

UCLA

UCLA Electronic Theses and Dissertations

Title

Deep Learning for Ultrasound Imaging: From Task-Specific Architectures to Foundation Models

Permalink

<https://escholarship.org/uc/item/6p88n9vs>

ISBN

9798189299292

Author

Radhachandran, Ashwath

Publication Date

2026-08-31

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Deep Learning for Ultrasound Imaging: From Task-Specific Architectures to Foundation
Models

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Bioengineering

by

Ashwath Radhachandran

ABSTRACT OF THE DISSERTATION

Deep Learning for Ultrasound Imaging: From Task-Specific Architectures to Foundation
Models

by

Ashwath Radhachandran

Doctor of Philosophy in Bioengineering

University of California, Los Angeles, 2026

Professor William F. Speier, Co-Chair

Professor Corey W. Arnold, Co-Chair

This dissertation investigates the application of deep learning to clinical ultrasound imaging, progressing from task-specific models for thyroid nodule diagnostics to large-scale self-supervised foundation models for ultrasound. The first part begins with a systematic review of AI methods in the thyroid nodule diagnostic pipeline, identifying problematic gaps in benchmarking rigor and external validation that serve as a driver for following work. Building on those findings, we develop a joint detection and segmentation framework that addresses nodule localization in a multitask model, and ThyGraph, a graph-based architecture for multi-view feature aggregation across an ultrasound image study to improve study-level malignancy classification. These task-specific approaches demonstrate strong performance, but external validation reveals a generalization gap and supervised, single-task training demonstrates representational limits. These observations motivate the second part of the dissertation, which proposes self-supervised pretraining as a path toward more informative ultrasound representations. We develop US-JEPA, a Joint Embedding Predictive Architecture adapted for ultrasound. In parallel, we curate the Open Ultrasound (OPUS) dataset, the largest public ultrasound corpus assembled to date, for model pretraining. The model achieves state-of-the-art performance on various classification tasks, and outperforms

competing models in low-label regimes and under test set image corruption. Finally, we introduce UltraBench 2.0, a standardized benchmark and evaluation protocol for ultrasound foundation models. Dataset coverage spans multiple anatomies and clinical tasks, addressing the absence of a common framework that has made meaningful evaluation across models difficult. Collectively, these contributions outline a grounded transition from task-specific deep learning to foundation model development and rigorous benchmarking, establishing both practical tools and conceptual standards for future ultrasound-based AI methods.

The dissertation of Ashwath Radhachandran is approved.

Michael W. Yeh

Holden H. Wu

William Hsu

Corey W. Arnold, Committee Co-Chair

William F. Speier, Committee Co-Chair

University of California, Los Angeles

2026

To my parents, whose silent sacrifices made this possible

Contents

Abstract	ii
List of Figures	viii
Acknowledgements	xi
1 Introduction	1
1.1 Clinical Motivation	2
1.2 Part 1: Task-Specific Deep Learning for Thyroid Ultrasound	4
1.3 Pivoting From Task-Specific to Foundation Models	5
1.4 Part 2: Self-Supervised Learning and the Ultrasound Foundation Model . . .	6
1.5 Contributions	7
1.6 Outline of the Dissertation	9
2 Background	11
2.1 Ultrasound Imaging	11
2.2 Thyroid Nodules and the Diagnostic Pipeline	14
2.3 Deep Learning for Medical Image Analysis	18
2.4 Self-Supervised Learning	24
3 From Bench to Bedside: How Artificial Intelligence Is Changing Thyroid Nodule Diagnostics	32
3.1 Introduction	33
3.2 Materials and Methods	34
3.3 Results	35
3.4 Discussion	42
4 A Multitask Approach for Automated Detection and Segmentation of Thyroid Nodules in Ultrasound Images	55
4.1 Introduction	55
4.2 Methods	58
4.3 Experiments	63
4.4 Results	66
4.5 Discussion	72
5 ThyGraph: A Graph-Based Approach for Thyroid Nodule Diagnosis from Ultrasound Studies	75

5.1	Introduction	76
5.2	Methods	78
5.3	Experiments	82
5.4	Results	83
5.5	Discussion	85
6	US-JEPA: A Joint Embedding Predictive Architecture for Ultrasound	87
6.1	Background	88
6.2	Introduction	92
6.3	Related Work	96
6.4	Preliminaries	98
6.5	US-JEPA	99
6.6	Experiments and Results	103
6.7	Discussion	113
6.8	Appendix	115
7	UltraBench 2.0: A Standardized Benchmark for Ultrasound Foundation Model Evaluation	132
7.1	Introduction	132
7.2	Related Work	134
7.3	Investigating “Saturation” in Downstream Tasks	141
7.4	Tasks	143
7.5	The UltraBench 2.0 Framework	144
7.6	Evaluations of Existing Models	146
7.7	Results	150
7.8	Future Directions	157
7.9	Discussion	158
8	Conclusions	161
8.1	Summary	161
8.2	Future Work	163
	Bibliography	165

List of Figures

2.1	A simple illustration of thyroid anatomy [noa11].	14
2.2	ACR TI-RADS scoring calculator [AvdL].	16
2.3	Linear probing versus full fine-tuning evaluation protocols. Adapted from [HCX ⁺ 21].	29
3.1	Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) Flow Diagram. Meta-analysis was not performed as part of this review. . . .	34
3.2	Intersection of study designs. This Venn diagram quantifies the different combinations of study designs captured in this review.	36
3.3	Clinical workflow with resultant data inputs and predicted outcomes. A diagnostic pipeline was abstracted from the papers starting with known clinical steps and culminating in five high-level predicted outcomes. If a certain input modality was included in a paper for one of the predicted outcomes, the corresponding data collection box had an arrow pointing out that outcome. .	36
3.4	Distribution of data inputs used to predict clinical outcomes. Each paper was categorized by data input and predicted clinical outcome. If a paper was multi-input or multi-outcome, it was counted for each input-outcome combination.	37
4.1	Example of an image with corresponding annotation mask from each dataset.	59
4.2	One example of AD being integrated with a segmentation architecture. This is an overview of the model architecture for MSUNet-AD, with an in-depth look at the operations in each multi-scale block (bottom left) and in the AD module (bottom right).	61
4.3	Box plot comparing segmentation performance by nodule size quartile between models with and without AD module. The plot represents the distribution of the DSC values (used to calculate DSC_+) for each quartile. A black asterisk above a comparison represents a statistically significant difference. The white crosshair marks the mean of the distribution.	68
4.4	Image-wide recall was calculated for each model for each of the nodule size quartiles. The difference in recall between a model without and with AD is plotted for each quartile.	69

4.5	Examples of the difference in segmentation performance between three of the models evaluated. First row: All models without the AD module conflate a carotid artery as a nodule leading to a false positive. When the AD module is added, none of the models contain a prediction since the AD output correctly eliminates any predicted segmentations. Second row: An example of all models except MSUNet struggling to segment a small nodule. Third row: An example of a smaller nodule where all three models and their counterparts with AD exhibit agreement and successfully segment the target. Fourth row: An example of all six models successfully segmenting the larger of two nodules. Fifth row: A relatively difficult example with three different small nodules. The majority of models consistently miss the far right nodule. Here, MSUNet-AD and SResUNet-AD perform at par with their non-AD baselines.	70
4.6	F1 performance across AD thresholds for the six models with AD. Dot marker represents the optimal threshold, where the best validation set F1 score is achieved for a given model.	71
5.1	Various spatial orientations and locations captured by frames within a thyroid US study. The two example images are labeled with superior (S), inferior (I), medial (M) and lateral (L) to help the viewer orient each image within the context of the 3-D coordinate system.	81
5.2	Overview of proposed graph construction and model architecture.	83
5.3	Visualization of attention heatmap overlaid on a thyroid mesh model for two patients. The attention weights are scaled between 0 and 1 for each patient. The biopsied nodule location, extracted from corresponding cytology report, is marked with a green X. As indicated, the x-axis stretches from right to left, the y-axis from inferior to superior and the z-axis from posterior to anterior.	85
6.1	SSL pretext task architectures illustrated on a lung image. (a) MAE reconstructs masked patches in pixel space. (b) I-JEPA predicts masked patch representations in latent space from visible context. (c) DINOv3 aligns student and teacher representations across local and global crops.	89
6.2	USrc-JEPA framework. URFM is the frozen teacher that extracts target embeddings. The student and predictor are jointly optimized with $\mathcal{L}_{\text{US-JEPA}}$ to align with the target.	100
6.3	Distribution of pretraining data. To characterize the dataset composition at the organ level, we report the distribution of (a) temporal sequences, including videos and volumes (n_v), and (b) individual static frames (n_f).	104
6.4	Results for few-shot scaling. We report the mean macro F1 score % over five seeds for each model’s probe to measure performance with 1% to 100% of training labels. <i>Note: each dataset is plotted on a unique y-axis scale to better highlight model-specific performance trends.</i>	109
6.5	Robustness to domain-specific corruption. Results show the mean macro F1 across five seeds for each model–dataset–corruption permutation. Linear probes were trained on full, uncorrupted training sets and evaluated on increasingly corrupted test sets to assess structural representation stability.	110

6.6	UMAP feature visualizations. UMAP projections of features extracted from US-JEPA, USrc-JEPA, and URFM across downstream datasets.	113
6.7	Examples of images and corresponding USrc mask.	123
6.8	Few-shot classification robustness across all datasets. Here the test set evaluation for linear probe tuned with varying percentages of training data are shown for all models.	126
6.9	OOD robustness results for remaining 4 downstream datasets.	128
6.10	Different corruptions at three severity levels applied to a liver ultrasound from AUL dataset.	130
6.11	Attention visualization for US-JEPA and USrc-JEPA for three downstream datasets.	131
7.1	Mean F1 score over all classification tasks for each model. The dashed line shows the evolution of the best model over time.	150
7.2	Mean positive Dice score over all segmentation tasks for each model. The dashed line shows the evolution of the best model over time.	150
7.3	Overall classification and segmentation performance per model. The size of each point represents the number of trainable parameters.	151
7.4	Per-task results for each classification task using linear probing.	152
7.5	Per-task results for each classification task using full fine-tuning.	153
7.6	Per-task results for each segmentation task using decoder-only training.	154
7.7	Per-task results for each segmentation task using full fine-tuning.	156

Acknowledgements

First, to all my friends and family, thank you for your unwavering support and love - I would not be here without you.

To my advisor, Dr. William Speier, thank you for your time, patience and mentorship over these last five years. Working with you has deeply influenced the way I approach science and allowed me to become a better problem solver. Our weekly 1-1s have never failed to make me happy and I will miss talking about sports, movies and anything under the sun, but leaving just enough time to cover research.

To my co-chair, Dr. Corey Arnold, for making the BAIR lab an amazing place to work in and allowing us to build a great lab culture.

To my committee members, Dr. Michael Yeh, Dr. Holden Wu, and Dr. William Hsu, for their commitment to this project and support over the years.

To all the members of the BAIR/CDx lab, past and present, with whom I have had the pleasure of working with - Katya, Saarang, Ved, Nathan, Carlos, Mara, Shreeram, Micah, Hongye, Santi, Dingyuan, Jen, Harry, Zichen, Alekhya, Vivek, Shawn, Kimberly, Melanie, Rose, among many others. Going to lab for the last five years has never felt like work because of how fun and welcoming everyone has always been.

To my friends from college, high school and every stage of life before, for always being in my corner. And a special thanks to the amazing people who have entered my life while living in LA. These past five years would not have been the same without you.

To my partner, Pooja, for always reminding me that I'm a capable scientist. She has been a steadfast source of support and affirmation, without which this journey would have been immeasurably more difficult. But more than this, it is her dedication to her medical career, daily work ethic and unlimited empathy for those around her that serves as a constant inspiration.

To my dad, Radha, thank you for teaching me that opportunities are not given, but earned. You've always pushed me to be the best version of myself. To my mom, Lavanya, thank you for the sacrifices you have made, and continue to make – you are the reason I can wake up every day and do what makes me happy. And lastly, to my sister, Aathmika, for always being my number one supporter. I'm so lucky to have a built-in friend for life.

Thank you to everyone who made this possible.

Vita

2017–2021	B.S. (Bioengineering), Minor (Computer Science), University of California, Berkeley
2020–2022	Machine Learning Intern, Dascena, Remote
2023–2024	Biodesign Discovery Fellow, University of California, Los Angeles
2025	Deep Learning Intern, Philips Research, Boston
2021–2026*	PhD (Bioengineering), University of California, Los Angeles

* - expected

Selected Publications (* denotes equal contribution)

Radhachandran, A.*, Tupper, A.*, et al. (2026). UltraBench 2.0: A Standardized Dataset and Evaluation Benchmark for Ultrasound Foundation Models. *In preparation*.

Radhachandran, A., Ivezić, V., Athreya, S., et al. (2026). US-JEPA: A Joint Embedding Predictive Architecture for Medical Ultrasound. arXiv:2602.19322.

Ivezić, V., Pleasure, M., **Radhachandran, A.**, Panchavati, S., Athreya, S., et al. (2026). Pretext Matters: An Empirical Study of SSL Methods in Medical Imaging. arXiv:2603.22649.

Lee, D., Amara, D., Beon, C., Swee, S., **Radhachandran, A.**, et al. (2026). Leveraging Large Language Models for Thyroid Nodule Information Extraction and Matching Across Medical Reports. AMIA Annual Symposium Proceedings 2025, 673–682.

Ivezic, V., **Radhachandran, A.**, Redekop, E., Athreya, S., et al. (2025). CytoFM: The first cytology foundation model. Proceedings of the IEEE/CVF Conference on Computer Vision

and Pattern Recognition, 4788–4796.

Athreya, S., Melehy, A., Suthahar, S. S. A., Ivezić, V., **Radhachandran, A.**, et al. (2025). Combining Ultrasound Imaging and Molecular Testing in a Multimodal Deep Learning Model for Risk Stratification of Indeterminate Thyroid Nodules. *Thyroid®*, 35(5), 590–594.

Radhachandran, A., Vittalam, A., Ivezic, V., et al. (2024). ThyGraph: A Graph-Based Approach for Thyroid Nodule Diagnosis from Ultrasound Studies. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, 753–763.

Radhachandran, A., Kinzel, A., Chen, J., et al. (2024). A multitask approach for automated detection and segmentation of thyroid nodules in ultrasound images. *Computers in Biology and Medicine*, 170, 107974.

Athreya, S., **Radhachandran, A.**, Ivezić, V., Sant, V. R., et al. (2024). Enhancing Ultrasound Image Quality Across Disease Domains: Application of Cycle-Consistent Generative Adversarial Network and Perceptual Loss. *JMIR Biomedical Engineering*, 9(1), e58911.

Sant, V. R.*, **Radhachandran, A.***, Ivezic, V., Lee, D. T., et al. (2024). From Bench-to-Bedside: How Artificial Intelligence is Changing Thyroid Nodule Diagnostics, a Systematic Review. *The Journal of Clinical Endocrinology & Metabolism*, 109(7), 1684–1693.

CHAPTER 1

Introduction

The promise of AI for rapid medical change has been long standing and eagerly awaited. Healthcare, spanning hospitals, clinics, payers and other health systems, generates more data than an individual practitioner can fully synthesize. Imaging studies accumulate faster than they can be read, diagnostic criteria differ between institutions and readers, and rare but consequential findings can go undetected. The gap between what AI could offer and what it has tangibly delivered in clinical practice has narrowed considerably in recent years, but it has not fully closed [Top19]. Standardized, data-driven AI methods offer a promising path toward closing this diagnostic gap. However, as we develop new AI methods for medical diagnostics, a few critical considerations must remain in mind: how involved should clinicians be with an AI tool in practice? Are AI tools truly better than clinicians for certain tasks? And how do we manage the interests of healthcare systems, physicians, and most importantly patients, when building clinical AI tools? These are difficult questions to answer, but it is essential that we continue to use them as a guide as we develop novel AI methods for problems in medicine and healthcare.

In the last ten years there has been a remarkable transformation in the way the research community has approached and addressed problems in this space. The long-standing tension between specialized models and generalist approaches has largely reached a middle ground through the rise of foundation models [B⁺21]. These large models pretrained on broad data and then fine-tuned have considerably blurred the line between these two paradigms. Researchers are starting to ask different questions such as, can a general base model, fine-

tuned carefully, beat a specialized model? And as more empirical and theoretical findings accumulate behind this approach, the question has expanded further: how much domain-specific adaptation does a foundation model actually need? The questions we ask from both these clinical and methodological lenses will dictate whether we can deliver on the promise of AI in biomedicine.

This dissertation frames these questions within the context of a focused slice in this expansive field: ultrasound imaging. First, we examine the limitations and pain points within a specific clinical domain: thyroid nodule diagnostics, and establish where AI methods stand with regard to making a tangible clinical impact. Based on those findings and guidance from collaborating physicians, we propose clinically motivated, task-specific models to address specific challenges. These solutions specifically revolve around improving detection, localization, and diagnosis of thyroid nodules. Next, in line with the technical advancements within the field, we begin asking bigger questions about the efficacy and generalizability of these models. We approach the challenge of modeling ultrasound from the technical lens and pursue methods to build better representations for downstream tasks. By pursuing foundation models for ultrasound and establishing better standards for evaluation and reproducibility through benchmarking, we demonstrate the promise of large-scale pretraining paradigms and the degree to which generalizable representations can serve as a viable path forward for AI models for ultrasound in clinical practice.

1.1 Clinical Motivation

1.1.1 Why Thyroid Cancer?

Thyroid nodules are among the most common incidental findings in clinical practice, with prevalence in the adult population being as high as 50%. Most nodules are incidentally detected by unrelated cross-sectional imaging or during a physical examination. The vast majority are benign, yet the malignancy rate has been reported to be as high as 15% –

high enough to warrant careful evaluation. Thyroid cancer itself has become one of the fastest-growing cancer diagnoses in the United States, although it is a trend driven in part by incidental detection on imaging studies. The clinical standard of care revolves around ultrasound. Nodules are imaged, and risk-stratified using structured reporting systems such as TI-RADS, and referred for fine-needle aspiration biopsy when specific features trigger concern [AC22].

This pipeline, while well-established, is far from perfect. Inter-reader variability in TI-RADS scoring is a well-documented problem, and high biopsy rates for ultimately benign nodules impose unnecessary burden on patients and on already-stretched radiology and endocrinology workflows [GSFD24]. These are exactly the kinds of high-volume diagnostic tasks where AI has the potential to introduce consistency and reduce unnecessary intervention. This point has made thyroid nodule imaging a natural and well-motivated starting point, which drives the first part of this dissertation.

1.1.2 Why Ultrasound?

Ultrasound is the most widely used medical imaging modality in the world. It is accessible, real-time, non-ionizing, and inexpensive relative to CT or MRI. Yet, it is also among the most challenging modalities for AI; images are corrupted by speckle noise and acoustic artifacts, acquisition is heavily operator-dependent, and the same anatomy can look substantially different across scanners, probe frequencies, and clinical settings. These properties lend an unfortunate brittleness to supervised models, as models trained on one institution’s data may not seamlessly transfer to another. In this sense, ultrasound is not only a clinically important modality, but also a particularly relevant subject for the generalization challenges that motivate the second half of this dissertation.

1.2 Part 1: Task-Specific Deep Learning for Thyroid Ultrasound

When this dissertation work began in 2021, supervised deep learning, primarily powered by convolutional neural networks, had firmly established its value in medical imaging. Key studies had demonstrated that CNNs could match or exceed clinician performance on well-scoped tasks: diabetic retinopathy screening from fundus photographs, skin lesion classification, and chest radiograph interpretation, among others [EKN⁺17, GPC⁺16]. These results were new and exciting, but they shared a common thread: strong performance on clean, curated, single-site datasets that did not necessarily reflect the messiness of real clinical deployment. Simultaneously, vision transformers had just emerged as a compelling alternative to CNNs for image-based tasks [DBK⁺21], and their applicability to medical imaging, where data was scarce and domain-specific, remained an open, but uniquely challenging question.

The dominant roadblocks were practical as much as architectural. Acquiring high-quality labeled data required clinical expertise and was expensive to scale. Models that performed well in-distribution often degraded significantly when evaluated on data from different institutions, scanners, or patient populations. Architectural innovation continued rapidly; however, the field was in the midst of understanding the gap between impressive quantitative improvements and reliable clinical deployment. The early work in this dissertation, covered in Chapters 3–5, is largely motivated by these trends from the technical side of this field.

The first part of this dissertation approaches the problem of thyroid nodule diagnostics through three complementary efforts. Chapter 3 provides a systematic review of the state of AI for thyroid nodule ultrasound, documenting the architectures, datasets, clinical tasks, and performance claims that had accumulated in the literature, and surfacing the gaps in benchmarking rigor and external validation that would motivate the subsequent work. Chapter 4 develops a joint detection and segmentation framework that simultaneously localizes nodules and delineates their boundaries, addressing a natural set of tasks that radiologists

perform together in practice. Curating the annotation dataset for this project required direct collaboration with radiologists and surgeons at UCLA, which gave us an early firsthand appreciation for how expensive and clinically demanding high-quality labeled data truly is to produce. Chapter 5 moves towards relational reasoning through ThyGraph, a graph-based architecture that models spatial and semantic relationships across multiple views of an ultrasound study, motivated by the way clinicians synthesize information across a full imaging session rather than a single frame.

Collectively, these projects confirmed the promise of task-specific deep learning while exposing its limits. Models trained on UCLA data performed well internally, but performance degraded when evaluated on external datasets, a pattern consistent with what broader literature had reported. ThyGraph also made clear that even with more intentionally engineered architectures, the field was approaching a ceiling due to the feature extractors themselves. Without richer, more generalizable representations of ultrasound images, further gains might just be incremental. These observations directly motivate the second part of this dissertation.

1.3 Pivoting From Task-Specific to Foundation Models

The limitations we encountered were not unique to our work as they reflected a larger problem with the task-specific approach. Separate models for each clinical task meant no shared representations, no transfer of learned knowledge across problems, and no efficient way to scale without repeating the annotation process for new institutions or modalities. The representations learned were, by design, tied to the distribution of the training data.

In parallel, the broader research community was undergoing a significant shift. NLP methods, such as GPT and BERT, had demonstrated that pretraining on scale followed by efficient fine-tuning could produce broadly capable models. This concept was now being extended to vision. Self-supervised methods such as SimCLR, DINO, and MAE showed that powerful visual

representations could be learned from unlabeled images [CKNH20a, CTM⁺21a, HCX⁺21]. The medical imaging community began translating these ideas, developing foundation models for radiology and pathology. However, ultrasound remained comparatively underexplored. As we observed this gap become more prominent, we began asking a new question: could we build a foundation model specifically for ultrasound, pretrained with self-supervised learning on a large and diverse corpus, that generalizes across anatomical domains and downstream clinical tasks? Such a model had potential in multiple use cases such as serving as a feature extractor for task-specific pipelines, being fine-tuned for specific clinical endpoints with far less labeled data, and potentially even sidestepping the need for institution- or task-specific retraining. The second part of this dissertation pursues these directions.

1.4 Part 2: Self-Supervised Learning and the Ultrasound Foundation Model

The second part of this dissertation addresses the challenge of building and evaluating a generalizable foundation model for ultrasound across two chapters. Chapter 6 first lays the conceptual groundwork by surveying how the broader field transitioned from task-specific models to SSL-pretrained foundation models across radiology, pathology, and ultrasound, and building a theoretical basis for why the choice of self-supervised pretext task matters. We start by defining a hierarchy of the three main families: 1) reconstruction-based methods (such as MAE), 2) joint embedding architectures (such as DINO), and 3) joint embedding predictive architectures (such as I-JEPA). This background argues that JEPA is particularly well-suited for a noisy, structurally-organized modality like ultrasound, where pixel-level reconstruction wastes capacity on irrelevant variance, and augmentation-based invariances are harder to construct compared to in natural images. Chapter 6 then presents US-JEPA, a Joint Embedding Predictive Architecture adapted for ultrasound. As a part of this contribution, we also propose OPUS, the Open Ultrasound Dataset, the largest curated corpus

of publicly available ultrasound images assembled to date, spanning multiple anatomies and pathologies. Rather than reconstructing masked pixel values, US-JEPA learns by predicting latent representations of masked spatial regions from visible context, a pretraining objective that encourages the model to learn spatial and structural relationships without modeling noise. Chapter 7 addresses an unresolved gap in this space with UltraBench 2.0, a comprehensive benchmark for ultrasound foundation model evaluation covering linear probing, segmentation probing, and full fine-tuning across multiple anatomical sites and clinical tasks. In developing US-JEPA, it became clear that despite a growing number of ultrasound foundation models being released in rapid succession, the field lacked the standardization needed to meaningfully compare them. Specifically, these gaps spanned the full model development pipeline:

1. pretraining corpora varied widely with minimal efforts to unify publicly available datasets;
2. proposed architectures were rarely evaluated against controlled pretraining data baselines;
3. downstream evaluation protocols differed in tuning modes and hyperparameters; and
4. there was no agreed-upon set of tasks and datasets for assessment, making it difficult to distinguish real model improvements from artifacts of benchmark saturation.

UltraBench 2.0 is designed to address these shortcomings directly, providing a community standard for future comparisons and proposing more challenging evaluation settings to guide the next generation of model development.

1.5 Contributions

The main contributions of this dissertation can be summarized in the following specific aims:

Aim 1. To comprehensively survey the state of AI research in the thyroid nodule diagnostic pipeline

- a. Conducted a systematic review of deep learning methods for thyroid nodule ultrasound, cataloguing architectures, datasets, clinical tasks, and reported performance across the literature
- b. Identified critical gaps in benchmarking rigor, external validation, and clinical translatability that motivate subsequent work in this dissertation

Aim 2. To develop novel deep learning approaches addressing limitations in thyroid nodule detection, segmentation, and classification

- a. Built a joint detection and segmentation framework that couples nodule localization and image-level detection in a single model
- b. Evaluated segmentation models on external datasets, characterizing performance degradation under distribution shift
- c. Improved study-level malignancy classification by modeling relational reasoning across multiple image views
- d. Proposed ThyGraph, a graph-based architecture that aggregates multi-view ultrasound features to capture spatial and semantic relationships between diagnostic attributes

Aim 3. To introduce a large-scale self-supervised pretraining framework for ultrasound foundation model development

- a. Curated the largest publicly available ultrasound pretraining corpus, spanning multiple anatomies and pathologies
- b. Developed US-JEPA, a Joint Embedding Predictive Architecture adapted for ultrasound
- c. Conducted the first comprehensive baseline comparison of ultrasound foundation models under a standardized evaluation protocol

- d. Characterized label-efficient learning and feature robustness to imaging corruptions across diverse downstream tasks

Aim 4. To establish the first standardized benchmark for ultrasound foundation model evaluation

- a. Defined a diverse evaluation suite spanning downstream datasets capturing multiple organs and clinical tasks
- b. Evaluated existing ultrasound foundation models and presented results in a public-facing leaderboard to promote transparency and reproducibility
- c. Released reusable implementations of linear probing and full finetuning protocols for community adoption
- d. Identified performance saturation in existing benchmarks and proposed more evaluation settings to guide future model development

1.6 Outline of the Dissertation

The chapters in this dissertation are organized as follows:

Chapter 1 (current) acts as a brief introduction and motivation behind the topics that will be covered in this dissertation.

Chapter 2 provides an overview of the technical concepts required to understand the subsequent chapters.

Chapter 3 is a systematic review of AI methods for thyroid nodule ultrasound diagnostics.

Chapter 4 presents a multitask detection and segmentation framework for thyroid nodules.

Chapter 5 presents ThyGraph, a graph-based model capturing relational diagnostic reasoning across multi-view ultrasound.

Chapter 6 provides a background on SSL pretext tasks for ultrasound and presents US-JEPA, a JEPA-based foundation model for ultrasound, pretrained on the largest curated public ultrasound corpus.

Chapter 7 presents UltraBench 2.0, a standardized dataset benchmark and evaluation protocol for ultrasound foundation models.

Chapter 8 concludes the dissertation with a summary of completed work, major takeaways, and future research directions.

CHAPTER 2

Background

2.1 Ultrasound Imaging

2.1.1 Physics of Ultrasound Imaging

Medical ultrasound imaging generates cross-sectional images of tissue using high-frequency sound waves, typically in the range of 2–18 MHz for clinical applications. A piezoelectric transducer in the probe converts electrical pulses into mechanical vibrations, transmitting brief bursts of acoustic energy into the body. As these waves propagate through tissue, they are partially reflected at acoustic impedance boundaries, which are interfaces between tissues of differing density and elasticity. The transducer then converts the returning echoes back into electrical signals. By measuring the time between transmission and reception, the depth of each reflecting boundary is inferred, while the amplitude of the return signal determines image brightness. This pulse-echo process is repeated along many scan lines across the probe face, assembling a two-dimensional image in real time [Sza14].

The dominant clinical modality is B-mode (brightness mode) imaging, in which each scan line’s depth profile is rendered as a column of pixels with intensity proportional to echo amplitude. The result is a grayscale cross-section of anatomy, acquired at frame rates typically between 15 and 60 Hz. This real-time nature distinguishes ultrasound from static modalities: structures can be visualized dynamically as the probe moves, and blood flow can be superimposed using Doppler techniques [Sza14].

The choice of probe directly shapes what is visible. Linear array transducers (5–15 MHz)

provide high spatial resolution at shallow depths and are standard for superficial structures such as the thyroid, breast, and peripheral vasculature. Curvilinear probes (2–6 MHz) trade resolution for penetration and field of view, making them suitable for abdominal and obstetric imaging. Phased array transducers use a small footprint with electronic beam steering to access structures through narrow acoustic windows, as in cardiac imaging (echocardiography). Beyond probe selection, the operator can control gain, time-gain compensation, focal depth, and dynamic range. Each of these features can modify the resulting image in ways that are not standardized across operators or institutions. Manufacturers apply proprietary signal processing pipelines (harmonic imaging, speckle reduction filters, adaptive beamforming) that further differentiate images across scanner platforms from vendors such as GE, Siemens, Philips, and Samsung.

2.1.2 Challenges in Ultrasound Interpretation

The process of ultrasound image formation introduces several properties that collectively make interpretation substantially harder than for other imaging modalities.

Speckle noise. The most pervasive artifact in B-mode images is speckle, a granular, "salt-and-pepper" texture arising from the coherent interference of echoes from microscopic tissue scatterers. Unlike additive Gaussian noise, speckle is multiplicative, meaning that its amplitude scales with local signal intensity. Moreover, speckle is fundamentally different from random noise; it arises from the way sound waves overlap and interfere, so it can't be reduced simply by averaging multiple images together. The resulting texture obscures soft-tissue boundaries, reduces contrast, and introduces patterns that vary with tissue microstructure, transducer frequency, and scan geometry [WSSL83].

Acoustic artifacts. The way sound waves travel through tissue can create structured artifacts through patterns that may look like, or hide, real pathology. Acoustic shadowing occurs behind strongly attenuating objects (weakens the reflected signal), such as calcifications, bone, and gas, thus producing dark regions that carry no tissue information. Posterior acoustic

enhancement occurs behind fluid-filled structures, such as cysts or vessels, creating artificially bright regions distal to the object. Reverberation artifacts, side-lobe contamination, and refraction all introduce bright or dark structures with no clear anatomical correlates. While experienced sonographers can interpret these artifacts diagnostically, they can also mislead less experienced interpreters and complicate systematic image review [KT86].

Operator and probe dependence. Unlike with CT, where the scanner and patient are fixed in place, producing consistent, repeatable images, ultrasound is performed by hand in real time. The image at any moment is a function of probe position, orientation, tilt, and applied pressure. All of these factors are controlled by the sonographer. Not to mention, there is no fixed coordinate frame so the same anatomical structure may appear at different depths, orientations, and scales across examinations. Inter-observer variability in ultrasound interpretation is well-documented [CKK⁺10, PKJ⁺10], and upstream acquisition variability is a contributing cause.

Scanner and protocol heterogeneity. Even for identical anatomy, images from different scanner manufacturers or protocol settings can be visually distinct. The absence of a standardized intensity scale means that pixel values carry institution-specific meaning. This heterogeneity compounds the operator-dependence problem.

2.1.3 Why Ultrasound is Important

Ultrasound is among the most widely used medical imaging modalities in the world [MC11]. The absence of ionizing radiation makes it safe for repeated examinations, pregnant patients, and pediatric populations. Portable and handheld devices bring diagnostic imaging to emergency departments, intensive care units, rural clinics, and field settings where CT or MRI infrastructure is unavailable. In low- and middle-income countries, ultrasound is frequently the only cross-sectional imaging modality accessible at the point of care, making it a critical component of global health infrastructure. Thus, despite the drawbacks listed above, ultrasound has proven to be an integral diagnostic tool.

2.2 Thyroid Nodules and the Diagnostic Pipeline

2.2.1 Anatomy and Nodule Formation

The thyroid gland is a butterfly-shaped endocrine organ situated in the anterior neck, wrapped around the trachea. Its two lateral lobes are connected by a narrow isthmus, and a small pyramidal lobe extending superiorly is present in some individuals (Figure 2.1). The gland synthesizes thyroxine (T4) and triiodothyronine (T3), hormones that regulate metabolic rate, growth, and development throughout the body [MS12].

Thyroid nodules are discrete lesions within the thyroid parenchyma that are sonographically and pathologically distinct from the surrounding gland tissue. They arise from a variety of processes: benign follicular adenomas and hyperplastic nodules account for the large majority, while cysts, colloid nodules, thyroiditis-related changes, and primary malignancies or metastatic deposits account for the rest. Nodules tend to be asymptomatic, with most being discovered incidentally during neck imaging performed for other indications [DGL⁺18].

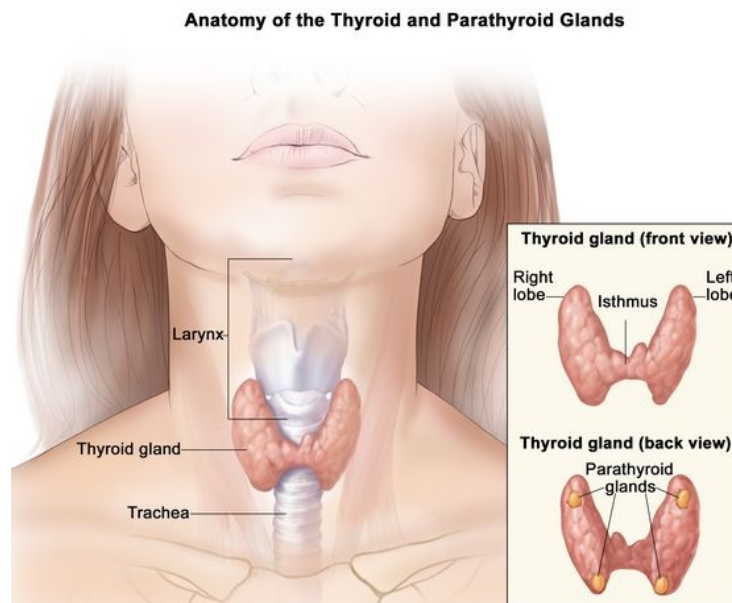


Figure 2.1: A simple illustration of thyroid anatomy [noa11].

2.2.2 Epidemiology

The prevalence of thyroid nodules is substantially higher than clinical intuition might suggest. Palpation detects nodules in roughly 5% of women and 1% of men in iodine-sufficient populations [WC97], but high-frequency ultrasound reveals nodules in 20–76% of adults screened, with higher rates in older individuals and women [E⁺94]. This discrepancy between palpable and ultrasound-detectable nodules reflects the resolution advantage of modern imaging. It also has important downstream consequences as a large fraction of the population has nodules that necessitate at least an initial radiological assessment.

