propagation is the linear and lossless wave equation:

$$\frac{\partial^2 p}{\partial t^2} = c_0^2 \Delta p, \tag{5.1}$$

in which p is the local pressure variation from baseline, c_0 is the local speed of sound, and t is time. While this is important starting point for ultrasound physics, this approximation neglects many phenomena relevant to tissue imaging that are captured by other models (Garrett, 2020 ; Panfilova et al., 2021). The Westervelt equation, for example, also accounts for nonlinear propagation (Westervelt, 1963), and can be written as:

$$\nabla^2 p - \frac{1}{c_0^2}\frac{\partial^2 p}{\partial t^2} = -\frac{\beta}{\rho_0 c_0^4}\frac{\partial^2 p^2}{\partial t^2}, \tag{5.2}$$

in which ρ_0 is the equilibrium density and the second-order pressure term $\frac{\partial^2 p^2}{\partial t^2}$ represents nonlinear propagation. In the Westervelt equation, $\frac{\partial^2 p^2}{\partial t^2}$ is multiplied by the nonlinear coefficient β, which describes the relative importance of second-order nonlinearity in the medium. The nonlinear coefficient can be defined as $= 1 + \frac{B}{2A}$, where $\frac{B}{A}$ is the nonlinear parameter. It is this latter term that we seek to predict in order to characterize the nonlinearity of tissue.

The Westervelt equation has been later extended to consider losses in a viscous fluid (Tjotta & Tjotta, 1981). This equation can be written as:

$$\nabla^2 p - \frac{1}{c_0^2}\frac{\partial^2 p}{\partial t^2} + \frac{\delta}{c_0^4}\frac{\partial^3 p}{\partial t^3} = -\frac{\beta}{\rho_0 c_0^4}\frac{\partial^2 p^2}{\partial t^2}, \tag{5.3}$$

with the new term including the sound diffusivity $\delta = \frac{2c_0^3 a}{\omega^2}$ to describe the loss of intensity of the waves. This is expressed in terms of an attenuation factor a expressed in Nepers that increases with the square of the angular frequency of the wave ω. In soft tissue, by contrast, the dependence of attenuation on frequency has been measured as being nearly linear (Goss, Frizzell, & Dunn, 1979).

Finally, the Khokhlov-Zabolotskaya-Kuznetsov (KZK) equation presents another representation that accounts for both nonlinear propagation and attenuation in very similar terms (Rozanova, 2007). It can be written in terms of pressure as:

$$\frac{\partial^2 p}{\partial z \partial \tau} = \frac{c_0}{2}\Delta_\perp p + \frac{\delta}{2c_0^3}\frac{\partial^3 p}{\partial \tau^3} + \frac{\beta}{2\rho_0 c_0^3}\frac{\partial^2 p^2}{\partial \tau^2}, \tag{5.4}$$

with $\tau = t - \frac{z}{c_0}$ being a retarded time variable re-centered around the moment the wavefront reaches each point in the direction of propagation, and $\Delta_\perp$ referring to the transverse Laplacian in the directions orthogonal to the axis of propagation (i.e. $\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$) accounting for diffraction effects that occur when the wave is not a perfect plane wave, as is the case for ultrasound transducer arrays (Kuc & Regula, 1984).

Other models, such as the well-known Burgers' equation (which presumes plane waves and does not account for diffraction), also exist, but the KZK has become the state-of-the-art. Examining these equations serves to illustrate the impact that attenuation, nonlinear propagation, and diffraction have on pulses transmitted into tissue and the RF signals received by ultrasound probes.

5.2.1.1 RF Data Acquisition *In Vivo*

Indeed, when RF data is acquired in patients with a real transducer, diffraction affects the signal amplitude as a function of frequency and distance from the probe (Kuc & Regula, 1984). This can make it more difficult to measure the properties of the medium, such as attenuation and nonlinear propagation, based on the signal that returns to the probe.

Scattering, which generates the signal returning to the ultrasound probe, can also impact RF data analysis. Scattering is local and tissue-dependent, arising when the ultrasound waves interact with structure of different acoustic impedance smaller than the ultrasound wavelength. It therefore also has an unknown spatial distribution which, in ultrasound imaging, is convolved with the effects of attenuation and nonlinear propagation, thereby further complicating the measurement of both the attenuation coefficient and of $\frac{B}{A}$. In addition, though the effect of scattering is often assumed not to vary within the bandwidth of an imaging probe, in principle this effect is also frequency-dependent.

5.2.2 The Attenuation Coefficient α

Attenuation of the signal amplitude affects both the forward-propagating and backscattered signals. As described in the Westervelt equation with losses and the KZK equation, the magnitude of the loss in amplitude depends on both local tissue properties as well as the frequency of the ultrasound waves. This is generally modeled as a power law relationship dependent on frequency (Brandner et al., 2021 ; Treeby & Cox, 2010), of the form

$$\frac{\mathcal{P}(z)}{\mathcal{P}(0)} = e^{-\alpha f^\epsilon z}, \tag{5.5}$$

where $\mathcal{P}(0)$ is the initial pressure amplitude of the wave, $\mathcal{P}(z)$ is the amplitude after propagating a distance z into the medium, f is the frequency of the wave, and αf^{ϵ} is the attenuation in $\mathrm{Np} \cdot \mathrm{cm}^{-1}$. In soft tissue, the frequency dependence of attenuation has been measured as being nearly linear (Goss et al., 1979), or:

$$\frac{\mathcal{P}(z)}{\mathcal{P}(0)} = e^{-\alpha f z}, \tag{5.6}$$

with α being expressed in units of $\mathrm{Np} \cdot \mathrm{MHz}^{-1} \cdot \mathrm{cm}^{-1}$.

5.2.2.1 Effects of Attenuation on Estimation of the Nonlinear Parameter

Because most techniques for the estimation of $\frac{B}{A}$ depend on the amplitude of the second harmonic waves at the frequency $2f_0$, the effects of attenuation are very important to the measurement of the nonlinear parameter. To begin with, the harmonic waves will be more severely attenuated than waves at the fundamental frequency f_0, making it more difficult to detect this signal. Furthermore, as harmonic waves arise progressively through nonlinear propagation of the fundamental pulse, they are simultaneously dissipated by attenuation. The amplitude spectrum of the harmonic pulse thus depends strongly on the values of both α and $\frac{B}{A}$ in the medium.

In addition, the fundamental pulse which generates the harmonic is also progressively attenuated as it travels into the tissue, leading to reduced cumulative harmonic generation. This means that the influence of attenuation on the manifestations of nonlinear propagation cannot be easily disentangled. Therefore, techniques for the estimation of $\frac{B}{A}$ should be robust to different attenuation characteristics in the medium.

5.2.3 The Nonlinear Parameter $\frac{B}{A}$

The definition of the nonlinear parameter $\frac{B}{A}$ comes from the Taylor series expansion of the adiabatic state equation relating the pressure and density of the propagation medium (Beyer, 1960 ; Panfilova et al., 2021):

$$p = A\left(\frac{\rho - \rho_0}{\rho_0}\right) + \frac{B}{2}\left(\frac{\rho - \rho_0}{\rho_0}\right)^2 + \ldots \tag{5.7}$$

in which

$$p = P - P_0 \tag{5.8}$$

where p represents the local pressure variation of the absolute pressure P from the baseline pressure P_0, ρ corresponds to the density of the medium, and the subscripts $0, s$ signify that the partial derivatives are evaluated at the equilibrium density $\rho = \rho_0$ and with constant entropy. The terms A and B are

$$A = \rho_0 \left(\frac{\partial P}{\partial \rho}\right)_{0,s} = \rho_0 c_0^2 \tag{5.9}$$

and

$$B = \rho_0^2 \left(\frac{\partial^2 P}{\partial \rho^2}\right)_{0,s}, \tag{5.10}$$

such that the ratio of the second order term to the first can be expressed as:

$$\frac{B}{A} = \frac{\rho_0}{c_0^2}\left(\frac{\partial c^2}{\partial \rho}\right)_{0,s} = 2\rho_0 c_0 \left(\frac{\partial c}{\partial P}\right)_{0,s} \tag{5.11}$$

with details provided in Appendix I.

5.2.3.1 Generation of Harmonic Waves

In terms of ultrasound wave propagation, the value of $\frac{B}{A}$ has important repercussions on the frequency spectrum of the ultrasound pulse. This is attested to by the following relationship which may be derived from the Taylor series expansion in Equation 5.7 (Beyer, 1973 ; Panfilova et al., 2021):

$$c \approx c_0 \left[1 + \frac{p}{\rho_0 c_0^2}\left(1 + \frac{B}{2A}\right)\right], \tag{5.12}$$

in which c is the local speed of sound and c_0 is the equilibrium speed of sound in the medium. This relationship indicates that the speed of sound is greater during compression than during rarefaction (Beyer, 1973). This means that the positive part of the pressure wave travels faster than the negative part, creating a distortion. As a result, the propagation of a wave with an initial center frequency f_0 will lead to the accumulation of harmonic components, i.e. at frequencies of nf_0 for positive integer values of n (Garrett, 2020).

In practice, only the second harmonic, at $2f_0$, is detectable, owing to the limited fraction of the wave energy being converted to higher-order harmonics, the increased attenuation at higher frequencies, and the limited bandwidth of ultrasound transducers (Garrett, 2020). This generation of waves at the second harmonic frequency has been used as the basis for multiple techniques for $\frac{B}{A}$ characterization of biological tissues (Fujii, Taniguchi, Akiyama, Tsao, & Itoh, 2004 ; Panfilova et al., 2021).

5.2.3.2 Biological Values of $\frac{B}{A}$

Beginning with pure water, the value of $\frac{B}{A}$ has been described in the range 5.1-5.2 (Dunn, Law, & Frizzell, 1981 ; Davies, Tapson, & Mortimer, 2000). Aqueous solutions of bovine serum albumin and hemoglobin were measured by Dunn et al. to have values of $\frac{B}{A}$ increasing linearly with concentration (Dunn et al., 1981). Porcine blood was measured by the same group as having a similar nonlinear parameter as a hemoglobin solutions of a similar concentration by dry weight, at approximately 6.3 (Dunn, Law, & Frizzell, 1982). Among solid tissues, $\frac{B}{A}$ values are consistently higher than for liquids (Panfilova et al., 2021). The most commonly studied organ in this regard appears to be the liver (Panfilova et al., 2021). Sehgal et al. measured healthy human livers as having $\frac{B}{A}$ values in the range of 6.5-7.3 (Sehgal et al., 1986). In cases of hepatic steatosis, with increased fat content in the liver, the same group found increased B/A values, ranging from 7.1 to about 8.8 (Sehgal et al., 1986).

Medium	$\frac{B}{A}$ Value(s)	Source(s)
Water (30°C)	5.1-5.2	(Dunn et al., 1981 ; Davies et al., 2000)
Porcine Blood (30°C)	6.3	(Dunn et al., 1982)
Normal Human Liver (30°C)	6.5-7.1	(Sehgal et al., 1986)
Fatty Human Liver (30°C)	7.1-8.8	(Sehgal et al., 1986)

Table 5.1: Values of the nonlinear parameter in various biological media reported in the literature. The values of $\frac{B}{A}$ represent the range of values in the cited sources rounded to the nearest tenth.

5.2.3.3 Characterization of $\frac{B}{A}$

Among the laboratory approaches for characterization of the nonlinear parameter, one popular category is referred to as finite element insert substitution, and consists of measuring the second harmonic signal generated by a given pulse through a medium of known $\frac{B}{A}$ (e.g. water), and comparing the difference in the signals when the reference medium is replaced by a known thickness of the sample material (Panfilova et al., 2021). As this typically involves measuring a signal transmitted through the tissue sample, such a method is not suited to clinical ultrasound. The echo-mode versions of this technique measuring a returning signal depend upon a reflective plate placed behind the sample (Panfilova et al., 2021). The thermodynamic method detects the change in the speed of sound with an isentropic change of pressure (Panfilova et al., 2021); despite its precision, however, its elaborate experimental setup also precludes its usefulness *in vivo* (Panfilova et al., 2021).

Only a few methods have been proposed that are compatible with *in vivo* ultrasound. Among these is the example of Fujii et al., who studied patients with healthy and fatty livers. Their estimation of a global parameter incorporating the value of $\frac{B}{A}$ allowed for discrimination between the two groups (Fujii et al., 2004). The key to their method lies in dividing the amplitude spectra of RF signals generated by two pulses transmitted at a frequency f_0 and at twice that frequency in order to compensate for the effects of attenuation and acoustic scattering (Fujii et al., 2004).

However, their work did not directly estimate the value of the nonlinear parameter and instead relied on comparative values of the composite parameter $h = \frac{2\pi f_0\left[\frac{B}{A}+2\right]}{4\rho_0 c_0^3}$ to distinguish between groups of patients (Fujii et al., 2004). While effective, the differences detected were no doubt also influenced by variations in density and speed of sound between health and fatty livers, rather than discriminating purely on the basis of $\frac{B}{A}$ (Fujii et al., 2004). In addition, their technique did not account for the effects of diffraction, which could also influence the relative intensities of their two signals, particularly in regions close to the probe's surface (Fujii et al., 2004 ; Kuc & Regula, 1984).

Toulemonde et al. more recently proposed an local-estimation approach using multitaper coherent plane wave compounding with a technique similar to that of the aforementioned insertion substitution methods by comparing the harmonic pressure to that of a reference medium (Toulemonde, Varray, Bernard, Basset, & Cachard, 2015). Their method allowed for estimating the value of $\frac{B}{A}$ within different regions of a phantom, albeit with substantial errors: regions of $\frac{B}{A}$ of 5 were estimated with a mean value of 5.1, regions of $\frac{B}{A}$ of 7 were estimated with a mean value of 9.6, and those with a $\frac{B}{A}$ value of 10 were estimated with a mean value of 8.5 (Toulemonde et al., 2015). The predictions for fluid-like values of 5 were accurate, but the mean errors for a region of soft-tissue like values of 7 and fat-like values of 10 were so large that they could obscure the difference reported elsewhere between healthy and diseased tissue (Sehgal et al., 1986). Furthermore, while the work utilized simulations with different scatterer densities, the attenuation in the media was assumed to be homogeneous, and multiple α values were not tested.

Therefore, there is still work to be done in accurately estimating the value of $\frac{B}{A}$ of tissues imaged with an ultrasound probe. The approach of Fujii et al., which compensates for some of the effects of attenuation and scattering on the estimation of harmonic generation is a promising starting point. Improving the direct estimation of $\frac{B}{A}$ via this technique in tissue-like media with variable values of the attenuation coefficient and accounting for the diffraction effects of the ultrasound probe will be a first step toward future practical *in vivo* applications.

5.3 Methods

Our goal was to develop a technique to estimate the mean value of the nonlinear parameter $\frac{B}{A}$ in a tissue-like medium from RF data generated from focused wave pulses from an ultrasound probe. We based our strategy on the pulse-division method utilized by Fujii et al. to eliminate the impact of scatterers, but tried to account for the effects of diffraction and attenuation on harmonic generation which are not fully addressed by the pulse division.

For this preliminary investigation, our approach was to process the single, central RF line received by the probe after transmission of focused waves, without beam steering. This allowed for acquisition of a single line of RF data with a simple time-depth relationship, at a lower dimension than for plane wave data, and with a straightforward connection to nonlinear propagation along that axis. The mean estimation from this RF line could suffice in theory to compare between a region of a thyroid lobe with and without a nodule, though future development could add spatial discrimination.

For $\frac{B}{A}$ estimation from this data, the pulse-division strategy of Fujii et al. (described in the next section) can, in theory, compensate for the effects of scatterers and of attenuation on the fundamental frequencies, but the accuracy of this approach for direct nonlinear parameter estimation is limited by the underlying approximations: (1) oversimplification of the influence of α values on harmonic generation, and (2) a coarse assumption about diffraction effects. Accounting analytically for the RF signal amplitude variations due to these medium-based and probe-based functions is indeed challenging.

Therefore, instead of attempting to precisely calculate the impact of the diffraction effects of the probe and the effects of α values on harmonic generation, we opted for a deep learning-based approach to implicitly learn these underlying functions. Neural networks have been used previously as universal function approximators in a variety of medical-imaging related applications (Najjar, 2023). However, the success of general-purpose networks has been limited in this domain, compared to networks trained for a precisely-defined and restricted task. For problems for which physical descriptions of the processes of interest exist, many recent strategies have therefore tried to incorporate this knowledge into a narrower prediction strategy (Karniadakis et al., 2021).

We attempted to learn a narrower and more specific function with a neural network to correct for the diffraction effects and the influence of variations in α on harmonic generation when using a pulse-division strategy to estimate the mean value of $\frac{B}{A}$ from the central RF lines. In order to train this network, we relied on simulations to generate a training dataset that mimics real tissues with variable attenuation coefficients and inclusions using k-Wave, an open-source toolbox that allows for GPU-accelerated nonlinear ultrasound simulation (Treeby, Jaros, Rendell, & Cox, 2012). For these simulations, we also optimized the frequency content of the pulses to prepare them for use with the pulse division strategy.