Despite this high prevalence, the risk of malignancy among detected nodules is relatively low. Among nodules referred for clinical evaluation, approximately 5–15% prove malignant on surgical pathology, and the majority of those that are tracked longitudinally without biopsy remain stable or regress [DCL⁺15]. Papillary thyroid carcinoma accounts for roughly 85% of thyroid malignancies and generally carries a favorable prognosis. Follicular, medullary, and anaplastic carcinomas make up most of the remainder, with anaplastic carcinoma being rare but highly aggressive [FW16]. The incidence of diagnosed thyroid cancer has risen sharply over recent decades in many countries [KSS20], a trend attributed largely to increased detection of small, indolent papillary tumors that may never progress to clinically significant disease. This trend raises an ongoing concern about overdiagnosis and its downstream consequences [V⁺16].

2.2.3 Clinical Workflow

Detection. Most thyroid nodules are discovered incidentally, typically on a neck ultrasound ordered for an unrelated indication, on CT or MRI of the neck or chest, or through alternative screening programs. A smaller fraction present through palpation on physical examination or as a result of tangential symptoms such as dysphagia or hoarseness.

Risk stratification: ACR TI-RADS. Once a nodule is identified on ultrasound, the primary clinical task is estimating its malignancy risk to guide further diagnostic decisions.

The ACR Thyroid Imaging, Reporting and Data System (TI-RADS) provides a structured reporting framework for this assessment [T⁺17]. Radiologists evaluate five sonographic features: 1) composition (proportion of cystic versus solid tissue), 2) echogenicity (pixel intensity relative to surrounding tissue), 3) shape (wider-than-tall versus taller-than-wide), 4) margin character (smooth, ill-defined, lobulated or irregular, or extrathyroidal extension), and 5) echogenic foci (comet-tail artifacts, macrocalcifications, peripheral calcifications, or punctate echogenic foci). They assign points to each feature and the summed score yields a TI-RADS level from TR1 (benign) to TR5 (highly suspicious) as outlined in Figure 2.2. FNA is recommended when a nodule’s TI-RADS level and maximum diameter together exceed defined thresholds, with higher-risk categories biopsied at smaller sizes.

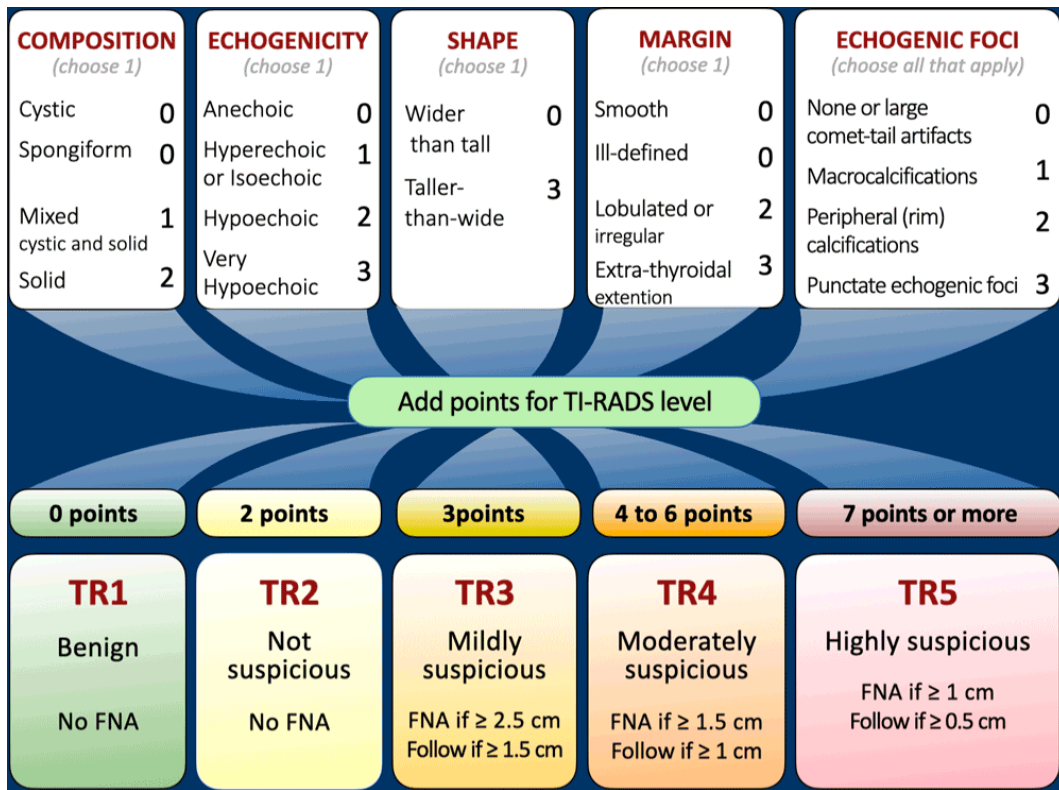


Figure 2.2: ACR TI-RADS scoring calculator [AvdL].

Biopsy: fine needle aspiration. Fine needle aspiration (FNA) biopsy is the standard diagnostic procedure for sonographically suspicious nodules [HGM⁺82]. A thin needle is passed into the nodule under ultrasound guidance, aspirating cellular material for cytopathological

analysis. Results are reported using the Bethesda System for Reporting Thyroid Cytopathology, which classifies specimens into six categories: I (nondiagnostic), II (benign), III (atypia of undetermined significance), IV (follicular neoplasm), V (suspicious for malignancy), and VI (malignant) [CA17, ABCP⁺23]. Categories II and VI are relatively straightforward in terms of clinical implications. Categories III and IV are diagnostically indeterminate, carrying estimated malignancy risks of 10–30% and 25–40%, respectively; however, cytological features alone do not reliably resolve them.

Further Management. Nodules classified as benign (Bethesda II) are managed with periodic ultrasound for surveillance. Malignant or suspicious cytology (Bethesda V–VI) typically leads to surgery, in the form of a hemithyroidectomy or total thyroidectomy, depending on clinical context [HAB⁺16, PYL⁺20]. Patient who undergo surgical procedure in the form of a total thyroidectomy typically require lifelong thyroid hormone replacement therapy, since they can no longer produce their own. Indeterminate nodules (Bethesda III–IV) present the greatest challenge in management as there are multiple options, including repeat FNA, molecular testing of the biopsy aspirate (gene expression classifiers or mutational panels), or diagnostic surgical resection. Active surveillance without immediate intervention is increasingly adopted for low-risk papillary microcarcinomas.

2.2.4 Clinical Challenges

Inter-reader variability. TI-RADS scoring requires assessing multiple sonographic features that are often subtle, context-dependent, and subject to individual interpretation. Significant inter-observer disagreement in feature assignment and risk categorization has been noted both across and within institutions [CKK⁺10, PKJ⁺10], meaning that the same nodule can receive different risk classifications and thus, different management recommendations depending on who performs the reading.

Biopsy burden. Because TI-RADS is designed to be sensitive rather than specific, a substantial proportion of biopsied nodules prove benign. Estimates suggest roughly 70–

80% of FNA procedures yield benign cytology [Ros06], representing a large volume of procedures carrying a real burden for patients such as procedural discomfort, anxiety, and rare complications, as well as health system costs [KWZ18]. The indeterminate Bethesda categories compound this problem further, as many patients with Bethesda III–IV results undergo diagnostic surgery for nodules that are ultimately benign on histological examination [PSW⁺21].

Diagnostic uncertainty. Even after FNA, the Bethesda III and IV categories leave a meaningful fraction of patients without a definitive diagnosis. Molecular tests have reduced unnecessary surgical rates for indeterminate nodules in settings where they are available, but they add cost, depend on adequate specimen cellularity, and are not universally accessible. Not to mention, they have high negative predictive values so do not confidently indicate which cases actually require surgical intervention [KZ20].

2.3 Deep Learning for Medical Image Analysis

This section provides a technical overview on the deep learning architectures and frameworks that form the methodological foundation of this dissertation. The following content should provide sufficient background to understand the architectural choices made in later chapters.

2.3.1 Convolutional Neural Networks

Convolutional neural networks (CNNs) are the dominant architecture for visual recognition tasks and remain central to medical image analysis. The core operation is the discrete convolution: a learned filter K of small spatial extent is slid across the input image I , computing a weighted sum at each location,

$$(I * K)[i, j] = \sum_m \sum_n I[i + m, j + n] \cdot K[m, n],$$

where (i, j) is the output spatial position and (m, n) indexes the kernel elements. Weight

sharing across spatial locations gives convolutions two important properties: 1) translational equivariance (a feature detector fires wherever the relevant pattern occurs) and 2) parameter efficiency (a single filter applied everywhere is far cheaper than a fully connected layer). In practice, multiple filters are learned simultaneously to detect different local patterns, such as edges or oriented textures, producing a multi-channel feature map.

Stacking convolutional layers allows the network to aggregate low-level features into higher-level representations. Neurons in early layers have smaller receptive fields, responding to fine-grained local patterns, while neurons in deeper layers integrate information from larger image regions and capture abstract, global structure. Pooling layers condense spatial resolution between convolutional stages, increasing the receptive field while building invariance to small spatial shifts.

Training very deep convolutional networks is complicated by vanishing gradients. Gradients propagated back through many layers become exponentially small, effectively immobilizing early layers. Residual networks (ResNets) [HZRS16] resolve this with shortcut connections that add the input of a block directly to its output,

$$\mathbf{y} = \mathcal{F}(\mathbf{x}, \{W_i\}) + \mathbf{x},$$

where $\mathbf{x}$ and $\mathbf{y}$ are the block’s input and output, $\mathcal{F}(\mathbf{x}, \{W_i\})$ is the residual mapping learned by the stacked convolutional layers, and $\{W_i\}$ are their weights. This ensures that gradients can flow through the identity path regardless of network depth. ResNets enable stable training at depths of 50, 101, and 152 layers, and ResNet variants are often used as feature extractors in image-based tasks.

The standard pipeline in supervised visual recognition begins with pretraining on ImageNet [DDS⁺09], a dataset of roughly 1.2 million natural images spanning 1,000 object categories. Weights learned during this pretraining can then be adapted, through fine-tuning, for the target task, typically by replacing the final classification layer and continuing training on the target dataset. The intuition behind this transfer learning paradigm is that early network

layers learn meaningful edges and textures that are broadly generalizable across visual domains. However, this intuition must be considered carefully, especially when understanding applications in medical imaging.

2.3.2 Segmentation

Semantic segmentation assigns a class label to every pixel in an image. In medical imaging, segmentation is used to delineate organs, lesions, and anatomical structures for downstream clinical use cases.

The standard training objective is the Dice loss,

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2|P \cap G|}{|P| + |G|},$$

where P is the predicted binary mask, G is the ground-truth mask, and $|\cdot|$ denotes the number of positive pixels. This measures the overlap between predicted and ground-truth masks and is designed for medical segmentation where target structures are often small relative to the image. In this medical domain, if we used pixel-wise cross-entropy loss, the error would be dominated by the majority background class.

The U-Net architecture [RFB15] introduced the design that remains the standard backbone for medical image segmentation. An encoder progressively reduces spatial resolution while expanding feature channels, extracting increasingly abstract representations. Then, a symmetric decoder restores pixel-level resolution for pixel-level classification. Skip connections transfer high-resolution spatial information directly from corresponding encoder stages to the decoder to help preserve fine structural detail that would otherwise be lost due to successive downsampling.

Several architectural variants extend the U-Net design. Attention U-Net [OSF⁺18] adds soft attention gates on skip connections to suppress spatially irrelevant activations. UNet++ [ZSTL18] replaces straight skip connections with dense sub-networks connecting all encoder-

decoder pairs at the same resolution level, bridging the semantic gap between encoder and decoder features. 3D U-Net extends the architecture to volumetric inputs by replacing all 2D convolutions with 3D convolutions, enabling voxel-level prediction from CT or MRI volumes.

2.3.3 Multiple Instance Learning

Multiple instance learning (MIL) is a framework designed for weakly supervised settings. Training data is organized into bags, each containing multiple instances, and only a bag-level label is assigned. The network is trained to classify bags while localizing which instances drive the bag prediction.

This paradigm translated naturally to medical image settings where annotation at the "instance"-level is unavailable or impractical. For example, predicting whether there is a malignancy somewhere across the images in a study, when only a study-level pathology label is present.

Attention-based MIL (AMIL) [ITW18] replaces equally weighted instance aggregation with a learned attention mechanism. A small attention network produces a scalar weight for each instance,

$$a_k = \frac{\exp\{w^\top \tanh(V h_k^\top)\}}{\sum_j \exp\{w^\top \tanh(V h_j^\top)\}},$$

where h_k is the embedding of instance k , and w and V are learned parameters. The bag-level embedding is then a weighted sum of instance embeddings,

$$z = \sum_k a_k h_k,$$

which is passed to a classifier trained with the bag label. The learned attention weights identify which instances are most informative, providing a form of interpretability without instance-level supervision.

2.3.4 Graph Neural Networks

Many problems in biomedicine, specifically within imaging, involve entities with an explicit relational structure. For example, lymph nodes forming anatomical chains, image patches related by spatial adjacency, or nodule observations from serial scans linked over time. Graphs provide a natural representation, where nodes hold feature vectors encoding entities, and edges encode relationships. Graph neural networks (GNNs) generalize the intuition behind image convolution to this non-Euclidean domain by defining operations that aggregate information from a node’s neighborhood.

Specifically, graph Convolutional Networks (GCNs) [KW17] update node representations by computing a normalized sum over neighbors,

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-1/2} \tilde{A} \tilde{D}^{-1/2} H^{(l)} W^{(l)}\right),$$

where $H^{(l)}$ is the matrix of node feature vectors at layer l , $\tilde{A} = A + I$ is the adjacency matrix with self-loops, $\tilde{D}$ is the corresponding diagonal degree matrix, $W^{(l)}$ is a learned weight matrix, and σ is a nonlinear activation function. Stacking layers propagates information across multi-hop neighborhoods, so a k -layer GCN incorporates context from the k -hop neighborhood of each node. Just as increasing image convolutional layers aggregate more global information, more graph convolutional layers learn overall graphical structure.

Graph Attention Networks (GATs) [VCC⁺18] replace the uniform normalized aggregation with learned, query-dependent attention weights,

$$\alpha_{ij} = \text{softmax}_j\left(\text{LeakyReLU}\left(a^\top [Wh_i \parallel Wh_j]\right)\right),$$

where h_i and h_j are the feature vectors of nodes i and j , W is a shared linear projection matrix, a is a learnable attention vector, $[\cdot \parallel \cdot]$ denotes concatenation, and α_{ij} is the normalized attention coefficient from node j to node i . This allows the network to upweight informative neighbors on a per-node basis.

For graph-level prediction (as opposed to node-level), the node embeddings must be pooled into a single fixed-size vector. Self-Attention Graph (SAG) pooling [LLK19] scores nodes via an auxiliary network, selects the top- k highest-scoring nodes, and creates a subgraph from the retained nodes. The procedure is summarized as

$$y = \sigma\left(\hat{A} X \Theta_{\text{att}}\right), \quad \text{select top-}k \text{ nodes by score } y,$$

where X is the node feature matrix, $\hat{A}$ is the symmetrically normalized adjacency matrix, Θ_{att} is a learnable projection vector, and y is the resulting scalar score for each node. The procedure can be repeated hierarchically until the graph collapses to a compact representation suitable for classification.

2.3.5 Vision Transformers

The transformer architecture [VSP⁺17], originally developed for sequence modeling in natural language processing, was adapted to image recognition by Dosovitskiy et al. [DBK⁺21] through the Vision Transformer (ViT). The key design choice is representing an image as a sequence of patch tokens. An image of resolution $H \times W$ is split into a grid of $N = HW/P^2$ non-overlapping patches of size $P \times P$; each patch is flattened and linearly projected to a d -dimensional token embedding via a projection matrix $E \in \mathbb{R}^{(P^2 \cdot C) \times d}$, where C is the number of image channels. A learnable positional encoding is added to inject spatial order, and a special classification token x_{cls} can be prepended,

$$z_0 = [x_{\text{cls}}; x_p^1 E; \dots; x_p^N E] + E_{\text{pos}},$$

where x_p^i is the flattened pixel vector of the i -th patch and E_{pos} is the learnable positional encoding matrix.

The token sequence is processed by a stack of transformer encoder blocks, each consisting of multi-head self-attention followed by a position-wise MLP, with layer normalization and

residual connections around each sub-layer. Multi-head self-attention computes, for each head,

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V,$$

where Q , K , and V are the query, key, and value matrices obtained by separate linear projections of the token embeddings, and d_k is the dimensionality of the key vectors (used to prevent dot products from growing large in magnitude).

The important property of self-attention is its global receptive field, where every token attends to every other token from the very first layer. This method contrasts with convolutions, which build up receptive fields gradually through depth. The absence of this spatial inductive bias makes ViTs more expressive, as they can model any pairwise relationship in the image, but are also more data-hungry. Thus, the advantage of ViTs typically appears at scale, when pretrained on large datasets or with self-supervised learning objectives.

2.4 Self-Supervised Learning

2.4.1 Core Motivation

Self-supervised learning (SSL) defines a pretext task that generates a training signal directly from unlabeled data, without requiring human annotation. Rather than predicting predetermined labels, a model is trained to predict some aspect of the data that can be derived heuristically, such as a masked region, a future patch, or a different augmented view of the same image. The intuition is that solving a "sufficiently" challenging pretext task forces the model to learn representations that capture the semantics of the data.

The motivation in medical imaging is quite clear. Expert annotation, for segmentation masks, lesion classifications, or pathology diagnoses, requires clinical training and significant time. This drawback creates a fundamental bottleneck on supervised learning. Most healthcare

systems accumulate large archives of unlabeled imaging data that supervised methods cannot exploit. SSL makes these datasets usable - a model pretrained on millions of unlabeled scans can then be fine-tuned on the small labeled set available for a specific task.

2.4.2 Three Training Families

Reconstruction methods. Masked autoencoders (MAE [HCX⁺21]) and related methods (BEiT [BDPW22]) train a model to reconstruct portions of the input that have been randomly masked out. A large fraction of the input patches are removed, and the model must predict the missing content from the remaining visible context. MAE reconstructs raw pixel values and BEiT instead predicts discrete visual tokens generated by a separate tokenizer. In either case, the reconstruction loss over masked patches $\mathcal{M}$ takes the form

$$\mathcal{L} = \frac{1}{|\mathcal{M}|} \sum_{p \in \mathcal{M}} \|x_p - \hat{x}_p\|_2^2,$$

where x_p is the original content of masked patch p and $\hat{x}_p$ is the model’s reconstruction. Masking forces the model to learn a component of the spatial structure and semantic context of the full image from partial observations. The high masking ratio means that nearby visible patches are insufficient to reconstruct the masked ones by simple interpolation so the model must develop an understanding of global structure. These methods are computationally efficient at inference because the encoder only processes the visible (unmasked) tokens.

Joint Embedding Architectures (JEAs). Rather than operating in pixel space, JEAs train the model in a latent embedding space. Two augmented views of the same image, generated by applying random crops, color jitter, blur, and other transformations, are encoded separately. The objective pulls their representations together. The intuition is that a good representation should be invariant to augmentations that preserve semantic content.

Contrastive methods (SimCLR [CKNH20a], MoCo [HFW⁺20]) make this concrete by additionally pushing apart embeddings of different images, preventing the trivial collapse

solution in which all inputs map to the same vector.

Non-contrastive methods (BYOL [GSA⁺20], DINO [CTM⁺21a]) avoid explicit negative pairs. Instead, they use asymmetric architectures, typically through an online network and a momentum (EMA) teacher network, combined with stop-gradient operations to prevent collapse without requiring negative samples. DINO trains a student network to match the output distribution of a teacher network, with the teacher’s outputs sharpened and centered to prevent collapse:

$$\mathcal{L}_{\text{DINO}} = - \sum_{x' \in \{x_1^t, x_2^t\}} P_t(x') \log P_s(x'),$$

where x_1^t and x_2^t are two augmented views passed through the teacher, x' indexes these views, and $P_t(\cdot)$ and $P_s(\cdot)$ are the sharpened softmax output distributions of the teacher and student networks evaluated on the given view.

Joint Embedding Predictive Architectures (JEPAs). Introduced with I-JEPA [ADM⁺23] and extended to video with V-JEPA [BGP⁺24], JEPAs draw from elements of masked modeling and joint embedding methods; however, they contain their own distinctions. Given a set of visible context patches, a context encoder produces a set of embeddings. A predictor then uses these embeddings to predict the latent representation of masked target patches in the output space of a separate target encoder. The key point here is that this prediction does not happen in pixel space. The training objective minimizes the prediction error in the representation space:

$$\mathcal{L}_{\text{I-JEPA}} = \frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} \|s_{\theta}(\bar{z}_{\tau}) - z_{\tau}\|_2^2,$$

where $\mathcal{T}$ is the set of target blocks, s_{θ} is the predictor, $\bar{z}_{\tau}$ is the context encoder’s output conditioned on the target’s position, and z_{τ} is the target encoder’s representation of the masked block. Because prediction targets are embeddings rather than raw pixels, the model is not penalized for failing to reconstruct low-level details that may be irrelevant to semantics.

JEPAs also avoid using explicit augmentation invariances (so they do not necessarily require train-time transformations), instead letting the model learn which variations to abstract over from the structure of the data. This design trait is particularly relevant for medical imaging, where domain-specific augmentations are more difficult to curate and generally underexplored compared to natural image settings.

2.4.3 Supporting Concepts

Several concepts are essential for understanding SSL in practice, and are referred to in this Introduction and later chapters.

Augmentation invariance: JEA methods train models to produce similar representations for different augmentations of the same image. These augmentations may present as random crops, color jitter, Gaussian blur, horizontal flipping. The specific augmentation set encodes an assumption about what variations should be treated as irrelevant to semantics. For example in natural imaging, color jitter is reasonable (a cat is a cat regardless of lighting); in medical imaging, modifying intensity values may unintentionally modify diagnostic meaning as this feature directly effects interpretation. Thus, depending on the type of input, augmentation choices must be made carefully.

Representation collapse: A typical source of failure during training is collapse, in which the model learns to map all inputs to the same representation. This solution may trivially satisfy any similarity objective, but prevents the model from learning any meaningful representations. Contrastive methods prevent collapse with negative pairs. Non-contrastive methods prevent it through asymmetric architectures (stop-gradient and/or EMA encoder), centering operations, and prediction tasks that are engineered to be non-trivially satisfied.

EMA (exponential moving average) encoder: Many SSL methods use a slowly evolving target network whose weights are an exponential moving average of the online (student) network,

$$\theta_t \leftarrow m \theta_t + (1 - m) \theta_s,$$

where $m \in [0, 1)$ is a momentum coefficient (typically 0.996–0.9999) and θ_s are the online network weights. The EMA encoder provides a stable prediction target that changes slowly as training progresses. This component is important because if both networks are updated simultaneously by gradient descent, the optimization landscape can become unstable. A prime example of this in practice is in the DINO and I-JEPA frameworks, where the teacher encoder is updated with an EMA of the student.

Linear probing: A standard evaluation protocol for SSL representations is linear probing (Figure 2.3). In this setup, the pretrained backbone is frozen, and a single linear classifier is trained on top of its output embeddings using labeled data. Because the backbone cannot adapt, linear probing isolates the quality of the learned representation. A high linear probing accuracy indicates that semantically meaningful information is linearly separable in the learned embedding space.

Fine-tuning. In fine-tuning evaluation, all backbone weights are updated starting from the pretrained initialization (Figure 2.3). This measures the representation’s usefulness as a starting point when adapted for task-specific situations. This method is typically the evaluation that matters most for real-world, downstream applications.

2.4.4 Foundation Models

A foundation model is a large model pretrained on broad data at substantial scale, designed to transfer to many downstream tasks with minimal task-specific supervision [B⁺21]. The defining properties, particularly from the lens of what makes them useful, can be defined by the following:

- *Breadth:* supports diverse downstream tasks, such as recognition, segmentation, detection, and retrieval, ideally without major task-specific architectural modifications.

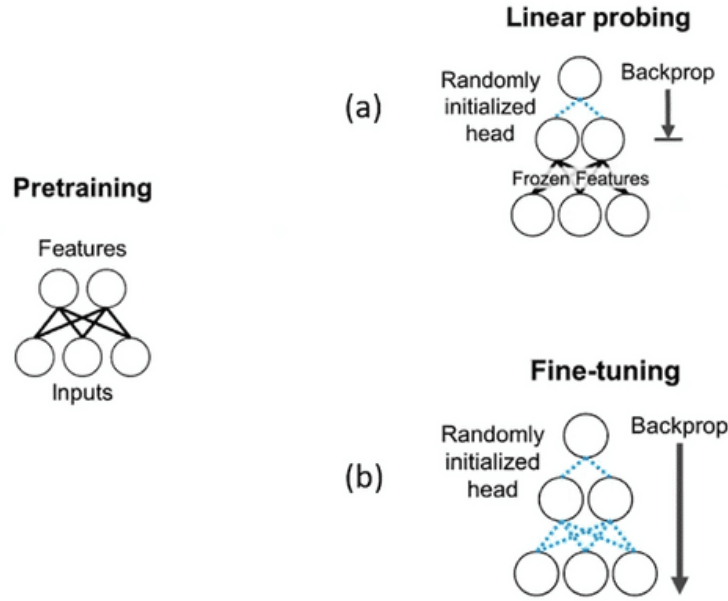


Figure 2.3: Linear probing versus full fine-tuning evaluation protocols. Adapted from [HCX⁺21].

- *Sample efficiency*: the pretrained representation reduces the labeled data needed for fine-tuning. A strong foundation model should achieve competitive performance with a small fraction of the labeled data required to train from scratch.
- *Robustness*: the representation generalizes to distribution shift better than narrow task-specific models. In medical imaging, this shift may present as images from different acquisition protocols, institutions, or populations.
- *Emergent Capability*: behaviors that arise from scale and were not explicitly optimized during pretraining. For example, the parameter scaling advancement from GPT-2 to GPT-3 allowed for in-context learning, where an LLM could be adapted to a downstream task with an additional prompt [BMR⁺20].

The foundation model paradigm was established in natural language processing. BERT [DCLT19] demonstrated that masked language modeling, in which randomly masked tokens are predicted from bidirectional context using a transformer encoder, yields general-purpose

language representations that can be fine-tuned to achieve state-of-the-art performance across diverse language benchmarks with minimal task-specific modification. GPT [RWC⁺19] and following advancements showed that autoregressive generation, predicting the next token, scales to very large models with increasingly broad and emergent capabilities, enabling few-shot and zero-shot task completion from natural language prompts. Together, these systems established the future template - scale plus self-supervision can yield representations that transfer broadly.

SSL is the dominant pretraining strategy for vision foundation models. We can leverage large unlabeled datasets to remove the annotation bottleneck, and self-supervised objectives yield robust representations. CLIP [RKH⁺21] uses contrastive pretraining on image-text pairs from the internet, learning aligned image and language embeddings. DINO [CTM⁺21a] and its successor DINOv2 [ODM⁺24] apply self-distillation with no labels on large curated image collections, producing features with strong spatial properties that support segmentation, detection and global image classification. The Segment Anything Model (SAM) [KMR⁺23] was trained on over one billion masks and implements a promptable segmentation interface. SAM represents the first general-purpose segmentation foundation model and has seen rapid adoption as a backbone for interactive annotation and downstream segmentation tasks.

Foundation models in medical imaging: SSL-based pretraining has seen rapid uptake across medical imaging. In computational pathology, models such as UNI [C⁺24] and Virchow [V⁺24] are pretrained on millions of whole-slide image patches and transfer broadly across cancer types, staining protocols, and downstream tasks. In radiology, vision-language models such as BioViL-T [BHL⁺23] and BiomedCLIP [ZXU⁺23] leverage paired imaging and clinical text, while modality-specific models such as CT-CLIP [H⁺24] use large CT archives for self-supervised pretraining.

Nevertheless, there are still largely unresolved pain points within this field of research. What constitutes a generalized medical imaging foundation model remains an open question. Strong performance on one task or organ does not reliably generalize across the full breadth of

medical imaging. Standardized evaluation benchmarks are needed to systematically compare pretraining strategies, and the relationship between pretraining scale, data diversity, and downstream performance is not well established in the medical domain. Questions regarding which SSL objectives are best suited to which imaging modalities is another underexplored avenue of research. In this dissertation, we begin answering such questions within the context of ultrasound imaging, while laying the methodological groundwork and questions for future research to tackle.

CHAPTER 3

From Bench to Bedside: How Artificial Intelligence Is Changing Thyroid Nodule Diagnostics

Associated Publications: Sant, V. R.*, **Radhachandran, A.***, Ivezic, V., Lee, D. T., et al. (2024). From Bench-to-Bedside: How Artificial Intelligence is Changing Thyroid Nodule Diagnostics, a Systematic Review. *The Journal of Clinical Endocrinology & Metabolism*, 109(7), 1684–1693.

Use of artificial intelligence (AI) to predict clinical outcomes in thyroid nodule diagnostics has grown exponentially over the past decade. The greatest challenge is in understanding the best model to apply to one’s own patient population, and how to operationalize such a model in practice. A systematic literature search of PubMed and IEEE Xplore was conducted for English language publications between January 1, 2015 and January 1, 2023 studying diagnostic tests on suspected thyroid nodules that utilized AI. A total of 61 studies were identified; all performed external validation, sixteen studies were prospective, and 33 compared a model to physician prediction of ground truth. A diagnostic pipeline was abstracted, yielding five high-level outcomes: (1) nodule localization, (2) ultrasound risk score, (3) molecular status, (4) malignancy, and (5) long-term prognosis. Models predominantly used ultrasound images to predict malignancy. Of four FDA-approved products, only S-Detect was extensively validated. Implementing an AI model locally requires data sanitization and re-validation to ensure appropriate clinical performance.

3.1 Introduction

Thyroid nodules are highly prevalent, affecting over 50% of people above the age of 50 [MWB55, WC97]. Increasingly, their discovery may be incidental, given the increased use of cross-sectional imaging [HLM⁺15]. Nodules may span the diagnostic spectrum from benign and asymptomatic, to symptomatic, to cancerous. Diagnostic workup often involves several steps, associated cost, and resultant patient anxiety [HGM⁺82, KWZ18, PSW⁺21].

Ultrasound (US) and fine needle aspiration (FNA) biopsy have been the gold standard for thyroid nodule assessment, but performance and interpretation are plagued by variability across physicians interpreting US findings, performing the FNA, and interpreting cytology, yielding frequent inadequate or indeterminate results [PKJ⁺10, CA17]. Assessment of a thyroid nodule represents an area where technology may offer opportunities for improvement [BPA⁺21]. With the development of new techniques and greater accessibility of computational power, artificial intelligence (AI) has been applied by thyroid researchers over the past decade to automate nodule localization and improve risk stratification, in order to decrease operator variability and the need for potentially unnecessary biopsies or diagnostic surgery [TT23].

Numerous AI models have been proposed in the literature, but the greatest challenge is in understanding a model’s applicability to one’s own patient population, and how to operationalize such a model in practice. Efforts have been made to identify available models. Toro-Tobon et al. systematically identify manuscripts applying AI to various thyroid conditions, reporting on model performance and stage of development. Similarly, Taha et al. review a selection of 18 AI models in thyroid diagnostics. Ludwig et al. report on a selection of 33 models utilizing AI to diagnose and classify thyroid nodules [TTLTD⁺23, TSTM⁺23, LLM⁺23]. However, these reviews do not systematically assess relative model strength and external validity, making assessment of future or external performance difficult. Our aim is to perform a systematic literature review to identify and compare the various AI models developed for thyroid nodule diagnostics, their relative strength and external validity, and

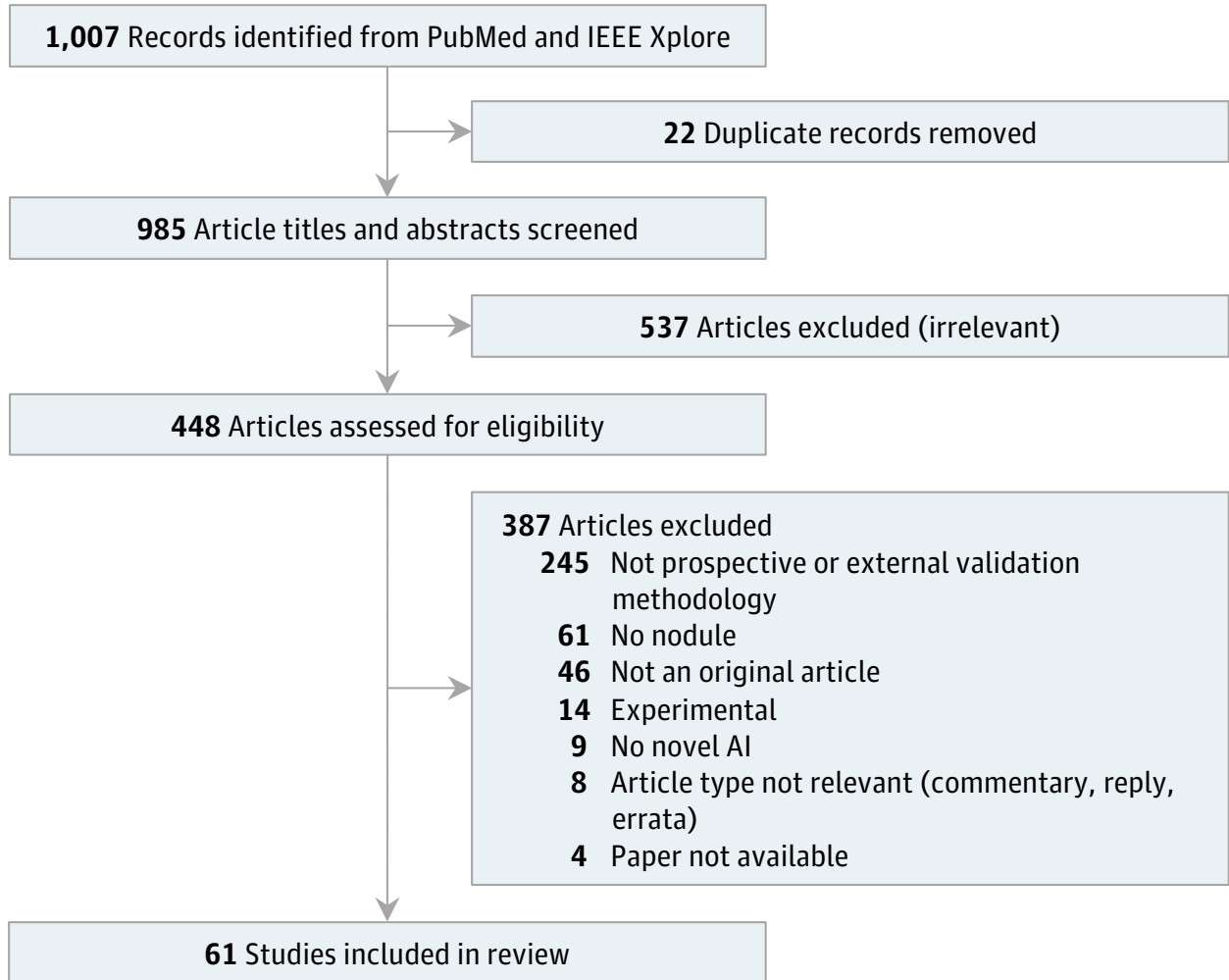


Figure 3.1: Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) Flow Diagram. Meta-analysis was not performed as part of this review.

the clinical inputs utilized for each outcome predicted.

3.2 Materials and Methods

3.2.1 Systematic Literature Search

An inclusive query was developed and then refined to ensure that known relevant articles appeared in the results. The PubMed and IEEE Xplore databases were queried with the following terms: (artificial intelligence OR machine learning OR deep learning OR neural network) AND (thyroid). The search was restricted to English language papers published on

or after January 1, 2015 through the search date, January 1, 2023. Titles and abstracts were screened for relevance (985 studies), followed by full-text review (448 studies) to determine inclusion. During full-text review, relevant articles identified in citations were reviewed for inclusion. The PRISMA flow diagram [PMB⁺21] denotes the number of studies excluded at each step (Figure 3.1).

3.2.2 Study Selection

We included original research studying diagnostic tests on suspected thyroid nodules that utilized AI. AI techniques included non-deep learning (e.g., regression, decision tree, random forest) and deep-learning (e.g., neural network) techniques. We excluded articles without prospective or external validation, non-primary literature, duplicates, focused on non-nodular thyroid conditions, not using AI, and those incidentally utilizing AI in support of an experimental diagnostic outside standard clinical practice (e.g., spectrometry, spectroscopy).

3.2.3 Data Abstraction and Analysis

Each paper was independently reviewed by two of three authors (V.S., A.R., and V.I.), first for inclusion, then for data extraction, and discrepancies were addressed by discussion. Details including sample size, study design, statistical methods, AI methodology, clinical inputs, and outcome predicted were tabulated and level of evidence was rated according to the Oxford Centre for Evidence-Based Medicine [OCE].

3.3 Results

In total, 61 studies were included (Table 3.1). Oxford level of evidence ranged from 2–3 (Table 3.3). In total, 61 studies met the inclusion criteria of including external validation. Of these, 33 also compared the model’s performance to physician assessment (physician comparison). Of the 61 studies, 16 were prospective in study design. The intersection of

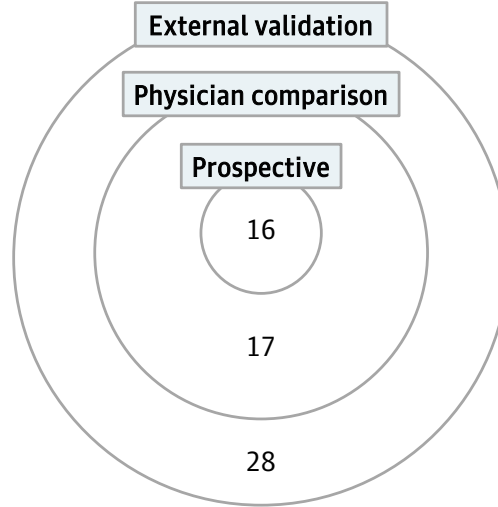


Figure 3.2: Intersection of study designs. This Venn diagram quantifies the different combinations of study designs captured in this review.

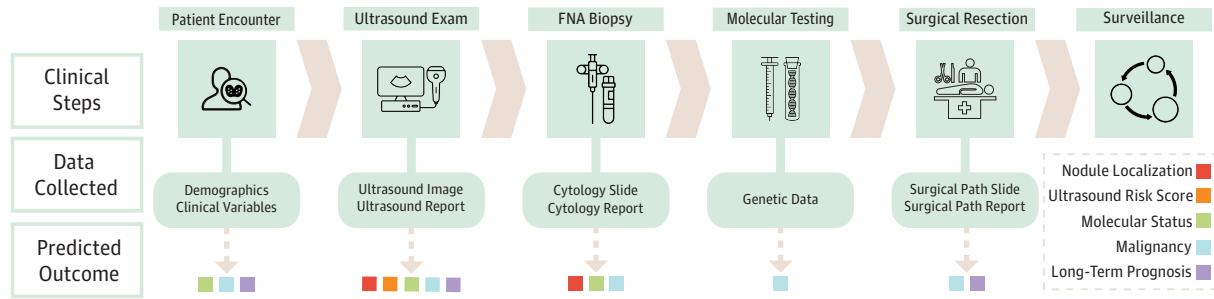


Figure 3.3: Clinical workflow with resultant data inputs and predicted outcomes. A diagnostic pipeline was abstracted from the papers starting with known clinical steps and culminating in five high-level predicted outcomes. If a certain input modality was included in a paper for one of the predicted outcomes, the corresponding data collection box had an arrow pointing out that outcome.

study designs is shown in Figure 3.2. Statistical validation of model stability was reported in 50 papers (Table 3.3).

Starting with known clinical steps, a diagnostic pipeline was abstracted from the papers, representing the clinical data collected and used in AI-models, and the resultant AI-predicted outcomes (Figure 3.3). The distribution of data inputs used to predict clinical outcomes is visualized in Figure 3.4. Additional details on the four FDA-approved thyroid AI diagnostic solutions are reported in Table 3.2. The following five high-level abstracted outcomes were used to organize reporting of the remaining results.

Predicted Clinical Outcome	Nodule localization	0	0	5	0	0	0	0
	US risk score	0	0	16	1	0	0	0
	Molecular status	0	0	0	0	0	0	2
	Malignancy	1	1	42	1	2	5	2
	Long-term prognosis	0	1	2	2	1	2	0
		Demographics	Clinical variables	US image/report	CT/MRI	Cytology slide/report	Genetic data	Surgical path slide/report
		Data Input						

Figure 3.4: Distribution of data inputs used to predict clinical outcomes. Each paper was categorized by data input and predicted clinical outcome. If a paper was multi-input or multi-outcome, it was counted for each input-outcome combination.

3.3.1 Nodule Localization

Localization of nodules in ultrasound images was assessed by five papers [NZT⁺22, CBP⁺17, KHH19, YYY⁺20, LHC⁺21]. Localization was performed through the tasks of detection and segmentation. Detection involves identifying an image region that contains the object of interest. To create training labels for the ultrasound studies, radiologists annotated nodules on ultrasound images with rectangular bounding boxes to serve as ground truth for model training [KHH19]. Segmentation involves identifying the specific contour of the object of interest. To create training labels, radiologists manually contoured nodules on ultrasound images to serve as ground truth for model training [NZT⁺22, YYY⁺20].

The goals of such models were two-fold: (1) to improve speed and standardization in determining whether an image had thyroid nodules present, and (2) to focus feature extraction on suspected nodule regions for use in downstream tasks. For instance, Yang et al. [YYY⁺20] found that a model trained to predict malignancy from ultrasound images performed better

when trained with nodule-focused features.

Prospective localization studies used Samsung’s commercial computer aided diagnosis (CAD) software, S-Detect, which first requests users to perform detection [CBP⁺17, WZW⁺20, BSBW⁺20, XYZ⁺19, JKH⁺19, YHC⁺18, HZL⁺22]. S-Detect then uses AI to propose segmentation masks for the user to choose from, before generating an AI-based US risk score and malignancy prediction. Despite the effort to improve standardization through semi-automated segmentation, Jeong et al. [JKH⁺19] found that when the CAD was used by experienced radiologists, it performed better than when used by less experienced radiologists. They concluded that persistent performance variability was likely due to user variability, even in choosing among segmentation candidates.

3.3.2 Ultrasound Risk Score

Sixteen papers predicted sonographic features associated with risk of malignancy or a composite score such as the American College of Radiology Thyroid Imaging Reporting and Data System (TIRADS) [CBP⁺17, KHH19, LHC⁺21, WZW⁺20, XYZ⁺19, JKH⁺19, YHC⁺18, HZL⁺22, CWB⁺17, GLJ⁺18, BCY⁺20, DDJ⁺20, ZLR⁺22, HHP21, CFS⁺23, STN⁺22]. Sonographic risk stratification systems aim to decrease variability in reporting nodule features and estimating malignancy risk, but inter- and intra-reporter variability persist [CBP⁺17]. These papers used AI to decrease variability further, reduce the burden on physicians manually assessing a score, and improve US risk score performance.

Six papers prospectively validated the commercial product S-Detect, which predicts sonographic features as well as malignancy [CBP⁺17, WZW⁺20, XYZ⁺19, JKH⁺19, YHC⁺18]. Wei et al. [WZW⁺20] found that less experienced radiologists benefited most from S-Detect assistance, bringing their TIRADS assessments in-line with those of experienced radiologists, which were considered as ground truth. Choi et al. [CBP⁺17] and Xia et al. [XYZ⁺19] found substantial agreement between S-Detect and experienced radiologists for all feature predictions apart from margin. Jeoung et al. [JKH⁺19] and Yoo et al. [YHC⁺18] reported

similar results with discrepancy in composition and spongiform features, respectively.

Cui et al. prospectively evaluated AI-TIRADS, a machine learning (ML) model that reassigned new values to constituent TIRADS features in an effort to preserve model explainability [CFS⁺23]. They compared its performance to ACR’s original (2017) version of TIRADS, and found a lower rate of unnecessary FNA (41.0% vs 47.8%) and missed cancer diagnosis (22.8% vs 27.5%).

3.3.3 Molecular Status

Testing tumors for genetic alterations and assessing differential RNA expression has proven helpful in predicting malignancy, tumor aggressiveness, and response to targeted therapy, but cost prohibits their routine use [CYZM21, BKK⁺20]. Two papers used AI to predict molecular status opportunistically from routinely obtained clinical inputs such as ultrasound, to provide an alternative: Anand et al. [AYK⁺21] and Dolezal et al. [DTL⁺21] analyzed cytology and surgical pathology slides to predict high-risk mutations. Dolezal et al. [DTL⁺21] took this one step further, externally validating their model to predict BRAF-RAS gene status (area under the receiver operating characteristic curve (AUC) 0.97) and non-invasive follicular thyroid neoplasm with papillary-like nuclear features (NIFTP), a liminal diagnosis with high inter-observer variability.

3.3.4 Malignancy

The majority of papers (51) predicted malignancy, aiming to improve speed and reliability of diagnosis and to avoid unnecessary biopsies for patients [CBP⁺17, KHH19, YYY⁺20, LHC⁺21, WZW⁺20, BSBW⁺20, XYZ⁺19, JKH⁺19, YHC⁺18, HZL⁺22, HHP21, CFS⁺23, STN⁺22, BKK⁺20, DTL⁺21, PHK⁺16, WLS⁺18, LZZ⁺19, PHS⁺19, SCM⁺19, KLH⁺20, LYLC20, SZZ⁺20, WGY⁺20, ZWCW20, ZDJ⁺20, ZJD⁺20, ZWT20, BTS⁺21, LCJ⁺21, PKOK21, PLL⁺21, ZZL⁺21, ZZY⁺21, ZAJ⁺21, ZJB⁺21, ZDJ⁺22, HCS⁺22, JMK⁺22, JPO⁺22, KLM⁺22, KCH⁺22, LDY⁺22, RSH⁺22, WWL⁺22, WXL⁺22, XWW⁺22, YPW⁺23, CYZM21]. Malignancy

nancy predictions used either of two sources as ground truth: cytology (FNA) and surgical pathology (resection). Several papers used both, although none to our knowledge included nodules that were not biopsied and presumed benign sonographically. Most studies excluded cytologically indeterminate (Bethesda III or IV) nodules from their analysis, with only 5 papers including them for evaluation [STN⁺22, KLM⁺22, JPO⁺22, LDY⁺22, YPW⁺23].

All 16 prospective studies assessed malignancy, including seven that evaluated the commercial product S-Detect. Barczynski et al. [BSBW⁺20] and Wei et al. [WZW⁺20] found S-Detect had greater sensitivity and specificity in predicting malignancy from ultrasound compared to a junior physician, but less compared to a senior physician. Performance of junior physicians improved significantly when assisted by AI, but senior physician performance did not. Choi et al. [CBP⁺17] found S-Detect to have similar sensitivity but lower specificity and AUC compared to an experienced physician. Xia et al. [XYZ⁺19] found similar results, and in subgroup analysis noted a shortfall in model performance specifically with malignant nodules (PTC or FTC). Jeong et al. [JKH⁺19] noted lower specificity compared to senior physicians only, as well as operator dependence of CAD performance as noted above. Yoo et al. [YHC⁺18] found that a CAD-assisted physician had higher sensitivity (92.0%) than either physician (84.0%) or CAD (80.0%) alone, and Huang et al. [HZL⁺22] found a similar result across all physician experience levels. Although there was no universal definition of experience level, “senior physicians” were typically those with > 9–20 years of experience with thyroid ultrasound.

AI-SONIC is a commercial CAD that completely automates nodule detection, but is not currently approved for use in the US. Zhang et al. [ZWCW20] prospectively studied AI-SONIC, finding that compared to a senior physician, the CAD had similar specificity (86.0% for both) but lower sensitivity (71.5% vs. 95.2%) and AUC (0.79 vs 0.91). However, they noted that sensitivity improved for both senior and junior physicians when assisted by CAD (95.2% vs 97.8% and 75.3% vs 88.2%, respectively). For non-papillary follicular-patterned lesions, Wang et al. [WXL⁺22] found that AI-SONIC had greater specificity (0.802) than

radiologists of all experience levels (0.549 for junior, 0.533 for mid-level, 0.478 for senior). However, the senior radiologist again outperformed the AI in terms of sensitivity (0.898 vs 0.729). Xu et al. [XWW⁺22] found a similar pattern.

Of the prospective studies, Kim et al. [KCH⁺22] and Yoo et al. [YHC⁺18] used cytology alone as malignancy ground truth, seven papers [BSBW⁺20, CFS⁺23, ZWCW20, ZZY⁺21, WWL⁺22, WXL⁺22, XWW⁺22] used surgical pathology alone, and the rest used a more balanced mix of both, reflecting the respective potential biases as described above. One paper reported receiving funding from Samsung to study S-Detect [CBP⁺17] but the remaining prospective studies reported no industry conflicts of interest.

3.3.5 Long-Term Prognosis

Long-term prognostic cancer outcomes were characterized by seven papers, including metastases [BKK⁺20, WYJ⁺22, YWL⁺22, AAMMAA⁺23, ZSS⁺22] and recurrence/survival [CYZM21, SLL⁺22a]. Thyroid cancer prognosis currently involves dynamic risk stratification, whereby initial risk is assessed from operative and histopathologic findings and modified over a subsequent time interval using clinical evidence of tumor response. AI models tried to predict tumor biology from the outset, with the aim of allowing the physician to better counsel the patient before surgery regarding anticipated cancer aggressiveness, lymph node metastases, and need for adjuvant treatment. Most models used as inputs imaging or genetic data alone, while Zou et al. additionally utilized clinical variables [ZSS⁺22].

Bhalla et al. [BKK⁺20] identified a 36 gene panel that could predict American Joint Committee on Cancer (AJCC) staging as I/II vs III/IV with PPV 84% and sensitivity 76%. Chen et al. [CYZM21] trained and externally validated a model using DNA methylation data and were able to better predict overall survival than could be done through tumor, node, metastases staging.

3.4 Discussion

The past decade has seen a major push to deliver on the promise of AI to improve physician workflow, patient outcomes, and healthcare delivery. Physicians may benefit from faster, standardized, and less burdensome processes to predict nodule localization, sonographic features, and malignancy risk, and can more appropriately counsel patients regarding extent of surgery and de-escalation of care. Patients may avoid unnecessary biopsies and benefit from improved diagnostic accuracy, decreased need for diagnostic surgery, and clearer expectations for long-term prognosis. Healthcare may be delivered at lower cost as information later in the diagnostic pipeline is predicted opportunistically from routinely obtained data, and task automation allows healthcare providers to practice at the top of their scope.

3.4.1 The AI-Empowered Physician

For physicians who wish to use a commercial AI solution, four CADs have been approved by the FDA: AmCAD-UT, S-Detect, Koios DS, and MEDO-Thyroid, as reviewed in further detail in Table 3.2. Although none are currently in widespread clinical use, S-Detect has undergone the most extensive validation.

Sonographic risk stratification and malignancy prediction, outcomes with significant clinical impact and the most readily available data, received the most attention in analyzed papers. Specifically with commercial CADs, S-Detect had difficulty with automated margin and composition prediction. AI-SONIC had comparable or better specificity but worse sensitivity than experienced radiologists. However, rather than trying to replace a physician, the greatest value appeared to be in democratizing care via the concept of the “AI-empowered physician.” Studies found that the AI-empowered physician performed better than physician or AI alone. Several studies found that inexperienced physicians with AI-assistance made malignancy predictions with similar accuracy as experienced physicians. This use of AI has tremendous potential to enable patients at non-specialized centers to receive similar diagnostic

care as patients presenting to tertiary-care centers. The biggest barrier to widespread adoption is the need for pilot testing in each local environment to identify how AI can provide the greatest benefit (e.g., time, staffing, or quality improvement).

How a paper chose their data source for malignancy prediction affected potential bias. Some papers used cytology alone, others used surgical pathology alone, and some used both. Using surgical pathology alone limited data to patients who required surgery, typically those with larger nodules or higher suspicion of malignancy. Papers that used cytology alone had more benign cases, but could not account for false negative biopsies, and typically excluded indeterminate results. Thus, caution should be used when extrapolating the findings of such papers beyond the population evaluated.

Although the literature is inundated with retrospective studies reporting successful nodule localization techniques, few validated these approaches prospectively or externally. In fact, most commercial CADs do not entirely automate nodule localization, reflecting that real-world segmentation is not easy to automate. Additionally, nodule localization is an important technical intermediary step but not a critical clinical endpoint, which may explain the dearth of prospective or external validation studies.

Molecular prediction was externally validated in fewer papers, reflecting the recency of molecular testing gaining mainstream utility. In most countries outside the US, molecular testing for indeterminate nodules has not entered routine practice. Although opportunistic molecular prediction has strong clinical potential, lack of data access may hold up practical application. Long-term outcome prediction suffered similar inattention. Although opportunities to develop such datasets are not limited by the same logistical and financial issues as molecular testing, maintaining long-term oncologic follow-up requires significant effort. This data curation issue likely drove researchers to assess other more readily available outcomes. Moreover, as a later stage in the clinical pipeline, it may have appeared comparatively lower-value in a disease with generally good prognosis.

3.4.2 The Ideal AI Model

Critical review of an AI model’s versatility requires assessment of several criteria including performance on prospective and external datasets, statistical validity, adequate training cohort size, and evaluation of bias in the training dataset. Utilizing a standardized methodology for model evaluation allows one to better distinguish clinically effective tools from poorly designed ones.

Models solely trained and evaluated on single retrospective datasets have uncertain external validity. Prospective evaluation and external validation promote exposure to unseen data, which can help uncover overfitting and bias. Beyond robust study design, reporting statistical validity through bootstrapping and cross-validation is important to understanding model stability. Of the reviewed papers, 82% included a confidence interval, standard error, or p-value with their metrics of model performance. Such reporting helps evaluate variation in model performance on different data subsets, giving greater confidence in a model’s performance on varied real-world data.

Next, sample size and dataset curation must be considered. A minimum sample size is typically necessary to ensure data diversity, although numbers are application-specific. However, even large datasets may yield limited diversity due to poor dataset curation and risk developing models that are overfit to idiosyncrasies unique to the training data. Validation of an AI model should be performed on a dataset that closely reflects the expected real-world use case. For example, most validation studies excluded cytologically indeterminate nodules, which comprise a significant portion of real-world thyroid nodule FNA results; thus, expected performance of existing AI models for such patients remains unclear. Similarly, given the low overall prevalence of malignancy in thyroid nodules, it is important that a validation dataset reflects this pattern as well. Some studies artificially balance the frequency of benign and malignant cases for model training, and use similar proportions during model validation, not recognizing that these validation results may not reflect performance in the expected target population.

Rigorous implementation of these methods is intensive but crucial to improving confidence in real-world validity of an AI model. While we see their utility in thyroid nodule diagnostics, they apply even more broadly in advancing AI-powered diagnostics in other disease domains.

3.4.3 Future Directions

We anticipate the next decade to yield both advancements in AI research applications as well as improvement in integration of AI solutions into mainstream clinical practice.

As seen in Figure 3.4, few studies leveraged rich data modalities such as clinical variables and cytopathology slides. Future technical innovations should aim to integrate multiple data sources in a multimodal approach that could emulate a physician’s synthesis of all available clinical data to provide the most accurate prediction. Along a similar vein, future methods leveraging ultrasound should explore the impact of multiple frames (through cine clips or sampling full image studies) in improving automated diagnosis performance. Given that in clinical practice physicians make use of multiple views of a nodule to make a diagnosis, this could address a limitation of most existing CAD systems that rely on manually-selected, single frame inputs. Several outcomes have also received minimal attention, such as prediction of molecular status and long-term outcomes, although they have high clinical value. Expanding data access through more effective data-sharing and publication of open-source datasets will enhance research replication and refinement of existing methodologies. Lastly, stronger collaboration between computational researchers and clinicians will ensure that the most important clinical problems are addressed, and that models are designed most effectively.

Regarding clinical integration, four commercial CADs are FDA-approved for use but no reports in the literature mention their routine clinical use, to our knowledge. Each takes slightly different inputs and has slightly different strengths and potential use cases. Building physician trust and adapting to local practice variation will be the two most important steps to enabling integration of AI solutions into mainstream clinical use. Deep learning models in particular can appear to some as a “black-box”, and in the absence of thorough

external validation, may engender physician mistrust in the predicted results. Tools such as Gradient-weighted Class Activation Mapping (Grad-CAM) to visualize influential regions in an input image may help overcome this issue [HCS⁺22]. Even without obvious explainability, rigorous prospective and external validation may help physicians eventually feel comfortable using a model.

The second important hurdle to overcome is local variation. Institutions use different ultrasound machines, serve patients with differing characteristics, and utilize different workflows. In contrast, training datasets used to train AI models are often “idealized” in that they lack artifacts and noise, furthering the need for effective, real-world validation. Integration of AI will require careful local stakeholder and workflow analysis to select the most impactful value proposition for that site, as well as infrastructure to locally pilot models.

3.4.4 Limitations

The included studies had significant heterogeneity in model technique, data sources, and outcomes predicted. As a result, direct comparison between studies was not possible. While papers were limited to English language, most models were trained on datasets from China, Korea, or India, where screening and surgical practices are notably different compared to the United States and Europe. This challenges model validity outside their respective environments. Even within a similar clinical environment, implementing a model locally requires data sanitization, re-validation, and potential retraining to ensure performance.

We identified 61 studies utilizing AI to predict thyroid nodule-focused outcomes, of which 16 were prospective, 61 performed external validation, and 33 performed physician-comparison. Models using ultrasound images to predict malignancy predominated, and one of the four FDA-approved CADs (S-Detect) was extensively validated. Further validation and integration into clinical workflow will help models achieve greater clinical utility.

Table 3.1: Included articles and outcomes assessed.

Source by outcome	Outcome ground truth
<i>Nodule localization</i>	
Nie et al., 2022 [NZT ⁺ 22]	Nodule contour
<i>Ultrasound risk score</i>	
Chi et al., 2017 [CWB ⁺ 17]	US risk score
Duan et al., 2020 [DDJ ⁺ 20]	US risk score
Zhang et al., 2022 [ZLR ⁺ 22]	US risk score
Gao et al., 2018 [GLJ ⁺ 18]	US risk score
Bai et al., 2020 [BCY ⁺ 20]	US risk score
<i>Molecular status</i>	
Anand et al., 2021 [AYK ⁺ 21]	Molecular test
<i>Malignancy</i>	
Pankratz et al., 2016 [PHK ⁺ 16]	Surgical pathology
Wang et al., 2018 [WLS ⁺ 18]	Surgical pathology
Song et al., 2020 [SZZ ⁺ 20]	Cytology
Zhou et al., 2020 [ZJD ⁺ 20]	Cytology
Zhou et al., 2020 [ZWT20]	Cytology
Böhlend et al., 2021 [BTS ⁺ 21]	Surgical pathology
Lee et al., 2021 [LCJ ⁺ 21]	Cytology
Park et al., 2021 [PKOK21]	Surgical pathology
Zhang et al., 2021 [ZZL ⁺ 21]	Cytology
Jia et al., 2022 [JMK ⁺ 22]	Surgical pathology
Jin et al., 2022 [JPO ⁺ 22]	Cytology, surgical pathology
Keutgen et al., 2022 [KLM ⁺ 22]	Surgical pathology
Liu et al., 2022 [LDY ⁺ 22]	Cytology
Randolph et al., 2022 [RSH ⁺ 22]	Surgical pathology
Li et al., 2019 [LZZ ⁺ 19]	Cytology, surgical pathology
Park et al., 2019 [PHS ⁺ 19]	Cytology, surgical pathology

Continued on next page

Table 3.1 – *continued from previous page*

Source by outcome	Outcome ground truth
Song et al., 2019 [SCM ⁺ 19]	Cytology, surgical pathology
Koh et al., 2020 [KLH ⁺ 20]	Cytology, surgical pathology
Wei et al., 2020 [WGY ⁺ 20]	Surgical pathology
Zhang et al., 2020 [ZDJ ⁺ 20]	Surgical pathology
Zhu et al., 2021 [ZAJ ⁺ 21]	Cytology, surgical pathology
Zhu et al., 2021 [ZJB ⁺ 21]	Cytology
Zhu et al., 2021 [ZDJ ⁺ 22]	Cytology, surgical pathology
Han et al., 2022 [HCS ⁺ 22]	Cytology, surgical pathology
Yang et al., 2022 [YPW ⁺ 23]	Cytology, surgical pathology
Barczyński et al., 2020 [BSBW ⁺ 20]	Surgical pathology
Liang et al., 2020 [LYLC20]	Cytology, surgical pathology
Zhang et al., 2020 [ZWCW20]	Surgical pathology
Peng et al., 2021 [PLL ⁺ 21]	Cytology, surgical pathology
Zhu et al., 2021 [ZZY ⁺ 21]	Surgical pathology
Kim et al., 2022 [KCH ⁺ 22]	Cytology
Wang et al., 2022 [WWL ⁺ 22]	Surgical pathology
Wang et al., 2022 [WXL ⁺ 22]	Surgical pathology
Xu et al., 2022 [XWW ⁺ 22]	Surgical pathology
<i>Long-term prognosis</i>	
Abbasian et al., 2022 [AAMMAA ⁺ 23]	Metastases
Yu et al., 2022 [YWL ⁺ 22]	Metastases
Zou et al., 2022 [ZSS ⁺ 22]	Metastases
<i>Multi-outcome</i>	
Bhalla et al., 2020 [BKK ⁺ 20]	Metastases, surgical pathology
Yang et al., 2020 [YYY ⁺ 20]	Nodule contour, surgical pathology
Chen et al., 2021 [CYZM21]	Surgical pathology, survival/recurrence
Dolezal et al., 2021 [DTL ⁺ 21]	Molecular test, surgical pathology
Swan et al., 2022 [STN ⁺ 22]	Surgical pathology, US risk score

Continued on next page

Table 3.1 – *continued from previous page*

Source by outcome	Outcome ground truth
Wu et al., 2022 [WYJ ⁺ 22]	Metastases, surgical pathology
Kim et al., 2019 [KHH19]	Cytology, nodule presence, surgical pathology, US risk score
Han et al., 2021 [HHP21]	Cytology, surgical pathology, US risk score
Liang et al., 2021 [LHC ⁺ 21]	Cytology, nodule contour, surgical pathology, US risk score
Stenman et al., 2022 [SLL ⁺ 22a]	Surgical pathology, survival/recurrence
Choi et al., 2017 [CBP ⁺ 17]	Cytology, nodule contour, surgical pathology, US risk score
Yoo et al., 2018 [YHC ⁺ 18]	Cytology, US risk score
Jeong et al., 2019 [JKH ⁺ 19]	Cytology, surgical pathology, US risk score
Xia et al., 2019 [XYZ ⁺ 19]	Cytology, surgical pathology, US risk score
Wei et al., 2020 [WZW ⁺ 20]	Cytology, surgical pathology, US risk score
Cui et al., 2022 [CFS ⁺ 23]	Surgical pathology, US risk score
Huang et al., 2022 [HZL ⁺ 22]	Cytology, surgical pathology, US risk score

Table 3.2: Assessment of commercially available CADs in the US.

Commercially Available CAD	Input	Output	Nodule Localization	Downstream Clinical Implications	Limitations	Clinical Workflow Integration
AmCAD-UT	Static ultrasound image	Nodule features (echogenicity, echogenic foci, texture, margin, tumor shape, anechoic area percentage); Ultrasound risk score (TI-RADS, ATA, other international systems); Structured summary report	Automated	Provides a malignancy risk for several risk stratification systems, along with quantitative and visual data on key nodule features, allowing for faster and more efficient patient triaging	Lack of extensive external validation in clinical setting; Expects ultrasounds with discrete nodules >1 cm [fda18, WTTZM22]	Windows-based platform intended for use on a PC or at a workstation [fda18, WTTZM22]

Continued on next page

Table 3.2 – *continued from previous page*

Commercially Available CAD	Input	Output	Nodule Localization	Downstream Clinical Implications	Limitations	Clinical Workflow Integration
S-Detect	Static ultrasound image	Nodule features (composition, shape, orientation, margins, echogenicity, and spongiform appearance); Binary recommendation of possibly malignant/benign; Structured summary report	Semi-automated (user marks middle of nodule)	Primarily improves diagnostic performance for junior radiologists [WZW ⁺ 20]	Primarily improves diagnostic performance for junior radiologists; Does not incorporate non-imaging patient data; Performance is dependent on choice of input image which is dependent on operator experience; Potential discrepancy in nodule margin and composition assessment; Only works with vendor-specific ultrasound machines	Developed for a vendor-specific ultrasound machine [WTTZM22]
Koios DS	2 static ultrasound images (axial and sagittal) [koia]; Location on thyroid	Nodule features (echogenicity, echogenic foci, tumor shape, composition, margin); Thyroid nodule dimensions; Ultrasound risk score (TI-RADS)	Manual		Lack of extensive external validation of this software in real-world clinical setting; Requires manual input in the form of ROI to localize nodule [koib, fda21]	Used at the PACS workstation [WTTZM22]
Exo AI (formerly MEDO-Thyroid)	Static ultrasound image(s)	Thyroid lobe and nodule dimensions; Ultrasound risk score (TI-RADS) (based on user input); Structured summary report	Semi-automated	Improves ease of reporting an ultrasound study for clinicians	Lack of extensive external validation of this software in real-world clinical setting; Requires manual input in the form of ROI to localize nodule; Only tested on Philips, GE and Siemens ultrasound devices [fda20]	Cloud-based software

Table 3.3: Included articles, methodology and level of evidence. Abbreviations: P = prospective, R = retrospective, EV = external validation, PC = physician comparison; ML = machine learning, NN = neural network (deep learning). Asterisk (*) represents number of images utilized if number of patients not reported.

Source by outcome	Cohort size	Study design	AI method	Inputs used	Outcome ground truth	Stat. valid.	Oxford LoE
<i>Nodule localization</i>							
Nie et al., 2022 [NZT ⁺ 22]	483	R, EV	NN	US image	Nodule contour	Y	2
<i>Ultrasound risk score</i>							
Chi et al., 2017 [CWB ⁺ 17]	592	R, EV	NN	US image	US risk score	N	2
Duan et al., 2020 [DDJ ⁺ 20]	3,933	R, EV	NN	US image	US risk score	N	2
Zhang et al., 2022 [ZLR ⁺ 22]	978	R, EV	NN	US image, CT/MRI	US risk score	Y	2
Gao et al., 2018 [GLJ ⁺ 18]	262	R, EV, PC	NN	US image	US risk score	Y	2
Bai et al., 2020 [BCY ⁺ 20]	14,991	R, EV, PC	NN	US image	US risk score	Y	2
<i>Molecular status</i>							
Anand et al., 2021 [AYK ⁺ 21]	529	R, EV	NN	Surgical pathology slides	Molecular test	Y	2
<i>Malignancy</i>							
Pankratz et al., 2016 [PHK ⁺ 16]	27	R, EV	ML	Genetic data	Surgical pathology	Y	2
Wang et al., 2018 [WLS ⁺ 18]	4,453	R, EV	NN	US image	Surgical pathology	N	2
Song et al., 2020 [SZZ ⁺ 20]	676	R, EV	NN	US image	Cytology	N	2
Zhou et al., 2020 [ZJD ⁺ 20]	1,734	R, EV	NN	US image	Cytology	N	2
Zhou et al., 2020 [ZWT20]	5,602	R, EV	NN	US image	Cytology	Y	2
Böhland et al., 2021 [BTS ⁺ 21]	289	R, EV	NN	Surgical pathology slides	Surgical pathology	Y	2
Lee et al., 2021 [LCJ ⁺ 21]	8,557	R, EV	NN	US image	Cytology	N	2

Continued on next page

Table 3.3 – *continued from previous page*

Source by outcome	Cohort size	Study design	AI method	Inputs used	Outcome ground truth	Stat. valid.	Oxford LoE
Park et al., 2021 [PKOK21]	763	R, EV	ML	Genetic data	Surgical pathology	Y	3
Zhang et al., 2021 [ZZL ⁺ 21]	132,277	R, EV	ML	US report	Cytology	Y	2
Jia et al., 2022 [JMK ⁺ 22]	3,002	R, EV	NN	US image	Surgical pathology	Y	2
Jin et al., 2022 [JPO ⁺ 22]	2,848	R, EV	ML	Clinical variables, demographics, US image	Cytology, surgical pathology	Y	2
Keutgen et al., 2022 [KLM ⁺ 22]	282	R, EV	NN	Cytology report, US image	Surgical pathology	Y	2
Liu et al., 2022 [LDY ⁺ 22]	3,176*	R, EV	NN	US image	Cytology	N	2
Randolph et al., 2022 [RSH ⁺ 22]	483	R, EV	ML	Genetic data	Surgical pathology	Y	2
Li et al., 2019 [LZZ ⁺ 19]	45,644	R, EV, PC	NN	US image	Cytology, surgical pathology	Y	2
Park et al., 2019 [PHS ⁺ 19]	5,205	R, EV, PC	NN	US image	Cytology, surgical pathology	Y	2
Song et al., 2019 [SCM ⁺ 19]	1,458	R, EV, PC	NN	US image	Cytology, surgical pathology	N	2
Koh et al., 2020 [KLH ⁺ 20]	16,009*	R, EV, PC	NN	US image	Cytology, surgical pathology	Y	2
Wei et al., 2020 [WGY ⁺ 20]	26,541	R, EV, PC	NN	US image	Surgical pathology	Y	2
Zhang et al., 2020 [ZDJ ⁺ 20]	6,185	R, EV, PC	NN	US image	Surgical pathology	Y	2
Zhu et al., 2021 [ZAJ ⁺ 21]	821	R, EV, PC	NN	US image	Cytology, surgical pathology	Y	2
Zhu et al., 2021 [ZJB ⁺ 21]	779	R, EV, PC	NN	US image	Cytology	Y	2
Zhu et al., 2021 [ZDJ ⁺ 22]	674	R, EV, PC	NN	US image	Cytology, surgical pathology	Y	2
Han et al., 2022 [HCS ⁺ 22]	4,792	R, EV, PC	NN	US image	Cytology, surgical pathology	Y	2
Yang et al., 2022 [YPW ⁺ 23]	139	R, EV, PC	NN	US image	Cytology, surgical pathology	Y	2

Continued on next page

Table 3.3 – continued from previous page

Source by outcome	Cohort size	Study design	AI method	Inputs used	Outcome ground truth	Stat. valid.	Oxford LoE
Barczyński et al., 2020 [BSBW ⁺ 20]	50	P, EV, PC	NN	US image	Surgical pathology	Y	2
Liang et al., 2020 [LYLC20]	124	P, EV, PC	NN	US image	Cytology, surgical pathology	Y	2
Zhang et al., 2020 [ZWCW20]	303	P, EV, PC	NN	US image	Surgical pathology	Y	2
Peng et al., 2021 [PLL ⁺ 21]	11,114	P, EV, PC	NN	US image	Cytology, surgical pathology	Y	2
Zhu et al., 2021 [ZZY ⁺ 21]	6,687	P, EV, PC	NN	US image	Surgical pathology	Y	2
Kim et al., 2022 [KCH ⁺ 22]	7,577	P, EV, PC	NN	US image	Cytology	Y	2
Wang et al., 2022 [WWL ⁺ 22]	607	P, EV, PC	NN	US image	Surgical pathology	Y	2
Wang et al., 2022 [WXL ⁺ 22]	342	P, EV, PC	NN	US image	Surgical pathology	Y	2
Xu et al., 2022 [XWW ⁺ 22]	607	P, EV, PC	NN	US image	Surgical pathology	Y	2
<i>Long-term prognosis</i>							
Abbasian et al., 2022 [AAMMAA ⁺ 23]	373	R, EV	NN	US image	Metastases	N	2
Yu et al., 2022 [YWL ⁺ 22]	199	R, EV	ML	CT/MRI	Metastases	Y	2
Zou et al., 2022 [ZSS ⁺ 22]	2,097	R, EV	ML	US image, clinical variables	Metastases	Y	2
<i>Multi-outcome</i>							
Bhalla et al., 2020 [BKK ⁺ 20]	578	R, EV	ML	Genetic data	Metastases, surgical pathology	Y	3
Yang et al., 2020 [YYY ⁺ 20]	1,130	R, EV	NN	US image	Nodule contour, surgical pathology	Y	2
Chen et al., 2021 [CYZM21]	653	R, EV	ML	Genetic data	Surgical pathology, survival/recurrence	Y	3
Dolezal et al., 2021 [DTL ⁺ 21]	115*	R, EV	NN	Surgical pathology slides	Molecular test, surgical pathology	Y	2
Swan et al., 2022 [STN ⁺ 22]	329	R, EV	NN	US image	Surgical pathology, US risk score	N	2

Continued on next page

Table 3.3 – *continued from previous page*

Source by outcome	Cohort size	Study design	AI method	Inputs used	Outcome ground truth	Stat. valid.	Oxford LoE
Wu et al., 2022 [WYJ ⁺ 22]	412	R, EV	ML	CT/MRI	Metastases, surgical pathology	Y	2
Kim et al., 2019 [KHH19]	106	R, EV, PC	ML	US image	Cytology, nodule presence, surgical pathology, US risk score	Y	2
Han et al., 2021 [HHP21]	372	R, EV, PC	Both	US image	Cytology, surgical pathology, US risk score	Y	2
Liang et al., 2021 [LHC ⁺ 21]	124	R, EV, PC	NN	US image	Cytology, nodule contour, surgical pathology, US risk score	N	2
Stenman et al., 2022 [SLL ⁺ 22a]	190	R, EV, PC	NN	Cytology slide	Surgical pathology, survival/recurrence	Y	3
Choi et al., 2017 [CBP ⁺ 17]	89	P, EV, PC	NN	US image	Cytology, nodule contour, surgical pathology, US risk score	Y	2
Yoo et al., 2018 [YHC ⁺ 18]	50	P, EV, PC	NN	US image	Cytology, US risk score	Y	2
Jeong et al., 2019 [JKH ⁺ 19]	76	P, EV, PC	ML	US image	Cytology, surgical pathology, US risk score	Y	2
Xia et al., 2019 [XYZ ⁺ 19]	171	P, EV, PC	Both	US image	Cytology, surgical pathology, US risk score	Y	2
Wei et al., 2020 [WZW ⁺ 20]	199	P, EV, PC	Both	US image	Cytology, surgical pathology, US risk score	Y	2
Cui et al., 2022 [CFS ⁺ 23]	2,374	P, EV, PC	ML	US image	Surgical pathology, US risk score	Y	2
Huang et al., 2022 [HZL ⁺ 22]	100	P, EV, PC	Both	US image	Cytology, surgical pathology, US risk score	Y	2

CHAPTER 4

A Multitask Approach for Automated Detection and Segmentation of Thyroid Nodules in Ultrasound Images

Associated Publications: **Radhachandran, A.**, Kinzel, A., Chen, J., et al. (2024). A multitask approach for automated detection and segmentation of thyroid nodules in ultrasound images. *Computers in Biology and Medicine*, 170, 107974.