5.3.1 Pulse Division Method for $\frac{B}{A}$ Estimation

Each point in the RF data corresponding to a particular position is generated by the forward-propagating signal interacting with local scatterers, and both the forward-propagating and backscattered signals are progressively attenuated as they travel through the tissue. Because in the frequency domain, convolution is equivalent to multiplication, this relationship was described by Fujii et al. (Fujii et al., 2004) as:

$$S_{f_0}(z) = \mathcal{P}_{f_0}(0) \cdot e^{-\int_0^z \alpha(f_0,\tilde{z})d\tilde{z}} \cdot \Gamma(f_0, z) \cdot e^{-\int_0^z \alpha(f_0,\tilde{z})d\tilde{z}}, \tag{5.13}$$

with $\mathcal{P}_{f_0}(0)$ being the amplitude of the pulse at frequency f_0 emitted by the probe at depth $z = 0$, $S_{f_0}(z)$ being the Fourier transform of the RF signal at the retarded time corresponding to a depth z. The first occurrence of the term $e^{-\int_0^z \alpha(f_0,\tilde{z})d\tilde{z}}$ represents the cumulative effects of attenuation on the forward path. $\Gamma(f_0, z)$ is the influence of local scatterers at a depth z and frequency f_0. This function determines the amplitude of the back-propagating signal originating that position. The second occurrence of $e^{-\int_0^z \alpha(f_0,\tilde{z})d\tilde{z}}$ represents the cumulative effects of attenuation on the backscattered signal during the return trip.

The attenuation terms in Equation 5.13 that depend on the attenuation coefficient α, (whose impact is expressed here in Nepers) represent a solution to an progressive attenuation relationship of the form $\frac{d\mathcal{P}}{dz} = -\alpha\mathcal{P}$. We also note that the formulation proposed by Fujii et al. in Equation 5.13 neglects depth-dependent diffraction effects from the probe. Given that the signal passes through the same distribution of the attenuation coefficient on the forward and backward trips, the attenuation terms are simplified from Equation 5.13 to give:

$$S_{f_0}(z) = \mathcal{P}_{f_0}(0) \cdot \Gamma(f_0, z) \cdot e^{-2\int_0^z \alpha(f_0,\tilde{z})d\tilde{z}} \tag{5.14}$$

5.3.1.1 Second Harmonic Signal

As previously discussed, the impact of nonlinear propagation on the RF signal is the distorsion of the frequency spectrum of the original pulse leading to the generation of harmonic waves. The accumulation of these harmonic signals depends on the value of $\frac{B}{A}$ in the tissue; in general, however, only a small percentage of the forward-propagating fundamental signal is transferred into harmonics, of which only the second harmonic ($2f_0$) is practically detectable (Garrett, 2020). In addition, because the backscattered pressures are orders of magnitude lower than the incident pressure, and because harmonics are generated a rate that is quadratic with respect to the fundamental pressure (Bjørnø, 1986), harmonic generation on the return path is negligible.

Therefore, we consider only the second harmonic generated by the forward propagating signal and backscattered to the probe. This has been described with the following relationship (Fujii et al., 2004):

$$S_{2f_0,harm}(z) \propto \mathcal{P}_{f_0}(z)^2 \cdot \int_0^z \frac{2\pi f_0 \left[\frac{B}{A}(\tilde{z}) + 2\right]}{4\rho_0 c_0^3} d\tilde{z} \tag{5.15}$$

Considering only the generation of the second harmonic RF signal at $2f_0$ in forward propagation, the harmonic RF signal received by the probe at $2f_0$ after sending out a pulse at f_0 can be described by substituting the expression for the pressure amplitude of the harmonic from Equation 5.15 in place of the fundamental pressure in Equation 5.13. Fujii et al. modified the attenuation terms from Equation 5.13 to apply to the fundamental signal during forward propagation and to the attenuation of the harmonic signal during back propagation to give:

$$\begin{aligned} S_{2f_0,harm}(z) = &\left[\mathcal{P}_{f_0}(0) \cdot e^{-\int_0^z \alpha(f_0,\tilde{z})d\tilde{z}}\right]^2 \cdot \int_0^z \frac{2\pi f_0 \left[\frac{B}{A}(\tilde{z}) + 2\right]}{4\rho_0 c_0^3} d\tilde{z} \\ &\cdot \Gamma(2f_0, z) \cdot e^{-\int_0^z \alpha(2f_0,\tilde{z})d\tilde{z}}, \end{aligned} \tag{5.16}$$

where the pulse signal is attenuated during forward propagation depending on the position-dependent $\alpha(f_0, z)$, the second harmonic is cumulatively generated during forward propagation as a function of $\frac{B}{A}(z)$, the harmonic signal is backscattered following $\Gamma(2f_0, z)$, and attenuated on the return path as a function of $\alpha(2f_0, z)$. Of note, Fujii et al. represent the harmonic being backscattered as soon as it is generated, rather than traveling part of the forward path.

Combining the attenuation terms allows for rewriting Equation 5.16 as:

$$\begin{aligned} S_{2f_0,harm}(z) =& \mathcal{P}_{f_0}(0)^2 \cdot \int_0^z \frac{2\pi f_0 \left[\frac{B}{A}(\tilde{z}) + 2\right]}{4\rho_0 c_0^3} d\tilde{z} \\ & \cdot \Gamma(2f_0, z) \cdot e^{-2\int_0^z \alpha(f_0,\tilde{z})d\tilde{z} - \int_0^z \alpha(2f_0,\tilde{z})d\tilde{z}} \end{aligned} \tag{5.17}$$

5.3.1.2 Pulse Division

Fujii et al. eliminated effects of scattering by dividing the harmonic signal generated from an acoustic pulse at the fundamental frequency f_0 by the fundamental signal of an acoustic pulse generated at the frequency $2f_0$. The latter should have the same pulse length to maintain the bandwidth. They described the RF signal generated in response to the second pulse as:

$$S_{2f_0}(z) = \mathcal{P}_{2f_0}(0) \cdot \Gamma(2f_0, z) \cdot e^{-2\int_0^z \alpha(2f_0,\tilde{z})d\tilde{z}}, \tag{5.18}$$

where $S_{2f_0}(z)$ is the RF signal with a frequency content comparable to that of $S_{2f_0,harm}(z)$, $\mathcal{P}_{2f_0}(0)$ is the amplitude of the pulse sent by the probe centered at the frequency $2f_0$ at $z = 0$, $\Gamma(2f_0, z)$ represents the scatterering at $2f_0$, and $e^{-2\int_0^z \alpha(2f_0,\tilde{z})d\tilde{z}}$ is the round-trip attenuation. This allows for eliminating the scattering term through division of the signal generated through nonlinear propagation by that generated directly at $2f_0$. (Fujii et al., 2004):

$$\frac{S_{2f_0,harm}(z)}{S_{2f_0}(z)} = \frac{\mathcal{P}_{f_0}(0)^2}{\mathcal{P}_{2f_0}(0)} \cdot \frac{\pi f_0}{2\rho_0 c_0^3} \int_0^z \left[\frac{B}{A}(\tilde{z}) + 2\right] d\tilde{z} \cdot e^{\int_0^z \alpha(2f_0,\tilde{z})d\tilde{z} - 2\int_0^z \alpha(f_0,\tilde{z})d\tilde{z}} \tag{5.19}$$

This can be rewritten to isolate $\frac{B}{A}$:

$$\int_0^z \left[\frac{B}{A}(\tilde{z}) + 2\right] d\tilde{z} = \frac{S_{2f_0,harm}(z)}{S_{2f_0}(z)} \cdot \frac{\mathcal{P}_{2f_0}(0)}{\mathcal{P}_{f_0}(0)^2} \cdot \frac{2\rho_0 c_0^3}{\pi f_0} \cdot e^{2\int_0^z \alpha(f_0,\tilde{z}d\tilde{z}) - \int_0^z \alpha(2f_0,\tilde{z})d\tilde{z}} \tag{5.20}$$

Taking the derivative with respect to z yields:

$$\frac{B}{A}(z) = \frac{d}{dz}\left[\frac{S_{2f_0,harm}(z)}{S_{2f_0}(z)} \cdot e^{2\int_0^z \alpha(f_0,\tilde{z}d\tilde{z}) - \int_0^z \alpha(2f_0,\tilde{z})d\tilde{z}}\right] \cdot \frac{2\rho_0 c_0^3}{\pi f_0} \cdot \frac{\mathcal{P}_{2f_0}(0)}{\mathcal{P}_{f_0}(0)^2} - 2 \tag{5.21}$$

This is further simplified with the assumption that the attenuation in soft tissue is nearly linearly dependent on frequency (Fujii et al., 2004), i.e.:

$$\alpha(2f_0, z) = 2\alpha(f_0, z) \tag{5.22}$$

Substituting Equation 5.22 into Equation 5.21 yields:

$$\frac{B}{A}(z) = \frac{d}{dz}\left[\frac{S_{2f_0,harm}(z)}{S_{2f_0}(z)}\right] \cdot \frac{2\rho_0 c_0^3}{\pi f_0} \cdot \frac{\mathcal{P}_{2f_0}(0)}{\mathcal{P}_{f_0}(0)^2} - 2 \tag{5.23}$$

Equation 5.23 reveals the proposed relationship between the value of the nonlinear parameter and the ratio between the RF signals received at the second harmonic frequency and from a signal generated by a pulse at $2f_0$.

5.3.1.3 Limitations of the Pulse Division Strategy

While the description proposed in Equation 5.23 by Fujii et al. eliminates the impact of scattering, some of its approximations limit its utility for $\frac{B}{A}$ estimation from the RF signal ratio. First, position- and frequency-dependent diffraction effects that could distort the RF signal amplitude are neglected (Kuc & Regula, 1984). This would mean that rather than changes in signal amplitude during propagation being governed solely by an attenuation relationship of the form $\frac{\partial \mathcal{P}}{\partial z} = -\alpha \mathcal{P}$, they would instead be described as $\frac{\partial \mathcal{P}}{\partial z} = -\alpha \mathcal{P} + \Theta(z)$, with $\Theta(z)$ being a function representing the diffraction effects. Second, the amplitude evolution of the harmonic, which occurs over time during forward propagation, would also depend on a relationship affected by diffraction effects in addition to progressive generation and attenuation at $2f_0$, with the form $\frac{\partial \mathcal{P}_{2f_0}}{\partial z} = -\alpha \mathcal{P}_{2f_0} + \Theta_{harm}(z) + g(\mathcal{P}_{f_0}^2, z)$, with $g(\mathcal{P}_{f_0}^2, z)$ corresponding to harmonic generation dependent on both depth and the square of the amplitude of the fundamental signal.

This would mean that attenuation terms of the form $e^{-\int_0^z \alpha(f_0, \tilde{z}) d\tilde{z}}$ for the fundamental signal and $e^{-\int_0^z \alpha(2f_0, \tilde{z}) d\tilde{z}}$ at the harmonic would no longer be adequate representations of changes in signal amplitude. These differences would also mean that division would not eliminate attenuation and diffraction effects on the amplitudes of the RF signals S_{2f_0} and $S_{2f_0}(z)$. Therefore, the accuracy of $\frac{B}{A}$ estimation from the derivative of their ratio would be limited even for known values of ρ_0, c_0, and $\frac{\mathcal{P}_{2f_0}(0)}{\mathcal{P}_{f_0}(0)^2}$.

Despite these limitations, the relationship proposed by Fujii et al. in Equation 5.23 provides a useful RF signal processing strategy to highlight the impact of $\frac{B}{A}$. Explicit compensation for the effects of attenuation on harmonic generation and of diffraction on RF signal amplitude is difficult, which is why we combine the pulse division strategy with a neural network trained to implicitly learn an approximate function that accounts for these additional factors. To do this, the network will need to learn from RF data corresponding to a wide range of combinations of α and $\frac{B}{A}$, and adapt to a specific transducer and pulse sequence to account for the values of $\frac{\mathcal{P}_{2f_0}(0)}{\mathcal{P}_{f_0}(0)^2}$ and the diffraction effects. The transducer, pulses, and tissue-like media used for training will therefore need to be carefully defined.

5.3.2 Simulation

In order to generate adequate training data with a realistic transducer and pulses, we used simulations with the k-Wave MATLAB toolbox (Treeby et al., 2012). This toolbox was designed to simulate high-frequency ultrasound in biological media with nonlinear propagation with GPU acceleration (Treeby et al., 2012). It allowed for the simulation of an ultrasound probe with a specific pulse sequence, as well as the definition of tissue-like media whose values of $\frac{B}{A}$ and α were randomized; the central RF data received after pulse transmission and backscattering in 3D therefore contained the information necessary to train the neural network.

5.3.2.1 Virtual Probe

The probe used for the simulations was based on a physical P4-1 ultrasound transducer (Philips, Amsterdam, the Netherlands). Although this 96-element probe has a low-frequency range, being more suitable for echocardiography, it has a well-characterized model developed by Blanken et al. (Blanken et al., 2024). It has an element width of 0.245 mm, a pitch size of 0.295 mm, and elevation length of 16 mm. We tested the model via laboratory comparisons with the real probe in a water bath with a hydrophone as depicted in Figure 5.1. The probe, operated using a Verasonics Vantage 256 system (Verasonics, Kirkland, Washington, USA), was programmed to transmit a focused wave generated from only the center third of its elements, giving an aperture of 9.44 mm. The probe was focused at 40 mm in direction of propagation, and 80 mm in the elevation direction.

Based on the calibration we used a driving voltage of 20V, which represents approximately 1 MPa peak pressure. These pressures allow for significant harmonic generation through nonlinear propagation, while remaining below the ultrasound mechanical index (MI) safety limit (MI ≈ 0.62 at 2MHz, with the upper limit of 1.9 (Kollman et al., 2013) being recommended).

5.3.2.2 Probe Calibration

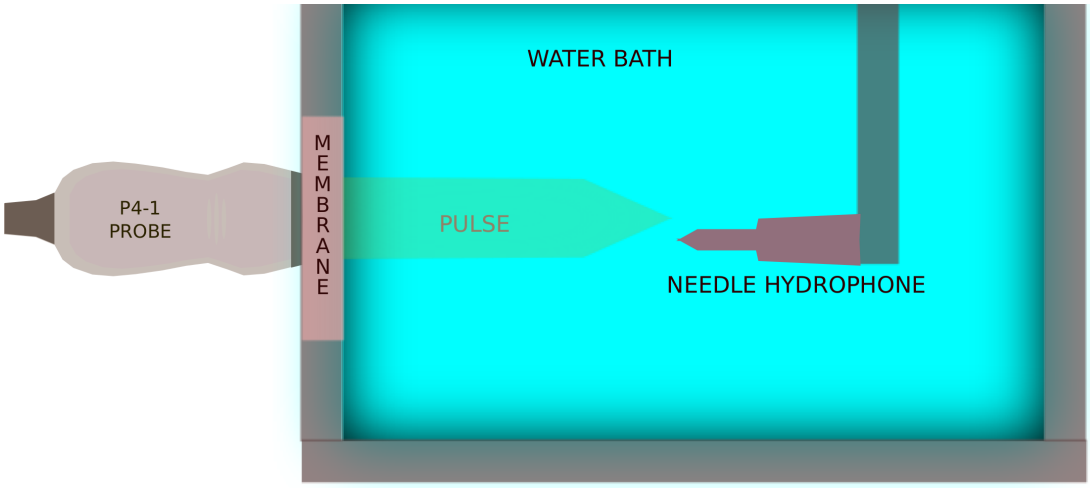

Figure 5.1 – Experimental setup to compare the P4-1 probe to the simulation: a water bath, with a needle hydrophone placed inside. The probe placed on the side of the tank emitted a pulse through a thin membrane into the tank, where it could be measured by the hydrophone at different positions. This generated a map of pressure recordings over time in a plane in front of the probe.

The probe was configured to repeatedly emit a 2 MHz, 2.5 cycle ultrasound pulse, and an optical hydrophone system (Precision Acoustics, UK) was synchronized to move throughout the area and measure the pressure over time by raster scanning on a regular grid with 1 mm spacing. This data was then compared to the virtual probe in k-Wave to verify that the signal propagating forth from the probe both at the fundamental frequency of 2 MHz and at the second harmonic frequency matched the simulations. Pressure at this frequency, calculated from Fourier transforms, are shown for the experimental and simulated data in Figure 5.2. The maximum difference be-

tween the normalized simulation and measured data was at 60 mm: about 18% at the fundamental and about 7% at the harmonic.

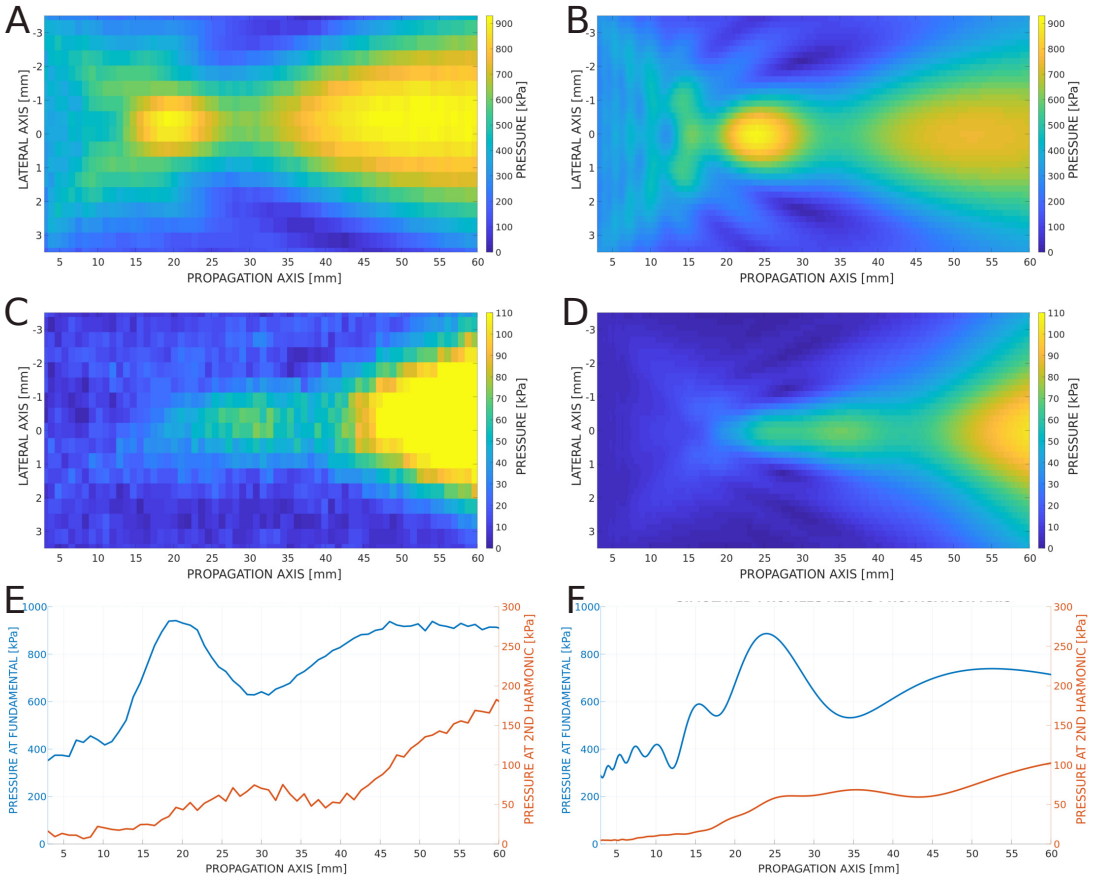

Figure 5.2 – Comparison of between experimental and simulation data with the P4-1 probe. (A) Measured pressure map at the fundamental frequency. (B) Simulated pressure map at the fundamental frequency. (C) Measured pressure map at the second harmonic frequency. D) Simulated pressure map at the second harmonic frequency. (E) Measured fundamental (blue) and second harmonic (red) pressure profiles along the central propagation axis. (F) Simulated fundamental (blue) and second harmonic (red) pressure profiles along the central propagation axis.

This data confirms that simulations with this model could generate data representative of the probe on the transmission of signals at the fundamental f_0 and harmonic $2f_0$ frequencies, including nonlinear propagation. Thus, upon definition of specific transmit pulses, this numerical model can be used to generate synthetic data and let a neural network learn to account for diffraction effects and the probe's beam profile.

5.3.2.3 Pulse Definition

With the probe properly characterized, the next step is to define the pulses to be transmitted by the probe. As in Fujii et al., these must be calibrated such that one would be at twice the frequency of the other, with the same duration to match the bandwidths (Fujii et al., 2004).

The ratio $\frac{S_{2f_0,harm}(z)}{S_{2f_0}(z)}$ could be readily calculated from fundamental and pure harmonic pulse with an infinitely short bandwidth, but in practice the amplitude spectra for these signals have a non-negligeable bandwidth and will vary due to features of the probe and medium. The probe creates a mismatch between the pulse at $2f_0$ and the harmonic pulse generated through nonlinear propagation, since the central frequency of the programmed pulses shift toward the center frequency of the probe upon transmission. This was compensated for by programming pulses waveforms that would attain the desired frequency characteristics after transmission. The properties of the medium, which are unknown, also shape the spectra, as the stronger attenuation associated with the higher-frequency end of spectra generates spectral shifts towards lower frequencies.

Since we can control only the probe and not the medium properties, we programmed pulses so as to overlap in their frequency spectra in a tissue-mimicking medium with a high value of $\frac{B}{A} = 11$ and a low value of the attenuation coefficient expressed in decibel terms as $\alpha_{dB} = 0.3\ \mathrm{dB} \cdot \mathrm{MHz}^{-1} \cdot \mathrm{cm}^{-1}$. The nonlinear coefficient corresponds to the upper bound of what can be expected for tissue, while the attenuation was chosen in the lower bound to minimize the influence from attenuation. This subjective choice was made so as to provide an initial guess of tissue effects on the pulses. The frequencies of the driving pulses were altered so that the harmonic amplitude spectra generated from the f_0 pulse would align with the fundamental spectrum of the $2f_0$ pulse. A sequence programmed for 2.5 cycles with a frequency of 1.4 MHz and another at 5 cycles with a frequency of 3.85 MHz yielded pulses with fundamental frequencies of 1.7 MHz and 3.4 MHz, respectively. These two pulses had the same temporal duration (see Figure 5.3).

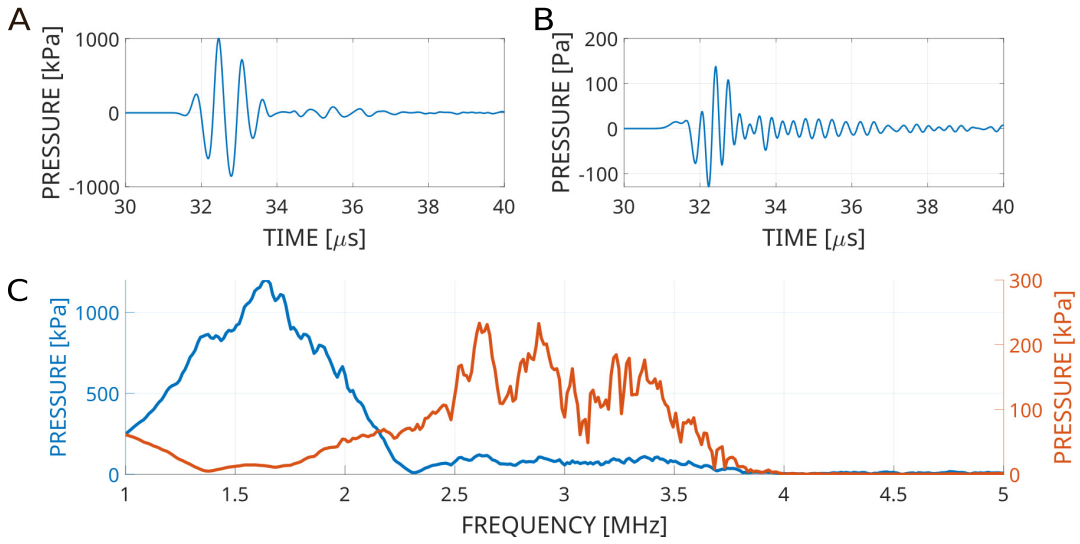

Figure 5.3 – Plots of the simulated forward-propagating signals in a homogeneous tissue-mimicking medium with $\frac{B}{A} = 11$ and $\alpha_{dB} = 0.3\ \mathrm{dB} \cdot \mathrm{MHz}^{-1} \cdot \mathrm{cm}^{-1}$. (A): The time-domain pressure generated 4 cm into the tissue by a pulse programmed for 2.5 cycles with a frequency of 1.4 MHz. (B): Time-domain pressure generated 4 cm into the tissue by a pulse programmed for 5 cycles with a frequency of 3.85 MHz. (C): Fourier transforms of the pressure signals in A (blue) and B (red).

However, given the low amplitude of the harmonic pulse generated through nonlinear propagation, it is difficult to avoid spectral contamination from the fundamental pulse into the harmonic pulse. In order to avoid spectral overlap between fundamental and harmonic signals, it was necessary to isolate the second harmonic signal generated from the pulse transmitted at 1.7 MHz. This is possible using the technique called pulse inversion. The basis of this technique is to transmit two temporally identical, albeit inverted pulses (Jiang, Mao, & Lazenby, 1998). Since the harmonics generated by nonlinear propagation are not inverted and keep the same polarity, summing the resulting RF signals cancels the fundamental and doubles the harmonics. This eliminated the issue of spectral overlap with the fundamental signal. The mean of the two pulses can therefore be used for $S_{2f_0,harm}$ in Equation 5.23.

The use of pulse inversion thus bypasses the need for frequency filtering, that is necessarily sub-optimal, since that approach cannot handle spectral leakage. In short, this allows for more practical isolation of the RF signal ratio of interest.

5.3.2.4 Simulated Tissue-Like Media

Now equipped with an accurate virtual probe and a suitable combination of pulses, the final requirement to train a neural network that can estimate the value of $\frac{B}{A}$ is to generate a training set with a sufficiently representative breadth of tissue-like media. Given the strong impact that the attenuation coefficient α has on the high-frequency second harmonic signal, it is necessary to represent tissues with different combinations of these two parameters. To this end, randomized tissue-like media were simulated in k-Wave with and without lesions. These were inclusions with differing values of both α and $\frac{B}{A}$.

The medium was 6.28 cm long in the direction of propagation (passing the probe's focus of 4 cm). The dimensions along the lateral axis of the imaging plane and in the elevation direction were 3.64 cm and 2.36 cm, respectively, to be just wider than the head of the probe. These media had a speed of sound of 1540 $\frac{\mathrm{m}}{\mathrm{s}}$ and a density of 1050 $\frac{\mathrm{kg}}{\mathrm{m}^3}$ to mimic soft tissue. In order to simulate the effects of scatterers, random variations of up to 15% in the speed of sound and density were introduced into the tissue. A uniform density of the random variations was used for this preliminary investigation of our strategy. The spatial resolution of the medium was 100 $\mu\mathrm{m}$ with a timestep of 16 ns in order to allow k-Wave to simulate waves of up to 6.545 MHz, greater than the harmonic and fundamental content of the 1.7 MHz and 3.4 MHz pulses.

Three pulses were used to generate the necessary signals: two pulses at 1.7 MHz to generate the pulse inversion sequence, and one at 3.4 MHz. The RF data for each pulse were acquired as the pressure time-series data of the central element of the ultrasound probe after the pulse was transmitted. Of the 527 simulated media, 211 were homogeneous, and 316 had a simulated lesion placed inside with different $\frac{B}{A}$ and α values. For media with a lesion, the lesion was centered in the lateral axis of the imaging plane as well as in the elevation axis, occupying half of the medium in these directions. These locations were not varied, as only the central RF line was used. In the axial depth direction, the length of the lesion was randomized, with the minimum being a single pulse length (2.3 mm), and the maximum being half the length of the medium (i.e. 3.14 cm). The depth at which the lesion was located was also randomized, with limits of at least 1 pulse length away from either end of the medium.

A range of approximately 6.0 to 8.1 for soft tissues has been reported in the literature, with lower values for fluids and higher values for fat (Panfilova et al., 2021); however, values for thyroid nodules of different compositions (e.g. cystic, solid, mixed) or of colloid have not been measured.

Therefore, the values of $\frac{B}{A}$ were randomly drawn an extended range of 1.0 to 11.0 so that the none of the values expected in thyroid tissue would fall outside of the sample distribution. The values of the attenuation coefficient (expressed in terms of decibels) α_{dB} were also randomly chosen within the range $0.3 - 1.3\ \mathrm{dB} \cdot \mathrm{MHz}^{-1} \cdot \mathrm{cm}^{-1}$ based on the literature (Brandner et al., 2021).

Simulation Parameter	Value(s)
Medium Length (Propagation Axis)	6.28 cm
Medium Width (Lateral Axis of Imaging Plane)	3.64 cm
Medium Height (Elevation Axis)	2.36 cm
Grid Size	100 μm
Time Step	16 ns
Simulation Time	79.92 μs
Lesion Length (Randomized)	2.3 mm - 32.8 mm
Lesion Width	1.82 cm
Lesion Height	1.18 cm
Lesion Start Position	> 2.3 mm
Lesion End Position	< 60.5 mm
$\frac{B}{A}$ Values (Randomized)	1.0-11.0
α_{dB} Values (Randomized)	0.3-1.3 $\frac{dB}{MHz^{1.005} \cdot cm}$
Frequency Exponent in Power Law	1.005*
Density	1050 $\frac{kg}{m^3}$
Speed of Sound	1540 $\frac{m}{s}$

Table 5.2: Parameters used for the k-Wave simulations. * A nearly linear power-law model was used for attenuation.

5.3.3 Nonlinear Parameter Estimation with a Neural Network

With this simulated data available, we can now train neural networks to predict the mean value of $\frac{B}{A}$ along the central RF line. The RF data signals, sampled at 62.5 MHz based on the sampling frequency of the Verasonics system, had 4652 pressure datapoints over a span of approximately 80 μs. These signals were associated with the $\frac{B}{A}$ profile and α profile along an axis in the direction of propagation centered within the medium, from which mean values were calculated.

The available samples were divided into a test set of 127 simulations and a set of 400 simulations for training the network with 4-fold cross validation, preserving the ratio of media with and without lesions (see Figure 5.4). Cross-validation was used to reduce the effect of random errors by calculating the ensembled average from the predictions of models trained on the different folds (Mohammed & Kora, 2023).

In order to represent more variations in $\frac{B}{A}$ and α values and allow the networks to learn their prediction functions, a strategy to generate new RF-data and acoustic parameter profiles was employed for training. Samples were truncated to generate new samples with an end point between half way through the medium up to the entire signal. By virtue of potentially dividing part of the lesion, this generated new pairs of RF signals and acoustic parameter maps with different mean $\frac{B}{A}$ values and α distributions, depending on how much of the lesion was included. These were zero-padded to the length of the full signals and maps.

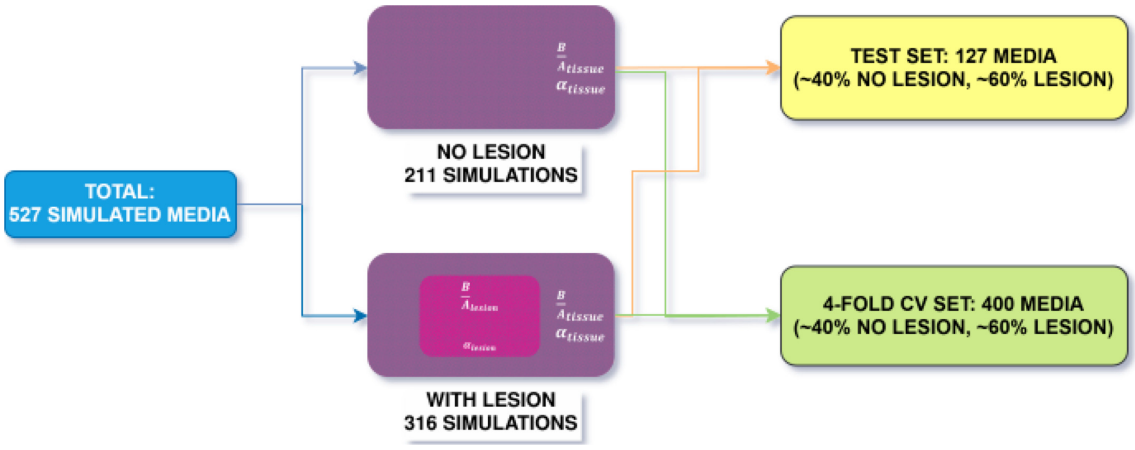

Figure 5.4 – Partition of simulation data between test and cross-validation groups. The proportions of simulations with and without lesions were preserved in the partition. The neural networks were trained in 4-fold cross validation, and the mean prediction of the four networks was evaluated on the test set.

A multi-layer perception was used to learn to predict the mean value of $\frac{B}{A}$, based on the frequency domain relationship given in Equation 5.23. This simple architecture was intended to test if neural networks could learn from a well-prepared signal, without trying to optimize the network structure itself. In order to assess the relative amplitude of the signals corresponding to the term $\frac{S_{2f_0,PI}(z)}{S_{2f_0}(z)}$, a ratio was calculated between the root mean square values of $S_{2f_0,PI}$ and S_{2f_0}. This was similar to calculating the local discrete signal energy; this approach was used since the frequency content of the pulse inversion and $2f_0$ signals had been designed to correspond one to the other in the frequency domain and therefore allowed for an approximation of the relative spectral amplitudes. While this would not be exactly equivalent to comparing the overall spectral content, it was also a form of comparative amplitude signal that could be used in future applications for localized estimation. The signal was calculated from the RF data as:

$$A(n) = \frac{\sqrt{\sum_{j=n-l}^{n+l} \left[S_{2f_0,PI}(j)\right]^2}}{\sqrt{\sum_{j=n-l}^{n+l} \left[S_{2f_0}(j)\right]^2}}, \tag{5.24}$$

in which $2l$ was the length of the window and n the integer index in the timeseries data. This amplitude ratio signal was used as input for the network. The output of the network was an estimate of the mean value of $\frac{B}{A}$ in the simulated medium along a line originating in the center of the ultrasound probe and moving along the axis of propagation. The window length was set to be equivalent to the pulse length used in the simulations.