An increase in the incidence and diagnosis of thyroid nodules and thyroid cancer underscores the need for a better approach to nodule detection and risk stratification in ultrasound (US) images that can reduce healthcare costs, patient discomfort, and unnecessary invasive procedures. However, variability in ultrasound technique and interpretation makes the diagnostic process partially subjective. Therefore, an automated approach that detects and segments nodules could improve performance on downstream tasks, such as risk stratification. This chapter presents a multitask approach centered on an anomaly detection (AD) module that can be integrated with any UNet architecture variant to improve image-level nodule detection, and evaluates this approach on a UCLA Health dataset of 9888 images from 280 patients as well as two external validation datasets.

4.1 Introduction

There has been a recent rise in thyroid cancer incidence, making it the fifth most common malignancy amongst women in the United States between 2015–2019.[noaa] By age 50, 50–

60% of the population have one or more thyroid nodules incidentally discovered on imaging performed for an unrelated indication.[TG97] Although $\sim 90\%$ of these detected nodules are actually clinically insignificant and benign, uncertainty in radiologic evaluation can lead to unnecessary biopsies.[DCL⁺15] Ultrasound (US) imaging is the primary imaging modality used to assess the thyroid along with the morphology and internal architecture of any present nodules.[CKK⁺09] When evaluating nodules, the first step involves detecting if and where a nodule is present on a series of US images. However, the presence of various technical limitations and artifacts in US images such as vague nodule boundaries, poor contrast, variable probe positioning, and non-standardized spatial resolution, makes the diagnostic process partially subjective. These limitations lead to high inter-radiologist variability[CKK⁺10] in risk stratification, and an increased need for invasive diagnostic procedures such as fine needle aspiration biopsies (FNABs).[Ros06] And oftentimes, FNABs can be inconclusive, leading to unnecessary healthcare costs and complications for patients.[PYL⁺20]

In recent years, an increase in the development of computer-aided diagnosis (CAD) methods using deep learning has proven useful for image analysis tasks in biomedicine.[RFB15, ÇAL⁺16, RSK⁺22] Deep learning can be especially useful in eliminating variability and enhancing diagnostic reliability in thyroid cancer risk assessment by improving nodule localization techniques.[SBD⁺21] Multiple groups have explored different deep learning architectures to address localization through the segmentation of thyroid nodules.[SLL⁺22b, YGC⁺22, PZL21, DHSN19] While these existing models demonstrate promising results, they are semi-automated because they are evaluated solely on images known to have nodules and do not assess ability to identify suspicious images. Thus, most current work tackles a more narrow task, which despite performing well, does not translate in terms of clinical applicability. For such a segmentation model to be implemented successfully, it would require a radiologist to first identify suspicious images, on which the model would make nodule segmentation predictions. However, in the clinical setting, when a radiologist evaluates a set of US images and identifies an image with a potential nodule, they are simultaneously deciding where the

nodule is located. Without incorporating a method for image-wide detection, a model that purely segments is unlikely to have a tangible clinical impact. Furthermore, such a model is more reliant on initial manual detection of images with nodules, which might not be as accurate if not done by someone without the expertise and training of a radiologist. The inclusion of image-wide detection may also extend the utility of a segmentation model to care providers without the experience of a radiologist. This is particularly relevant when extending this model beyond the United States, where radiologists may not be sole physicians performing imaging and subsequent diagnostics.

As a standalone algorithm, image-wide detection of thyroid nodules on ultrasound has not been extensively studied. Moreover, such an algorithm would require a secondary purpose such as nodule localization to augment the clinical workflow.[SLL⁺19] In order to improve detection of anomalous images and elevate the clinical utility of a segmentation model, this work introduces a multitask approach through an anomaly detection (AD) module to be used in conjunction with any UNet architecture[RFB15] variant. Previously, Jain et al.[JSO⁺21] proposed a two-stage model that performed AD to filter out non-suspicious capsule endoscopy images and then segmented the remaining images. In a different study, Chen et al.[CDL⁺19] integrated a mask classification component into their dermoscopic image segmentation network to determine if a lesion was present. In that work, the mask classification output was only used to improve the training of model weights. These studies have demonstrated the value of integrating the domain knowledge that images can have empty segmentations, which in turn suppresses incidental false positive nodule predictions. Since previous thyroid nodule segmentation studies have not measured image-wide detection, there is no guarantee that those models would effectively predict empty masks and not need a specific detection module. The proposed AD module expands on previous work by not only making a binary classification of nodule presence, but multiplying the AD output with the nodule segmentation to post-process the prediction. Hence the AD module is integrated with a variety of state-of-the-art segmentation architectures to understand its impact on

performance. By concurrently determining which images may have nodules and segmenting suspicious regions, our multitask approach allows for a clinically translatable model that aptly parallels the workflow for thyroid nodule assessment.

In summary, the main contributions of this work are as follows:

1. To more closely parallel the distribution of images seen in medical practice and gather a more clinically relevant metric, segmentation performance is measured on images with and without nodules. In contrast to previous research, negative images (those without nodules) are included during evaluation.
2. The AD module is integrated with the segmentation model architecture to assist in image-wide nodule detection. Thus, the proposed multitask method steps towards an automated segmentation method that does not rely on guidance from radiologists.
3. Various state-of-the-art architectures for nodule segmentation are compared with and without the AD module.
4. The evaluated architectures are trained and tested on a novel thyroid ultrasound dataset. Furthermore, they are externally validated on two publicly available thyroid ultrasound datasets, assessing model stability and robustness to data variability.

4.2 Methods

4.2.1 Dataset Descriptions

UCLA Dataset

UCLA Health has implemented a standardized protocol for the acquisition of thyroid US images. Over the past three years, patients who have had an US examination following this protocol have had their imaging collected and aggregated in the UCLA Thyroid RadPath research dataset. Consent was waived because images were acquired through standard clinical

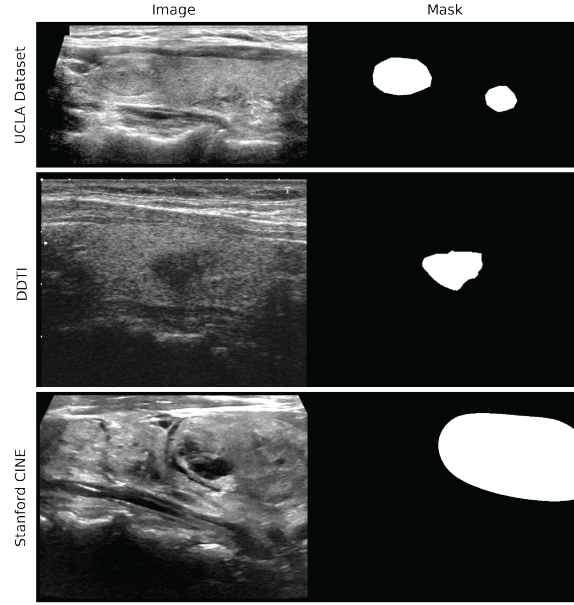


Figure 4.1: Example of an image with corresponding annotation mask from each dataset.

care and analyzed retrospectively without direct patient contact. IRB (IRB#19-001535) of University of California, Los Angeles gave ethical approval for this work on 9/26/2019. This study consists of 280 patients from this dataset, where each patient has a set of 20–40 US images that span a variety of anatomic regions and ultrasound probe orientations. Each US image has been manually segmented by 2 radiologists and validated by a more senior radiologist at UCLA.

DDTI

DDTI is a public, open access dataset from the IDIME Ultrasound Department, one of Colombia’s largest diagnostic imaging centers.[PVN⁺15] This dataset contains 480 images from 290 patients. Each of these images is accompanied by a nodule annotation made by a radiology resident. Preprocessing of the original 480 images resulted in 610 images (images that contained two frames in one were split and considered separately).

Stanford CINE

The Stanford CINE dataset contains 192 cine clips from 167 patients with biopsy-confirmed thyroid nodules.[noab] The cine clips in this dataset included radiologist-annotated segmentations and captured a single nodule. On average, each cine clip was composed of 90 frames, but the total number of frames was downsampled by a factor of 10 due to redundancy in image content. The frames from the cine clips were extracted and treated as independent images.

4.2.2 AD Module

The central innovation of this approach is the integration of the AD module with any variant of the U-Net architecture, such as the multi-scale UNet (MSUNet),[SZLC21] to enable a multitask model for thyroid nodule segmentation on ultrasound images (Figure 4.2). The contracting path of a UNet based architecture is engineered to extract a high-level representation of an input image. The AD module flattens this feature map and passes it through a four-layer fully connected network (FCN) that performs a binary classification of nodule presence. The FCN includes Dropout layers, to prevent overfitting and LeakyReLU activation, to maintain a non-zero gradient during training. Then, the FCN output is rescaled between 0 and 1 using a sigmoid function. The AD module leverages the feature rich image encoding, without requiring unnecessary feature engineering or processing steps, to make a streamlined classification of nodule presence with minimal overhead. The expansive path of the U-Net architecture also takes the image encoding from the contracting path, but reconstructs a predicted segmentation mask. Thus, for every input image, MSUNet with the AD module (MSUNet-AD) will predict a nodule mask and AD output. During evaluation, an empirically determined threshold (details under Sensitivity Analysis in Results) is used to binarize the AD output, which corresponds to MSUNet-AD’s prediction of whether or not the input image contains a nodule. The AD output is then multiplied with the binary mask to either maintain the segmentation prediction or convert it into an empty mask.

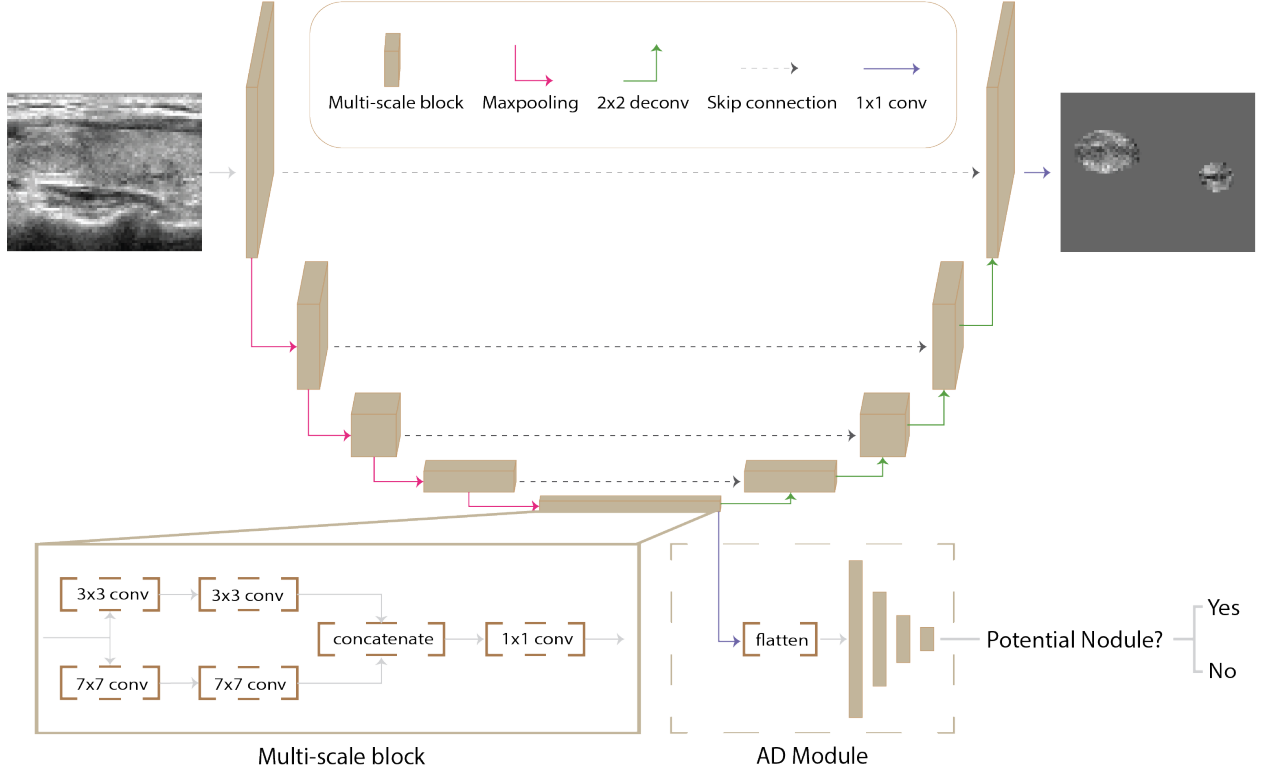


Figure 4.2: One example of AD being integrated with a segmentation architecture. This is an overview of the model architecture for MSUNet-AD, with an in-depth look at the operations in each multi-scale block (bottom left) and in the AD module (bottom right).

4.2.3 Baseline Segmentation Architectures

The impact of AD is evaluated by comparing architectures with the module against their baseline counterparts that do not integrate the module for image-wide detection. These baselines include various architectures that are commonly considered in segmentation papers, such as the UNet,[RFB15] UNet++ (Nested UNet),[ZSTL18] UNet with pretrained ResNet50 encoder (SResUNet) and Attention UNet (AttUNet).[OSF⁺18] Other baseline architectures tailored towards medical image or thyroid nodule segmentation are also considered, with their specific design being explained in the subsequent sections.

Multi-scale UNet (MSUNet)

The UNet has been a standard architecture used for medical image segmentation. The encoding pathway consists of multiple layers, each of which contain a sequence (block) of 3×3

convolutions that extract relevant image features. However, using the same 3×3 kernel size for encoding pathway convolutions will lead to a fixed receptive field size, which can cause shortcomings in feature extraction during the encoding pathway and feature recovery during the decoding pathway. Small receptive fields typically excel at segmenting small objects, but struggle at localization, whereas large receptive fields are more conducive to segmenting and localizing larger objects.[Bao19] In response to this tradeoff, Su et al. introduced MSUNet[SZLC21] and the idea of the multi-scale block, which incorporates convolution sequences with different kernel sizes, allowing for the extraction of more diverse features. The multi-scale block consists of parallel paths each applying a double convolution, one with kernel size 3×3 and one with kernel size 7×7 . The outputs of these layers are concatenated, a 1×1 convolution is applied and the resulting matrix is fed into the next layer.

Semantic Guided UNet (SGUNet)

SGUNet[PZL21] is a UNet inspired architecture designed specifically for thyroid nodule segmentation in ultrasound. Their primary innovation is the SGM module, which reduces noise interference inherent to ultrasound imaging that may be propagated in the encoding layer convolutions as well as in the skip connections to the decoder layers. A SGM module is included after each decoding layer to convert the upsampled decoding into a one channel feature map. This one channel map is then concatenated with the previous decoding layer’s SGM module output. A 1×1 convolution is applied to the resulting two channel map to result in that decoding layer’s one channel semantic map. SGUNet was re-implemented in PyTorch, trained on the UCLA dataset and used as a baseline to compare against the proposed architecture.

MICCAI 2020 TN-SCUI Challenge Winner (Wang et al.)

The Wang et al. model[Wan23] is a two stage cascaded framework that was developed for single-target segmentation of thyroid nodules. The stage 1 network is trained to predict a

nodule given an ultrasound image, while the stage 2 network is trained on an image crop containing the region of interest derived from the ground truth annotation. Thus, during evaluation the two stages can be used sequentially, where stage 1 makes an initial prediction to localize the nodule, and stage 2 performs a finer segmentation of the identified nodule. Pretrained weights for stage 1 and stage 2 from training on the TN-SCUI challenge dataset were used to compare against Wang et al. when trained and evaluated on the UCLA dataset. Lastly, the proposed AD module could not be integrated with the Wang et al. model. Instead, a similar AD module was trained in parallel and during evaluation, its binary output was used to post-process the predicted mask. The AD output is multiplied with the mask to either maintain the prediction or convert it to an empty mask.

4.3 Experiments

4.3.1 Model Training

An 80:20 train-test split was applied at the patient level to the UCLA dataset. 10% of the training set patients made up the validation set. Images with certain artifacts, such as transducer information, vascular flow assessments (Doppler images) or caliper markings, were removed. Each of the remaining images went through removal of protected health information, contrast enhancement and whitening. Image counts after applying this preprocessing pipeline are presented in Table 4.1.

Image inputs for model training, validation, and testing were resized to 256×256 . Image augmentation was performed through random flipping, rotation, and addition of Gaussian noise. The Adam optimizer was used with an initial learning rate of 0.0005 and batch size of 64. The learning rate was decreased by a factor of 10 if the validation loss did not improve for 5 epochs. All models were developed using PyTorch 1.11.0 and trained on a NVIDIA DGX-1.

Table 4.1: Patient and image counts for train, validation and test splits. Images with at least one ground truth nodule are referred to as positive images. The images without nodules are negative images. The percentage of positive images are included in the last column.

Split	Patients	Images	Positive Images
train	202	7097	2892 (40.7%)
validation	22	793	332 (41.9%)
test	56	1998	766 (38.3%)

4.3.2 Loss Function

The segmentation models without AD were trained using DiceBCE loss. A value of 0.5 is empirically assigned to w so that both the Dice loss and pixel-level BCE loss (P-BCE) have equal weighting. Upon integration of the AD module with the segmentation model, an additional BCE loss is used for the AD classification task (AD-BCE). Equation 4.1 shows how these losses are used in concordance to train the multitask model.

$$L_{\text{total}} = w \cdot (L_{\text{P-BCE}} + L_{\text{Dice}}) + (1 - w) \cdot L_{\text{AD-BCE}} \quad (4.1)$$

A sensitivity analysis was conducted to determine the optimal value for w in the loss function used to train each multitask model. Each of the evaluated segmentation architectures with AD were separately trained using values of 0.1, 0.25, 0.5, 0.75 and 0.9 for w . The optimal value of w was selected by picking the permutation that had the highest DSC_{all} on the validation set.

4.3.3 Evaluation Metrics

Various evaluation metrics were used to quantify segmentation performance. The image-wide metrics are Dice similarity coefficient (DSC_{all}), precision, recall and F1 score. Precision, recall, and F1 score are calculated using a slightly different definition for true positives (non-empty prediction that overlaps with ground truth positive), true negatives (empty prediction for a ground truth negative), false positives (non-empty prediction for a ground truth negative) and

false negatives (empty prediction for a ground truth positive or non-empty prediction that does not overlap with ground truth positive). DSC_{all} is the average Dice similarity coefficient (DSC) across all evaluated images. Because the Dice coefficient is undefined in cases where no nodule is present, Laplace smoothing ($\lambda=1e-6$) is applied to allow for calculation across all images. Thus, when DSC is calculated for a negative image, an empty predicted mask will lead to a DSC of 1, indicating a true negative. The positive image-level metrics are DSC_+ and Intersection over Union (IoU). These metrics are averaged across only the positive images in the test set. When calculating DSC, as appears in Equation 4.2, true positives (TP), false positives (FP) and false negatives (FN) are determined using the traditional pixel-level definition.

$$DSC = \frac{2 \cdot TP + \lambda}{2 \cdot TP + FP + FN + \lambda} \quad (4.2)$$

Experiments were also conducted to understand trends in DSC_+ based on nodule size. Nodule size was determined by calculating the proportion of pixels a nodule(s) occupied on the image. The nodule sizes for all the positive images in the test set were aggregated and split into quartiles, where Q1 (smallest quartile) represents the smaller nodules and Q4 (largest quartile) represents the larger nodules. These quartiles are used to stratify the positive images in the test set and perform a more robust analysis.

The Wilcoxon rank-sum test and Mann-Whitney U test was used to compare model performance based on DSC. Since the test set was determined based on patient-level split, there could be overlap in anatomical regions captured by a patient’s US images, which could unintentionally inflate the statistical power of the results. To address this shortcoming, the mean DSC for each patient’s positive images were calculated and the test statistic was calculated using these patient-level DSC averages instead of DSC_{all} or DSC_+ . For the Stanford CINE dataset, since adjacent frames tended to capture related anatomical views, DSC was averaged across the frames for each cine clip. For DDTI, although some patients had multiple images, they typically were of varying anatomical views so result aggregation

was unnecessary.

4.4 Results

4.4.1 Segmentation Performance

The baseline architectures were evaluated on the holdout test set with and without the AD module. Of the models with AD, SResUNet-AD had the highest DSC_{all} of 0.808 and F1 of 0.839 when compared against the other models with AD. MSUNet-AD had the highest DSC_+ of 0.655, while the other models with AD ranged between 0.502 and 0.631. To further assess the effect of the AD module, all the baseline architectures were trained and evaluated without AD (Table 4.2). A central pattern was that when AD was integrated into a model architecture, there were increases in DSC_{all} and F1. For example, MSUNet saw an increase in F1 from 0.683 to 0.837 and in DSC_{all} from 0.581 to 0.797. Another observed trend with the addition of AD was a drop in DSC_+ . For example, the Wang et al. model had the highest DSC_+ of 0.757, but dropped to 0.609 with the addition of AD. All models exhibited a similar drop in DSC_+ , with only SResUNet-AD and SGUNet-AD not showing significant decreases ($P > 0.16$). The remaining models with AD demonstrated significant decreases ($P < 0.017$).

To investigate whether nodule size played a role in this pattern of decreasing DSC_+ , further analyses were conducted. A closer look at performance across the nodule size quartiles revealed a larger drop in DSC_+ on the smallest quartile (Q1) than on the largest quartile (Q4), after addition of AD. Across all architectures, mean DSC_+ on Q1 dropped from values between 0.347 and 0.648 to values between 0.215 and 0.404. However, DSC_+ dropped on Q4 from values between 0.808 and 0.851 to values between 0.68 and 0.829. Only UNet, AttUNet and Wang et al. demonstrated significant DSC_+ differences in Q1 and Q2 when AD was added ($P < .001$). On the other hand, MSUNet-AD and SResUNet-AD had statistically similar DSC_+ distributions to their non-AD baselines in Q1 ($P = .006$, $P = 0.34$) and Q2

Table 4.2: Performance of models with and without AD on hold-out test set. Bolded values highlight the model with the best performance on a certain metric.

Model Name	Image-Wide				Average Across Positive Images	
	DSC_{all}	Precision	Recall	F1	DSC_+	IoU
UNet-AD	0.724±0.384	0.775	0.764	0.769	0.502±0.337	0.401±0.295
NestedUNet-AD	0.782±0.354	0.81	0.841	0.825	0.629±0.34	0.537±0.319
SResUNet-AD	0.808±0.341	0.861	0.817	0.839	0.631±0.363	0.55±0.342
Wang et al.-AD	0.755±0.385	0.755	0.768	0.761	0.609±0.379	0.533±0.355
SGUNet-AD	0.773±0.36	0.812	0.822	0.817	0.598±0.349	0.506±0.324
AttUNet-AD	0.719±0.387	0.75	0.791	0.77	0.53±0.335	0.428±0.297
MSUNet-AD	0.797±0.344	0.821	0.852	0.837	0.655±0.335	0.565±0.319
UNet	0.544±0.445	0.512	0.946	0.665	0.711±0.276	0.610±0.277
NestedUNet	0.539±0.444	0.508	0.944	0.661	0.710±0.271	0.606±0.271
SResUNet	0.259±0.369	0.369	0.941	0.53	0.673±0.277	0.563±0.271
Wang et al.	0.299±0.402	0.377	0.961	0.541	0.757±0.248	0.660±0.258
SGUNet	0.442±0.435	0.458	0.920	0.611	0.621±0.289	0.506±0.273
AttUNet	0.547±0.444	0.515	0.945	0.667	0.707±0.277	0.606±0.279
MSUNet	0.581±0.442	0.535	0.944	0.683	0.726±0.271	0.628±0.273

($P = .009$, $P = 0.34$). Furthermore, when AD was added, decreases in DSC_+ on larger quartiles were not significant across most models. An example of this can be seen with MSUNet; the addition of AD dropped DSC_+ from 0.779 to 0.742 on Q3 ($P = 0.18$) and from 0.851 to 0.829 on Q4 ($P = 0.31$). Lastly, there was less variability in DSC_+ on the largest quartile compared to the smallest quartile. Specifically for the models with AD, this is observed with IQR ranging between from 0.141 to 0.254 on Q4 and 0.444 and 0.83 for Q1. Specific segmentation examples that reflect the aforementioned patterns across nodule sizes can be found in Figure 4.5.

Ultimately, SResUNet-AD and MSUNet-AD performed best across all baselines with AD in terms of the three key evaluation criteria of DSC_+ , DSC_{all} and F1 score. The addition of AD, despite leading to a drop in DSC_+ , did not lead to statistical differences in DSC_+ performance when stratified by nodule size for these two best performing models. More importantly, AD contributed to an increase in DSC_{all} and F1, improving capability of concurrent image-wide nodule detection and segmentation.

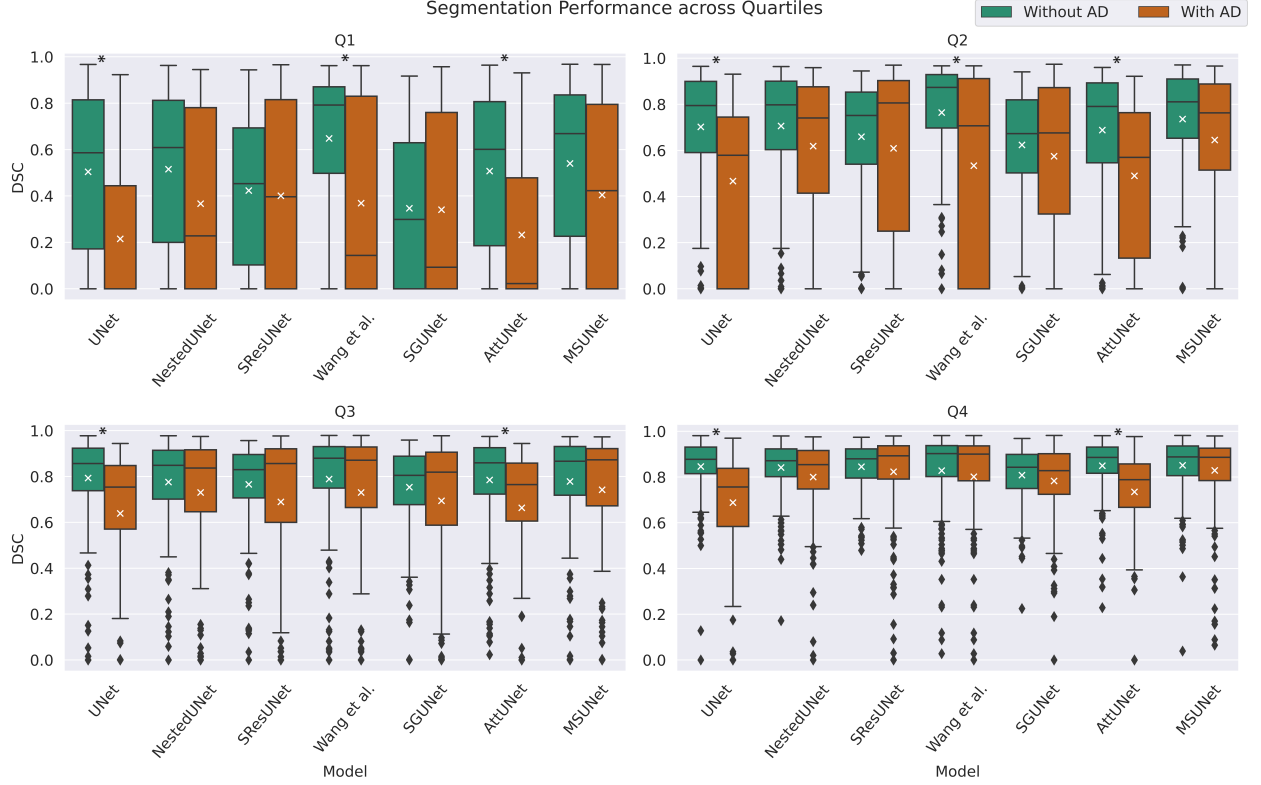


Figure 4.3: Box plot comparing segmentation performance by nodule size quartile between models with and without AD module. The plot represents the distribution of the DSC values (used to calculate DSC_+) for each quartile. A black asterisk above a comparison represents a statistically significant difference. The white crosshair marks the mean of the distribution.

Wang et al. Evaluation Results

The Wang et al. architecture was trained on the TN-SCUI dataset and evaluation on the TN-SCUI test set showed an increase in DSC when the stage 2 network was used with stage 1. However, when trained and tested on the UCLA dataset, there was a decrease in DSC_+ from 0.757 to 0.635 when stage 1 and stage 2 were used in conjunction as opposed to just stage 1. Thus, the results for the Wang et al. model (Table 4.2) are from the best performing permutation of the model architecture, which is when only stage 1 was used. Of the models without AD, the Wang et al. model had one of the lowest F1 scores of 0.541 and DSC_{all} of

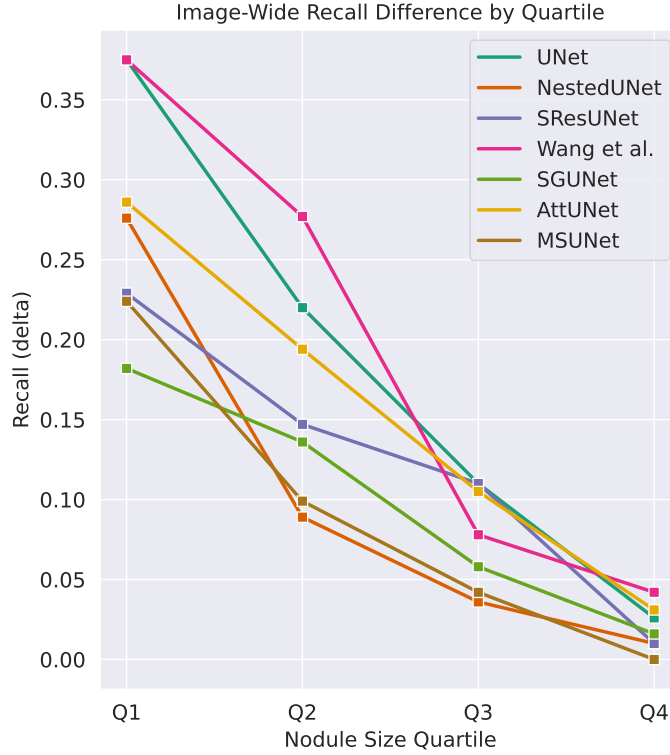


Figure 4.4: Image-wide recall was calculated for each model for each of the nodule size quartiles. The difference in recall between a model without and with AD is plotted for each quartile.

0.299, but the highest DSC_+ of 0.757. When AD was incorporated, the Wang et al. model had a DSC_{all} of 0.755, but the lowest F1 score of 0.761 and a drop in DSC_+ to 0.609.

4.4.2 Sensitivity Analysis

A sensitivity analysis was conducted to determine the best threshold for binarizing AD output. Various thresholds between 0 and 0.75 in increments of 0.05 were tested. Each network with AD was evaluated on the validation set and image-wide F1 scores were recorded using each of these different thresholds (Figure 4.6). All the models with the AD module had an optimal AD threshold between 0.1 and 0.5, where they demonstrated best F1 performance on the validation set. There is a consistent plateau in F1 scores between 0.1 and 0.5 for all models,

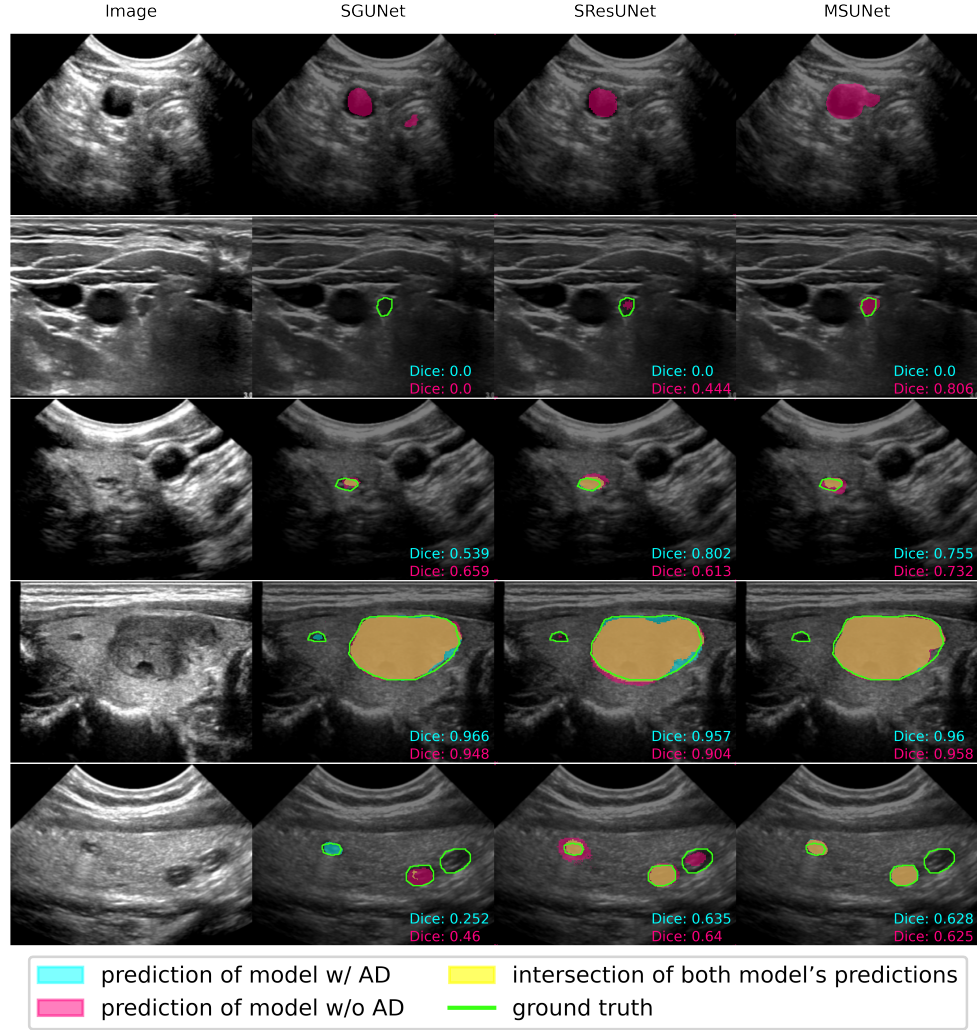


Figure 4.5: Examples of the difference in segmentation performance between three of the models evaluated. First row: All models without the AD module conflate a carotid artery as a nodule leading to a false positive. When the AD module is added, none of the models contain a prediction since the AD output correctly eliminates any predicted segmentations. Second row: An example of all models except MSUNet struggling to segment a small nodule. Third row: An example of a smaller nodule where all three models and their counterparts with AD exhibit agreement and successfully segment the target. Fourth row: An example of all six models successfully segmenting the larger of two nodules. Fifth row: A relatively difficult example with three different small nodules. The majority of models consistently miss the far right nodule. Here, MSUNet-AD and SResUNet-AD perform at par with their non-AD baselines.

indicating that performance would not differ too drastically for most AD thresholds in that range.

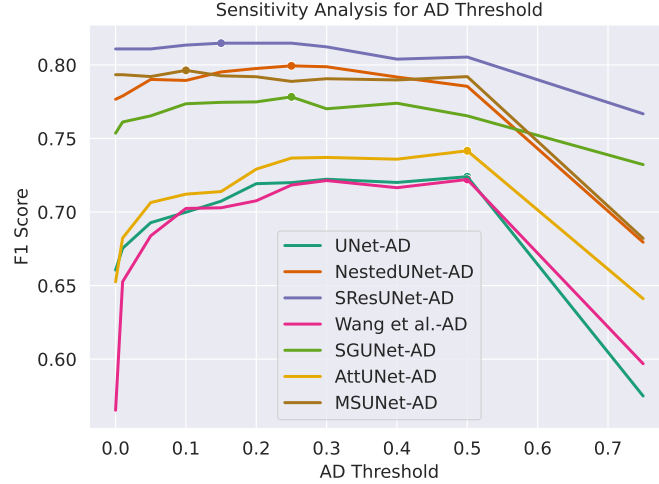


Figure 4.6: F1 performance across AD thresholds for the six models with AD. Dot marker represents the optimal threshold, where the best validation set F1 score is achieved for a given model.

4.4.3 External Validation

The models with the AD module were evaluated on two external validation (EV) datasets, DDTI and Stanford CINE dataset. One difference between the UCLA dataset and these two datasets is that the latter do not have any negative images. Despite the EV datasets only containing positive images, the capability of the AD module in identifying these images could be further evaluated. The lack of public EV datasets with both positive and negative images required the use of the available datasets, which strictly had positive images. Despite this challenge, the models with AD could still be evaluated on their ability to identify and segment images with nodules, albeit on a biased dataset.

SResUNet-AD had the highest DSC_+ of 0.402 and 0.592 on DDTI and Stanford CINE. MSUNet-AD had a comparable DSC_+ of 0.391 on DDTI ($P = 0.71$) and of 0.581 on Stanford CINE ($P = 0.73$). DSC_+ for MSUNet-AD on the UCLA dataset was 0.655, which was higher than its DSC_+ on DDTI ($P < .001$) and on Stanford CINE ($P < .001$). For SResUNet-AD,

Table 4.3: Performance of models on DDTI and Stanford CINE datasets. Bolded values highlight the model with the best performance on a certain metric.

Model Name	DDTI		Stanford CINE	
	Avg. Across Positive Images		Avg. Across Positive Images	
	DSC_+	IoU	DSC_+	IoU
UNet-AD	0.211±0.314	0.162±0.255	0.411±0.354	0.325±0.302
NestedUNet-AD	0.358±0.378	0.293±0.327	0.559±0.356	0.468±0.329
SResUNet-AD	0.402±0.384	0.333±0.341	0.592±0.369	0.51±0.345
Wang et al.-AD	0.301±0.39	0.255±0.34	0.466±0.413	0.404±0.377
SGUNet-AD	0.282±0.348	0.223±0.293	0.527±0.361	0.438±0.328
AttUNet-AD	0.289±0.345	0.225±0.284	0.456±0.354	0.365±0.307
MSU-Net-AD	0.391±0.384	0.324±0.341	0.581±0.363	0.495±0.338

DSC_+ on the UCLA dataset was 0.631, which was higher than its DSC_+ on DDTI ($P < .001$). However, SResUNet-AD exhibited a DSC_+ on UCLA which was comparable to its DSC_+ on Stanford CINE ($P = 0.08$).

4.5 Discussion

The AD module improved F1 score and DSC_{all} across the board when integrated with the baseline architectures, with SResUNet-AD performing best in these two metrics. This increase can be attributed to the decrease in false positives at the image-wide level. Prior to the addition of AD, the Wang et al. model performed the best in terms of DSC_+ ; however, the low DSC_{all} and precision, observed across all architectures without AD, demonstrated that these baselines were prone to false positive predictions. Although AD boosted precision, it also caused an increase in false negatives, and hence a decrease in recall and DSC_+ . Further investigation into this pattern of decreasing DSC_+ revealed that this drop was mainly due to an inability to segment smaller nodules (Figure 4.3). This was seen across most models, as exemplified in the quantitative disparities between Q1 and the other quartiles representing larger nodules. This drop in segmentation performance can be specifically explained by the greater frequency of false negatives in the smaller quartiles, as evidenced by the greater

delta in image-wide recall between a model without and with AD in Q1 compared to in Q4 (Figure 4.4). However, smaller nodules are less likely to be clinically significant,[HAB⁺16] thus reducing the clinical consequences of the increase in false negatives due to AD.

In regards to the EV experiments, there were notable advantages and disadvantages to the datasets used. DDTI was a favorable dataset because it contained multi-nodule images, which is comparable to real-world data and it followed an image collection protocol different from UCLA’s protocol, demonstrating the model’s independence from a specific image collection method. The main drawback of the Stanford CINE dataset was that it only contained clips of single nodules, which does not reflect real-world frequency. The lack of public EV datasets with both positive and negative images required the use of the available datasets, which strictly had positive images. Despite this challenge, the models with AD could still be evaluated on their ability to identify and segment images with nodules, albeit on a biased dataset. As detailed in the results, SResUNet-AD and MSUNet-AD also outperformed the majority of other architectures with AD on the EV datasets. More importantly, statistical analysis revealed that SResUNet-AD displayed similar DSC_+ performance between the UCLA and Stanford CINE datasets, supporting the model’s generalizability. Despite both models falling short on the DDTI dataset, the results were promising in that there was no severe overfitting to patterns unique to ultrasounds collected using UCLA’s standard protocol. Although the clinical implications of models with AD need to be evaluated to understand how they would augment practice, these results indicate that such a multitask method is versatile and could have clinical applicability. Other groups have explored 3D volume ultrasound techniques (PIUR imaging, IMFUSION, etc.) or developed CAD-based solutions (KOIOS, S-Detect, etc.) for localization and diagnosis. These methodologies have been well-received, but still lack widespread adoption, permitting continued innovation in this space. Also, 2D ultrasound is still the primary method for nodule detection and diagnosis, and presents advantages in its ease of use and relative affordability.

The proposed multitask approach advances current thyroid nodule segmentation methods

with an automated technique that is designed to parallel clinical practice, and in doing so augment a radiologist’s workflow. The experimental results demonstrate that existing architectures such as SResUNet and MSUNet, upon the addition of AD, excel in terms of image-wide detection and pixel-level nodule segmentation. Furthermore, the architectures with AD were evaluated on two external validation sets, DDTI and Stanford CINE, to assess model stability and robustness to data variability. This technique can be extended to have further clinical application by extracting nodule features from segmented regions to develop a network that can perform an improved TI-RADS evaluation or automated biopsy prediction. Future work will aim to integrate the proposed multitask model into an end-to-end network that is also capable of nodule malignancy classification and further risk stratification. With rigorous evaluation, it would be possible to reduce the subjectivity in radiological evaluation and perceive certain clinical impacts such as an improvement in triaging the need for biopsy.

CHAPTER 5

ThyGraph: A Graph-Based Approach for Thyroid Nodule Diagnosis from Ultrasound Studies

Associated Publications: **Radhachandran, A.**, Vittalam, A., Ivezic, V., et al. (2024). ThyGraph: A Graph-Based Approach for Thyroid Nodule Diagnosis from Ultrasound Studies. Medical Image Computing and Computer Assisted Intervention – MICCAI 2024, 753–763.

Improved thyroid nodule risk stratification from ultrasound (US) can mitigate overdiagnosis and unnecessary biopsies. Previous studies often train deep learning models using manually selected single US frames; these approaches deviate from clinical practice where physicians utilize multiple image views for diagnosis. This chapter introduces ThyGraph, a novel graph-based approach that improves feature aggregation and correlates anatomically proximate images, by leveraging spatial information to model US image studies as patient-level graphs. Graph convolutional networks are trained on image-based and patch-based graphs generated from 505 US image studies to predict nodule malignancy. Self-attention graph pooling is introduced to produce a node-level interpretability metric that is visualized downstream to identify important inputs. The best performing model demonstrated an AUROC of 0.866 ± 0.019 and AUPRC of 0.749 ± 0.043 across five-fold cross validation, significantly outperforming two previously published attention-based feature aggregation networks.

5.1 Introduction

Over the past two decades, thyroid cancer incidence in the United States has nearly tripled [KSS20]. Ultrasound (US) imaging is the primary modality used by radiologists to assess detected nodules, but this process can be subjective and time-consuming. Moreover, over 90% of detected nodules are clinically insignificant or benign, so uncertainty in radiologic evaluation can lead to unnecessary fine-needle aspiration biopsies (FNABs) [DCL⁺15, Ros06]. With the fundamental role of US in thyroid nodule diagnostics and the need for a more deterministic, non-invasive diagnostic alternative to biopsies, research has focused on developing deep learning methods that enhance nodule malignancy prediction from US [TT23]. In recent years, convolutional neural networks (CNNs) have become popular for automating diagnostic tasks from imaging [LCA⁺17, WLC⁺18]. Most current approaches for thyroid nodule diagnosis select an optimal US displaying a nodule and develop a CNN-based classifier with these images [ZCY⁺22, SWZ⁺23]. This approach is limited because it requires a physician to manually select a specific frame from an image study. More importantly, physicians typically use multiple image views of a nodule to make effective diagnoses. In this context, a view is defined as a specific perspective of a nodule/anatomy captured through an ultrasound image. Given the clinical context for this diagnostic task, an improved information aggregation technique across US studies could improve automated malignancy classification.

Methodology that combines information from various imaging perspectives has been explored in tangential domains, such as lung CT, chest radiographs, and breast ultrasound [ZF21, WCC⁺20, KLHZ17]. But to the best of our knowledge, there has been minimal research within the space of thyroid nodule diagnosis in leveraging features across multiple US images. Two prior studies demonstrated how cine clips could be used to improve risk stratification of thyroid nodules [YKA⁺22, WLC⁺22]. However, the cines for these studies have highly correlated frames from a narrow anatomical region and are clipped to contain a consistent view of a nodule. Huang et al. used a transverse and longitudinal image of the thyroid, both containing a view of the nodule, to develop a model [HDJ⁺22]. Despite

demonstrating potential, this approach would require the user to choose two input images and does not leverage the remaining image study to provide additional physiological context. Wang et al. also attempt to leverage whole image studies using an attention-based feature aggregation method [WZZ⁺20]. Most recently, Zhuang et al. proposed an ensembled attention multiple instance learning (AMIL) approach that combines US features at various image scales [ZIF⁺24]. Despite using whole US studies and exhibiting promising results, the approaches proposed by Wang and Zhuang are trained in an AMIL framework, where each US is treated independently, disregarding their inter-image relationships.

Graphs provide an effective alternative for modeling US. By representing image features as nodes and accounting for inter-image relationships as edges, spatial dependencies and contextual information can be learned through the application of graph convolutional networks (GCNs). GCNs layer convolution operations to learn better node representations and incorporate local and global graph structure information [KW17]. Yin et al. proposed an end-to-end framework using a joint CNN-GCN to diagnose kidney US studies [YPL⁺19]. In this case, feature similarity between deep features extracted from US frames was used as a metric to determine connectivity. Yan et al. were the first to use GCNs in the domain of thyroid US. In their study, a dataset-level graph was created to represent every patients' nodules and correlate images with similar features [YYD⁺21]. Despite the effectiveness of these methods, an approach that integrates anatomical context during graph construction and generates unique patient-level graphs could be more effective. To take advantage of untapped opportunities with thyroid US, this chapter presents ThyGraph, a novel graph-based approach to model image studies and improve thyroid nodule diagnosis. The main contributions are as follows:

1. ThyGraph, a patient-level graph construction method that models US studies by correlating anatomically related images to underscore inter-image dependencies.
2. Implementation of multi-scale graphs that use information from both full frames and localized patches.

3. A visualization method that leverages attention-based graph pooling to correlate attention weights with image orientation and anatomical location, and plot the weights for better model explainability.

5.2 Methods

5.2.1 Data Collection and Preprocessing

A retrospective, dataset of 505 US image studies from unique patients was collected from an academic medical center. Institutional Review Board (IRB# 19-001535) of University of California, Los Angeles gave ethical approval for this study. During an US examination, 20-80 images are taken that span a variety of anatomic regions and spatial orientations. Each image study includes a transverse or longitudinal view of the thyroid isthmus, and transverse (inferior, middle, and superior) and longitudinal (lateral, middle, and medial) views of the right and left thyroid lobes.

The main inclusion criterion was a fine-needle aspiration biopsy (FNAB) performed within one year of the US examination. The corresponding cytology report was acquired and following the Bethesda System for Reporting Thyroid Cytopathology, the categorization for risk of malignancy was extracted. Consistent with clinical practice, biopsies with benign cytology were labeled as benign, while those with malignant/suspicious for malignancy cytology were deemed malignant [ABCP⁺23]. In cases of multiple biopsied nodules, the one with the highest Bethesda categorization determined the image study’s label: benign if all nodules were benign and malignant if at least one nodule was malignant. This resulted in 371 benign and 134 malignant image studies. Across these samples, there were 17,626 US images. Ultrasounds were passed through a data preprocessing pipeline, which involved deidentification, negative space and artifact removal, intensity rescaling and image whitening.

5.2.2 State-of-the-Art Baselines

The first baseline, proposed by Wang et al. (Wang model), is an attention-based aggregation network that fuses features across an US study for thyroid nodule diagnosis [WZZ⁺20]. The architecture is similar to a MIL framework, but unique in that each image study is randomly subsampled or oversampled each training iteration to ensure a consistent input size across all studies. The second baseline, proposed by Zhuang et al., is the multi-scale, attention-based multiple-instance learning model (MS-AMIL) ensemble model [ZIF⁺24]. By aggregating features across whole frames and image patches, multi-scale information regarding the nodule was extracted. Their best performing model, an ensemble across features from 128x128 (MS-AMIL-patch128) and 256x256 (MS-AMIL-patch256) frame patches was replicated in our experiments.

5.2.3 ThyGraph

Image-Based Graph Construction

The dataset can be defined as $D: \{(s_n, l_n); n=1, \dots, N\}$ for N image studies. l_n is the label, signifying benign or malignant, for image study s_n . There are M_n ultrasound frames in image study $s_n: \{f_{nm}; m = 1, \dots, M_n\}$. Each frame, f_{nm} has unique dimensions $R^{w \times h \times 1}$, where w and h represent the width and height of the frame.

Every frame within an image study also has embedded text representing that frame’s location and spatial orientation within the thyroid (ex. Left Transverse Superior). An optical character recognition (OCR) algorithm was developed to process each frame and extract this text. Any frames with views of surrounding lymph nodes were not included. Figure 5.1 provides a representation of the different locations and spatial orientations a frame can have. For each frame, f_{nm} , its OCR output is used to determine a 2-D coordinate, (x, y) . Specifically, the x-coordinate determines whether the frame captures the isthmus, or a lateral, middle or medial view of the left or right lobes. The y-coordinate determines whether the

frame is in an inferior, middle or superior position. Thus, each frame’s OCR output, which equates to the frame’s anatomical location, is mapped to a 2-D coordinate that allows us to estimate each frame’s anatomical distance from one another.

Next, for each image study, an undirected graph is constructed in which each node is represented by a frame-level feature vector. This feature vector, $R^{1 \times 6016}$, is a concatenation of the final fully connected layer of a pretrained ResNet101, DenseNet201 and ResNeXt101. Since this ensemble model demonstrated promising results in a previous study for thyroid nodule diagnosis, it was repurposed for this task [PLL⁺21]. Euclidean distance is then calculated between the 2-D coordinates for every pair of frames in the study following

$$d(f, f') = \|f - f'\|. \quad (5.1)$$

The adjacency matrix, A , to represent the graph’s edges is constructed to capture the weights associated with each edge

$$A_{ij} = \max(0, \alpha - d(f_i, f_j)), \quad (5.2)$$

where i and j represent the row and column indices of the adjacency matrix, corresponding to two different graph nodes. This definition permits the graph to assign more weight to edges between nodes whose frames share proximate coordinates, an indicator of anatomical proximity.

Patch-Based Graph Construction

Next, the image-based graph construction method is extended to accommodate image patches. Each frame, f_{nm} , with an OCR-extracted location is tiled into nine patches, represented as p_{nmk} for $k \in [1, 9]$, each with dimension $R^{224 \times 224 \times 1}$. The patch extraction algorithm maximizes frame coverage by generating patches within the frame boundaries, avoiding negative space. Each patch is given an additional z-dimension, extending the coordinate representation of its source frame, resulting in a three-dimensional coordinate, (x, y, z) . The

z-coordinate represents the patch’s position along the posterior-anterior axis. By leveraging the anatomical context provided by OCR-extracted location, each patch is represented within the 3-D coordinate system. Now, for each image study, it is possible to build a patch-based undirected graph, where each node is represented by a patch-level feature vector, $R^{1 \times 2048}$ which is extracted using a ResNet50 model, pretrained on ImageNet. Equation (5.1) is adapted to calculate Euclidean distance between all pairs of patch coordinates. Then, Equation (5.2) determines the graph’s adjacency matrix, with the patch coordinates as inputs and a different empirically determined threshold selected accordingly. Thus, similar to the image-based graphs, each node’s neighborhood inherently estimates that patch’s anatomical proximity to others.

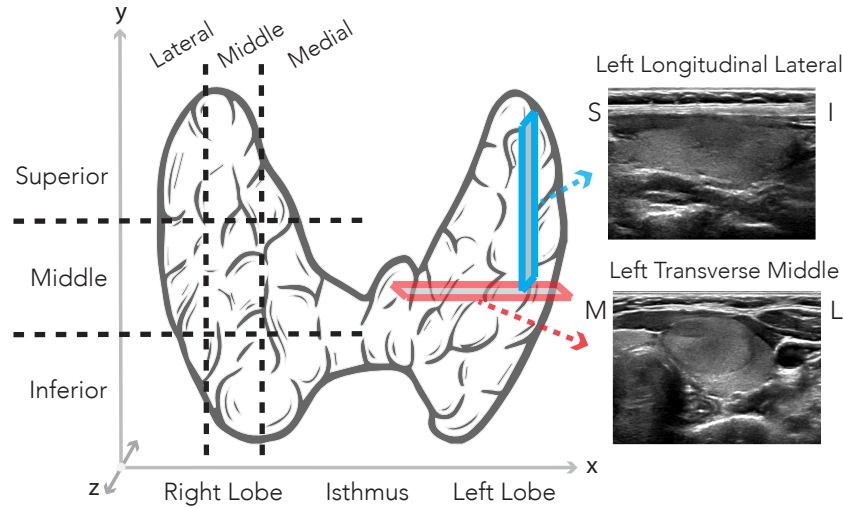


Figure 5.1: Various spatial orientations and locations captured by frames within a thyroid US study. The two example images are labeled with superior (S), inferior (I), medial (M) and lateral (L) to help the viewer orient each image within the context of the 3-D coordinate system.

GCN with Self-Attention Graph Pooling

The non-Euclidean nature of graphs can be leveraged through GCNs to promote learned correlations. A two-layer GCN is trained on the image-based (image-GCN) and patch-based (patch-GCN) graphs to serve as baselines. Furthermore, we explore the effect of introducing

self-attention graph (SAG) pooling [LLK19] to the patch-GCN. The resulting patch-SAG-GCN model includes SAG pooling after each graph convolution. SAG pooling calculates node-level attention scores using an auxiliary graph network and then creates an attention mask to retain a subset of nodes the network deems to be of higher value. This joint feature-based and topology-based pooling method allows the architecture to progressively identify more relevant nodes and capture localized graph information.

During evaluation, for each patient-level graph passed through patch-SAG-GCN, we extract the attention scores generated by the first SAG pooling operation. Each score is associated with a graph node, which via our graph construction method, represents a different image patch. Since each patch has a 3-D coordinate mapping within the system in Figure 5.1, a coordinate-level attention can be assigned. Moreover, in a given graph, certain patches can be mapped back to the same coordinate; in this case, the aggregated attention scores are averaged to create a mean coordinate-level attention.

5.3 Experiments

All architectures, including the Wang model and MS-AMIL baselines, are trained and evaluated using stratified 5-fold cross validation. Evaluation metrics include area under the receiver operating characteristic curve (AUC), area under the precision-recall curve (AUPRC), accuracy, precision and recall. The results are reported in terms of mean and standard deviation across the five cross validation folds. All architectures, except the Wang model, were trained for 200 epochs with early stopping based on validation AUC and a patience value of 25 epochs. The Adam optimizer is used with a learning rate of $1e-4$. Weighted cross-entropy loss is used to train the models. The Wang model was trained using different hyperparameters in line with the original paper. The statistical significance of the proposed models relative to the Wang model and MS-AMIL was compared at significance level 0.05 using DeLong test for AUC and Wilcoxon signed-rank test for accuracy. All experiments are

implemented in PyTorch with a NVIDIA DGX-1.

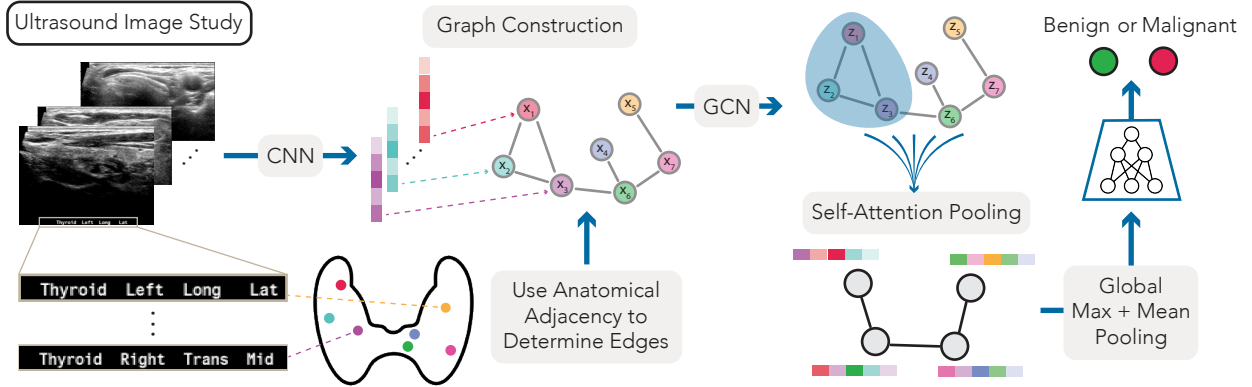


Figure 5.2: Overview of proposed graph construction and model architecture.

5.4 Results

First, the impact of ThyGraph graph construction method is evaluated. The image-GCN and the patch-GCN demonstrated 0.713 and 0.847 AUC, respectively. The patch-SAG-GCN on its own also achieved 0.847 AUC, but when ensembled with the image-GCN and patch-GCN models (GCN-Ensemble), demonstrated 0.853 AUC. The MS-AMIL-patch128 and MS-AMIL-patch256 branches demonstrated a 0.83 and 0.824 AUC, respectively, and when ensembled had a 0.837 AUC. With a larger patient cohort, we were able to demonstrate improved performance for MS-AMIL patch ensemble from the published results of 0.785 ± 0.104 AUC. Despite this improvement, the patch-GCN still outperformed the MS-AMIL-patch128, MS-AMIL-patch256 and the patch ensemble.

The MS-AMIL framework, although designed to independently analyze and aggregate features across images, can still identify important input instances through its attention mechanism. Despite the framework not considering dependencies between inputs, the patterns it learns are still valuable in nodule malignancy prediction. To incorporate the contributions of the MS-AMIL model with ThyGraph, we assessed an ensemble of MS-AMIL-patch128, MS-AMIL-patch256, image-GCN, patch-GCN and patch-SAG-GCN (Full-Ensemble), which

demonstrated an 0.866 AUC, significantly improving upon MS-AMIL ($p=0.03$) and Wang model ($p<0.001$). The Wang model achieved a 0.632 AUC on the in-house dataset, but was significantly outperformed by all proposed models ($p<0.001$). The performance discrepancy with the Wang model can be attributed to the use of a different training dataset, and the published results using smaller and more variably sized images studies.

Table 5.1: Model performance across the five cross validation folds comparing the proposed approach with state-of-the-art baselines (*statistically significant compared to MS-AMIL at significance level 0.05).

Model	AUC	AUPRC	Accuracy	Precision	Recall
Wang model [WZZ ⁺ 20]	0.63 ± 0.04	0.39 ± 0.04	0.63 ± 0.09	0.41 ± 0.06	0.67 ± 0.15
MS-AMIL [ZIF ⁺ 24]	0.84 ± 0.03	0.68 ± 0.06	0.83 ± 0.04	0.68 ± 0.08	0.73 ± 0.07
image-GCN	0.71 ± 0.06	0.52 ± 0.12	0.68 ± 0.07	0.46 ± 0.06	0.73 ± 0.19
patch-GCN	0.85 ± 0.03	0.71 ± 0.09	0.81 ± 0.05	0.63 ± 0.1	0.81 ± 0.06
patch-SAG-GCN	0.85 ± 0.03	0.70 ± 0.07	0.85 ± 0.03	0.71 ± 0.05	0.75 ± 0.06
GCN-Ensemble	0.85 ± 0.03	0.72 ± 0.06	0.85 ± 0.04	0.73 ± 0.08	0.73 ± 0.08
Full-Ensemble	$0.87 \pm 0.02^*$	0.75 ± 0.04	$0.84 \pm 0.04^*$	0.69 ± 0.08	0.78 ± 0.09

5.4.1 Attention-Based Interpretability

The coordinate-level attention scores extracted from patch-SAG-GCN are visualized within a 3-D thyroid mesh [Thi]. Each 3-D coordinate is mapped to a corresponding location within the mesh and the coordinate-level attention scores are used to create a volumetric heat map. Hot spots indicate that patch-SAG-GCN is focusing on imaging features from those thyroid regions when making a certain prediction. Figure 5.3 illustrates the value of our attention-based visualization technique in enhancing model interpretability. In case a), the model places greater weight on features from the right lobe, which is also the location of the biopsied malignant nodule. Similarly, in case b), the model focuses on features from the left lobe, the location of the biopsied benign nodule. However, it also directs attention to the right inferior lobe, which according to the cytology report, is the location of a second, unbiopsied nodule.

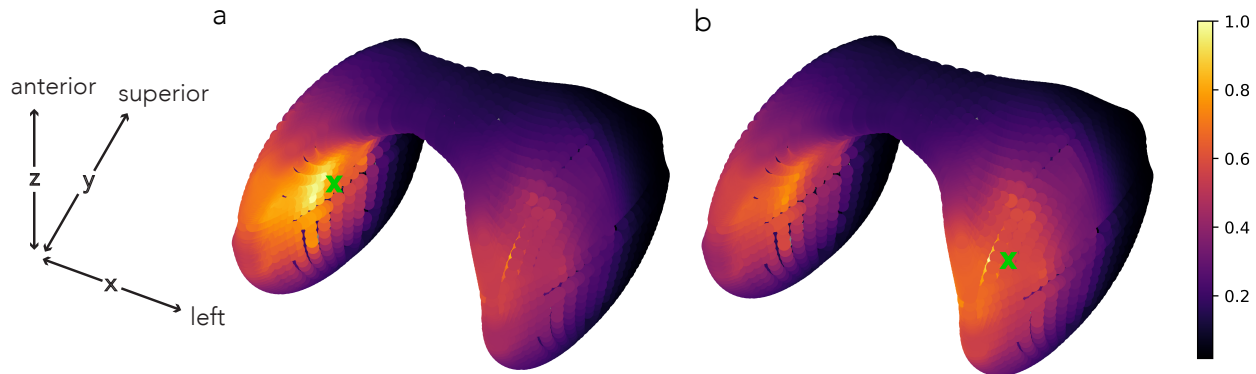


Figure 5.3: Visualization of attention heatmap overlaid on a thyroid mesh model for two patients. The attention weights are scaled between 0 and 1 for each patient. The biopsied nodule location, extracted from corresponding cytology report, is marked with a green X. As indicated, the x-axis stretches from right to left, the y-axis from inferior to superior and the z-axis from posterior to anterior.

5.5 Discussion

ThyGraph presents a novel approach to aggregate information across thyroid US image studies. The majority of past work required manual intervention to select an optimal frame to train, evaluate and prospectively use deep learning-based thyroid nodule diagnostic methods. Previous approaches seeking to fuse multi-view information typically treated each image independently, through MIL-inspired approaches, disregarding inter-image dependencies. ThyGraph not only effectively models an US study despite its non-uniform nature but specifically integrates spatial location information to better represent correlated ultrasounds. Subsequently, the patch-GCN and image-GCN are able to learn multi-scale information across the image study, and the patch-SAG-GCN promotes model explainability, allowing us to localize suspicious regions within the thyroid. Moving forward, different patch- and frame-level feature extractors can be considered, such as vision transformers to improve on the current CNN-based technique. Another limitation of this paper is that further external validation datasets from other academic institutions could help highlight this method’s strengths and weaknesses. Ultimately, the novel graph construction method of ThyGraph and model ensemble of MIL and GCN techniques, presents notable advantages in terms

of automation and interpretability, enhancing our ability to effectively risk stratify thyroid nodules.

CHAPTER 6

US-JEPA: A Joint Embedding Predictive Architecture for Ultrasound

Associated Publications: **Radhachandran, A.**, Ivezić, V., Athreya, S., et al. (2026). US-JEPA: A Joint Embedding Predictive Architecture for Medical Ultrasound. arXiv:2602.19322.

Ultrasound (US) imaging poses unique challenges for representation learning due to its inherently noisy acquisition process. The low signal-to-noise ratio and stochastic speckle patterns hinder standard self-supervised learning methods relying on a pixel-level reconstruction objective. Joint-Embedding Predictive Architectures (JEPAs) address this drawback by predicting masked latent representations rather than raw pixels. However, standard approaches depend on hyperparameter-brittle and computationally expensive online teachers updated via exponential moving average. We propose US-JEPA, a self-supervised framework that adopts the Static-teacher Asymmetric Latent Training (SALT) objective. By using a frozen, domain-specific teacher to provide stable latent targets, US-JEPA decouples student–teacher optimization and pushes the student to expand upon the semantic priors of the teacher. In addition, we provide the first rigorous comparison of all publicly available state-of-the-art ultrasound foundation models on UltraBench, a public dataset benchmark spanning multiple organs and pathological conditions. Under linear probing for diverse classification tasks, US-JEPA achieves performance competitive with or superior to domain-specific and universal vision foundation model baselines. Our results demonstrate that masked latent prediction provides a stable and efficient path toward robust ultrasound representations.

6.1 Background

The task-specific models developed in Chapters 4 and 5 demonstrated that deep learning has the potential to improve thyroid nodule diagnostics in controlled settings, while also exposing two persistent limitations. First, performance gains did not transfer cleanly outside the distribution of training data, a pattern documented in the systematic review of Chapter 3 and confirmed in our own external validation of the multitask detection and segmentation model from Chapter 4. Second, the annotation cost required to build even modestly sized labeled datasets was substantial. Radiologist time, surgical correlation for malignancy ground truth, and institutional collaboration were all necessary, and these resources do not scale to new clinical tasks without repeating the process from scratch.

Both of these limitations reflect the same underlying problem. Task-specific models couple representation learning to labeled data, offering no mechanism for knowledge sharing across tasks, anatomical sites, or institutions. This section examines how the medical imaging research community has responded to this problem, and develops the theoretical rationale for the specific approach pursued in the remainder of this chapter. Rather than re-introducing the SSL frameworks or the foundation model paradigm, both of which are covered in Chapter 2, the goal here is two-fold: 1) to briefly survey the field’s transition to SSL-pretrained foundation models, and 2) to argue that in a noisy, spatially-organized modality like ultrasound, the choice of self-supervised pretext task is not a secondary implementation detail but a fundamental design decision that shapes what the model learns to represent.

6.1.1 How Pretext Tasks Shape What Models Learn

The mechanisms of MAE, joint embedding architectures (JEAs), and joint embedding predictive architectures (JEPAs) are covered in Section 2.4.2, including their respective loss formulations and training objectives. The focus here is not on how they are trained, but on what each one shapes the model to learn, and why those differences matter for a modality

with ultrasound’s specific properties. Figure 6.1 depicts a representative example for each pretext task, using MAE [HCX⁺21] for reconstruction, DINOv3 [SVS⁺25] for JEAs, and I-JEPA [ADM⁺23] for JEPAs.