The network consisted of fully connected layers with Leaky ReLU activation functions. The first layer took in as input the amplitude ratio signals, and successive layers divided the number of neurons by two. The output was a prediction of the mean nonlinear parameter value that was compared to the mean value of the $\frac{B}{A}$ profile from the center of the medium. An L1 loss function was

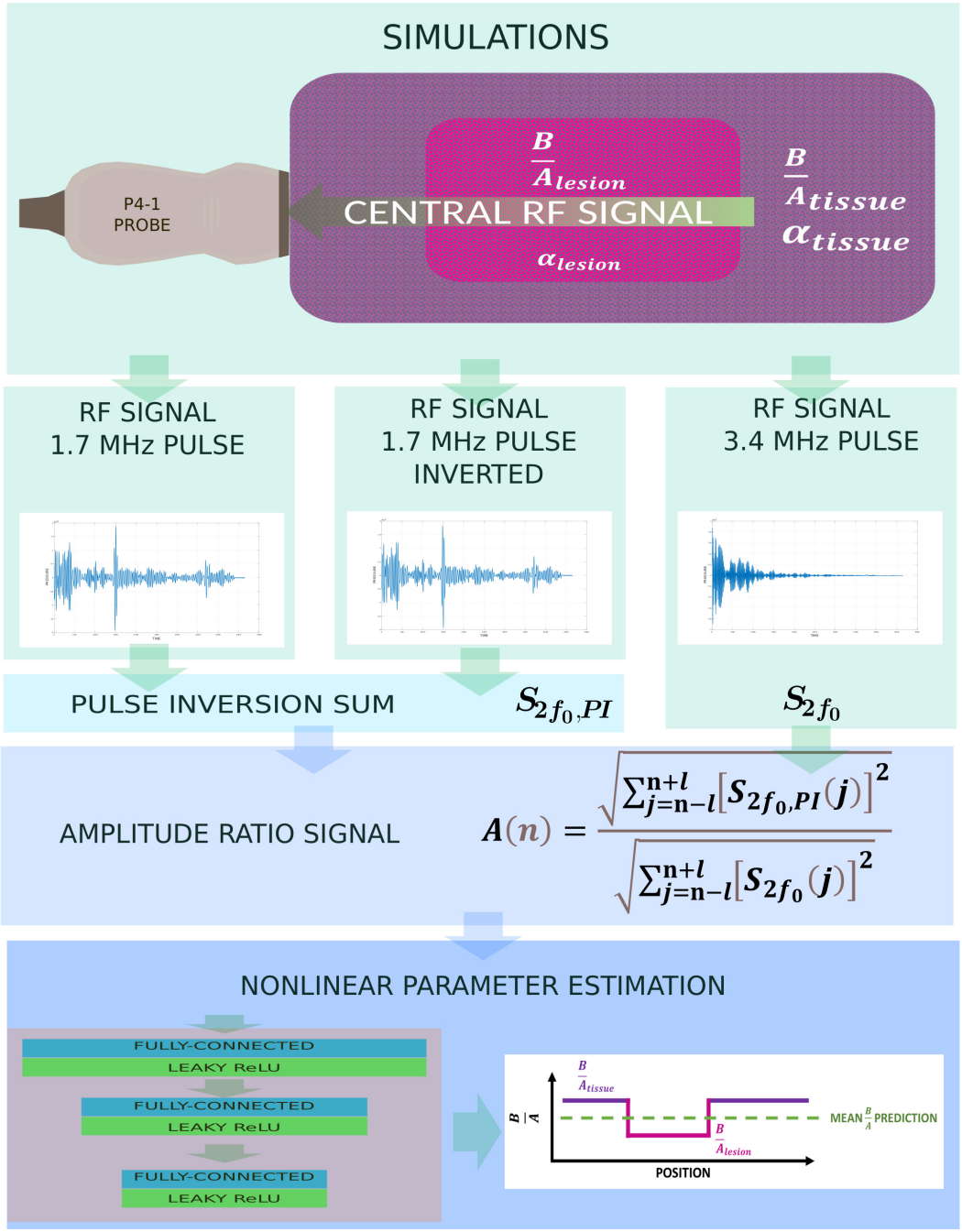

Figure 5.5 – Training strategy for $\frac{B}{A}$ and α prediction. The RF signals from the k Wave simulations were used as input data, being combined into an input signals for simple multi-layer perceptron with Leaky ReLU activation functions. The predicted mean $\frac{B}{A}$ value was compared with a smooth L1 loss to the mean value from the simulated medium.

used for training, in order to reduce the effects from outliers if some $\frac{B}{A}$ and α profiles generated differences in harmonic signals that were too subtle to detect.

5.4 Results

The predicted values of the mean $\frac{B}{A}$ on the test set are presented in Figure 5.6, compared to the reference values used in the simulated media. The predicted values of B/A increased with the reference values, suggesting that a relevant relationship was learned by the networks. However, the predictions consistently overestimated the lower values of $\frac{B}{A}$ (from around 1 to 4), while slightly underestimating higher values (from 8 to 11). Given this trend, a global average of the prediction error may be less informative than analysis at different regions of the $\frac{B}{A}$ range. In order to examine the prediction errors for different reference values of $\frac{B}{A}$, a boxplot of errors grouped within bins of 1 unit length of $\frac{B}{A}$ is also presented in Figure 5.6, with the mean errors and standard deviations of the errors in Table 5.3.

At $\frac{B}{A} = 5$, corresponding to water, the mean error was an overestimation of 1.13, greater than the standard deviation of the error of 0.45, though there were few samples in this range. Moving up to the range of blood at $\frac{B}{A} = 6$, the mean error was only 0.55, but standard deviation was more significant. This was also true for soft-tissue like values of $\frac{B}{A} = 7$ and $\frac{B}{A} = 8$; the mean error was low, but the standard deviations were around 1.09 and 0.85 for each bin, respectively. Moving up to the highest range, like fat with $\frac{B}{A}$ values from 9-11, mean underestimation became prominent, still with a high degree of random error.

Reference $\frac{B}{A}$ Range $\pm$ 0.5	**Mean Prediction Error**	**Std. Dev. of Error**	**Number of Samples**
5	1.13	0.45	4
6	0.55	0.80	23
7	0.41	1.09	19
8	-0.34	0.84	22
9	-0.75	0.85	20
10	-0.90	1.23	16
11	-0.94	0.81	8

Table 5.3: Means and standard deviation of $\frac{B}{A}$ prediction errors on the test set.

For a point of reference, the differences between healthy and diseased tissue samples that have been described in the laboratory setting (Sehgal et al., 1986) can be subtle enough that prediction errors that are consistently greater in magnitude than 1.0 could be too imprecise. While the mean error of predictions in soft-tissue range fall below this threshold, the random errors in this range are still substantial. The overall trend of overestimation of low values and underestimation of high values also reduces the usefulness of the technique, as this trend might tend to obscure actual differences between tissues.

Another basis for comparison is with the predictions made directly from Equation 5.23, using the values of the pulse amplitude, frequencies, density, and speed of sound values from the simulations. These results were not in a physically plausible range, with values ranging from about -0.7 to 5.9 for the test set, as seen in Figure 5.7. In the same figure, these results are also presented with a correction for the mean error on the test set, in order to appreciate the scale of random errors. The overestimation at lower values is more pronounced than with our strategy, and the variations in the range of soft tissue around $\frac{B}{A} = 5$ to $\frac{B}{A} = 7$ is also greater. This suggests that our strategy was able to learn useful corrections for $\frac{B}{A}$ prediction.

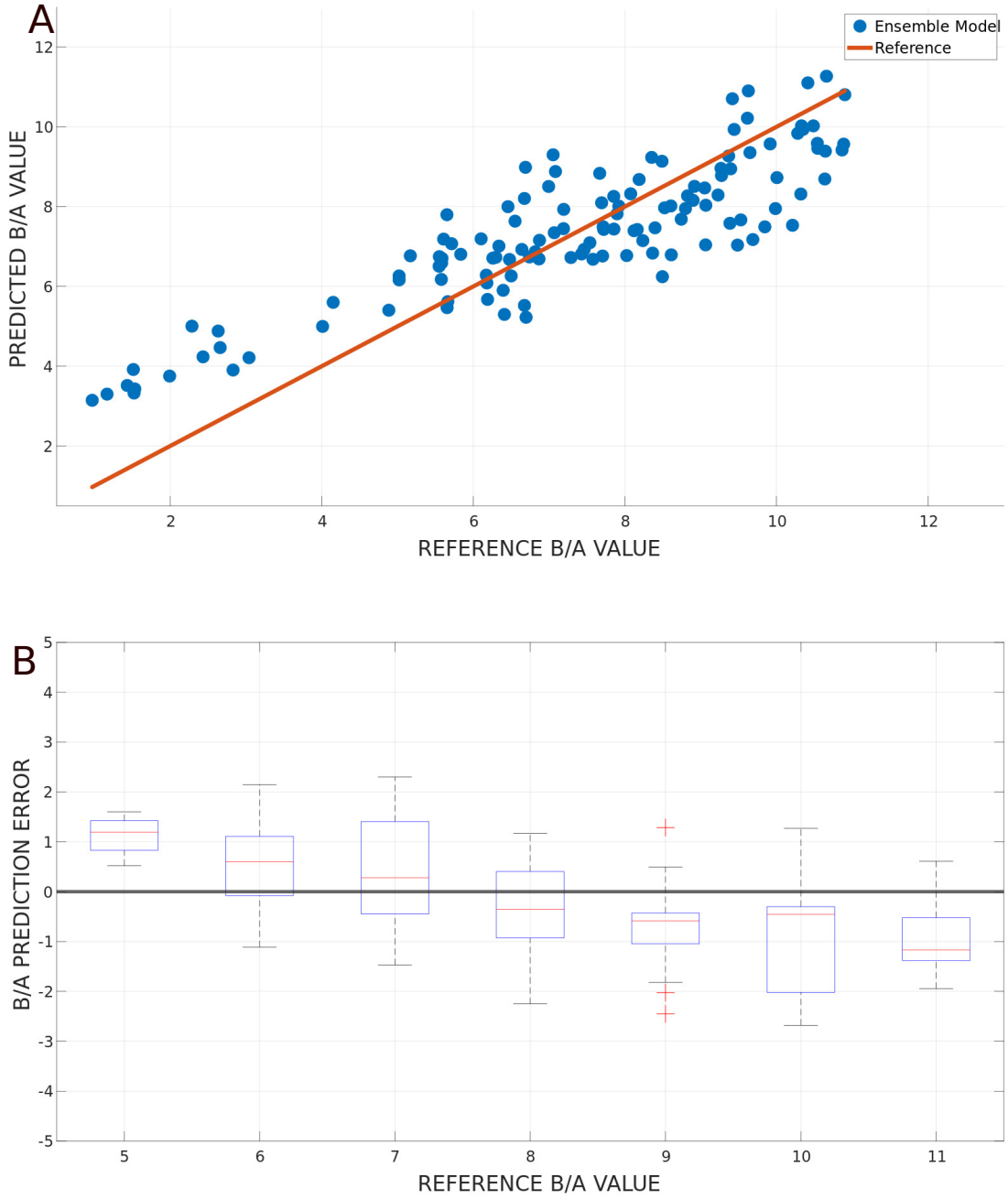

Figure 5.6 – (A) Scatter plot of reference $\frac{B}{A}$ values (x axis) vs. the average predictions (y axis) made by the ensemble of trained networks on the test set. The line in red indicates a perfect match between prediction and reference values. (B) Box plot of the errors in test set $\frac{B}{A}$ predictions, grouped by reference $\frac{B}{A}$ value. The horizontal line represents zero error, i.e. a perfect prediction. The mean value of each group is shown as a red line through the box plot, and outliers are presented as red crosses.

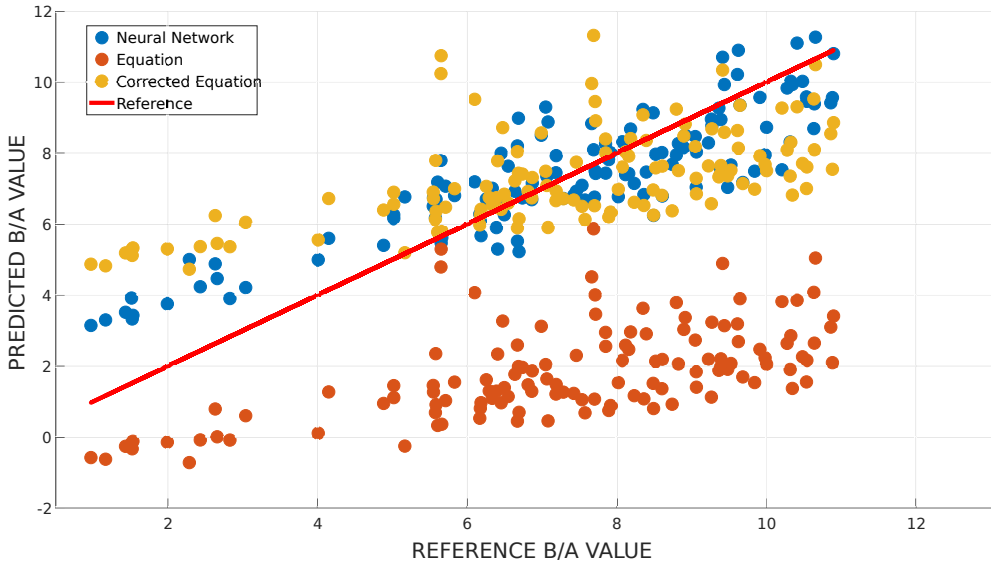

Figure 5.7 – Scatter plot of reference $\frac{B}{A}$ values (x axis) vs. the predicted values (y axis) made by the neural network strategy (in blue) and by Equation 5.23 (in red) on the test set. The equation results were also corrected by their mean error on the test set for comparison (in yellow).

5.5 Discussion

Learning to predict the values of $\frac{B}{A}$ from simulated RF data with simple neural network architectures is a difficult task, though one that can be made easier by shaping signals to have a clear connection to the parameters of interest. This allowed the network to learn to implicitly correct for the diffraction effects of the P4-1 probe and the residual effects of attenuation. This was evidenced with an improvement in accuracy compared to prediction with the equation alone, and reduced variability around predictions corresponding to soft-tissue values.

The mean prediction error for $\frac{B}{A}$ values from 5 to 10 were similar in magnitude to those of Toulemonde et al.. In that publication, mean errors were 0.1 at $\frac{B}{A} = 5$, 2.6 at $\frac{B}{A} = 7$, and -1.5 at $\frac{B}{A} = 10$ (Toulemonde et al., 2015). This suggests that our technique showed slightly more accurate prediction at higher values corresponding to soft-tissue (6-7), but was less accurate for fluid (around 5). Noticeably, in both approaches there was a marked overestimation of low $\frac{B}{A}$ values, and a tendency to underestimation of high $\frac{B}{A}$ values. In terms of standard deviations, Toulemonde et al. gave standard deviations of were 1.5 at $\frac{B}{A} = 3$, 1.4 at $\frac{B}{A} = 5$, 1.3 at $\frac{B}{A} = 7$, and 1.5 at $\frac{B}{A} = 10$ (Toulemonde et al., 2015). Our standard deviations were slightly tighter, and were calculated across a range of different α values.

Overall, however, the magnitude of prediction errors would need to be improved for future practical applications. One mechanism for this, considering the interactions between attenuation and harmonic generation, would be by providing the network with attenuation information to further narrow the scope of the correction functions it would need to learn. The α coefficient of the medium could be estimated from the RF data, as we were able to do using the $2f_0$ pulse data with a similar MLP architecture on the same data used for $\frac{B}{A}$ estimation. The prediction results on the test set for α are shown in Figure 5.8.

Future applications of machine learning techniques to estimate $\frac{B}{A}$ values in nodules would likely need to build upon this approach and incorporate local estimation, as in Toulemonde et al., albeit with sufficient precision to reliably detect small diferences in parameter value. Considering the challenges of local estimation highlights some of the limitations of this preliminary prediction strategy.