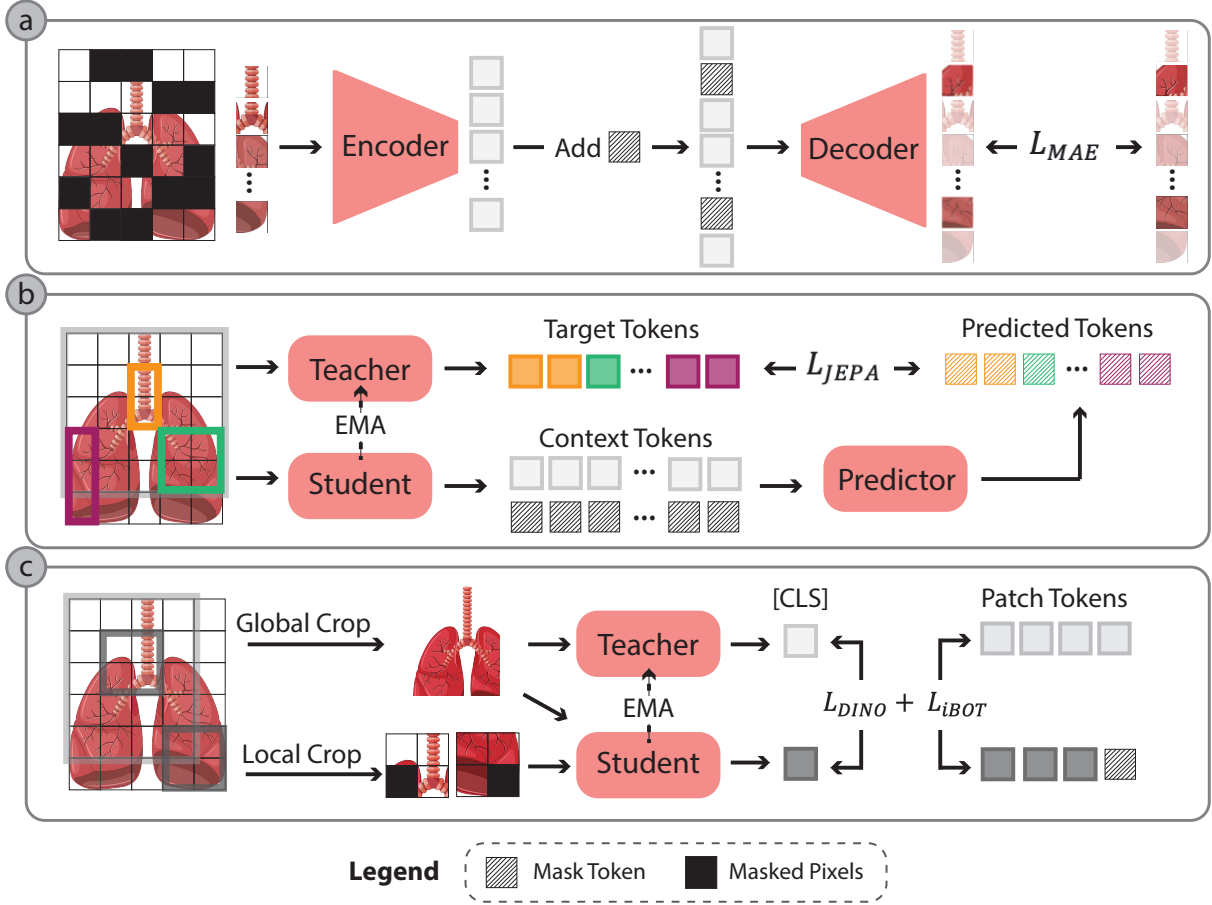


Figure 6.1: SSL pretext task architectures illustrated on a lung image. (a) MAE reconstructs masked patches in pixel space. (b) I-JEPA predicts masked patch representations in latent space from visible context. (c) DINOv3 aligns student and teacher representations across local and global crops.

Reconstruction-based methods optimize a pixel-level objective, meaning the encoder must produce representations sufficient to recover the specific pixel values of masked patches. This is a concrete learning signal, but one with an important bias. Reconstruction objectives allocate model capacity toward features that explain the most observed pixel variance, since minimizing mean squared error over raw pixels essentially asks the model to capture dominant directions of variation in the input. In cases where the dominant variance sources are not

correlated with the most relevant features for a downstream task, the model learns the wrong things. Balestrierio and LeCun demonstrated this formally, demonstrating that the subspace of features explaining the most observed pixel variance is not in general the subspace most useful for downstream recognition tasks [BL24]. In ultrasound, the dominant variance sources are modality-specific, such as speckle noise, acoustic artifacts, and scanner-specific intensity gradients, all of which carry high pixel-level magnitude but limited diagnostic meaning. A encoder pretrained with pixel reconstruction is forced to represent these phenomena well, potentially, at the cost of the anatomical and textural features that matter for clinical tasks.

Joint embedding architectures transition prediction to representation space, decoupling the model from the requirement to account for pixel-level variance. By training the encoder to align representations of two differently augmented views of the same image, JEAs can discard low-level details that vary across augmentations and retain only what is consistent. Theoretical analysis further confirms this advantage. Compared to reconstruction, both JEAs and JEPAs preferentially learn high-influence features, while suppressing noisy features during training [AIB⁺25, LSA⁺24]. However, JEA-based approaches introduce a different assumption that the applied augmentations define what variations are treated as semantically irrelevant. In natural imaging, random cropping, color jitter, and Gaussian blur are reasonable choices. In ultrasound, however, the appropriate augmentation family is far less straightforward. For example, echogenicity, which in thyroid nodule diagnostics is directly used for clinical risk stratification, reflects relative tissue brightness. Thus, intensity-based augmentation could disregard information that carries diagnostic meaning. Spatial cropping may remove the surrounding anatomical context that gives a local finding its clinical significance. Constructing effective augmentation invariances for ultrasound requires domain expertise that is still being systematically characterized.

Joint embedding predictive architectures retain the robustness to noise that comes with latent space prediction, but replace augmentation-defined invariances with a different pretext task entirely. Given the visible context patches of an image, a predictor network

must predict the latent representations of masked target patches at given spatial positions. Rather than learning to be invariant to a set of augmentations, the model learn structural reasoning about what is in a masked region given the visible context. To do this well, the encoder must develop representations that capture spatial composition, structural continuity, and how anatomical regions relate to one another across the image. JEAs do not naturally encourage learning these properties since matching representations from different “views” does not force the model to represent spatial arrangement. JEPAs, by construction, do not require hand-crafted augmentations during pretraining, and therefore impose no priors about which variations are correlated with semantic invariance. In a modality where the appropriate augmentation family is poorly characterized, this is a meaningful practical advantage.

The distinction in what each framework encourages the model to represent has significant consequences for downstream tasks. JEA-based models, optimized for local-to-global alignment through augmentation-matched views, tend to focus on specific salient regions. This focus is especially useful when diagnostic signal is localized - think fine-grained cellular analysis in histopathology. JEPA-based models, optimized for spatial prediction, develop representations that capture structural organization across an image. This focus is more relevant when diagnostic meaning depends on relative arrangement of structures. In ultrasound, where clinically meaningful features include organ boundaries, the spatial relationship between a lesion and surrounding tissue, and whole-organ texture patterns, this capacity for structural representation is directly relevant. This hypothesis is directly explored through an empirical study [IPR⁺26], which is not a core component of this dissertation work, but the findings directly motivate the sections that follow.

Ultimately, the takeaway is that the pretext task is a crucial inductive bias: it determines what the model is incentivized to represent and is strongly tied to the input modality’s noise distribution.

6.1.2 JEPAs as a Way Forward for Ultrasound

The argument built in this section is based on the observed limits of our task-specific thyroid work, the field’s unresolved landscape of competing SSL approaches for ultrasound, and the theoretical analysis of what each pretext task learns. The culmination motivates a specific design choice: JEPA-based pretraining as the objective for a general-purpose ultrasound foundation model.

This is not a claim that JEA-based or reconstruction-based approaches are without value. The Pretext Matters study finds that JEA-based objectives outperform JEPA on certain tasks where fine-grained local features dominate, and both outperform reconstruction across ultrasound tasks. However, given what we know about ultrasound, such as its high noise magnitude, spatially organized anatomy, and a lack of established augmentation protocols, JEPA provides the most principled starting point for building a foundation model that generalizes broadly across anatomical sites and clinical tasks. The remainder of this chapter directly tackles this hypothesis through the design and pretraining of US-JEPA. Chapter 7 addresses a lack of evaluation standardization in this research through UltraBench 2.0, a tool intended to help the community with reproducibility and benchmarking.

6.2 Introduction

Self-supervised learning (SSL) has revolutionized the ability to leverage vast unlabeled datasets to build robust foundation models. In particular, masked image modeling (MIM) has emerged as a dominant approach, where models are tasked to reconstruct masked image patches to learn image semantics [HCM⁺25]. Although pixel-level reconstruction objectives have proven effective in natural imaging, their utility partially depends on the assumption that local pixel intensities correlate strongly with underlying structural representations. However, when applied to domains characterized by low signal-to-noise ratio and acquisition artifacts, this assumption begins to break down [XGH⁺24].

The ultrasound (US) imaging domain is a prime example of this representational bottleneck. US is a critical diagnostic tool in clinical settings, as it is portable, radiation-free, and provides real-time bedside feedback. Yet, image interpretation is difficult due to inherent visual graininess (speckle noise), operator dependence, and substantial variability in acquisition protocols and patient anatomy. In a traditional MIM framework, a generative, pixel-reconstruction objective may force the model to use its representational capacity to model these uninformative, acquisition-dependent features. Since these features, such as blur, acoustic shadow, and pixel-level contrast, vary significantly across clinical environments, a model trained for pixel reconstruction may overfit to these specific noise sources. The resulting significant gap between manually curated training data and real-world settings raises critical concerns about model robustness in out-of-distribution (OOD) scenarios. Without a pretraining objective that prioritizes semantic invariance, US foundation models may fail when subjected to diverse, yet realistic, ultrasound image corruptions.

The challenge of ensuring representational robustness is further compounded by the scarcity of high-quality clinical labels, which are expensive to acquire and require specialized medical expertise. Consequently, there is a need to train foundation models that can extract robust US feature representations and achieve high performance on downstream tasks using minimal supervision. Recent efforts to build US foundation models have attempted to navigate these bottlenecks through domain-specific modifications [JZL⁺24, MDE⁺25, ZWZ⁺25]. Despite their innovations, these methods ultimately remain anchored to the pixel reconstruction paradigm, potentially underrepresenting the critical global features that distinguish different anatomical regions and pathologies in US.

6.2.1 Ultrasound Joint Embedding Predictive Architecture (US-JEPA)

To address these challenges, we introduce the Ultrasound Joint-Embedding Predictive Architecture (US-JEPA). Based on Image-based JEPA (I-JEPA), US-JEPA goes beyond the

generative pixel-filling paradigm. Instead, it operates entirely in a latent embedding space, predicting the representations of masked target regions from a context block within the same image. By performing feature reconstruction rather than raw pixel reconstruction, US-JEPA can focus on learning global anatomical dependencies and tissue textures.

The I-JEPA framework [ADM⁺23] relies on an online teacher updated via Exponential Moving Average (EMA) to guide a gradient-based student with smooth, yet improving targets. This coupling is computationally expensive and sensitive to hyperparameter selection. A core distinction of our framework is the adoption of the Static-teacher Asymmetric Latent Training (SALT) objective [LHL⁺25], in which a frozen teacher replaces an EMA-updated teacher. US-JEPA uses a frozen, domain-specific teacher, the Ultrasound Representation Foundation Model (URFM) [KLG⁺25], to provide stable targets.

URFM previously established the effectiveness of MIM for feature-level reconstruction in ultrasound, performing knowledge distillation from BiomedCLIP [ZXU⁺25], a general-purpose medical vision encoder. URFM’s reconstruction-based objective transfers rich semantic priors to the ultrasound domain, but remains tied to local reconstruction objectives as opposed to global representation consistency. By using URFM as a static teacher, US-JEPA takes a step further by leveraging the SALT latent prediction objective to refine URFM’s ultrasound-specific representations. Rather than just reconstructing missing features, US-JEPA learns to align representations across views within an image. This approach encourages the model to learn anatomical context and global structure within US imaging, providing an improved latent space for downstream clinical tasks.

To ensure the model pretrained with US-JEPA is exposed to a wide range of human physiology, we aggregate one of the largest collections of publicly available US imaging to date. After thoroughly searching public repositories, our pretraining dataset encompasses approximately 5.12 million frames from 50 datasets covering 22 distinct anatomies.

6.2.2 Standardizing Evaluation: UltraBench

A significant challenge in current US foundation model research is the lack of a standardized evaluation protocol. Currently, most studies evaluate on disparate, often private, downstream datasets with non-standardized splits, making rigorous iteration and objective comparison nearly impossible. To promote reproducibility and establish a gold standard for future work, our paper adopts and extends UltraBench [TG], a publicly available benchmark for US frame-level evaluation. While the original UltraBench provides a strong foundation, we expand its scope to enhance anatomical diversity. Specifically, we include eight classification tasks spanning various organs and pathologies, adding two new tasks for thyroid and breast pathology.

Beyond the dataset benchmark itself, we address a critical gap in previous work: the lack of side-by-side comparisons of available models. To the best of our knowledge, this work is the first to perform an exhaustive linear probing evaluation across all published, publicly available US foundation models. Previous works compare against a limited set of US-specific baselines or rely solely on full fine-tuning, making it difficult to isolate the intrinsic quality of the learned representations. By tuning just a linear head for the classification tasks with a frozen backbone, we provide a standardized assessment of the learned latent space. This rigorous benchmarking ensures that improvements from US-JEPA are measured against a standard and relevant set of baselines, providing a clear map of the current state-of-the-art.

This work takes a significant step forward in ultrasound foundation modeling with the following contributions:

- **JEPA-based US Foundation Model:** We introduce US-JEPA, the first frame-level US foundation model built on JEPA principles.
- **Label-Efficient Representations:** US-JEPA achieves strong linear probing performance with fewer labeled samples than competing baselines.
- **Robustness to Domain-Specific Image Corruption:** Learned representations

exhibit increased invariance to ultrasound-specific perturbations in image quality.

- **Comprehensive Benchmarking:** We conduct the most comprehensive linear-probing evaluation to date across publicly available ultrasound foundation models on eight clinical tasks from UltraBench.
- **Largest Public Pretraining Corpus:** We propose the OPUS (OPen UltraSound) dataset, the largest pretraining dataset to date for ultrasound composed exclusively from publicly available sources.

6.3 Related Work

6.3.1 Universal Vision Foundation Models

While early SSL focused on instance discrimination and contrastive objectives (e.g. MoCo [HFW⁺20] and SimCLR [CKNH20b]), the field has expanded to self-distillation and predictive modeling to capture higher-level semantics. DINO [CTM⁺21b] demonstrated that student-teacher distillation with Vision Transformers (ViTs) can yield semantically rich features without labels. This paradigm has been scaled by successors, DINOv2 [ODM⁺24] and DINOv3 [SVS⁺25], to set new benchmarks in universal representation learning. In parallel, I-JEPA [ADM⁺23] proposed moving beyond methods that predict in pixel or token space. By predicting missing latent representations from a visible context, I-JEPA learns the underlying structural logic of an image rather than superficial textures. Despite these universal models excelling on natural images, their direct application to US is limited by modality-specific artifacts, leading research to investigate specialized anatomical and task-based solutions.

6.3.2 Task-Specific Ultrasound Foundation Models

A significant branch of US research focuses on adapting universal architectures for specific clinical tasks, most notably through the specialization of the Segment Anything Model (SAM)

[KMR⁺23]. Works such as SAMUS [LXY⁺24] and UltraSAM [MMZ⁺25] utilize architectural modifications or large-scale fine-tuning on large segmentation datasets to enable interactive, zero-shot segmentation. Beyond segmentation, SSL has proven useful for US representation learning within specific anatomical systems. In fetal US, UltraDINO [AML⁺25] outperforms general vision models and USFM by training DINOv2 from scratch on fetal-specific corpora. In the cardiac domain, EchoNet-Dynamic and EchoFM [OHG⁺20, KJS⁺25] use large-scale US video corpora to learn temporal cardiac dynamics and apply these representations to clinical cardiac tasks. In parallel, others have pursued latent space prediction, with Ellis et al. introducing JEPAs to the domain by adopting V-JEPA for ultrasound video to prioritize temporal information over noisy pixels [EMB⁺25]. This approach was followed by EchoJEPA [MFS⁺26], which demonstrated that JEPAs can develop a functional understanding of cardiac cycles in echocardiography. These successes demonstrate the power of domain-specific SSL; however, they often rely on privately curated, task-specific datasets, motivating the need for a robust, general-purpose US foundation model.

6.3.3 General Ultrasound Foundation Models

The evolution of US foundation models has been defined by a shift from standard pixel-level reconstruction toward domain-aware representation learning. USFM [JZL⁺24] was the first to define the field by training on more than 2 million images (public and private) using a dual spatial-frequency MIM objective, specifically designed to tackle the low spatial resolution and noise inherent in US. Building on this precedent, USF-MAE [MDE⁺25] demonstrated that pre-training exclusively on public US data (using the 370k-image OpenUS-46 corpus) combined with noise-filtering preprocessing could significantly surpass models initialized on natural images.

More recent efforts have integrated generative refinements and structural priors. For instance, D²MAE [KGZ⁺] uses diffusional deblurring in the MIM framework to address the low SNR inherent to US. EchoCare [ZWZ⁺25] leverages a 4.5 million image dataset and

introduces a hierarchical classification objective to teach the model nuanced anatomical relationships.

To build on these previous approaches, US-JEPA evolves the feature reconstruction paradigm introduced by URFM. By utilizing this domain-specific teacher and an asymmetric latent objective, US-JEPA is designed to encode the structural logic within US, learning an improved latent space for downstream clinical tasks.

6.4 Preliminaries

6.4.1 I-JEPA

The I-JEPA [ADM⁺23] architecture consists of a context encoder f_θ , a target encoder $\bar{f}_\theta$, and a predictor g_ϕ . Given an input image x , the model operates on a sequence of non-overlapping patches. During training, we sample a context block B_c and a set of T target blocks $\{B_i\}_{i=1}^T$, where each block is a subset of patches. From these blocks, we derive disjoint masks M_c and M_i to ensure there is no overlap and prevent a non-trivial prediction task.

The target encoder processes the full image x to produce patch-level representations $\mathbf{s}_y = \bar{f}_\theta(x)$. The targets for the loss function are defined as the subset of these representations corresponding to the i -th target mask, $\mathbf{y}_i = \mathbf{s}_y[M_i]$. Simultaneously, the context encoder processes only the visible image patches defined by M_c to produce context embeddings $\mathbf{c} = f_\theta(x[M_c])$. The predictor g_ϕ then utilizes these context embeddings, conditioned on mask tokens p_i , which indicate the target-block locations. The mask tokens are parameterized by a learnable vector with an added positional embedding:

$$\hat{\mathbf{y}}_i = g_\phi(\mathbf{c}, p_i) \quad (6.1)$$

The training objective is to minimize the Smooth L_1 distance between the predicted $\hat{\mathbf{y}}_i$ and actual $\mathbf{y}_i$ target representations. To prevent representational collapse, I-JEPA employs

an asymmetric update rule where the target encoder parameters $\bar{\theta}$ are an EMA of the context encoder parameters θ .

6.4.2 Static-teacher Asymmetric Latent Training (SALT)

The SALT paradigm [LHL⁺25] helps modulate the parameter selection in the JEPA objective by decoupling the student and teacher optimization. SALT demonstrates that a high-performing student can emerge by predicting the latent space of a static teacher, provided the teacher possesses reasonably sufficient semantic priors.

Here, the target encoder h_ψ (e.g. an existing foundation model or encoder pretrained via a different objective such as pixel reconstruction) is frozen. The target embeddings $\mathbf{y}_i$ become static conditioned on the input image. Gradients propagate solely through the context encoder f_θ and predictor g_ϕ :

$$\mathcal{L}_{\text{SALT}} = \frac{1}{T} \sum_{i=1}^T \|g_\phi(f_\theta(x[M_c]), p_i) - h_\psi(x)[M_i]\|_1 \quad (6.2)$$

This approach eliminates the need for the EMA update, thus stabilizing the training dynamics and reducing computational overhead.

6.5 US-JEPA

US-JEPA (Fig. 6.2) relies on the masked latent objective from I-JEPA to learn the spatial semantics within ultrasound imaging. By forcing the student network to predict missing anatomical structures in a latent space with the guidance of a domain-specific teacher, the model is encouraged to develop a robust feature representation of tissue textures and organ morphology that is invariant to pixel-level noise.

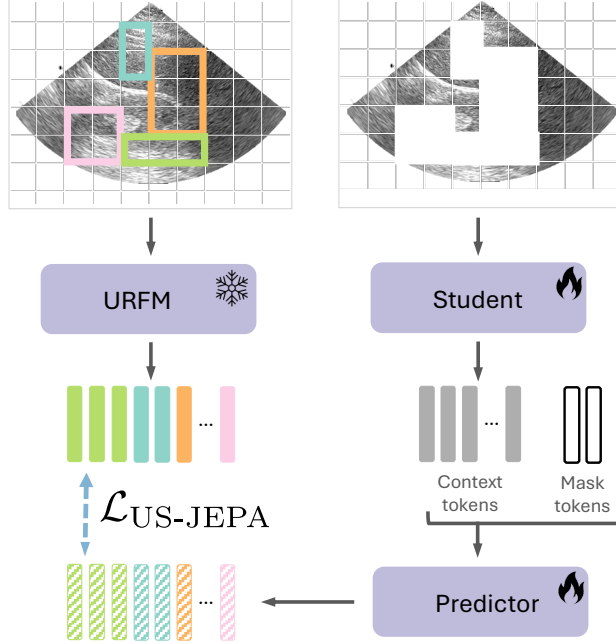


Figure 6.2: **USrc-JEPA framework.** URFM is the frozen teacher that extracts target embeddings. The student and predictor are jointly optimized with $\mathcal{L}_{\text{US-JEPA}}$ to align with the target.

6.5.1 Self-Distillation via SALT

While the SALT framework has shown that strong students can emerge even from sub-optimal teachers [LHL⁺25], our goal was to leverage the most robust US representations to establish a new state-of-the-art. To identify the optimal target provider, we performed an extensive benchmarking of existing US foundation models across UltraBench. Our preliminary results (detailed in Section 6.6) demonstrate that URFM consistently outperformed other baselines. We select URFM to serve as our frozen teacher and investigate whether the masked latent objective of US-JEPA can achieve better downstream performance, label efficiency in few-shot linear probing, and improved stability against domain-specific corruptions.

In our framework, the student context encoder and predictor are optimized to minimize the Smooth L_1 distance between the predicted target embeddings and the URFM target embeddings.

6.5.2 Ultrasound Region-Conditioning (USrc)

A significant challenge in training on diverse, public ultrasound datasets is the presence of non-anatomical artifacts. Ultrasound frames typically contain peripheral noise, including transducer metadata, intensity scales, patient information, and large black borders. Previous masked imaging approaches rely on raw frames, which is suboptimal since it allows non-ultrasound content to be sampled. We hypothesize that uniform random masking in I-JEPA might inadvertently task the model with predicting these meaningless regions, wasting representational capacity. To test this, we introduce USrc as a spatial prior to isolate the anatomical signal.

Let $x \in \mathbb{R}^{H \times W}$ be the input ultrasound image and $R \in \{0, 1\}^{H \times W}$ be a binary region mask, where $R_{ij} = 1$ denotes valid ultrasound signal and $R_{ij} = 0$ denotes background. We tile the image into non-overlapping patches and define the set of valid patches $\mathcal{P}_{valid}$ as those containing at least one pixel within R .

Target Sampling: Targets correspond to representations of image blocks. We feed x through the frozen target encoder h_ψ to obtain patch-level representations $\mathbf{s}_y = h_\psi(x)$. We sample T candidate blocks $\{B_i\}_{i=1}^T$ based on scale and aspect ratio constraints. We employ a rejection sampling strategy where a candidate block is accepted only if its intersection with $\mathcal{P}_{valid}$ exceeds a threshold τ :

$$M_i = B_i \cap \mathcal{P}_{valid}, \quad \text{subject to } |M_i| \geq \tau \quad (6.3)$$

The final targets are the representations $\mathbf{y}_i = \mathbf{s}_y[M_i]$.

Context Sampling: We sample a context block B_c with specific scale and aspect ratio constraints. To ensure a non-trivial task, we remove any regions overlapping with the target blocks and apply our overlap constraint:

$$M_c = B_c \cap \mathcal{P}_{valid}, \quad \text{subject to } |M_c| \geq \tau \quad (6.4)$$

The masked context $x[M_c]$ is processed by the context encoder f_θ to obtain $\mathbf{c} = f_\theta(x[M_c])$.

Optimization: The training objective is minimized strictly over these anatomical intersections. The predictor aims to minimize the Smooth L_1 distance between the predicted student embeddings and the frozen teacher embeddings:

$$\mathcal{L}_{\text{US-JEPA}} = \frac{1}{T} \sum_{i=1}^T \text{Smooth L1}(g_\phi(\mathbf{c}, p_i), \mathbf{y}_i) \quad (6.5)$$

By constraining both input and output to $\mathcal{P}_{\text{valid}}$, US-JEPA with USrc (USrc-JEPA) engineers the inputs to only model tissue texture and anatomical structure.

6.5.3 Training Pipeline

To facilitate standardization across our aggregated corpus of US frames, we implement a standardized preprocessing and sampling pipeline for training.

Preprocessing: Images are first converted to grayscale. Regions with minimal colored artifacts or annotations are inpainted as long as they occupy less than 5% of the total image area. The image is cropped to remove any negative space around the ultrasound content. To normalize for varying transducer gain settings across datasets, we perform intensity rescaling by mapping the 2nd and 98th percentiles of the pixel distribution to the full dynamic range. The USrc mask is used to condition the intensity rescaling only within the US content.

Data Balancing Strategy: A significant challenge in aggregating public medical datasets is the extreme variance in dataset size, which risks biasing the model toward the larger datasets. We mitigate this through a weighted dataset-level sampling strategy, where we define an upper bound sampling threshold N_t of 50,000 to limit the effective size of each dataset per training epoch. For each dataset D_i with actual size $|D_i|$, the effective count is $\min(|D_i|, N_t)$, and the sampling probability is:

$$P(D_i) = \frac{\min(|D_i|, N_t)}{\sum_j \min(|D_j|, N_t)} \quad (6.6)$$

This strategy ensures that datasets larger than N_t contribute equally during training, while smaller datasets contribute proportionally to their true size, balancing data diversity against over-representation of the largest sources.

6.5.4 US-JEPA Architecture

Both the student context encoder and the frozen URFM teacher employ a ViT-B/16 backbone with patch size 16 and embedding dimension 768, processing 224×224 inputs. The predictor is a narrower transformer with output dimension 384, followed by a linear projection head to align with the teacher’s embedding space prior to Smooth L_1 loss computation. Per image, we sample one large context block (scale $[0.85, 1.0]$) and four target blocks (scale $[0.075, 0.125]$, aspect ratio $[0.75, 1.5]$), enforcing non-overlapping masks to ensure the prediction task is non-trivial. We train for 100 epochs with batch size 128 using AdamW. Full hyperparameters are provided in Appendix 6.8.3.

6.6 Experiments and Results

6.6.1 Data

Pretraining Dataset

We introduce OPUS, to our knowledge the largest aggregation of publicly available ultrasound data, comprising 5,123,697 frames curated from 50 distinct publicly available datasets (Fig. 6.3). Both US-JEPA and USrc-JEPA are pretrained on OPUS. The data contains three different US acquisition types, with a majority (79.7%) of the frames derived from temporal video sequences and the remaining from volumetric (3D) scans or static 2D images. The dataset also spans a wide range of human anatomy, including major organ systems and important ancillary structures. Cardiac data represents the largest subset (23,904 videos), largely sourced from high-volume video datasets like EchoNet-Dynamic and EchoNet-LVH.

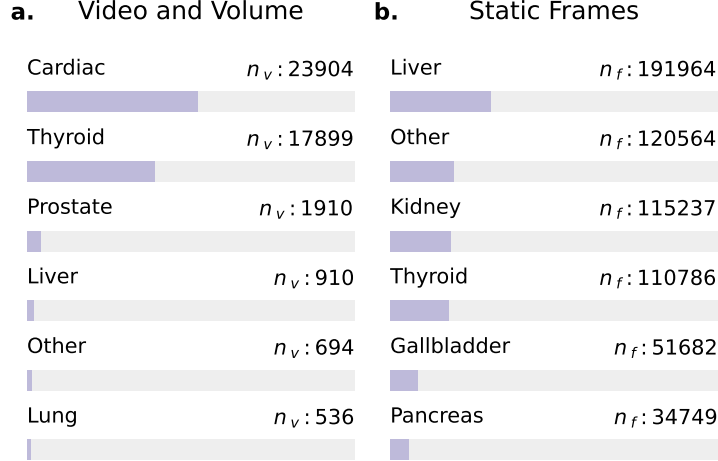


Figure 6.3: **Distribution of pretraining data.** To characterize the dataset composition at the organ level, we report the distribution of (a) temporal sequences, including videos and volumes (n_v), and (b) individual static frames (n_f).

This is complemented by significant representations of key organs, including the liver (318K frames), thyroid (264K frames), and prostate (227K frames), as well as breast (79K frames) and lung (55K). The data originates from a diverse array of clinical contexts, ranging from large-scale institutional repositories (e.g. Stanford, TCIA, RadImageNet) to more specialized, curated datasets, providing the model with exposure to varying image qualities, transducer frequencies, and pathological presentations. Further details regarding the OPUS pretraining dataset are included in Appendix 6.8.1.

UltraBench Downstream Dataset

To ensure standardized and reproducible evaluation, we leverage UltraBench [TG], a publicly available benchmark encompassing diverse ultrasound tasks across segmentation, detection, and classification. We add implementation for BUSBRA and TN5000 to UltraBench to expand anatomical coverage for breast and thyroid. Our evaluation focuses on eight classification tasks, comprising both binary and multi-class endpoints (Table 6.1), to assess the discriminative ability and representation quality of our models. Five of the eight datasets focus primarily on cancer detection. FATTY LIVER and POCUS have non-oncological applications in fatty

liver disease detection and lung pathology detection (healthy, pneumonia, and COVID), respectively. Lastly, BUTTERFLY is a general multi-organ detection task. In Section 6.6.3 and Section 6.6.4, our figures specifically show results for four datasets: FATTY LIVER and POCUS (non-cancer), MMOTU (challenging 8-class task), and BUSBRA (cancer-specific performance).

Table 6.1: Summary of downstream datasets used for classification tasks, including the target organ, number of classes (Y), and total image counts. All datasets are public.

Dataset	Organ	Total Images
AUL ($Y=3$)	Liver	735
BUSBRA ($Y=2$)	Breast	1064
BUTTERFLY ($Y=9$)	Multi-Organ	41076
FATTY LIV. ($Y=2$)	Liver	550
GBCU ($Y=3$)	Gallbladder	1255
MMOTU ($Y=8$)	Ovary	1469
POCUS ($Y=3$)	Lung	2064
TN5000 ($Y=2$)	Thyroid	5000

6.6.2 UltraBench Evaluation

To effectively evaluate the feature quality of US-JEPA and USrc-JEPA, we compare their performance against baselines using the standard SSL practice of linear probing [AB18]. We identified six US foundation models with publicly available pretrained weights: USFM [JZL⁺24], URFM [KLG⁺25], USF-MAE [MDE⁺25], EchoCare [ZWZ⁺25], UltraSAM [MMZ⁺25], and SAMUS [LXYY24]. UltraSAM and SAMUS are foundation models specifically developed for segmentation tasks, while the remaining are general ultrasound foundation models. Furthermore, we evaluate against two pretrained, state-of-the-art universal vision foundation models, DINOv3 and I-JEPA, which serve as out-of-domain benchmarks. To better isolate the contribution of our pretraining dataset, we use OPUS to train a ViT-Base under a standard MAE framework (MAE*) and retrain URFM from scratch (URFM*) using its defined architecture and hyperparameters. The final baseline is a ViT-Base trained with

I-JEPA (I-JEPA*), using the original EMA-updated teacher setup to better understand the impact of SALT’s frozen-teacher approach. Using the image encoder from each framework as a frozen feature extractor, we train linear probes for each downstream task across five random seeds to measure stability (Table 6.2 and 6.3).

Table 6.2: **Linear probe on downstream datasets.** We compare results of linear probes trained using our proposed method against state-of-the-art baselines. We report the mean macro F1 score % over five seeds for all eight downstream datasets. Bold indicates the best result and underlining indicates the second best.

Model	AUL	BUSBRA	BUTTERFLY	FATTY LIV.	GBCU	MMOTU	POCUS	TN5000
DINOv3	64.3±0.6	70.9±1.7	91.7±0.4	55.8±5.5	61.7±0.5	37.2±0.6	91.4±0.4	67.5±0.4
I-JEPA	61.5±1.1	71.2±4.0	90.5±0.6	54.8±1.6	53.7±0.4	35.3±0.6	88.1±0.4	68.9±0.2
UltraSAM	62.6±3.1	70.2±3.1	89.6±2.4	66.9±3.3	43.5±4.9	39.7±1.8	87.3±2.1	63.9±2.0
SAMUS	40.2±0.9	65.9±0.3	91.5±0.0	42.1±0.0	48.8±0.3	20.4±0.2	76.2±0.1	51.7±0.0
EchoCare	49.2±2.4	64.4±0.0	84.1±0.7	42.1±0.0	36.2±0.5	21.1±0.1	73.8±3.9	49.8±3.8
USF-MAE	58.1±1.4	62.9±0.5	91.1±0.3	42.1±0.0	45.9±0.3	28.7±0.3	90.1±0.0	56.3±1.1
USFM	61.6±1.2	<u>74.6±0.5</u>	<u>92.4±0.3</u>	73.6±8.8	<u>67.4±0.6</u>	33.8±0.3	85.7±0.5	65.0±2.6
URFM	<u>69.5±0.4</u>	<u>72.7±0.3</u>	92.7±0.4	71.8±1.9	55.4±0.8	41.9±0.1	92.1±0.0	78.1±0.3
US-JEPA	69.6±1.5	73.8±1.1	90.8±0.3	<u>82.5±1.1</u>	67.0±1.4	52.2±0.2	93.1±0.0	<u>73.1±0.7</u>
USrc-JEPA	67.6±0.5	76.0±1.2	91.5±0.6	89.2±0.9	70.2±0.5	<u>46.8±0.2</u>	<u>92.5±0.1</u>	70.8±1.3

Table 6.3: **Linear probe on downstream datasets — controlled pretraining comparison.** We compare our models against frameworks retrained on our pretraining corpus. MAE* denotes a ViT-Base trained under a standard masked autoencoder framework on our data; I-JEPA* denotes I-JEPA with an EMA-updated teacher trained on our data; URFM* denotes URFM retrained on our data using its original hyperparameters. We report the mean macro F1 score % over five seeds for all eight downstream datasets. Bold indicates the best result and underlining indicates the second best.

Model	AUL	BUSBRA	BUTTERFLY	FATTY LIV.	GBCU	MMOTU	POCUS	TN5000
MAE*	64.5±0.0	70.4±0.5	89.0±0.2	82.4±0.0	62.5±0.0	41.0±0.4	<u>92.5±0.2</u>	68.5±0.7
I-JEPA*	57.8±1.3	69.2±1.3	82.8±0.6	83.6±7.2	52.2±1.2	29.8±0.2	83.8±0.5	55.5±3.4
URFM*	72.1±0.2	71.2±0.2	93.6±0.1	89.9±0.0	62.9±0.2	45.9±0.5	90.9±0.0	<u>71.0±0.3</u>
US-JEPA	<u>69.6±1.5</u>	<u>73.8±1.1</u>	90.8±0.3	82.5±1.1	<u>67.0±1.4</u>	52.2±0.2	93.1±0.0	73.1±0.7
USrc-JEPA	<u>67.6±0.5</u>	76.0±1.2	<u>91.5±0.6</u>	<u>89.2±0.9</u>	70.2±0.5	<u>46.8±0.2</u>	<u>92.5±0.1</u>	70.8±1.3

Direct Comparison

US-JEPA and USrc-JEPA each achieve state-of-the-art performance on three of the eight tasks (AUL, MMOTU and POCUS for US-JEPA, and BUSBRA, FATTY LIV. and GBCU

for USrc-JEPA). They each also fall in second place on two tasks (FATTY LIV. and TN5000 for US-JEPA, and MMOTU and POCUS for USrc-JEPA). BUTTERFLY is the only task that our models underperform against the top 2 baselines, URFM and USFM. However, even in this case, the difference is less than 2% macro F1, and for a task where all models achieve competitive performance indicating it may be a relatively trivial task. Notably, our models excel in the most challenging scenarios where baseline methods struggle. The MMOTU dataset entails an eight-class ovarian tumor classification task. Average baseline performance drops below 40% on MMOTU, while US-JEPA establishes a new benchmark of 52.2%, surpassing the best baseline, URFM, by 10.3%. While URFM performs strongly on TN5000, our models still offer superior performance on the remaining downstream tasks, and clearly beat general-purpose vision models (DINOv3 and I-JEPA) and the majority of the domain-specific baselines.

Controlled Pretraining Comparison

Table 6.3 isolates the contributions of our pretraining corpus and architectural design by comparing baselines retrained on OPUS. MAE* and I-JEPA* underperform our models across all tasks, confirming that the pretraining data alone does not account for the gains from our framework. URFM*, which retrains the URFM architecture on OPUS, outperforms the original URFM on 5 of 8 downstreams, providing direct empirical support that OPUS yields stronger representations. The notable difference is with TN5000, where URFM achieves a macro F1 of 78.1 whereas URFM* drops to 71.0. Lastly, despite URFM* being a strong data-controlled baseline, US-JEPA outperforms URFM* on 5 of 8 tasks, demonstrating that our architectural contributions provide complementary gains beyond pretraining data alone.

6.6.3 Few-Shot Scaling for Linear Probe

To evaluate the efficiency and transferability of our models’ representations, we measure downstream performance with varying amounts of supervision. For each dataset, we train

linear probes (across five seeds to measure stability) on every model’s frozen features using stratified subsets of the training data: {1%, 5%, 10%, 50%, 100%}. By preserving the original UltraBench splits and class prevalence through subsampling, we quantify how effectively pretrained features aid learning in low-shot regimes. Superior representation quality is evidenced by higher quantitative performance across all label densities, and by a less severe performance drop as label density decreases.

Fig. 6.4 illustrates macro F1 performance across four datasets under varying label quantities for linear probing. We compare degradation against URFM and USFM. This pattern is most clearly seen with the FATTY LIVER downstream. When probes are trained with less than 50% of labels, our models achieve an average macro F1 of 19% higher than the URFM and USFM baselines. For the POCUS task, a similar trend is observed where the mean macro F1 degrades more severely for the baselines compared to our models. As probes are trained with fewer labels, even for BUSBRA, MMOTU, and GBCU (Appendix 6.8.4), our models are at least on par, if not better than URFM and USFM. Complete results for all downstream tasks are included in Appendix 6.8.4.

6.6.4 Robustness to Domain-Specific Corruption

To evaluate the stability of US-JEPA and USrc-JEPA under distribution shifts, we perform an incremental stress test consisting of three synthetic, domain-specific image corruptions: Gaussian blur, contrast depletion, and correlated speckle noise. The first motivation for this is that in real-world clinical environments, ultrasound image quality is frequently compromised by hardware constraints, operator variability, and suboptimal acquisition settings. The second is to demonstrate that the JEPA-based pretraining objective encourages the model to learn global, structural relationships, and not expend model capacity on pixel-level variance.

For each corruption type, we evaluate downstream performance on test sets across three levels of increasing severity ($\epsilon \in \{1, 2, 3\}$). While blur and contrast are implemented via standard intensity transforms, we model speckle as a multiplicative, spatially-correlated

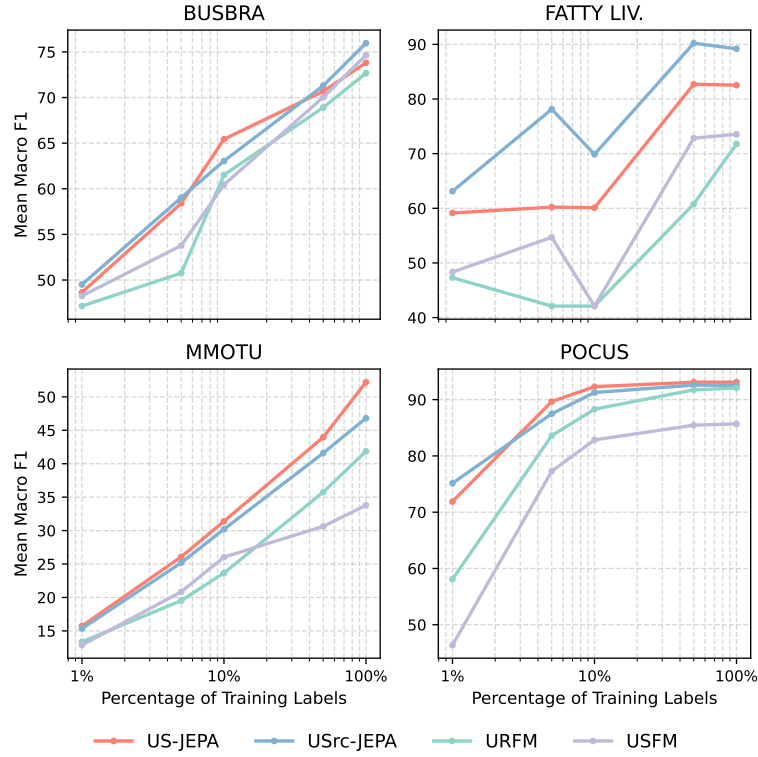


Figure 6.4: **Results for few-shot scaling.** We report the mean macro F1 score % over five seeds for each model’s probe to measure performance with 1% to 100% of training labels. *Note: each dataset is plotted on a unique y-axis scale to better highlight model-specific performance trends.*

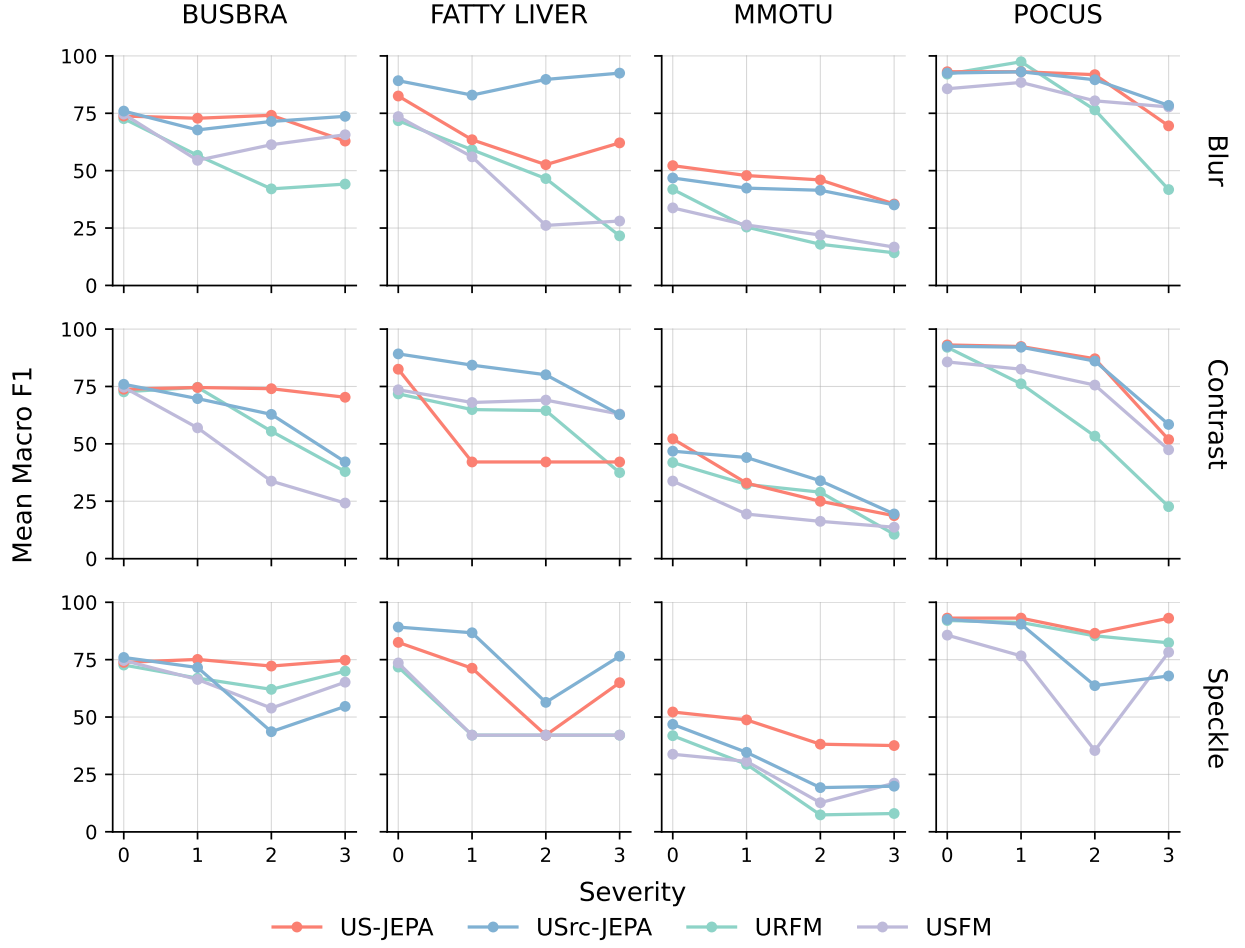


Figure 6.5: **Robustness to domain-specific corruption.** Results show the mean macro F1 across five seeds for each model–dataset–corruption permutation. Linear probes were trained on full, uncorrupted training sets and evaluated on increasingly corrupted test sets to assess structural representation stability.

process to approximate the “grainy” texture inherent to ultrasound scans. More detailed implementations for these corruptions are provided in Appendix 6.8.5.

Fig. 6.5 visualizes the performance degradation for a subset of downstream tasks, with the remaining tasks detailed in Appendix 6.8.5. Most notably, both US-JEPA and USrc-JEPA demonstrate significant resilience to the blur corruption. On the POCUS dataset, URFM degrades to less than half its uncorrupted performance, dropping from 92.1 to 41.8 F1 at maximum blur severity. In contrast, US-JEPA and USrc-JEPA are much more resilient, with F1 only dropping to 69.5 and 78.4 respectively at maximum blur severity. Similarly, on the BUTTERFLY dataset (Appendix 6.8.5), URFM drops to 17.8 F1 under severe blur, while US-JEPA and USrc-JEPA maintain F1 performance at 80.8 and 70.9. In these severe blur cases, the second baseline USFM drops to 77.8 and 59.8 F1 for POCUS and BUTTERFLY, exhibiting less degradation than URFM but still performing worse than both our models for BUTTERFLY.

Regarding texture and intensity corruptions, our models match or exceed URFM and USFM for most downstream tasks. One unique weakness is observed on GBCU (Appendix 6.8.5) where URFM maintains consistent performance with F1 only decreasing by 6.2% from normal to the highest contrast severity, while our models drop by 47%.

For speckle corruption, our models demonstrate remarkable stability. Specifically, on BUTTERFLY, US-JEPA and USrc-JEPA only drop by 0.6% and 9.8% under severe speckle noise ($\epsilon = 3$), whereas USFM and URFM drop by 25% and 44.1%. Notably, across all downstream tasks US-JEPA outperforms or matches URFM and USFM under peak speckle severity (USrc-JEPA does so for 6/8 downstreams), demonstrating robustness to this prevalent ultrasound-specific noise. While USFM shows competitive results against USrc-JEPA on four tasks, it remains inferior to USrc-JEPA on the remaining, often by substantial margins, reaching a performance gap of up to 34.4% on FATTY LIVER.

Ultimately, this stress test quantifies the invariance of our learned representations under significant corruption relative to the current state-of-the-art. By evaluating against domain-

specific synthetic noise, we provide a practical validation of US-JEPA and USrc-JEPA’s feature stability, a critical requirement given the high variance in US image quality across different manufacturers, operators, and clinical environments. The observed performance under corruption suggests that our approach is able to better capture structural semantics in US instead of overfitting to superficial, pixel-level features.

6.6.5 Visualizing Learned Representations

UMAP Feature Maps: The US-JEPA and USrc-JEPA models produce generally distinct and separable clusters for each downstream dataset (Fig. 6.6), whereas the URFM baseline demonstrates more overlap between representations for non-related anatomical structures (e.g., gallbladder and lung). We also note that our models adjacently cluster samples from the AUL (liver) and GBCU (gallbladder, with views of the liver) datasets, indicating that the features may capture overlapping anatomical regions. Lastly, for the multi-organ BUTTERFLY dataset, US-JEPA and USrc-JEPA form distinct clusters, which are correlated with their unique organ label.

Attention Maps: Appendix 6.8.6 contains examples of attention maps to illustrate that our models consistently prioritize global structure and anatomical context. This trend is evidenced by specific observations across key datasets. In GBCU, our models capture liver and gallbladder boundaries along with surrounding peripheral tissue. In BUSBRA, our models attend to lesion contours, internal architecture, and stromal tissue. In AUL, our models maintain attention toward liver boundaries and broader tissue structures.

We observe similar patterns across the remaining datasets, where our models consistently capture the global structure and anatomy relevant to the task. These visualizations lend qualitative support to our claim that the JEPA objective encourages the model to learn the spatial continuity and relationships inherent to ultrasound rather than just focusing on clinically insignificant intensity differences or high-contrast regions.

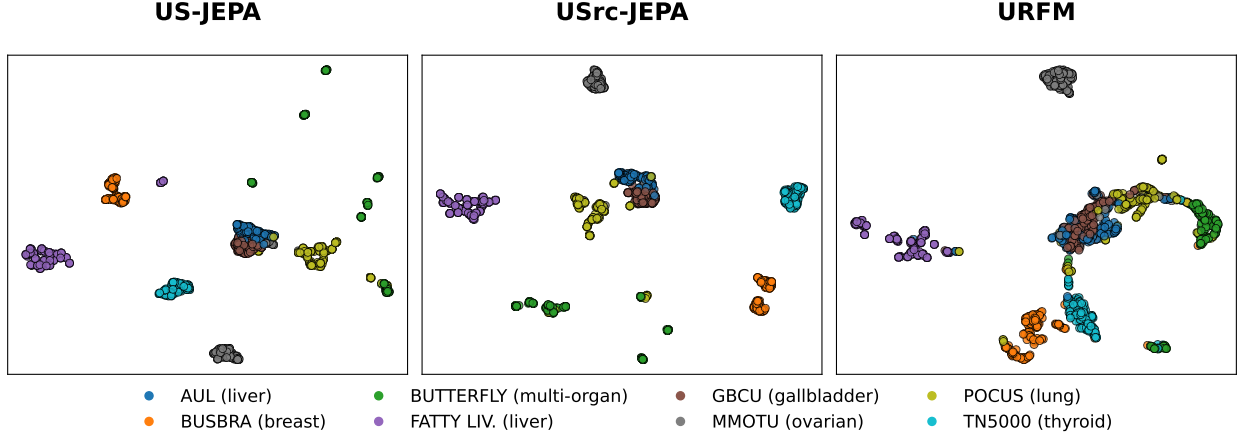


Figure 6.6: **UMAP feature visualizations.** UMAP projections of features extracted from US-JEPA, USrc-JEPA, and URFM across downstream datasets.

6.7 Discussion

We present US-JEPA, a novel approach that demonstrates that training using a JEPA framework with a robust, static teacher yields stable, data-efficient representations for ultrasound. By adopting the SALT paradigm, we decouple the optimization of the student and teacher and leverage the semantic priors of a domain-specific teacher. Furthermore, with USrc-JEPA we show that adding USrc to the framework also achieves state-of-the-art performance on certain downstream tasks. This contribution is highlighted in our downstream stress tests: our models remain robust in low-data regimes for linear probing and maintain performance during evaluation with domain-specific corruptions.

This work also advances standardization in ultrasound foundation model evaluation through three key contributions: 1) a rigorous linear probing comparison across all published models to establish standardized baselines for future benchmarking, 2) an expanded UltraBench incorporating TN5000 and BUSBRA for broader anatomical coverage and 3) the aggregation of OPUS, the largest publicly defined corpus of US imaging to date. By pretraining US-JEPA with OPUS, we demonstrate superior results against previous US foundation models on a majority of downstream tasks across relevant clinical domains.

Clinical Implications: First, our investigation into learning label-efficient representations

proved insightful for understanding which ultrasound foundation models are sample efficient. We observed that both US-JEPA and USrc-JEPA models were able to achieve as good, if not better, performance compared to the strongest baselines, USFM and URFM, in lower label regimes. These empirical results better position our proposed models to be more practically useful in resource-limited clinical settings, where annotated data may be difficult to acquire for fine-tuning, but a task-specific model is still necessary. Second, the experiments with domain-specific corruption provide a deeper understanding of how robust models are to synthetic, pixel-level corruptions. This experiment is an attempt to simulate real-world variability in visual appearance (e.g., transducer types, operators, etc.). More importantly, our models perform as well, or better, than baselines, USFM and URFM, under a majority of corruption-severity combinations.

Limitations: Our robustness evaluation relies on synthetic corruptions (Gaussian blur, contrast depletion, and speckle) as reproducible proxies for clinical artifacts. While these stress tests assess a model’s reliance on global structure over pixel-level texture, they do not fully capture complex ultrasound-specific phenomena such as acoustic shadowing or motion artifacts, whose occurrence depends on anatomical context and acquisition dynamics that are difficult to simulate without ground-truth maps. Future work will validate robustness against real-world distribution shifts, such as heterogeneous scanner settings, to more directly assess generalizability of models.

A second limitation concerns sensitivity to organ-level data density in pretraining. The performance gap between URFM and URFM* on thyroid nodule classification (TN5000), dropping from 78.1 to 71.0 macro F1, is likely attributable to the substantially lower thyroid prevalence in OPUS (5.2%) relative to the original URFM pretraining corpus (44.2%). This observation lends evidence to dataset scale alone being insufficient; anatomical or pathological balance in pretraining data composition may also meaningfully shape the quality of learned representations [WZG⁺26]. More broadly, underrepresented organs such as the thyroid and gallbladder may constitute failure modes for models pretrained on large-scale datasets

without sufficient heterogeneity. Building a more principled understanding of how organ- and pathology-level density affects downstream representations remains an open challenge, and future work will extend our dataset-weighted sampling strategy to explicitly account for organ- and pathology-level prevalence during pretraining.

Future Directions: Our work opens a new research direction in ultrasound foundation models by demonstrating that JEPA-based latent prediction is a viable and effective pretraining objective for this domain. While considerable effort has gone into frame-level foundation models, ultrasound video and volumetric imaging remain largely underexplored. Given the inherent spatiotemporal and volumetric structure in these modalities, masked latent reconstruction in the representation space offers a well-suited inductive bias for learning meaningful features. Beyond pretraining, broader task-specific downstream evaluation, spanning segmentation, detection, and regression alongside classification, would help surface application areas where current models struggle. And lastly, more systematic comparisons with task- and anatomy-specific foundation models would clarify how pretraining data composition and architectural design jointly shape learned representations and downstream performance.

Our findings confirm the utility of the JEPA framework for ultrasound self-supervised learning and provide an argument that joint-embedding predictive architectures could be a powerful way forward in learning stronger features from ultrasound imaging.

6.8 Appendix

6.8.1 Dataset Details

OPUS: Dataset Composition

For complete transparency, we have included the name of each dataset and associated metadata in OPUS.

Table 6.4: Comprehensive breakdown of datasets used, categorized by anatomical region, frame type, and total frame count.

Dataset Name	Anatomy	Frame Type	Frame Count
ultrasoundcases.info	Abdominal	Static	3199
abdominal_US [VOIL20]	Abdominal	Static	617
ultrasoundcases.info	Appendix	Static	781
ultrasoundcases.info	Bladder	Static	450
USAnotAI [KA24a]	Bladder	Static	61
RadImageNet [MLR ⁺ 22]	Bladder	Static	758
USAnotAI	Bowel	Static	61
ultrasoundcases.info	Brain	Static	466
ReMIND [JDK ⁺ 23]	Brain	Video	18734
ultrasoundcases.info	Breast	Static	4956
BUSI [ADGKF19]	Breast	Static	780
BUS_UC [Iqb23b]	Breast	Static	811
Breast-Lesions-USG [P��D ⁺ 24]	Breast	Static	256
BUSC [Iqb23a]	Breast	Static	250
DatasetA [JRAE ⁺ 24]	Breast	Static	250
BUV [LLZ ⁺ 22]	Breast	Video	25026
BUID [AAMMAAA23]	Breast	Static	205
ABUS-TDSC [LXC ⁺ 25]	Breast	Volume	44474
Breast_S1 [GDWG21]	Breast	Static	201
UDIAT [YPM ⁺ 18]	Breast	Static	163
BUSI_WHU [HZZ ⁺ 23]	Breast	Static	927
BUS_UCLM [VBD ⁺ 24]	Breast	Static	683
cardiacCCAUS [Aga22]	Cardiac	Static	1100
ultrasoundcases.info	Cardiac	Static	1443
EchoNet-LVH [DCY ⁺ 22]	Cardiac	Video	1957928
Carotid US Boundary Study [Mei22]	Cardiac	Static	500
CAMUS [LSP ⁺ 19]	Cardiac	Video	19232
cardiacUDC [YDZ ⁺ 23]	Cardiac	Video	38071

Dataset Name	Anatomy	Frame Type	Frame Count
EchoNet: Tee-View-Classifer [SCG ⁺ 24]	Cardiac	Video	55261
EchoNet Dynamic [OHG ⁺]	Cardiac	Video	1770636
RadImageNet	Cardiac	Static	19062
RadImageNet	Fibroid	Static	2095
GB1 [TMOB ⁺ 25]	Gallbladder	Static	10692
USAnotAI	Gallbladder	Static	61
RadImageNet	Gallbladder	Static	39743
ultrasoundcases.info	Gallbladder	Static	1186
GIST_EUS [HBL ⁺ 23]	Gastrointestinal	Static	514
ultrasoundcases.info	Gastrointestinal	Static	2618
USAnotAI	Kidney	Static	61
TRUSTED [NFT ⁺ 25]	Kidney	Volume	7280
OpenKidney [SRH ⁺ 22]	Kidney	Static	500
RadImageNet	Kidney	Static	111252
ultrasoundcases.info	Kidney	Static	3924
RadImageNet	Liver	Static	78535
ultrasoundcases.info	Liver	Static	2714
USAnotAI	Liver	Static	61
SYSU-CEUS-FLL [LCHL14]	Liver	Sampled Video	110654
B-mode-and-CEUS-Liver [ELW21]	Liver	Video	126543
COVID-Blues [BWC ⁺ 21]	Lung	Video	31696
COVIDx-US [EXM ⁺ 22]	Lung	Video	22697
ultrasoundcases.info	Lung	Static	714
ultrasoundcases.info	Lymph node	Static	320
ultrasoundcases.info	Muscle	Static	17459
Mus-V [GMC ⁺ 24]	Muscle	Sampled Video	3114
MuscleUS [Mei21]	Muscle	Static	8169
UBPD [DYW ⁺ 22]	Nerve	Sampled Video	955
ultrasoundcases.info	Nerve	Static	458
NerveUS [MSH ⁺ 16]	Nerve	Static	11143
ovarianUS [BSS ⁺ 25]	Ovarian	Static	301

Dataset Name	Anatomy	Frame Type	Frame Count
PCOS_US [CK]	Ovarian	Static	3841
RadImageNet	Ovarian	Static	3595
ultrasoundcases.info	Pancreas	Static	1506
LEPset [LZW ⁺ 23]	Pancreas	Static	11493
105US [HBH ⁺ 17]	Pancreas	Static	105
RadImageNet	Pancreas	Static	21645
RadImageNet	Portal Vein	Static	769
ProstateTRUS [BSM ⁺ 23]	Prostate	Volume	5037
Prostate-MRI-US-Biopsy [NPM ⁺ 20]	Prostate	Volume	219484
MUPSD [SB24]	Prostate	Volume	2910
ultrasoundcases.info	Reproductive	Static	3446
RadImageNet	Spleen	Static	6520
ultrasoundcases.info	Spleen	Static	1207
USAnotAI	Spleen	Static	61
RadImageNet	Thyroid	Static	92598
SegThy [KED ⁺ 22]	Thyroid	Volume	136243
ultrasoundcases.info	Thyroid	Static	2011
Thyd2 [HHZ ⁺ 24]	Thyroid	Static	8489
DDTI [PVN ⁺ 15]	Thyroid	Static	610
TG3K [gon26]	Thyroid	Static	3585
TN3K [gon26]	Thyroid	Static	3493
Stanford Thyroid Cine [YKA ⁺ 22]	Thyroid	Video	17412
RadImageNet	Uterus	Static	13312
Yang [Tia23]	Uterus	Static	1973

Downstream Datasets

Below, we provide a more detailed description of each downstream dataset, summarized in Table 6.5. We use the publicly available UltraBench tool to set up the datasets and corresponding training, validation and test splits. We contribute two more datasets, BUSBRA

and TN5000, to UltraBench for a more comprehensive downstream evaluation.

Annotated Ultrasound Liver (AUL): The Annotated Ultrasound Liver (AUL) dataset is a collection of 2D liver ultrasound images introduced for research on automated liver lesion analysis and mass characterization [XZL⁺23]. It comprises images acquired from distinct patients and includes expert annotations delineating the liver boundary and, when present, lesion contours. Each image is additionally assigned a clinical label reflecting the presence and type of focal liver mass, enabling both region-aware and image-level learning. The dataset spans a wide range of image resolutions and visual appearances, capturing realistic variability in ultrasound acquisition and anatomy. We use AUL for a supervised image-level mass classification downstream task with three classes: malignant mass, benign mass, and normal (no mass). The full dataset contains 735 images in total. We follow a fixed split consisting of 529 images for training, 59 for validation, and 147 for testing, ensuring that all splits remain patient-independent and preserve the original class distribution.

Butterfly: The Butterfly dataset was released by Butterfly Network for the 2018 MIT Grand Hack to promote research in point-of-care ultrasound understanding [But18]. It comprises ultrasound images acquired using the Butterfly iQ device from multiple anatomical regions across 31 patients, reflecting the diversity and variability encountered in real-world bedside imaging. Each image is labeled according to the anatomical site being examined, enabling supervised learning for anatomical recognition in ultrasound. We use the Butterfly dataset for a supervised organ classification downstream task with nine classes: Morison’s pouch, bladder, heart (parasternal long-axis view), heart (four-chamber view), heart (two-chamber view), inferior vena cava, carotid artery, lungs, and thyroid. The dataset contains a total of 41,076 images, with 28,053 images used for training, 6,272 for validation, and 6,751 for testing. All splits are patient-independent to prevent data leakage and ensure fair evaluation.

Fatty Liver: The Fatty Liver dataset consists of B-mode liver ultrasound images introduced for studying automated detection of hepatic steatosis and non-alcoholic fatty liver

disease (NAFLD) [BSS⁺18]. The dataset includes images acquired from multiple patients with expert clinical labeling based on the presence of fatty infiltration in the liver parenchyma. It has been widely used to evaluate representation learning and transfer learning methods for liver ultrasound analysis, particularly in low-data clinical settings. We use this dataset for a supervised image-level fatty liver discrimination downstream task with two classes: normal liver and NAFLD. The dataset contains a total of 550 images, drawn from 55 patients. We use 390 images for training, 50 for validation, and 110 for testing. All splits are patient-independent to ensure a fair assessment of generalization performance.

GBCU: The Gallbladder Cancer Ultrasound (GBCU) dataset was introduced to support research on automated gallbladder lesion characterization from abdominal ultrasound images [BGR⁺22]. It consists of expert-annotated images collected from a diverse patient cohort and labeled according to underlying pathology, enabling supervised learning for clinically relevant gallbladder cancer assessment. The dataset captures a range of normal and diseased presentations and has been used to benchmark deep learning methods for malignancy discrimination in ultrasound. We use GBCU for a supervised image-level lesion malignancy classification downstream task with three classes: normal, benign, and malignant. The full dataset contains 1,255 images in total. We use 1,019 images for training, 114 for validation, and 122 for testing. The original dataset provides a fixed train–test split with no patient overlap; the validation set is created by further splitting the training data, as patient identifiers are not available to enforce strict separation.

MMOTU: The Multi-Modality Ovarian Tumor Ultrasound (MMOTU) dataset was introduced to facilitate research on ovarian tumor analysis using ultrasound and contrast-enhanced ultrasound imaging [ZLB⁺23]. It provides expert annotations and diagnostic labels for a diverse set of ovarian pathologies, supporting both localization and classification tasks. In this work, we consider only the 2D ultrasound images and associated labels, focusing on image-level learning for ovarian tumor characterization. We use MMOTU for a supervised multi-class image-level tumor classification downstream task with eight classes: chocolate

cyst, serous cystadenoma, teratoma, theca cell tumor, simple cyst, normal ovary, mucinous cystadenoma, and high-grade serous cystadenocarcinoma. The dataset contains 1,469 images in total. We use 800 images for training, 200 for validation, and 469 for testing. The original dataset provides a fixed train–test split, and the validation set is obtained by further splitting the training data, as patient identifiers are not available to enforce strict patient-level separation.

POCUS: The Point-of-Care Ultrasound (POCUS) dataset was introduced to support automated diagnosis of lung pathologies, particularly COVID-19, from bedside ultrasound imaging [BWC⁺21]. It aggregates lung ultrasound data acquired using convex and linear probes from multiple sources, reflecting real-world variability in point-of-care acquisition. The dataset has been widely used to benchmark deep learning methods for pulmonary disease recognition in ultrasound under heterogeneous data conditions. We use POCUS for a supervised image-level lung pathology classification downstream task with three classes: healthy, pneumonia, and COVID-19. Following the preprocessing protocol described in the original work, we use 29 convex-probe images and extract frame-level samples from 124 convex-probe videos while grouping frames by video to avoid data leakage across splits. The resulting dataset contains 2,064 images in total, with 1,444 images used for training, 177 for validation, and 443 for testing, ensuring strict separation at the video level.

BUSBRA: The BUS-BRA (Breast Ultrasound Brazil) dataset was introduced to support the development and standardized evaluation of computer-aided diagnosis systems for breast ultrasound imaging, with an emphasis on clinically meaningful annotations and reproducible benchmarking [GFGCCdAP24]. It consists of anonymized breast ultrasound images acquired from female patients during routine clinical examinations and includes expert-provided lesion delineations and biopsy-confirmed pathology labels. In addition to image-level diagnostic labels, the dataset incorporates BI-RADS assessments assigned by an experienced ultrasonographer, enabling research across multiple levels of breast cancer risk stratification and lesion characterization. The dataset has been designed to facilitate fair comparison of learning-

based methods by providing well-defined partitions and comprehensive annotations commonly required in breast CAD pipelines. We use BUSBRA for a supervised image-level tumor malignancy classification downstream task with two classes: benign (722) and malignant (342). The dataset contains 1,064 images in total. We use 765 images for training, 86 for validation, and 213 for testing.

TN5000: The TN5000 dataset was introduced to enable large-scale research on automated thyroid nodule analysis from ultrasound imaging, with a particular emphasis on malignancy assessment [ZLH⁺25]. It comprises expertly curated ultrasound images of thyroid nodules collected through a rigorous process of data selection and annotation. Each image is paired with structured annotations following a PASCAL VOC-compatible format, where nodules are explicitly labeled and localized, allowing both region-aware modeling and image-level learning. Diagnostic labels are derived from clinical assessment, and the dataset was intentionally constructed to reduce bias by maintaining a relatively balanced representation of benign and malignant cases at scale. TN5000 has been widely adopted for benchmarking deep learning approaches for thyroid cancer detection due to its size, standardized organization, and clear labeling scheme. We use TN5000 for a supervised image-level thyroid nodule malignancy classification downstream task with two classes: benign and malignant. The dataset contains a total of 5,000 images, including 3,572 malignant nodules and 1,428 benign nodules. We use 3,500 images for training, 500 for validation, and 1,000 for testing.

Table 6.5: Comprehensive statistics of the downstream datasets, including organ, task description, number of classes and data splits.

Dataset	Organ	Task Description	Classes	Total	Train	Val	Test
TN5000	Thyroid	Nodule Malig.	2	5000	3500	500	1000
FATTY LIV.	Liver	Fatty Liver Dis.	2	550	390	50	110
POCUS	Lung	Pneumonia, COVID	3	2064	1444	177	443
BUTTERFLY	Multi	Organ Detection	9	41076	28053	6272	6751
GBCU	Gallbladder	Lesion Malig.	3	1255	1019	114	122
AUL	Liver	Mass Malig.	3	735	529	59	147
MMOTU	Ovary	Tumor Malig.	8	1469	800	200	469
BUSBRA	Breast	Tumor Malig.	2	1064	765	86	213

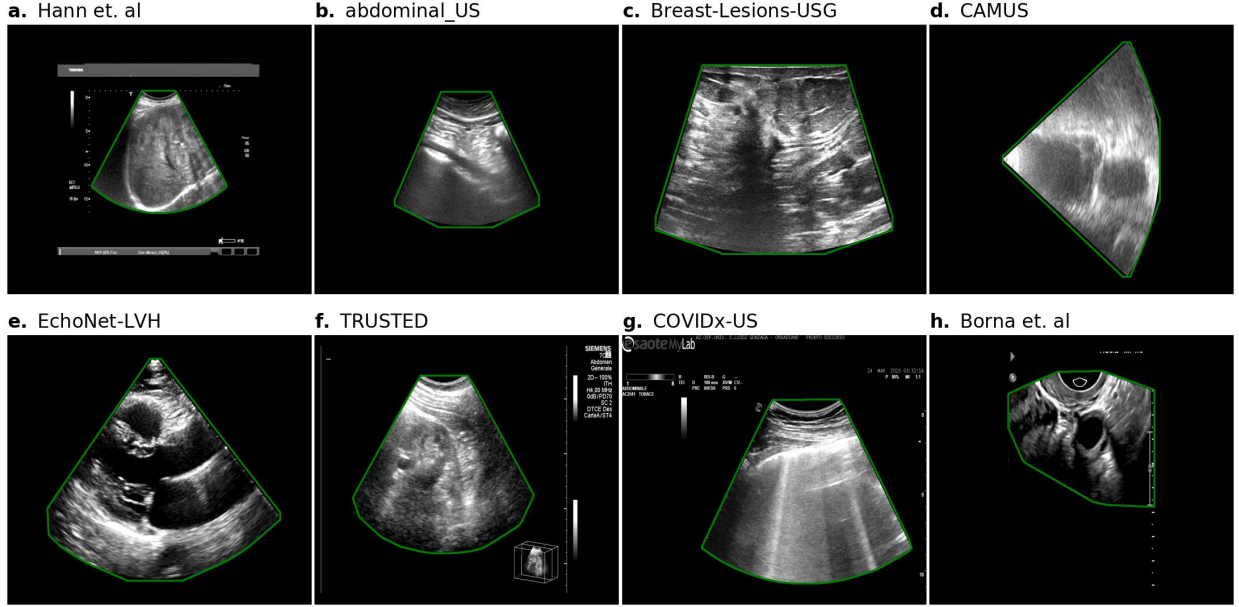


Figure 6.7: Examples of images and corresponding USrc mask.

6.8.2 US Region Conditioning (USrc)

Image and USrc Mask Examples

In Fig. 6.7, we have picked 8 datasets to include examples of images and the corresponding USrc mask. The USrc mask generated by our image processing algorithm is outlined in green and demonstrates the quality of the USrc mask generation.