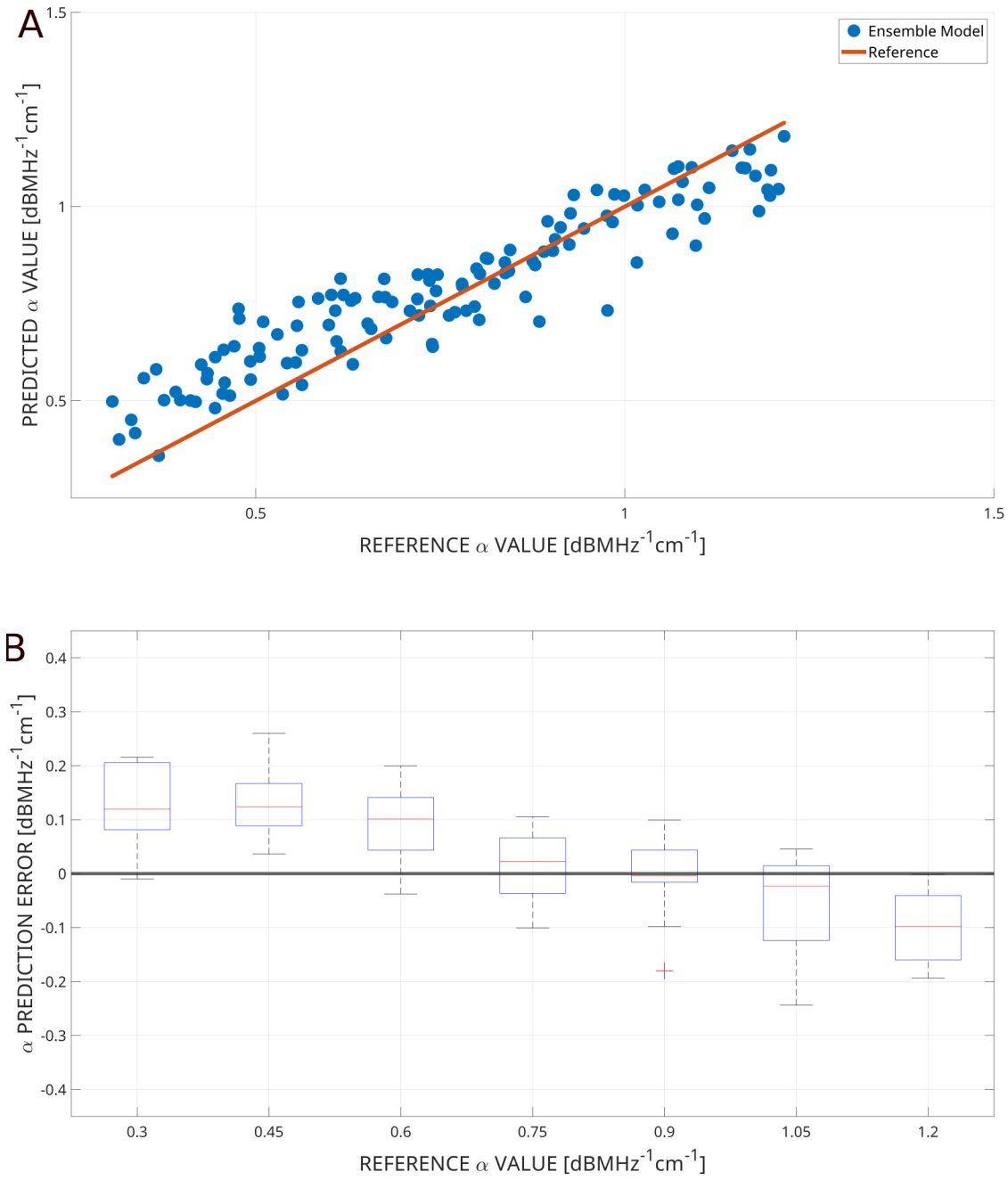

Figure 5.8 – (A) Scatter plot of reference α values in $\mathrm{dB} \cdot \mathrm{MHz}^{-1} \cdot \mathrm{cm}^{-1}$ (x axis) vs. the average predictions (y axis) made by the ensemble of trained networks on the test set. The line in red indicates a perfect match between prediction and reference values. (B) Box plot of the errors in test set α predictions, grouped by reference α value. The horizontal line represents zero error, i.e. a perfect prediction. The mean value of each group is shown as a red line through the box plot, and outliers are presented as red crosses.

5.5.1 Limitations

While the simulations were designed to mimic tissue, the heterogeneity of real thyroid may be more pronounced. Skin and muscle, in addition to zones of inflammation within the thyroid, would create multiple tissue layers, perhaps with variations in density and speed of sound which would impact $\frac{B}{A}$ estimation. Therefore, future training strategies should generate sufficiently complex tissue-like media with known $\frac{B}{A}$ profiles. Laboratory experiments with tissue-mimicking phantoms and tissue from cadavers or animal models could also provide an adequate source of training data with characterizable ground-truth values.

There were also assumptions in the creation of our training simulations and the pulse division strategy that would need to be adjusted for translation. Notably, the distribution and density of scatterers might vary between tissue types or within nodules; this would need to be reflected in the training data for the network. The scatterer density used here was high; it could also have impacted the results obtained. In addition, the assumption of linearity of attenuation dependence on frequency could also be a source of error, and would require further simulation adjustment as well as force the network to learn a more complex relationship between the RF data and $\frac{B}{A}$. This could involve using networks to estimate two parameters to define a power law as in Equation 5.6.

Finally, the neural networks used here were not specially adapted to learn from the data at hand, as the focus of this initial strategy was on generation and processing of the RF signals. However, more specialized architectures and training strategies, such as physics-informed neural networks, could better exploit dynamics present in the RF data.

5.6 Conclusion

We saw in Chapters 2 and 3 that thyroid nodule ultrasound suffers from inter-operator and inter-reader variability, and could benefit from more objective measures related to quantifiable aspects of tissue. Here we have presented a preliminary strategy for the estimation of the nonlinear $\frac{B}{A}$ by combining a neural network trained on realistic simulation data. A pulse-division strategy was used to eliminate the effects of scatterers and compensate some of the effects of attenuation. The network then implicitly learned corrections for the residual impact of attenuation on the harmonic signal and the diffraction effects of a specific probe to make mean $\frac{B}{A}$ predictions. The results of the predictions showed a capacity to estimate $\frac{B}{A}$, albeit with some limitations in accuracy particularly for low and high tissue values of these parameters. This strategy combining machine learning with input data pre-processed to compensate for known physical phenomena relevant to the quantity of interest could be a possible route toward more robust characterization of nodules in the future. This is especially promising when compared to existing subjective measures such as echogenicity. However, far more work is required for these techniques to be able to contribute to thyroid nodule ultrasound.

CHAPTER 6

Conclusion

Throughout the chapters of this thesis, we have seen the limitations of thyroid ultrasound for the evaluation of nodules, as well as the limitations for machine learning methods applied to these images. Despite this, we cannot lose sight of the profound utility of thyroid ultrasound; it is an accessible, non-invasive tool for evaluation of a soft tissue lesion that is common around the world. Machine learning tools hold the potential to improve the reproducibility of thyroid ultrasound and extend its utility to non-expert practitioners. The contributions of this thesis seek to add to advancing these benefits.

6.1 Contributions

The main contributions of this thesis are listed below. These represent responses to different challenges in thyroid nodule ultrasound and machine learning applications to improve it.

Creation of a Multicentric French Thyroid Ultrasound Dataset
Evaluation of machine learning applications to thyroid ultrasound in France requires as a starting point an understanding of French clinical practice. As thyroid ultrasound acquisition in France is directly performed by the interpreting radiologist or endocrinologist, the quality and nature of the images may vary between practictioners. In order to facilitate the study of EU-TIRADS evaluation in France, and with the support of the Association Francophone de Thyroïdologie (AFTHY) to assemble a dataset of 303 real clinical thyroid ultrasound images acquired during routine clinical practice by four different French experts, each in their own practice setting and each with their own ultrasound system. All four experts also contributed descriptions of EU-TIRADS score, composition, echogenicity, shape, margin, and the presence of echogenic foci on each image, according to their clinical experience. This dataset provided a unique opportunity for further study of thyroid ultrasound in France.

Inter-Expert Variability Study on French Thyroid Ultrasound
While recent studies have examined inter-expert variability in EU-TIRADS evaluations among European experts (Solymosi et al., 2023), no study to our knowledge has done this specifically with French experts, to analyze the specific biases arising from historial thyroid ultrasound practices in this country. Using the images from the data set above, we examined French inter-expert variability in EU-TIRADS scores. We also identified differences in the characterization of composition, echogenicity, shape, margin, and the presence of echogenic foci that were associated with disagreements in those scores. This provided insights into inter-expert scoring differences, and the sonographic features whose identification could be standardized with machine learning tools.

One feature in particular, nodule echogenicity, was of particular interest, as many expert disagreements about EU-TIRADS score were linked to disagreements about echogenicity labels. To further explore this, features, we analyzed expert disagreement about the labels of hyper-/isoechogenicity and hypoechogenicity, and asked the experts for their observations about the factors that made this distinction difficult. Quantitative differences between nodule and thyroid parenchyma within images were studied to look for associations with expert agreement or disagreement, finding less expert agreement in nodules whose interior regions showed more heterogeneity. This further highlighted the difficulties in subjective ultrasound image analysis for nodule characterization.

Test of Active Learning Strategies on Real Ultrasound Data
Active learning strategies hold promise as methods to reduce the annotation burden for training machine learning algorithms on thyroid ultrasound data. However, the practical effectiveness of active learning methods must be evaluated on real clinical data to confirm whether they consistently outperform random selection of images to annotate. Indeed, the effects of the initial random set have a strong impact on the performance of active learning results, so we tested multiple active learning strategies on a thyroid ultrasound dataset from a French hospital, in addition to two other external medical image datasets. These results suggested that many active learning strategies had difficulty performing any better than random selection, with the magnitude of differences between most strategies being outweighed by the variability from initial random set selection.
In addition to testing existing strategies, we also proposed a new strategy that combined the power of random sampling with active learning criteria, as done by Gaillochet et al. (Gaillochet et al., 2023). This was tested on the same dataset to see if it could outperform random sampling of images for annotation. There was a limited improvement over random selection when using semi-supervised learning; however, the magnitude of this difference might be too small have a meaningful impact on annotation budgets.

Initial Steps to a Strategy for Nonlinear Parameter Estimation
In light of the inter-expert variability in expert evaluation of thyroid ultrasound images, quantitative ultrasound techniques may be a promising means to develop objective measures for nodule evaluation. One possible target of quantitative ultrasound is the nonlinear parameter $\frac{B}{A}$; estimation of the value of this parameter in tissue is complicated because of the effects of attenuation, scatterers, and diffraction. As an initial step toward the implementation of a practical strategy for $\frac{B}{A}$ estimation, we proposed a signal acquisition and processing approach that follows physical intuitions about acoustic propagation to facilitate training a neural network to estimate the values of these parameters. This improved the accuracy of $\frac{B}{A}$ estimation in simulated tissue-mimicking media, and further progress with this strategy could provide a route towards developing more practical techniques in the future.

6.2 Future Directions

Many future directions are possible to improve applications of machine learning to thyroid nodule ultrasound.

Study of Inter-Expert Variability in Practice

The evidence that French experts vary in their evaluation of images in Chapter 1 has important implications for the risk stratification of thyroid nodules. However, thorough investigation of these differences should also account for differences in acquisition, since French practitioners examine patients and not still images. Such an investigation of expert variability in practice would entail having the same set of patients examined by multiple practitioners in their hospital center or clinics, using their standard procedures. The value of this study would lie in identifying real differences in practice, and could suggest means of standardization. In addition, confirming which of the sonographic features studied in this thesis on still images show similar inter-expert variability in real practice would be necessary to establish the predictive value of those labels; after all, the associations established between these features and malignancy are only useful if the features can be reproducibly described. In addition, this future study ought to also examine the interdependence between sonographic feature labels. In our analysis, it was evident that most experts used a few combinations of features to describe most nodules; a more thorough investigation could yield further links to see whether the presence of certain features made experts more or less likely to assign labels for other features. This information would also be useful to train practitioners to perform evaluations more consistently.

Thyroid Ultrasound Standardization with Quantitative Measures

Given the variability seen between experts and even within repeated evaluations by individual experts, quantitative definitions tied to the sonographic features associated with malignancy could make thyroid nodule ultrasound more reliable. This would require studies using multiple ultrasound systems and operators on the same nodule, with histopathologic confirmation of malignancy. Given the difficulty of obtaining biopsies from cases that most likely benign, a cadaveric study could be employed. In this way, quantitative image features that could be reliably calculated across a number of different ultrasound systems could be correlated with biopsy-proven malignancy. In addition, comparison among the images acquired on the same nodules could be used to develop quality control standards to establish whether an image was acquired with parameters such as contrast, field of view, and time-gain compensation that are suitable for analysis. These advances would turn thyroid nodule ultrasound into a tool that could be widely and reliably used by even non-expert medical practitioners.

Rigorous Evaluation of Machine Learning Tools for Thyroid Nodule Ultrasound

As reviewed in Chapter 2, many machine learning tools now exist for the automation of thyroid nodule ultrasound. In order to assess their clinical reliability, however, rigorous testing is necessary. The variability seen in Chapter 1 in expert labels might suggest that some of these algorithms have been trained with noisy labels; even those algorithms validated by biopsy might not have adequate training on benign samples as these would be less likely to be subjected to FNA or tissue biopsy. In a similar fashion to the description of quantifiable features for nodule characterization, this evaluation could be conducted with a cadaveric study. It would also require the participation of multiple operators, in order to verify that a proposed algorithm would be robust to differences in acquisition. This study would form a foundation for proper evaluation of which models hold real promise to improve clinical ultrasound.

Implementation of Practical $\frac{B}{A}$ Measurements

As with the utility of standardized quantitative measures of thyroid ultrasound, measurement of

an acoustic parameter like $\frac{B}{A}$ could be a more reliable marker to try to associate with malignancy. Given the limitations in our approach, much work is necessary to advance to a clinically-useful technique. However, the first steps that could be pursued in that direction would be to try to associate the value of $\frac{B}{A}$ in nodules to specific histologic differences in malignant lesions. Further improved characterization *in vivo* could be achieved with a physics-informed approach to RF signal acquisition and processing.

Bibliography

Abu-Zidan, F. M., Hefny, A. F., & Corr, P. (2011). Clinical ultrasound physics. *Journal of Emergencies, Trauma, and Shock*, *4*(4), 501-503.

Alghanimi, G. B., Aljobouri, H. K., & Al-shimmari, K. A. (2024). CNN and ResNet50 model design for improved ultrasound thyroid nodules detection. In *2024 ASU international conference in emerging technologies for sustainability and intelligent systems (ICETSIS)* (p. 1000-1004).

Barczyński, M., Stopa-Barczyńska, M., Wojtczak, B., Czarniecka, A., & Konturek, A. (2020). Clinical validation of S-Detect (TM) mode in semi-automated ultrasound classification of thyroid lesions in surgical office. *Gland Surgery*, *9*(S2).

Beyer, R. T. (1960). Parameter of nonlinearity in fluids. *The Journal of the Acoustical Society of America*, *32*(6), 719-721.

Beyer, R. T. (1973). Nonlinear acoustics. *American Journal of Physics*, *41*(9), 1060-1067.

Bjørnø, L. (1986). Characterization of biological media by means of their non-linearity. *Ultrasonics*, *24*(5), 254-259.

Blanken, N., Heiles, B., Kuliesh, A., Versuis, M., Jain, K., Maresca, D., & Lajoinie, G. (2024). PROTEUS: A physically realistic contrast-enhanced ultrasound simulator—Part I: Numerical methods. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*.

Bonavita, J. A., Mayo, J., Babb, J., Bennett, G., Oweity, T., Macari, M., & Yee, J. (2009). Pattern recognition of benign nodules at ultrasound of the thyroid: Which nodules can be left alone? *Am J Roentgenol*, *193*(1), 207-213.

Boozari, B., Potthoff, A., Mederacke, I., Hahn, A., Reising, A., Rifai, K., ... Gebel, M. (2010). Evaluation of sound speed for detection of liver fibrosis: Prospective comparison with transient dynamic elastography and histology. *J Ultrasound Med*, *29*(11), 1581-1588.

Brandner, D. M., Cai, X., Foiret, J., Ferrara, K. W., & Zagar, B. G. (2021). Estimation of tissue attenuation from ultrasonic B-mode images—spectral-log-difference and method-of-moments algorithms compared. *Sensors*, *21*(7).

Buda, M., Wildman-Tobriner, B., Hoang, J. K., Thayer, D., Tessler, F. N., Middleton, W. D., & Mazurowski, M. A. (2019). Management of thyroid nodules seen on us images: Deep learning may match performance of radiologists. *Radiology*, *292*(3), 695–701.

Budd, S., Robinson, E. C., & Kainz, B. (2021). A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis*, *71*, 102062.

Chambara, N., Liu, S. Y. W., Lo, X., & Ying, M. (2021). Diagnostic performance evaluation of different TI-RADS using ultrasound computer-aided diagnosis of thyroid nodules: An experience with adjusted settings. *PLOS ONE*, *16*(1).

Chammas, M. C., de Araujo Filho, V. J. F., Moysés, R. A., Brescia, M. D. G., Mulatti, G. C., Brandão, L. G., ... Ferraz, A. R. (2008). Predictive value for malignancy in the finding of microcalcifications on ultrasonography of thyroid nodules. *Head & Neck*, *30*(9), 1206-1210.

Chen, H., Song, S., Wang, X., Wang, R., Meng, D., & Wang, L. (2021). LRTHR-Net: A low-resolution-to-high-resolution framework to iteratively refine the segmentation of thyroid nodule in ultrasound images. *Segmentation, Classification, and Registration of Multi-Modality Medical Imaging Data*, 116–121.

Davies, J., Tapson, J., & Mortimer, B. (2000). A novel phase locked cavity resonator for B/A measurements in fluids. *Ultrasonics*, *38*(1), 284-291.

de Monchy, R., Rouyer, J., Destrempes, F., Chayer, B., Cloutier, G., & Franceschini, E. (2018). Estimation of polydispersity in aggregating red blood cells by quantitative ultrasound backscatter analysis. *The Journal of the Acoustical Society of America*, *143*(4), 2207-2216.

Dunn, F., Law, W., & Frizzell, L. (1981). Nonlinear ultrasonic wave propagation in biological materials. In *1981 Ultrasonics Symposium* (p. 527-532).

Dunn, F., Law, W. K., & Frizzell, L. A. (1982). Nonlinear ultrasonic propagation in biological media. *The British Journal of Cancer. Supplement*, *5*, 55-58.

Durante, C., Hegedüs, L., Na, D. G., Papini, E., Sipos, J. A., Baek, J. H., ... Tessler, F. N. a. (2023). International expert consensus on US lexicon for thyroid nodules. *Radiology*, *309*(1), e231481.

Fleiss, J. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378—382.

Fujii, Y., Taniguchi, N., Akiyama, I., Tsao, J.-W., & Itoh, K. (2004). A new system for in vivo assessment of the degree of nonlinear generation using the second harmonic component in echo signals. *Ultrasound in Medicine & Biology*, *30*(11), 1511-1516.

Gaillochet, M., Desrosiers, C., & Lombaert, H. (2023). Active learning for medical image segmentation with stochastic batches. *Medical Image Analysis*, *90*, 102958.

Garrett, S. L. (2020). Nonlinear acoustics. In *Understanding Acoustics: An Experimentalist's View of Sound and Vibration* (p. 701-753). Springer International Publishing.

Giovanella, L. (Ed.). (2023). *Integrated Diagnostics and Theranostics of Thyroid Diseases*. Springer Cham.

Gong, H., Chen, G., Wang, R., Xie, X., Mao, M., Yu, Y., ... Li, G. (2021). Multi-task learning for thyroid nodule segmentation with thyroid region prior. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)* (p. 257-261).

Gong, H., Chen, J., Chen, G., Li, H., Li, G., & Chen, F. (2023). Thyroid region prior guided attention for ultrasound segmentation of thyroid nodules. *Computers in Biology and Medicine*, *155*, 106389.

Goss, S., Frizzell, L., & Dunn, F. (1979). Ultrasonic absorption and attenuation in mammalian tissues. *Ultrasound in Medicine & Biology*, *5*(2), 181-186.

Grani, G., Lamartina, L., Cantisani, V., Maranghi, M., Lucia, P., & Durante, C. (2018). Interobserver agreement of various thyroid imaging reporting and data systems. *Endocrine Connections*, *7*(1), 1–7.

Ha, E. J., Na, D. G., & Baek, J. H. (2021). Korean thyroid imaging reporting and data system: Current status, challenges, and future perspectives. *Korean J Radiol*, *22*(9), 1569-1578.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. *arXiv Preprint arXiv:1512.03385*.

Henrichsen, T. L., Reading, C. C., Charboneau, J. W., Donovan, D. J., Sebo, T. J., & Hay, I. D. (2010). Cystic change in thyroid carcinoma: Prevalence and estimated volume in 360 carcinomas. *Journal of Clinical Ultrasound*, *38*(7), 361-366.

Hoang, J. K., Lee, W. K., Lee, M., Johnson, D., & Farrell, S. (2007). US features of thyroid malignancy: Pearls and pitfalls. *RadioGraphics*, *27*(3), 847-860.

Hu, S.-Y., Wang, S., Weng, W.-H., Wang, J., Wang, X., Ozturk, A., ... Samir, A. E. (2020). Self-supervised pretraining with DICOM metadata in ultrasound imaging. In F. Doshi-Velez et al. (Eds.), *Proceedings of the 5th Machine Learning for Healthcare Conference* (Vol. 126, pp. 732–749). PMLR.

Huang, K., Huang, J., Wang, W., Xu, M., & Liu, F. (2022). A deep active learning framework with information guided label generation for medical image segmentation. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (p. 1562-1567).

Jeon, S. K., Lee, J. M., & Joo, I. (2021). Clinical feasibility of quantitative ultrasound imaging for suspected hepatic steatosis: Intra- and inter-examiner reliability and correlation with controlled attenuation parameter. *Ultrasound in Medicine & Biology*, *47*(3), 438-445.

Jiang, P., Mao, Z., & Lazenby, J. (1998). A new tissue harmonic imaging scheme with better fundamental frequency cancellation and higher signal-to-noise ratio. In *1998 IEEE Ultrasonics Symposium. Proceedings* (Vol. 2, p. 1589-1594).

Kant, R., Davis, A., & Verma, V. (2020). Thyroid nodules: Advances in evaluation and management. *American Family Physician*, *102*(5), 298-304.

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021). Physics-informed machine learning. *Nature Reviews Physics*, *3*(6), 422-440.

Kelly, F. C. (1961). Iodine in medicine and pharmacy since its discovery—1811–1961. *Proceedings of the Royal Society of Medicine*, *54*(10), 831-836.

Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C., Liang, H., Baxter, S. L., ... et al. (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, *172*(5).

Kim, B. M., Kim, M. J., Kim, E.-K., Kwak, J. Y., Hong, S. W., Son, E. J., & Kim, K. H. (2008). Sonographic differentiation of thyroid nodules with eggshell calcifications. *Journal of Ultrasound in Medicine*, *27*(10), 1425-1430.

Kim, E.-K., Park, C. S., Chung, W. Y., Oh, K. K., Kim, D. I., Lee, J. T., & Yoo, H. S. (2002). New sonographic criteria for recommending fine-needle aspiration biopsy of nonpalpable solid nodules of the thyroid. *American Journal of Roentgenology*, *178*(3), 687-691.

Kim, H. L., Ha, E. J., & Han, M. (2019). Real-world performance of computer-aided diagnosis system for thyroid nodules using ultrasonography. *Ultrasound in Medicine & Biology*, *45*(10), 2672–2678.

Koike, E., Noguchi, S., Yamashita, H., Murakami, T., Ohshima, A., Kawamoto, H., & Yamashita, H. (2001). Ultrasonographic characteristics of thyroid nodules: Prediction of malignancy. *Archives of Surgery*, *136*(3), 334-337.

Kollman, C., ter Haar, G., Dolezal, L., Hennerici, M., Salvesen, K. A., & Valentin, L. (2013). Ultrasound output: Thermal (TI) and mechanical (MI) indices. *Ultraschall in der Medizin*, *34*(5), 422-434.

Krönke, M., Eilers, C., Dimova, D., Köhler, M., Buschner, G., Schweiger, L., ... Wendler, T. (2022). Tracked 3D ultrasound and deep neural network-based thyroid segmentation reduce interobserver variability in thyroid volumetry. *PLOS ONE*, *17*(7), e0268550.

Kuc, R., & Regula, D. P. (1984). Diffraction effects in reflected ultrasound spectral estimates. *IEEE Transactions on Biomedical Engineering*, *BME-31*(8), 537-545.

Küpper, F. C., Feiters, M. C., Olofsson, B., Kaiho, T., Yanagida, S., Zimmermann, M. B., ... Kloo, L. (2011). Commemorating two centuries of iodine research: An interdisciplinary overview of current research. *Angewandte Chemie International Edition*, *50*(49), 11598-11620.

Landis, J. R., & G., K. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159-174.

Lebrun, L., & Salmon, I. (2024). Pathology and new insights in thyroid neoplasms in the 2022 who classification. *Current Opinion in Oncology*, *36*(1), 13-21.

Li, T., Jiang, Z., Lu, M., Zou, S., Wu, M., Wei, T., ... Liao, J. (2020). Computer-aided diagnosis system of thyroid nodules ultrasonography. *Medicine*, *99*(23).

Liu, D., Yang, K., Zhang, C., Xiao, D., & Zhao, Y. (2024). Fully-automatic detection and diagnosis system for thyroid nodules based on ultrasound video sequences by artificial intelligence. *Journal of Multidisciplinary Healthcare*, *17*, 1641–1651.

Liu, Z., Zhong, S., Liu, Q., Xie, C., Dai, Y., Peng, C., ... Zou, R. (2021). Thyroid nodule recognition using a joint convolutional neural network with information fusion of ultrasound images and radiofrequency data. *European Radiology*.

Lu, J., Ouyang, X., Liu, T., & Shen, D. (2021). Identifying thyroid nodules in ultrasound images through segmentation-guided discriminative localization. *Segmentation, Classification, and Registration of Multi-modality Medical Imaging Data*, 135–144.

Lu, Y., Shi, X. Q., Zhao, X., Song, D., & Li, J. (2019). Value of computer software for assisting sonographers in the diagnosis of thyroid imaging reporting and data system grade 3 and 4 thyroid space-occupying lesions. *Journal of Ultrasound in Medicine*, *38*(12), 3291–3300.

Luijten, B., Chennakeshava, N., Eldar, Y. C., Mischi, M., & van Sloun, R. J. (2023). Ultrasound signal processing: From models to deep learning. *Ultrasound in Medicine & Biology*, *49*(3), 677-698.

Ma, Y., Huo, X., Kong, S., Xu, W., Zhao, W., & Zhu, M. (2023). A review about C-TIRADS, ACR-TIRADS, and K-TIRADS combined with real-time tissue elastography to diagnose thyroid nodules. *Discovery Medicine*, *35*(174), 1—10.

Malhi, H., Beland, M. D., Cen, S. Y., Allgood, E., Daley, K., Martin, S. E., ... Grant, E. G. (2014). Echogenic foci in thyroid nodules: Significance of posterior acoustic artifacts. *American Journal of Roentgenology*, *203*(6), 1310-1316.

Mena, G., Montalvo, A., Ubidia, M., Olmedo, J., Guerrero, A., & Leon-Rojas, J. E. (2023). Elastography of the thyroid nodule, cut-off points between benign and malignant lesions for strain, 2D shear wave real time and point shear wave: A correlation with pathology, ACR TIRADS and alpha score. *Frontiers in Endocrinology*, *14*.

Mohammed, A., & Kora, R. (2023). A comprehensive review on ensemble deep learning: Opportunities and challenges. *Journal of King Saud University - Computer and Information Sciences*, *35*(2), 757-774.

Moon, W.-J., Jung, S. L., Lee, J. H., Na, D. G., Baek, J.-H., Lee, Y. H., ... Lee, D. H. (2008). Benign and malignant thyroid nodules: US differentiation—multicenter retrospective study. *Radiology*, *247*(3), 762-770.

Munjal, P., Hayat, N., Hayat, M., Sourati, J., & Khan, S. (2022). Towards robust and reproducible active learning using neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (p. 223-232).

Na, D. G., Kim, D. S., Kim, S. J., Ryoo, J. W., & Jung, S. L. (2016). Thyroid nodules with isolated macrocalcification: Malignancy risk and diagnostic efficacy of fine-needle aspiration and core needle biopsy. *Ultrasonography*, *35*(3), 212-219.

Najjar, R. (2023). Redefining radiology: A review of artificial intelligence integration in medical imaging. *Diagnostics*, *13*(17).

Panfilova, A., van Sloun, R. J., Wijkstra, H., Sapozhnikov, O. A., & Mischi, M. (2021). A review on B/A measurement methods with a clinical perspective. *The Journal of the Acoustical Society of America*, *149*(4), 2200–2237.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Édouard Duchesnay (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825-2830.

Piticchio, T., Russ, G., Radzina, M., Frasca, F., Durante, C., & Trimboli, P. (2024). Head-to-head comparison of American, European, and Asian TIRADSs in thyroid nodule assessment: Systematic review and meta-analysis. *Eur Thyroid J.*, *13*(2), e230242.

Reverter, J. L., Vázquez, F., & Puig-Domingo, M. (2019). Diagnostic performance evaluation of a computer-assisted imaging analysis system for ultrasound risk stratification of thyroid nodules. *American Journal of Roentgenology*, *213*(1), 169–174.

Rozanova, A. (2007). The Khokhlov–Zabolotskaya–Kuznetsov equation. *Comptes Rendus. Mathématique*, *344*(5), 337-342.

Russ, G., Bonnema, S. J., Erdogan, M. F., Durante, C., Ngu, R., & Leenhardt, L. (2017). European Thyroid Association guidelines for ultrasound malignancy risk stratification of thyroid nodules in adults: The EU-TIRADS. *European Thyroid Journal*, *6*(5), 225–237.

Sajisevi, M., Caulley, L., Eskander, A., Du, Y. J., Auh, E., Karabachev, A., ... Davies, L. (2022). Evaluating the rising incidence of thyroid cancer and thyroid nodule detection modes: A multinational, multi-institutional analysis. *JAMA Otolaryngol Head Neck Surg*, *148*(9), 811-818.

Sehgal, C., Brown, G., Bahn, R., & Greenleaf, J. (1986). Measurement and use of acoustic nonlinearity and sound speed to estimate composition of excised livers. *Ultrasound in Medicine & Biology*, *12*(11), 865-874.

Settles, B. (2009). *Active Learning Literature Survey* (Computer Sciences Technical Report N^{o} 1648). University of Wisconsin–Madison.

Shen, X., Ouyang, X., Liu, T., & Shen, D. (2021). Cascaded networks for thyroid nodule diagnosis from ultrasound images. *Segmentation, Classification, and Registration of Multi-modality Medical Imaging Data*, 145–154.

Shingare, A., Maldar, A. N., Chauhan, P. H., & Wadhwani, R. (2023, 12). Use of ultrasound elastography in differentiating benign from malignant thyroid nodules: A prospective study. *Journal of Diabetes & Metabolic Disorders*, *22*(2), 1245-1253.

Shui, C., Zhou, F., Gagné, C., & Wang, B. (2020). Deep active learning: Unified and principled method for query and training. In S. Chiappa & R. Calandra (Eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics* (Vol. 108, pp. 1308–1318). PMLR.

Smailagic, A., Costa, P., Noh, H. Y., Walawalkar, D., Khandelwal, K., Galdran, A., ... et al. (2018). MedAL: Accurate and robust deep active learning for medical image analysis. *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*.

Solymosi, T., Hegedűs, L., Bonnema, S. J., Frasoldati, A., Jambor, L., Karanyi, Z., ... Nagy, E. V. (2023). Considerable interobserver variation calls for unambiguous definitions of thyroid nodule ultrasound characteristics. *European Thyroid Journal*, *12*(2), e220134.

Spanhol, F. A., Oliveira, L. S., Petitjean, C., & Heutte, L. (2016). A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering*, *63*(7), 1455-1462.

Sreedhar, H., Lajoinie, G. P. R., Raffaelli, C., & Delingette, H. (2024). Active learning strategies on a real-world thyroid ultrasound dataset. In Y. Xue, C. Chen, C. Chen, L. Zuo, & Y. Liu (Eds.), *Data Augmentation, Labelling, and Imperfections Workshop* (pp. 127–136). Cham : Springer Nature Switzerland.

Szczepanek-Parulska, E., Wolinski, K., Dobruch-Sobczak, K., Antosik, P., Ostalowska, A., Krauze, A., ... et al. (2020). S-Detect software vs. EU-TIRADS classification: A dual-center validation of diagnostic performance in differentiation of thyroid nodules. *Journal of Clinical Medicine*, *9*(8), 2495.

Tang, Z., & Ma, J. (2021). Coarse to fine ensemble network for thyroid nodule segmentation. In N. Shusharina, M. P. Heinrich, & R. Huang (Eds.), *Segmentation, Classification, and Registration of Multi-modality Medical Imaging Data* (pp. 122–128). Cham : Springer International Publishing.

Tessler, F. N., Middleton, W. D., Grant, E. G., Hoang, J. K., Berland, L. L., Teefey, S. A., ... et al. (2017). ACR Thyroid Imaging, Reporting and Data System (TI-RADS): White paper of the ACR TI-RADS committee. *Journal of the American College of Radiology*, *14*(5), 587–595.

Tjotta, J. N., & Tjotta, S. (1981). Nonlinear equations of acoustics, with application to parametric acoustic arrays. *The Journal of the Acoustical Society of America*, *69*(6), 1644-1652.

Toulemonde, M., Varray, F., Bernard, A., Basset, O., & Cachard, C. (2015). Nonlinearity parameter B/A of biological tissue ultrasound imaging in echo mode. *AIP Conference Proceedings*, *1685*(1), 040016.

Treeby, B. E., & Cox, B. T. (2010). Modeling power law absorption and dispersion for acoustic propagation using the fractional Laplacian. *The Journal of the Acoustical Society of America*, *127*(5), 2741-2248.

Treeby, B. E., Jaros, J., Rendell, A. P., & Cox, B. T. (2012). Modeling nonlinear ultrasound propagation in heterogeneous media with power law absorption using a k-space pseudospectral method. *The Journal of the Acoustical Society of America*, *131*(6), 4324-4336.

Uppal, N., Collins, R., & James, B. (2023). Thyroid nodules: Global, economic, and personal burdens. *Front. Endocrinol.*, *14*.

Wang, D., & Shang, Y. (2014). A new active labeling method for deep learning. In *2014 International Joint Conference on Neural Networks (IJCNN)* (p. 112-119).

Wang, M., Yuan, C., Wu, D., Zeng, Y., Zhong, S., & Qiu, W. (2021). Automatic segmentation and classification of thyroid nodules in ultrasound images with convolutional neural networks. *Segmentation, Classification, and Registration of Multi-modality Medical Imaging Data Lecture Notes in Computer Science*, 109–115.

Wei, Q., Zeng, S.-E., Wang, L.-P., Yan, Y.-J., Wang, T., Xu, J.-W., ... et al. (2020). The value of S-Detect in improving the diagnostic performance of radiologists for the differential diagnosis of thyroid nodules. *Medical Ultrasonography*, *22*(4), 415–423.

Westervelt, P. J. (1963, 04). Parametric acoustic array. *The Journal of the Acoustical Society of America*, *35*(4), 535-537.

Wu, M.-H., Chen, C.-N., Chen, K.-Y., Ho, M.-C., Tai, H.-C., Wang, Y.-H., ... Chang, K.-J. (2016). Quantitative analysis of echogenicity for patients with thyroid nodules. *Scientific Reports*, *6*.

Xu, W., Jia, X., Mei, Z., Gu, X., Lu, Y., Fu, C.-C., ... Zhou, J. a. (2023). Generalizability and diagnostic performance of AI models for thyroid US. *Radiology*, *307*(5), e221157.

Yamaguchi, T. (2021). Basic concept and clinical applications of quantitative ultrasound (QUS) technologies. *Journal of Medical Ultrasonics*, 1–12.

Yang, L., Zhang, Y., Chen, J., Zhang, S., & Chen, D. Z. (2017). Suggestive annotation: A deep active learning framework for biomedical image segmentation. *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017 Lecture Notes in Computer Science*, 399–407.

Ye, F.-Y., Lyu, G.-R., Li, S.-Q., You, J.-H., Wang, K.-J., Cai, M.-L., & Su, Q.-C. (2021). Diagnostic performance of ultrasound computer-aided diagnosis software compared with that of radiologists with different levels of expertise for thyroid malignancy: A multicenter prospective study. *Ultrasound in Medicine & Biology*, *47*(1), 114–124.

Zhan, X., Wang, Q., Huang, K.-h., Xiong, H., Dou, D., & Chan, A. B. (2022). A comparative survey of deep active learning. *arXiv preprint: 2203.13450*.

Zhang, Y., Lai, H., & Yang, W. (2021). Cascade UNet and CH-UNet for thyroid nodule segmentation and benign and malignant classification. *Segmentation, Classification, and Registration of Multi-modality Medical Imaging Data*, 129–134.

Zhou, Z., Gao, R., Wu, S., Ding, Q., Bin, G., & Tsui, P.-H. (2024). Scatterer size estimation for ultrasound tissue characterization: A survey. *Measurement*, *225*, 114046.

Zhou, Z., Shin, J., Feng, R., Hurst, R. T., Kendall, C. B., & Liang, J. (2019). Integrating active learning and transfer learning for carotid intima-media thickness video interpretation. *Journal of Digital Imaging*, *32*(2), 290–299.

Pages web

Thyroid nodule segmentation and classification in ultrasound images challenge 2020. (2020). https://tn-scui2020.grand-challenge.org/. International Conference on Medical Image Computing & Computer Assisted Intervention.

List of Figures

2.1 Representative B-mode ultrasound images of a thyroid nodule, indicated by the orange square. (Left) Image acquired in an axial view. (Right) Image acquired in a sagittal view. 9

2.2 (Left) Simplified illustration of the axial-view anatomy of the region explored by thyroid ultrasound, based on the image in Figure 2.1. (Right) Illustration of the ultrasound view of the image. The fluid-filled vessels and the air-filled trachea appear anechoic, though artifacts may create the appearance of internal structures. The nearby muscles are hypoechoic compared to the thyroid gland. Due to attenuation, deeper structures cannot be seen, especially at higher ultrasound frequencies. 10

2.3 Illustration of the EU-TIRADS algorithm, as proposed in (Russ et al., 2017). The score of EU-TIRADS 1 corresponds to the absence of a nodule; the other scores stratify the risk of malignancy with different indications for FNA. 11

2.4 Illustration of different nodule compositions. Solid nodules appear to be composed of solid tissue, while cystic lesions are dominated by large, dark, fluid-filled spaces. Some nodules may have similar proportions of solid and cystic components, and be described as being of mixed composition. Spongiform nodules are unique, in that they are composed of many tiny cystic spaces throughout the entire lesion. 12

2.5 Example of an axial-view ultrasound image containing a nodule judged as solid by four expert practitioners. The solid nodules may contain very small cystic spaces. 13

2.6 Example of an axial-view ultrasound image containing a thyroid nodule judged as cystic by four expert practitioners. These nodules are primarily composed of large, fluid-filled spaces. 13

2.7 Example of an axial-view ultrasound image containing a thyroid nodule judged as mixed cystic and solid by four expert practitioners. 13

2.8 Example of an axial-view ultrasound image containing a thyroid nodule judged as spongiform by four expert practitioners. These nodules are composed of numerous tiny cystic spaces. 13

2.9 Illustration of different lesion echogenicities. Hyperechoic nodules are brighter than the surrounding normal thyroid tissue, while isoechoic nodules have a similar level of intensity to their surroundings. Hypoechoic nodules are darker than the surrounding thyroid tissue, while very hypoechoic nodules are even darker than the adjacent muscles. Anechoic nodules are cystic, filled with fluid, and appear dark like the blood vessels near the thyroid. 15

2.10 Example of an axial-view ultrasound image containing a thyroid nodule judged as hyperechoic or isoechoic by four expert practitioners. This description is made relative to the echogenicity of nearby thyroid tissue. 16

2.11 Example of an axial-view ultrasound image containing a thyroid nodule judged as hypoechoic by four expert practitioners. This description is made relative to the echogenicity of nearby thyroid tissue. 16

2.12 Example of an axial-view ultrasound image containing a thyroid nodule judged as very hypoechoic by four expert practitioners. This description is made relative to the echogenicity of nearby muscles. 16
2.13 Example of an axial-view ultrasound image containing a thyroid nodule judged as anechoic by four expert practitioners. This description corresponds to fluid-filled lesions. 16
2.14 Illustration of different nodule shapes, in an axial or transverse view. A wider-than-tall, or oval, shape is defined as a nodule whose anteroposterior diameter is less than its transverse diameter. A taller-than-wide shape, having the opposite ratio of dimensions, is more associated with malignancy (Russ et al., 2017 ; Tessler et al., 2017). 18
2.15 Example of an axial-view ultrasound image containing a thyroid nodule judged as wider than tall by four expert practitioners. 19
2.16 Example of an axial-view ultrasound image containing a thyroid nodule judged as taller than wide by four expert practitioners. 19
2.17 Illustration of different lesion margins. Smooth margins are clearly visible demarcations between the nodule and the surrounding thyroid parenchyma. Ill-defined margins are not readily distinguishable from the thyroid parenchyma. Spiculated and lobulated margins are both considered irregular; the former have sharp, angular protrusions while the latter have smooth, round bumps. Finally, extra-thyroidal extension is an important feature of margins, and describes when a nodule appears to extend beyond the thyroid capsule into adjacent structures. 20
2.18 Example of an axial-view ultrasound image containing a thyroid nodule judged to have a smooth margin by four expert practitioners. 21
2.19 Example of an axial-view ultrasound image containing a thyroid nodule judged to have an ill-defined margin by four expert practitioners. 21
2.20 Example of an axial-view ultrasound image containing a thyroid nodule judged to have an irregular margin by four expert practitioners. 21
2.21 Example of an axial-view ultrasound image containing a thyroid nodule judged to show extra-thyroidal extension by three out of four expert practitioners. 21
2.22 Illustration of different echogenic foci. Punctate echogenic foci are small hyperechoic spots. These may correspond to benign signs such as colloid crystals (which are associated with large comet-tail artifacts) or the back walls of small cysts. However, they may also be associated with microcalcifications, which are associated with malignancy (and do not show large comet-tail artifacts). Macrocalcifications are larger, and generate acoustic shadows behind them. Peripheral or rim or eggshell calcifications are located around the margin of the nodule. (Russ et al., 2017 ; Tessler et al., 2017) 22
2.23 Example of an axial-view ultrasound image containing a thyroid nodule judged to have punctate echogenic foci without significant comet-tail artifacts by three out of four expert practitioners. 23
2.24 Example of an axial-view ultrasound image containing a thyroid nodule judged to have macrocalcifications by three out of four expert practitioners. These generate posterior acoustic shadows because they reflect much of the ultrasound signal. . . 23

2.25 Example of an axial-view ultrasound image containing a thyroid nodule judged to have peripheral calcifications by three out of four expert practitioners. These calcifications can obscure the interior of the nodule. 23
2.26 Evaluation inventory used by the experts. Evaluation began with a subjective evaluation based on an initial impression, followed by an inventory of sonographic features, before culminating in an EU-TIRADS score. 26
2.27 Percentage agreement among the four experts on EU-TIRADS labels for the 303 images. 28
2.28 Mean number of overall labels and strong consensus labels assigned by EU-TIRADS score for the 303 images. The bar on the left for each score represents the average number of times the score was assigned across all four experts. On the right, cases of strong consensus for that score are shown, with the bottom bar representing unanimous consensus. The remaining stacked segments of the right-hand bars are cases on which 3 out-of-four experts agreed on the score, sorted by the expert who was the lone dissenter. 29
2.29 Agreements and disagreements among the four experts on EU-TIRADS labels for the 303 images. The $\binom{4}{2} = 6$ pairwise comparisons for each image yield a total of 1818 agreements or disagreements. 31
2.30 Percentage agreement among the four experts on composition labels for the 303 images. 32
2.31 Mean number of labels and strong consensus labels assigned by composition category for the 303 images. The strong consensus labels also include 3/4 consensus cases sorted by the expert who was the lone dissenter. 33
2.32 The composition disagreements most commonly associated with disagreements in EU-TIRADS label. 34
2.33 Percentage agreement among the four experts on echogenicity labels for the 303 images. 35
2.34 Mean number of labels and strong consensus labels assigned by echogenicity category for the 303 images. The strong consensus labels also include 3/4 consensus cases sorted by the expert who was the lone dissenter. 36
2.35 The echogenicity disagreements most commonly associated with disagreements in EU-TIRADS label. 36
2.36 Percentage agreement among the four experts on shape labels for the 303 images. 37
2.37 Mean number of labels and strong consensus labels assigned by shape category for the 303 images. The strong consensus labels also include 3/4 consensus cases sorted by the expert who was the lone dissenter. 38
2.38 The shape disagreements most commonly associated with disagreements in EU-TIRADS label. 39
2.39 Percentage agreement among the four experts on margin labels for the 303 images. 40
2.40 Mean number of labels and strong consensus labels assigned by margin category for the 303 images. The strong consensus labels also include 3/4 consensus cases sorted by the expert who was the lone dissenter. 41
2.41 Percentage agreement among the four experts on echogenic foci labels for the 303 images. 43

2.42 Mean number of labels and strong consensus labels assigned for punctuate echogenic foci for the 303 images. The strong consensus labels also include 3/4 consensus cases sorted by the expert who was the lone dissenter. 43
2.43 Mean number of labels and strong consensus labels assigned for peripheral calcifications for the 303 images. The strong consensus labels also include 3/4 consensus cases sorted by the expert who was the lone dissenter. 44
2.44 Mean number of labels and strong consensus labels assigned for macrocalcifications for the 303 images. The strong consensus labels also include 3/4 consensus cases sorted by the expert who was the lone dissenter. 44
2.45 The echogenic foci disagreements most commonly associated with disagreements in EU-TIRADS label. PEF - Punctate echogenic foci. 45
2.46 The frequency of use of repeated combinations (used by the same expert on multiple images) of sonographic feature labels by each expert presented as a boxplot. The total number of unique evaluations per expert and the total number of repeated combinations per expert are presented as well. 47
2.47 The number of unique feature combinations used by each expert are presented on the left, along with the subset of combinations that were reused for multiple images. These are further filtered down to combinations of feature labels for which the expert did not always apply the same EU-TIRADS score in all cases. Finally, the images corresponding to these categories are separated into those which were labeled in accordance with the majority EU-TIRADS score for their feature combination, and the smaller subset that differed. This latter subset represents potential inconsistencies in EU-TIRADS score attribution on the basis of sonographic feature labels. 48
2.48 The shared combinations of echographic features used by all four experts. The combinations for which the four experts were not in unanimous agreement about the EU-TIRADS score most often assigned is shown, along with the small number of cases across all shared combinations that disagreed from the majority EU-TIRADS score. 49
2.49 The confusion matrices of the four experts' EU-TIRADS scores, as compared with the guideline-based EU-TIRADs score calculated from the sonographic feature inventory. It was not possible to calculate an EU-TIRADS score for some images with indeterminate composition or echogenicity, thus totals are not the same across all experts. 50
2.50 Expert 1's departures from a guideline EU-TIRADS algorithm. The default nodule labels are given at the top, with decisions about labels proceeding downwards. On the left are listed the scores that were assigned in accordance with the guideline, while departures from that guideline are in red on the right. 52
2.51 Expert 2's departures from a guideline EU-TIRADS algorithm. The default nodule labels are given at the top, with decisions about labels proceeding downwards. On the left are listed the scores that were assigned in accordance with the guideline, while departures from that guideline are in red on the right. 53
2.52 Expert 3's departures from a guideline EU-TIRADS algorithm. The default nodule labels are given at the top, with decisions about labels proceeding downwards. On the left are listed the scores that were assigned in accordance with the guideline, while departures from that guideline are in red on the right. 55

2.53 The top half of Expert 4's departures from a guideline EU-TIRADS algorithm. The default nodule labels are given at the top, with decisions about labels proceeding downwards. On the left are listed the scores that were assigned in accordance with the guideline, while departures from that guideline are in red on the right. Continues in Figure 2.54. 56
2.54 The bottom half of Expert 4's departures from a guideline EU-TIRADS algorithm. The default nodule labels are given at the top, with decisions about labels proceeding downwards. On the left are listed the scores that were assigned in accordance with the guideline, while departures from that guideline are in red on the right. Continues from a negative Taller-than-Wide determination in Figure 2.53. 57

3.1 Summary of different automated evaluation tasks for thyroid nodule ultrasound. Detection and localization of nodules within the thyroid gland go together. This is often a prerequisite for the task of nodule segmentation, identifying the border of the nodule within the image; this can be useful for characterizing nodule size and margin properties. Finally, nodule characterization can be a prediction of the risk of malignancy, or a classification according to various sonographic features. . . . 68
3.2 Example images for which the experts were not in agreement on the echogenicity description. 72
3.3 (Left) Example of an axial-view image with a nodule labeled as hyper-/isoechoic by all four experts. (Right) Example of an axial-view image with a nodule labeled as hypoechoic by all four experts. 73
3.4 (Top) Sample partition for cross-validation training and testing of ResNet50 for binary classification between hyper-/isoechoic and hypoechoic nodules. Training partitions were randomized with proportionate representations of both classes. (Bottom) Images for training within 3-fold cross validation were treated with data augmentation techniques, and used with class labels based on expert consensus with a ResNet50 architecture. 74
3.5 Examples of predictions on images, with the reference and predicted values for the hyper-/isoechoic labels listed. 76
3.6 Process followed for the consensus meetings. All participants secretly re-evaluated the image. A discussion among all four experts then proceeded sequentially along areas of the evaluation for which they did not reach a perfect consensus: composition, echogenicity, shape, margin, echogenic foci, and EU-TIRADS score. Finally, the images were re-evaluated again, secretly. 77
3.7 Example image for which three out of four experts changed their assigned echogenicity label. 79
3.8 Illustration of nodule echogenicity labels, which are determined by comparing the brightness or intensity of a nodule to that of a nearby reference area. Hyperechoic nodules are brighter than the surrounding normal thyroid parenchyma, while isoechoic nodules have a similar level of intensity to their surroundings. Hypoechoic nodules, by contrast, are darker than the surrounding thyroid parenchyma. 80

3.9 (Left) Illustration of ultrasound image with masks selecting the nodule region and available healthy thyroid parenchyma in the ipsilateral lobe as a reference area. These zones were filtered with pixel intensity thresholds to exclude cystic areas and echogenic foci such as microcalcifications. (Right) Illustration of pixel intensity plots being considered; for each image, the distributions of nodule pixel intensity were compared between the nodule and reference area to look for differences associated with the expert echogenicity labels. 82
3.10 (Top) The entire image and the zones selected as nodule and reference regions. (Bottom) The pixel intensity distributions for the nodule and reference regions. In this case, the nodule zone's distribution seems to have lower pixel intensities than the reference zone, corresponding with the experts' unanimous hypoechoic label. 82
3.11 (Left) Example of an image in which the difference between the mean pixel intensities in the nodule and reference regions is on the lower end of the observed range, denoting a more hypoechoic nodule. (Right) Example of an image in which the difference between the mean pixel intensities between the same regions is at the higher end of the observed range, denoting a hyper-/isoechoic nodule. 83
3.12 Violin plot of mean pixel intensity distribution differences between the nodule and thyroid parenchyma, separated by the proportion of experts who assigned a hyper-/isoechoic or hypoechoic label. Median values of the distributions are indicated. In some cases, one expert applied a different echogenicity label, so the experts are counted out of three rather than four. 84
3.13 (Left) Example of an image in which the standard deviation of pixel intensities in the nodule is on the lower end of the observed range, denoting a more homogeneous nodule in terms of echogenicity. (Right) Example of an image in which the standard deviation of pixel intensities in the nodule is greater, denoting a more heterogenous nodule in terms of echogenicity. 85
3.14 Violin plot of the standard deviations of pixel intensity distributions of nodule regions within ultrasound images, separated by the proportion of experts who assigned a hyper-/isoechoic or hypoechoic label. Median values of the distributions are indicated. In some cases, one expert applied a different echogenicity label, so the experts are counted out of three rather than four. 85
3.15 (Left) Example of an image in which the standard deviation of pixel intensities in the reference zone is on the lower end of the observed range, denoting a more homogeneous echogenicity. (Right) Example of an image in which the standard deviation of pixel intensities in the reference is greater, denoting a more heterogeneous echogenicity. 86
3.16 Violin plot of the standard deviations of pixel intensity distributions of reference tissue regions within ultrasound images, separated by the proportion of experts who assigned a hyper-/isoechoic or hypoechoic label. Median values of the distributions are indicated. In some cases, one expert applied a different echogenicity label, so the experts are counted out of three rather than four. 86
3.17 (Left) Example of an image in which the ratio of reference pixels to nodule pixels is small. Note that other nodules present are not included in the area of the reference zone. (Right) Example of an image in which the ratio is much larger. . 87

3.18 Violin plot of the ratio of number of pixels within the reference region to the number within the nodule region, separated by the proportion of experts who assigned a hyper-/isoechoic or hypoechoic label. Median values of the distributions are indicated. In some cases, one expert applied a different echogenicity label, so the experts are counted out of three rather than four. An outlier is excluded from the all experts hypoechoic group. 88

4.1 The basic cycle of pool-based active learning: an initial set of images is randomly chosen for annotation, and used for training. In subsequent iterations, further images are chosen for annotation from the unlabeled image pool to retrain the algorithm. The unlabeled images can also be used for semi-supervised strategies. 94
4.2 The two main categories of active learning criteria: uncertainty and diversity. Uncertainty sampling chooses cases for which classification is difficult, and may select a subset of images, such as those that are similar between two classes. Diversity sampling attempts to represent more varied samples. 95
4.3 Example ultrasound images from the dataset, with and without nodules, annotated by the non-expert reader. 97
4.4 Median AUC values for different active learning strategies with supervised learning on the ultrasound dataset. 101
4.5 Median AUC values for different active learning strategies with supervised learning on the PneumoniaMNIST dataset. 102
4.6 Median AUC values for different active learning strategies with supervised learning on the BreaKHis dataset. 102
4.7 Median AUC values for different active learning strategies with semisupervised learning on the ultrasound dataset. 104
4.8 Median AUC values for different active learning strategies with semisupervised learning on the PneumoniaMNIST dataset. 104
4.9 Median AUC values for different active learning strategies with semisupervised learning on the BreaKHis dataset. 105
4.10 Violin plots of classification AUC values on the at different label budgets with the rigged draw strategy at $\alpha = 25$ on the ultrasound dataset. 107
4.11 Violin plots of classification AUC values on the at different label budgets with the rigged draw strategy at $\alpha = 25$ on the PneumoniaMNIST dataset. 107
4.12 Violin plots of classification AUC values on the at different label budgets with the rigged draw strategy at $\alpha = 25$ on the BreaKHis dataset. 108

5.1 Experimental setup to compare the P4-1 probe to the simulation: a water bath, with a needle hydrophone placed inside. The probe placed on the side of the tank emitted a pulse through a thin membrane into the tank, where it could be measured by the hydrophone at different positions. This generated a map of pressure recordings over time in a plane in front of the probe. 124

5.2 Comparison of between experimental and simulation data with the P4-1 probe. (A) Measured pressure map at the fundamental frequency. (B) Simulated pressure map at the fundamental frequency. (C) Measured pressure map at the second harmonic frequency. D) Simulated pressure map at the second harmonic frequency. (E) Measured fundamental (blue) and second harmonic (red) pressure profiles along the central propagation axis. (F) Simulated fundamental (blue) and second harmonic (red) pressure profiles along the central propagation axis. 125
5.3 Plots of the simulated forward-propagating signals in a homogeneous tissue-mimicking medium with $\frac{B}{A} = 11$ and $\alpha_{dB} = 0.3\ \mathrm{dB} \cdot \mathrm{MHz}^{-1} \cdot \mathrm{cm}^{-1}$. (A): The time-domain pressure generated 4 cm into the tissue by a pulse programmed for 2.5 cycles with a frequency of 1.4 MHz. (B): Time-domain pressure generated 4 cm into the tissue by a pulse programmed for 5 cycles with a frequency of 3.85 MHz. (C): Fourier transforms of the pressure signals in A (blue) and B (red). . . 126
5.4 Partition of simulation data between test and cross-validation groups. The proportions of simulations with and without lesions were preserved in the partition. The neural networks were trained in 4-fold cross validation, and the mean prediction of the four networks was evaluated on the test set. 129
5.5 Training strategy for $\frac{B}{A}$ and α prediction. The RF signals from the k Wave simulations were used as input data, being combined into an input signals for simple multi-layer perceptron with Leaky ReLU activation functions. The predicted mean $\frac{B}{A}$ value was compared with a smooth L1 loss to the mean value from the simulated medium. 130
5.6 (A) Scatter plot of reference $\frac{B}{A}$ values (x axis) vs. the average predictions (y axis) made by the ensemble of trained networks on the test set. The line in red indicates a perfect match between prediction and reference values. (B) Box plot of the errors in test set $\frac{B}{A}$ predictions, grouped by reference $\frac{B}{A}$ value. The horizontal line represents zero error, i.e. a perfect prediction. The mean value of each group is shown as a red line through the box plot, and outliers are presented as red crosses. 133
5.7 Scatter plot of reference $\frac{B}{A}$ values (x axis) vs. the predicted values (y axis) made by the neural network strategy (in blue) and by Equation 5.23 (in red) on the test set. The equation results were also corrected by their mean error on the test set for comparison (in yellow). 134
5.8 (A) Scatter plot of reference α values in $\mathrm{dB} \cdot \mathrm{MHz}^{-1} \cdot \mathrm{cm}^{-1}$ (x axis) vs. the average predictions (y axis) made by the ensemble of trained networks on the test set. The line in red indicates a perfect match between prediction and reference values. (B) Box plot of the errors in test set α predictions, grouped by reference α value. The horizontal line represents zero error, i.e. a perfect prediction. The mean value of each group is shown as a red line through the box plot, and outliers are presented as red crosses. 136

List of Tables

2.2 Total labels assigned by each expert for each EU-TIRADS category. The mean value across all experts is presented in the final row. 28
2.3 Fleiss' kappa scores among the four experts for each subcategory of the sonographic feature inventory. The number of possible classes for each category from the sonographic feature inventory is also provided. 30
2.4 Total labels assigned by each expert for each composition category. The mean value across all experts is presented in the final row. 32
2.5 Total labels assigned by each expert for each echogenicity category. The mean value across all experts is presented in the final row. 35
2.6 Total labels assigned by each expert for the two shape categories. The mean value across all experts is presented in the final row. 37
2.7 Total labels assigned by each expert for each margin category. The mean value across all experts is presented in the final row. 39
2.8 Total labels assigned by each expert for each echogenic foci category. The mean value across all experts is presented in the final row. 42
2.9 Expert differences from guideline-based EU-TIRADS score, calculated on the basis of expert-assigned features. Not all images could be used to calculate a guideline-based EU-TIRADS score, if the composition or echogenicity had been assigned a "Cannot Determine" label. 51

3.1 AUROC and test set composition for the binary classification network across different sample partitions. 75
3.2 Intra-expert variability in echogenicity label. The first and second labels assigned individually by experts are listed, with cases of altered labels during the re-evaluation being highlighted in orange. 78

4.1 Active learning test sizes for each dataset. 99
4.2 Active learning test repetitions with a different initial set for each strategy and dataset. 100
4.3 Supervised learning AUBC values. Values closer to 1 indicate a more effective strategy. Rand = Random. LC = Least Certain. KM = KMeans. RD = Rigged Draw (ours) with $\alpha = 25$. 103
4.4 Supervised learning AUBC values for different Rigged Draw weights, with * indicating p-values < 0.05 when compared to random selection. 103
4.5 Semi-supervised learning AUBC values. Values closer to 1 indicate a more effective strategy, with * indicating p-values < 0.05 when compared to random selection. Rand = Random. LC = Least Certain. KM = KMeans. RD = Rigged Draw (ours) with $\alpha = 25$. 105
4.6 Semi-supervised learning AUBC values for different Rigged Draw weights, with * indicating p-values < 0.05 when compared to random selection. 106

5.1 Values of the nonlinear parameter in various biological media reported in the literature. The values of $\frac{B}{A}$ represent the range of values in the cited sources rounded to the nearest tenth. 118
5.2 Parameters used for the k-Wave simulations. * A nearly linear power-law model was used for attenuation. 128
5.3 Means and standard deviation of $\frac{B}{A}$ prediction errors on the test set. 132

Appendix

Appendices

A Appendix I: Definition of B/A

The nonlinear parameter $\frac{B}{A}$ comes from the definitions of the first and second order coefficients in the Taylor expansion of the adiabatic state equation between the pressure and density of the propagation medium

$$P = P_0 + \rho_0 \left(\frac{\partial P}{\partial \rho}\right)_{0,s} \left(\frac{\rho - \rho_0}{\rho_0}\right) + \frac{\rho_0^2}{2} \left(\frac{\partial^2 P}{\partial \rho^2}\right)_{0,s} \left(\frac{\rho - \rho_0}{\rho_0}\right)^2 + ... \tag{A.1}$$

where P represents the absolute local pressure, P_0 is the baseline hydrostatic pressure, ρ corresponds to the density of the medium, and the subscripts $0, s$ signify that the partial derivatives are evaluated at the equilibrium density $\rho = \rho_0$ and with constant entropy. This can be rewritten as

$$p = \rho_0 \left(\frac{\partial P}{\partial \rho}\right)_{0,s} \left(\frac{\rho - \rho_0}{\rho_0}\right) + \frac{\rho_0^2}{2} \left(\frac{\partial^2 P}{\partial \rho^2}\right)_{0,s} \left(\frac{\rho - \rho_0}{\rho_0}\right)^2 + ... \tag{A.2}$$

, in which $p = P - P_0$ is the local pressure variation from baseline. The first and second order terms can be written as

$$p = A \left(\frac{\rho - \rho_0}{\rho_0}\right) + \frac{B}{2} \left(\frac{\rho - \rho_0}{\rho_0}\right)^2 + ... \tag{A.3}$$

where

$$A = \rho_0 \left(\frac{\partial P}{\partial \rho}\right)_{0,s} \tag{A.4}$$

and

$$B = \rho_0^2 \left(\frac{\partial^2 P}{\partial \rho^2}\right)_{0,s} \tag{A.5}$$

Using the relationship

$$\frac{\partial P}{\partial \rho_0} = c_0^2, \tag{A.6}$$

we can simplify Equation A.4 to

$$A = \rho_0 c_0^2, \tag{A.7}$$

which in turn means that

$$\frac{B}{A} = \frac{\rho_0^2 \left(\frac{\partial^2 P}{\partial \rho^2}\right)_{0,s}}{\rho_0 c_0^2} = \frac{\rho_0}{c_0^2} \left(\frac{\partial^2 P}{\partial \rho^2}\right)_{0,s} \tag{A.8}$$

Substituting in once more the relationship from Equation A.6 yields:

$$\frac{B}{A} = \frac{\rho_0}{c_0^2} \left(\frac{\partial c^2}{\partial \rho}\right)_{0,s} \tag{A.9}$$

Then applying the chain rule since c is a function of p which is in turn a function of ρ, this yields

$$\begin{aligned}
\frac{B}{A} &= \frac{\rho_0}{c_0^2}\left(2c_0\frac{\partial c}{\partial \rho}\right)_{0,s} \\
&= \frac{2\rho_0}{c_0}\left(\frac{\partial c}{\partial \rho}\right)_{0,s} \\
&= \frac{2\rho_0}{c_0}\left(\frac{\partial c}{\partial P}\right)_{0,s}\left(\frac{\partial P}{\partial \rho}\right)_{0,s} \\
&= \frac{2\rho_0}{c_0}\left(\frac{\partial c}{\partial P}\right)_{0,s}\left(c_0^2\right) \\
&= 2\rho_0 c_0\left(\frac{\partial c}{\partial P}\right)_{0,s}
\end{aligned} \tag{A.10}$$

Thyrosonics

L'apprentissage automatique pour la détection et classification des nodules thyroïdiens dans les images échographiques

Hari SREEDHAR

Résumé

L'échographie est une technique indispensable pour l'évaluation du risque de malignité des nodules thyroïdiennes. Malgré son utilité, l'échographie thyroïdienne reste limitée par sa dépendance à l'expérience de l'opérateur, autant pour l'acquisition que pour l'interprétation. C'est pourquoi des algorithmes d'apprentissage automatique, ayant connu de grands succès sur des images naturelles et médicales, ont été proposés aussi pour l'interprétation des images échographiques thyroïdiennes.

L'intérêt suscité dans ce domaine par la promesse de l'IA a mené à un grand nombre de publications proposant des algorithmes pour la détection, segmentation, et classification de nodules, ainsi qu'à la création de plusieurs produits commerciaux pour la pratique clinique. Malgré tous ces outils, l'impact réel sur la pratique des endocrinologues et radiologues français reste faible ; cette limitation correspond dans une large mesure au fait que la majorité de ces algorithmes ne prennent pas en compte le contexte clinique de l'échographie thyroïdienne en France.

L'objet de cette thèse est donc d'explorer les particularités de l'échographie thyroïdienne en France, afin d'identifier les possibles pistes d'amélioration en utilisant les méthodes de l'apprentissage automatique.

Le premier chapitre consiste à examiner la variabilité inter-expert en évaluation de l'échographie thyroïdienne. Une étude multicentrique utilisant des images échographiques acquises au fil de l'eau de la pratique clinique de quatre experts français donnent une indication des points de difficulté pour les médecins. Les résultats permettent d'identifier les caractéristiques échographiques des nodules thyroïdiens dont la description génère des différences significatives entre les praticiens, et entraîne des conséquences sur la prise en charge des patients.

Le deuxième chapitre entre plus dans le détail de l'une des caractéristiques échographiques utilisées par les experts : l'échogénicité. En continuité du chapitre précédent, la possibilité de se servir d'un outil d'apprentissage automatique pour aider les praticiens non-experts à distinguer entre des nodules hyper-/isoéchogènes et nodules hypoéchogènes est explorée. Ensuite, les différences quantitatives entre les images sont étudiees pour évaluer la robustesse de la vérité terrain, et la reproductibilité de l'examen échographique.

Le troisième chapitre s'intéresse à la difficulté d'obtenir des annotations expertes pour l'entraînement et le raffinement d'algorithmes d'apprentissage automatique en échographie thyroïdienne. Á partir des résultats précédents, il est évident que l'obtention d'un consensus sur les étiquettes des experts pour entraîner des algorithmes demanderait un temps considérable. Afin de réduire ce coût pour le développement des algorithmes, des stratégies d'apprentissage actif pour entraîner des réseaux de neurones avec moins d'annotations sont explorées. Ce chapitre présente les limitations de ces stratégies sur des vraies données cliniques, et propose aussi une technique d'apprentissage actif qui mélange des critères de sélection classiques avec la représentativité de l'échantillonnage au hasard.

Le dernier chapitre explore l'échographie quantitative comme piste future pour améliorer l'évaluation des nodules thyroïdiens. En utilisant des simulations numériques de tissus mous et d'une vraie sonde échographique, des réseaux de neurones sont entrainés pour estimer le paramètre non linéaire d'un milieu de propagation à partir du signal brut reçu au niveau de la sonde. La stratégie utilise une combinaison de pulses pour créer un signal plus apte à être traité par le réseau. Les contributions de cette thèse cherchent à mieux contextualiser l'utilisation de l'apprentissage automatique dans l'échographie thyroïdienne, afin de permettre ces techniques d'avancer vers des applications ayant un vrai impact durable sur la pratique clinique.