USrc Mask Generation

To ensure reproducibility, we detail the deterministic image processing pipeline used to generate the binary mask R . We first apply a global threshold at a pixel intensity of $15/255$, followed by a 5×5 morphological closing to remove gaps in the mask. To adaptively distinguish between different scan types, we identify external contours and rank them by area. We dynamically include the second-largest connected domain only if its area exceeds 50% of the primary contour. A convex hull transformation is then applied to ensure the mask remains

continuous despite internal acoustic shadows, followed by a 25x25 morphological close and a gaussian blur ($\sigma = 7$) for final edge smoothing. The USrc mask generation process does not require manual tuning or specialized preprocessing for individual datasets. The same pipeline was applied to 49 of the 50 datasets, with only the prostate US TCIA dataset requiring a prefixed circular geometry.

6.8.3 Training Details

Pretraining Parameters

We exclusively hold out 5% of the entire pretraining dataset to use as validation. For our downstream experiments, we evaluate the epoch checkpoint that gets the lowest validation loss during pretraining. All pretraining parameters are detailed in Table 6.6.

Downstream Parameters

We perform downstream evaluation by training a randomly initialized linear layer on top of the frozen encoder backbone. We utilize early stopping based on the validation loss with a patience of 15 epochs to prevent overfitting. The remaining parameters are detailed in Table 6.7.

6.8.4 Few-Shot Scaling for Linear Probe

Extended Results

In Fig. 6.8 we include the comprehensive few-shot scaling experiments. For all of our baselines we perform few-shot scaling across five random seeds to see how performance changes on each downstream. In the main paper, we only compare against URFM and USFM because they were the most competitive baselines.

Table 6.6: Shared pre-training hyperparameters for US-JEPA and USrc-JEPA. The models differ only in their use of ultrasound region conditioning.

Configuration	Value
<i>Architecture</i>	
Encoder Backbone	ViT-Base
Predictor Depth	12
Predictor Embedding Dim	384
Predictor Target Dim	768
<i>Data & Augmentation</i>	
Input Size	224×224
Crop Scale	$[0.6, 1.0]$
Augmentations	H/V Flip, Blur, Artificial Speckle, Contrast
Batch Size	128
<i>Masking (JEPA)</i>	
Patch Size	16
Context Masks (Num, Scale)	1, $[0.85, 1.0]$
Target Masks (Num, Scale)	4, $[0.075, 0.125]$
Aspect Ratio	$[0.75, 1.5]$
Min. Patches Kept	10
<i>Optimization</i>	
Optimizer	AdamW
Total Epochs	100
Warmup Epochs	10
Base Learning Rate (LR)	5.0×10^{-5}
Start $\rightarrow$ Final LR	$5.0 \times 10^{-6} \rightarrow 5.0 \times 10^{-7}$
LR Schedule	Cosine Decay with Linear Warmup
Weight Decay $\rightarrow$ Final WD	$0.04 \rightarrow 0.4$

Table 6.7: Hyperparameters for downstream linear probing on UltraBench datasets.

Configuration	Value
Optimizer	AdamW
Base Learning Rate	1.0×10^{-3}
Weight Decay	1.0×10^{-4}
Batch Size	32
Input Size	224×224
Loss Function	Cross-Entropy
LR Schedule	Cosine Annealing
Max Training Epochs	150
Early Stopping Patience	15 Epochs

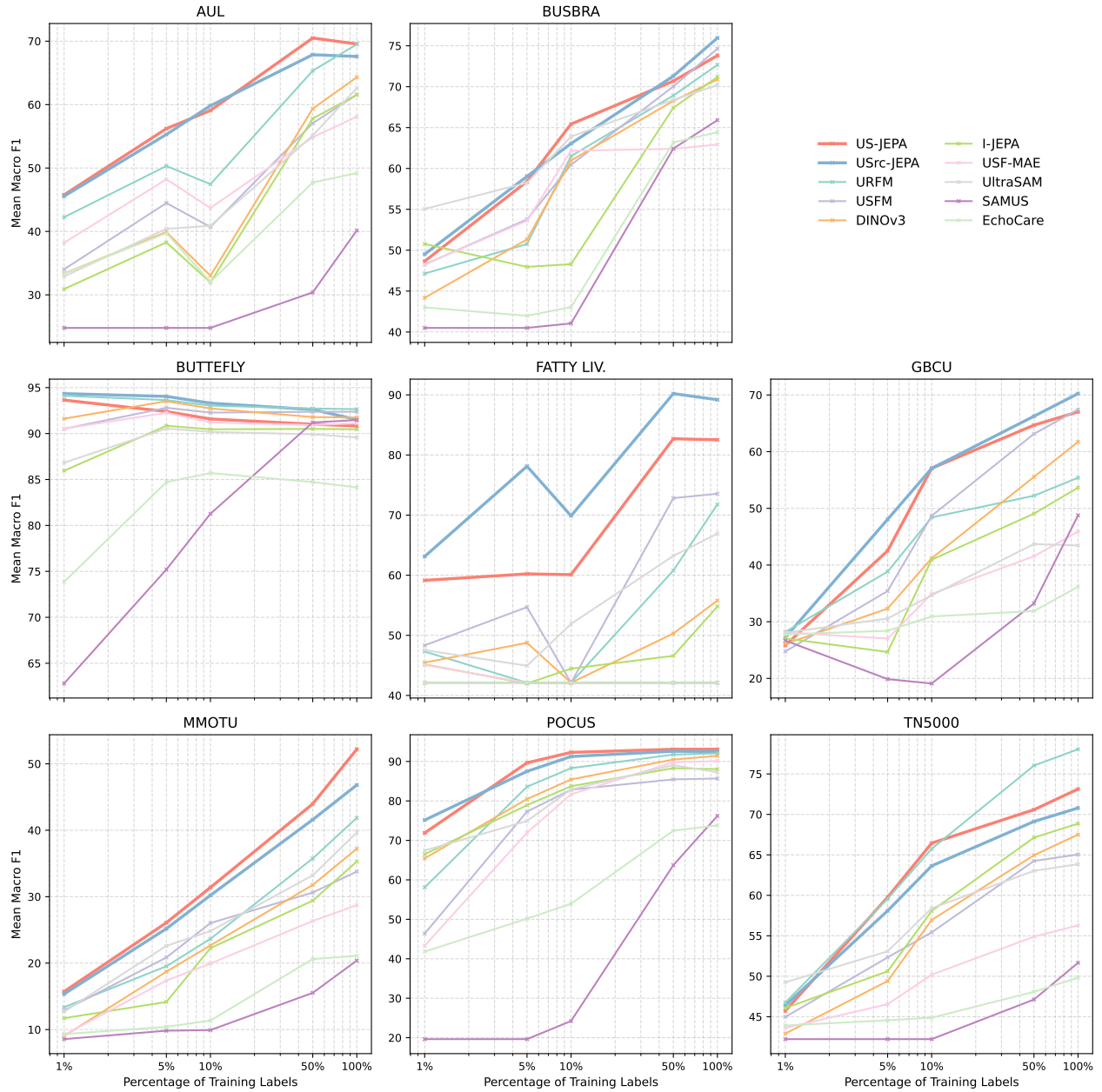


Figure 6.8: Few-shot classification robustness across all datasets. Here the test set evaluation for linear probe tuned with varying percentages of training data are shown for all models.

6.8.5 Robustness to Domain-Specific Corruption Details

Extended Results

In Fig. 6.9 we include the complete results for the OOD robustness experiment. We show the performance degradation for US-JEPA, USrc-JEPA, URFM and USFM with worsening levels of blur, contrast reduction and artificial speckle corrupting downstream test set images during probe evaluation.

Formulas for Corruptions

To simulate ultrasound imaging artifacts, we define three corruption functions $C(I, \epsilon)$ where I is the input image and ϵ is the severity level ranging from 1 to 3.

Gaussian Blur: Gaussian blur is modeled as the convolution of the image I with a Gaussian kernel G , where the standard deviation σ is controlled by the severity ϵ :

$$I_{blur} = I * G_{\sigma} \quad \text{where } \sigma = \epsilon \quad (6.7)$$

The discrete kernel size for G is adjusted to $2\lfloor 2\epsilon \rfloor + 1$ to ensure the Gaussian distribution is properly captured.

Contrast Depletion: This corruption shrinks intensities toward the median brightness μ_{med} of the pixels defined by the USrc mask:

$$I_{contrast} = \mu_{med} + \alpha(\epsilon) \cdot (I - \mu_{med}) \quad (6.8)$$

where the depletion factor α decreases as severity increases ($\alpha \in \{0.7, 0.5, 0.3\}$), effectively making anatomical boundaries less visible.

Correlated Speckle Noise: Speckle is modeled as multiplicative noise. The raw

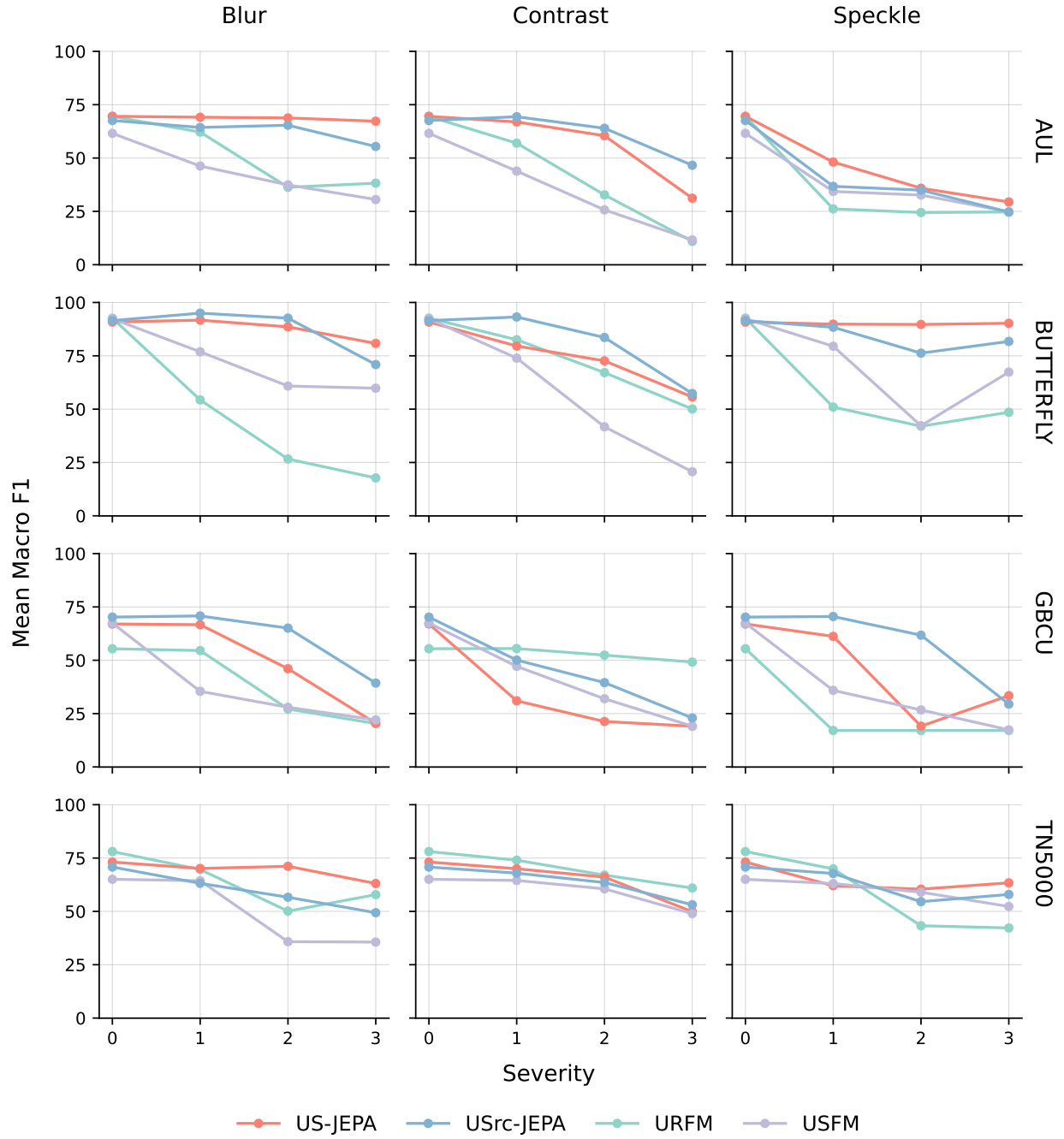


Figure 6.9: OOD robustness results for remaining 4 downstream datasets.

Gaussian noise η is spatially correlated via a smoothing kernel K before being applied.

$$I_{speckle} = I \cdot (1 + \eta_{corr}) \quad (6.9)$$

The noise characteristics are functions of severity ϵ :

- Magnitude: $\eta \sim \mathcal{N}(0, \sigma_{noise}^2)$ where $\sigma_{noise} = 0.35\epsilon$.
- Correlation: $\eta_{corr} = \eta * K_{size}(\epsilon)$, where the kernel size K_{size} increases with ϵ to simulate larger speckle grains at higher severity.

Qualitative Examples

In Fig. 6.10, we have included a qualitative example of the varying severities for each corruption on one example image from the AUL downstream dataset.

6.8.6 Attention Visualizations

In Fig. 6.11, we show the attention maps from three heads for each US-JEPA and USrc-JEPA across three distinct datasets.

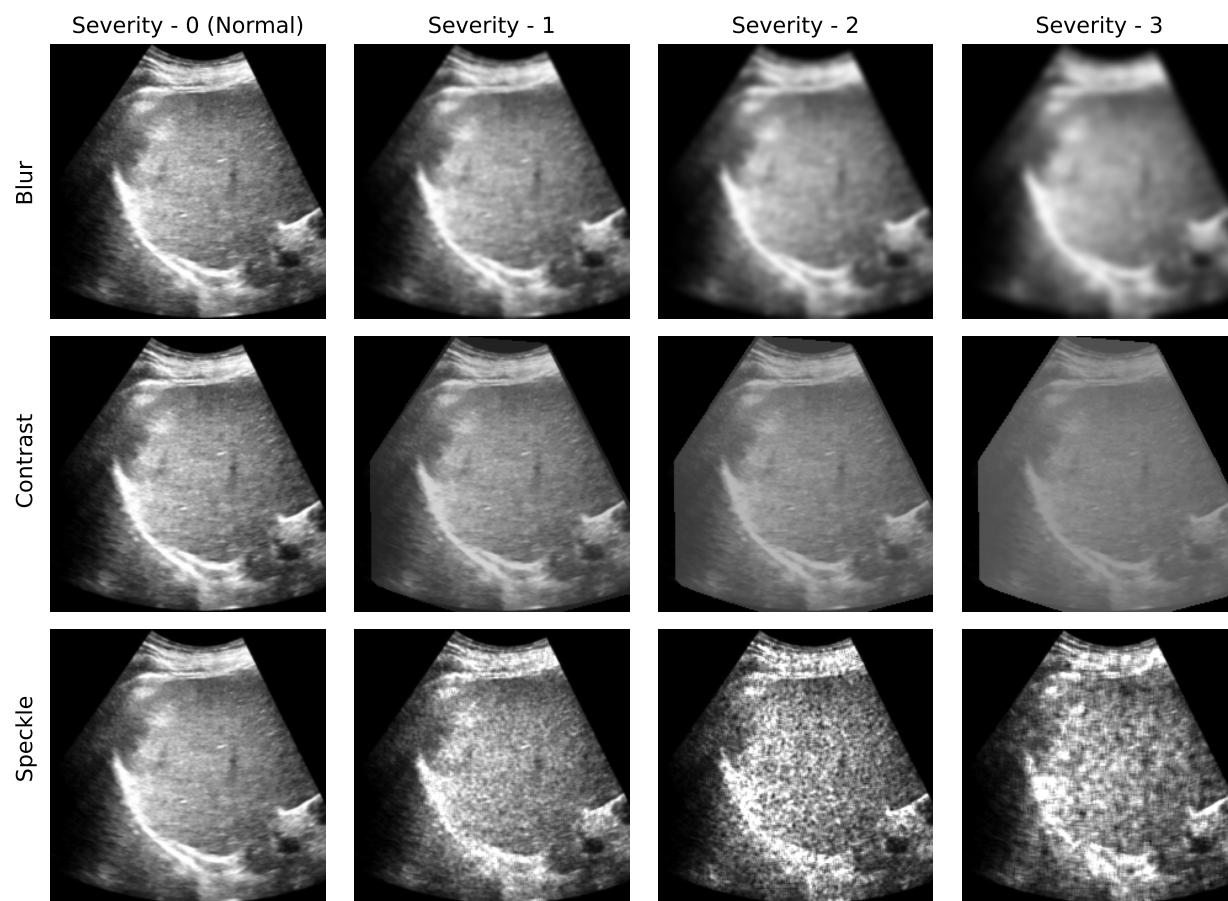


Figure 6.10: Different corruptions at three severity levels applied to a liver ultrasound from AUL dataset.

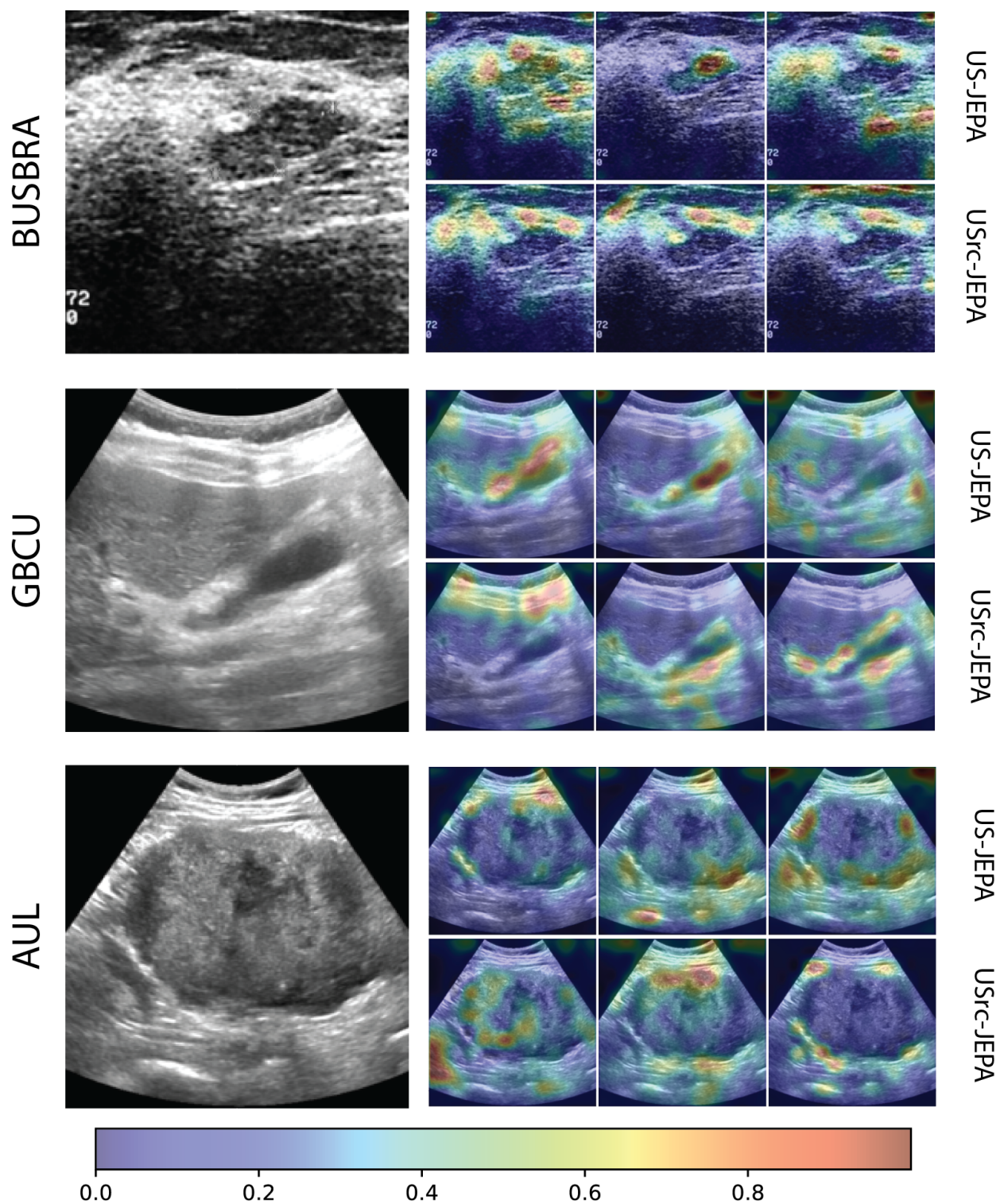


Figure 6.11: Attention visualization for US-JEPA and USrc-JEPA for three downstream datasets.

CHAPTER 7

UltraBench 2.0: A Standardized Benchmark for Ultrasound Foundation Model Evaluation

Associated Publications: **Radhachandran, A.***, Tupper, A.*, et al. (2026). UltraBench 2.0: A Standardized Dataset and Evaluation Benchmark for Ultrasound Foundation Models. *In preparation.*

7.1 Introduction

Ultrasound is a frontline diagnostic tool across cardiology, obstetrics, radiology, and oncology, valued for its low cost, portability, real-time capability, and absence of ionizing radiation [Sza14]. Automated analysis is complicated by speckle noise, low signal-to-noise ratio, and high variability across probe types, operators, and acquisition settings [WSSL83, PKJ⁺10], which has historically limited the generalizability of supervised deep learning models across clinical contexts [ACB⁺19].

The success of self-supervised pretraining and vision transformers in natural and medical imaging [HCX⁺21, DBK⁺21, LCT⁺23] has spurred development of ultrasound-specific foundation models (FMs). General-purpose models, such as USFM [JZL⁺24], URFM [KLG⁺25], USF-MAE [MDE⁺25], and EchoCare [ZWZ⁺25], adopt masked image modeling backbones pretrained on large ultrasound corpora, while segmentation-focused models, such as SAMUS [LXY⁺24] and UltraSAM [MMZ⁺25], adapt the Segment Anything Model. However, each model is evaluated on a different subset of tasks with inconsistent preprocessing, splits, and

metrics, making direct comparison infeasible and progress difficult to measure.

Benchmarks are crucial to capturing the attention of researchers and driving progress on problems we care about [HW24]. Despite the steady development of new ultrasound foundation models, the development of well-designed benchmarks to evaluate them has lagged behind. Benchmarking is an increasingly critical part of research in machine learning and the domains, such as healthcare, where it is applied. The core principle behind foundation models is that they are general-purpose models that can be adapted to a wide variety of tasks with limited examples, far fewer than are required for training a model from scratch. Foundation models are trained on large quantities of diverse data, often using self-supervised learning techniques, to learn the general knowledge and structure of a domain. However, to accurately assess the flexibility these models, we require equally diverse benchmarks to evaluate their claims.

Prior to foundation models, models were evaluated on single-task benchmarks, such as ImageNet for the classification of natural images, or CheXpert [IRK⁺19] for the interpretation of chest X-rays. While many benchmarks have been introduced in recent years for other medical imaging domains, or multi-domain benchmarks, in which ultrasound tasks make up only a few tasks at most, there is a lack of a broad, multi-task image analysis benchmark for ultrasound. This is highlighted in Table 7.1. Prior dataset benchmark collections such as MedMNIST v2 [YSW⁺21], UniUS-Bench [MJL⁺25], UltraBench [TG], US-43d [MMZ⁺25], UltraFedFM [JFR⁺25], and MedSegBench [KA24b], collectively cover many ultrasound tasks but individually lack standardized preprocessing pipelines, auto-download interfaces, permissive open licensing, or integrated evaluation frameworks. Furthermore, several commonly used datasets show near-ceiling reported performance across models, limiting their ability to discriminate between model improvements.

In this chapter, we propose UltraBench 2.0, the first comprehensive benchmark for ultrasound foundation models that standardizes dataset preprocessing and model evaluation in an easy to use package. This benchmark builds upon the foundation of UltraBench,

expanding the set of tasks, improving usability, and standardizing evaluation. UltraBench 2.0 makes four key contributions:

- all datasets are hosted on Hugging Face with a unified auto-download interface and fixed train/test splits
- a shared preprocessing and augmentation pipeline is applied uniformly across all models and tasks
- systematic evaluation covers linear probing, decoder probing, and full fine-tuning across previous ultrasound foundation models on classification and segmentation tasks
- a public leaderboard and code base intended to help future models benchmark performance and for the community to use as a tool to accelerate development and evaluation

7.2 Related Work

7.2.1 Benchmarks in Medical Imaging

Several works have compiled ultrasound datasets or benchmarks for evaluating medical image analysis models, but most of them fall short of providing a complete, accessible evaluation framework. A common limitation is the gap between dataset collection and benchmark infrastructure: datasets are listed and download links are provided, but reproducible preprocessing pipelines, fixed train/test splits, and integrated evaluation scripts are absent or only partially available.

Good examples of multi-domain biomedical imaging benchmarks that have stood the test of time are MedMNIST v2 [YSW⁺21], MedSegBench [KA24b], and MedIMeta [WJB25]. These benchmarks satisfy many of the important qualities described in the introduction. They standardize the format of the different individual datasets and the preprocessing applied to the images (e.g., resizing), define evaluation metrics, and provide easy-to-use interfaces

Benchmark	Ultrasound Tasks		Anatomical Regions	Standardized Preprocessing	Train/Test Splits	Download Interface	Permissive Licensing	Standardized Evaluation/Leaderboard
	Cls.	Seg.						
MedMNIST v2	1	0	1	Yes	Yes	Yes	Yes	No
UniUS-Bench	8	10	15	No	No	No	Mixed	No
UltraBench	7	7	11	Code	Code	No	Mixed	No
MedSegBench	0	5	5	Yes	Yes	Yes	Yes	No
UltraFedFM	10	5	14	No	No	No	Mixed	No
US-43d	0	43	20	Code	No	No	Mixed	No
MedIMeta	2	0	1	Yes	Yes	Yes	Yes	No
UltraBench 2.0	11	10	11	Yes	Yes	Mixed	Mixed	Yes

Table 7.1: A comparison between ultrasound-specific and general medical image analysis benchmarks and dataset collections. Cls. = Classification; Seg. = Segmentation. UltraBench 2.0 is the only benchmark with standardized downloads, preprocessing and evaluation, diverse coverage of ultrasound tasks and anatomical regions, and permissive licensing.

for downloading and working with the data. However, these datasets include few tasks for ultrasound.

The few existing ultrasound-specific benchmarks [JFR⁺25, MJL⁺25, TG] are each lacking in some areas, as highlighted in Table 7.1. While they all have good anatomical region and task coverage, depending on their focus, they suffer when it comes to reproducibility and usability. Regarding reproducibility, previous benchmarks either do not provide fixed train and test splits, or only the code to do so (i.e., with fixed random seeds). Likewise, they either do not specify the preprocessing applied to the images or only provide the code to do so.

These factors also impact the usability of the benchmark, adding extra steps for researchers who wish to follow the same evaluation protocol. Additionally, existing benchmarks provide only links back to the original dataset files. This means that researchers must download each individual dataset and transform it from its original format to a usable format.

MedMNIST v2 [YSW⁺21] is among the most usable collections in medical imaging, offering a standardized MedMNIST-style interface with auto-download, fixed splits, and published baselines. However, it includes only one ultrasound classification task (breast lesion detection), making it insufficient as a comprehensive ultrasound benchmark.

UltraBench [TG], the direct predecessor to this work, introduced 7 classification and 7 segmentation tasks with standardized preprocessing and ImageNet-pretrained baselines. However, datasets must be downloaded manually from external sources, train/test splits require running custom scripts, and the evaluation pipeline does not support foundation model probing and fine-tuning.

UniUS-Bench [MJL⁺25], introduced alongside the TinyUSFM model, provides the broadest task coverage among existing collections (8 classification and 10 segmentation tasks across 15 anatomical domains) and includes FM baselines. Nevertheless, it provides no code for train/val/test splitting, no data download interface beyond links to external sources, and no standardized preprocessing pipeline (only bare PyTorch dataset wrappers). Reproducing its results from scratch requires sourcing each dataset independently and implementing the full preprocessing workflow.

MedSegBench [KA24b] adopts a MedMNIST-style interface with auto-download and fixed splits for 5 ultrasound segmentation tasks, but includes only segmentation and trains baseline models from scratch rather than evaluating pretrained foundation models. **UltraFedFM** [JFR⁺25] covers the largest number of tasks (10 classification, 5 segmentation) but is primarily a federated learning framework rather than a benchmark: datasets are not centrally accessible, preprocessing is not standardized, and significant engineering effort would be required to adapt its code for standard centralized evaluation of new models.

Taken together, existing collections lack permissive open licensing, standardized auto-download, fixed evaluation splits, and an integrated foundation model evaluation pipeline. UltraBench 2.0 directly addresses these limitations by providing all four: auto-download via Hugging Face, fixed splits published alongside the data, and an end-to-end pipeline covering linear probing, decoder probing, and full fine-tuning, allowing researchers to benchmark a new model with minimal overhead.

7.2.2 Ultrasound Foundation Models

General FMs

USFM: USFM [JZL⁺24] was developed by researchers at Fudan University and collaborating institutions, published in Medical Image Analysis in 2024. The framework introduces a spatial-frequency dual masking scheme built on a masked image modeling (MIM) backbone to improve representational robustness from low-quality ultrasound images. Pretraining was conducted on the 3M-US database, the first multi-organ, multi-center, and multi-device ultrasound collection at the time. This database had over two million images from 12 human organs. Evaluation was conducted via full fine-tuning of USFM as a pretrained backbone encoder for segmentation and classification tasks. Downstream experiments spanned three task types: segmentation of key structures (brachial plexus, thyroid nodule, breast tumor, and fetal abdomen), image classification (breast tumor malignancy, multi-organ presence, and fetal planes) and image enhancement.

URFM: URFM [KLG⁺25] was introduced by researchers at West China Hospital, Sichuan University, and collaborating institutions, published in iScience in June 2025. The model proposes a representation-based masked image modeling framework that has the ViT encoder reconstruct high-level semantic features provided by BiomedCLIP, avoiding a potentially inefficient pixel-reconstruction objective [IPR⁺26] and addressing the low signal-to-noise ratio of ultrasound. The pretraining dataset had over one million ultrasound images ($\sim 50\%$ was proprietary) spanning 15 major anatomical organs. Evaluation was based on full fine-tuning

the backbone encoder for classification tasks. Ten downstream ultrasound clinical applications were evaluated across nine organs.

USF-MAE: USF-MAE [MDE⁺25] was developed by Megahed et al. at the University of Ottawa and affiliated institutions, and released as a preprint in October 2025. The architecture follows a standard masked autoencoder (MAE) framework adapted for ultrasound, using a ViT backbone as the encoder and a lightweight transformer decoder for image reconstruction from masked patches. The model was pretrained on the OpenUS-46 dataset, which has about 370,000 2D and 3D ultrasound images from 46 open-source datasets. Evaluation was performed exclusively via fine-tuning of the pretrained encoder for three downstream classification datasets: BUS-BRA (breast cancer), MMOTU-2D (ovarian tumors), and GIST514-DB (gastrointestinal stromal tumors).

EchoCare: EchoCare [ZWZ⁺25] was introduced by researchers at the Centre for Artificial Intelligence and Robotics (CAIR), Hong Kong Institute of Science and Innovation, and Chinese Academy of Sciences, and released as a preprint in September 2025. The core framework extends the standard self-supervised MAE architecture with a novel dual-branch design: an image encoder and image decoder perform pixel-level masked reconstruction, while a second anatomy-classifier decoder jointly learns hierarchical anatomical relationships. Pretraining was conducted on EchoCareData, a public dataset of 4.5 million ultrasound images encompassing 9 major body regions and 52 anatomical organs. Evaluation was performed via full fine-tuning the backbone encoder for 10 tasks in a downstream suite spanning disease diagnosis classification, lesion segmentation, organ detection, fetal brain landmark prediction, cardiac ejection fraction regression, image quality enhancement, and report generation.

OpenUS: OpenUS [ZCR⁺25] was developed by Zheng et al. at the Queen Mary University of London, and released as a preprint in November 2025. The framework is an ultrasound foundation model built exclusively on publicly available data. OpenUS uses a Vision Mamba (VMamba) backbone, which captures both local and long-range spatial dependencies via selective state-space modeling. A key novelty is the self-adaptive masking strategy during

training, which combines a teacher attention map with a student reconstruction loss into a unified Adaptive Learning Priority (ALP) score, directing the masking toward the most clinically informative regions. Pretraining was conducted on 308,584 images from 38 publicly available ultrasound datasets covering 12 organs. Evaluation was performed via fine-tuning a task head (keeping the encoder frozen) for two downstream classification datasets (BUSI breast cancer and Fetal Planes) and two segmentation datasets (BUS-BRA breast lesions and TN3K thyroid nodules).

UltraFedFM: UltraFedFM [JFR⁺25] was introduced by Jiang et al., a collaboration spanning 17 institutions across 9 countries led by researchers at The Chinese University of Hong Kong, and first released as a preprint in November 2024. The main contribution is a federated pretraining framework, where 16 geographically distinct clients each pretrain a local model (masked autoencoder) on their private ultrasound data, and only model parameters are shared for aggregation. To handle cross-site image differences, the framework introduces structured masking with adaptive thresholding (SMAT), which adjusts masking patterns to organ- and lesion-specific texture features, and a multi-scale image corruption (MIC) branch for improving robustness to low-quality ultrasound inputs. The pretraining dataset has 1,015,754 unlabeled images covering 19 organs. Evaluation was performed via full fine-tuning of the encoder for disease diagnosis tasks across 8 kinds of organs and 6 modalities and lesion segmentation across 5 datasets. This paper also conducts a clinical comparison on 8 systemic malignant diseases, demonstrating that UltraFedFM outperforms mid-level sonographers (4–8 years of experience) and matches expert-level sonographers (>10 years of experience).

DeblurMIM: DeblurMIM [KGZ⁺] was proposed by Kang et al. from Sichuan University and the Shanghai Artificial Intelligence Laboratory, and published in Medical Image Analysis in July 2024. As one of the first applications of masked image modeling to ultrasound, the work is motivated by standard masked autoencoder pretraining failing to account for ultrasound’s high noise-to-signal ratio and the importance of fine-grained diagnostic details. DeblurMIM extends the MAE objective by adding a blurring step, where inputs are first

blurred using one of several blur operators before random masking. Then, the model is trained to reconstruct the original sharp image from the masked and blurred input. The encoder is a multi-scale hybrid convolution-transformer based on ConvMAE. Pretraining was conducted on approximately 280,000 thyroid ultrasound images. Downstream evaluation is performed with full fine-tuning on tasks spanning classification of thyroid nodules and Hashimoto’s thyroiditis, and segmentation of thyroid nodules.

Segmentation-focused FMs

SAMUS: SAMUS [LXY24] was proposed by Lin et al. from Huazhong University of Science and Technology and collaborating groups, and published at MICCAI 2024. The framework adapts SAM for ultrasound segmentation by introducing a parallel CNN branch alongside SAM’s ViT encoder, injecting local feature information via cross-branch attention, and further incorporating a position adapter and feature adapter to retune the model from large-scale natural image inputs (1024×1024) to more practical input sizes (256×256). The model was trained and evaluated on a comprehensive ultrasound dataset of approximately 30,000 images and 69,000 masks covering six anatomical object categories from publicly available sources. Downstream evaluation was conducted in a full fine-tuning paradigm. Tasks were focused exclusively on interactive and automatic segmentation across the six ultrasound object categories in the US30K dataset.

UltraSAM: UltraSam [MMZ⁺25] was developed by Meyer et al. from the University of Strasbourg / IHU-Strasbourg, first released as a preprint in November 2024 and published in IJCARS in 2025. The framework fully fine-tunes the Segment Anything Model on a large ultrasound dataset under a prompt-conditioned segmentation setup, supporting both point- and bounding-box prompts. Pretraining data consisted of US-43d, a compilation of 43 open-access ultrasound segmentation datasets totaling over 280,000 image-mask pairs across 20 diverse clinical applications. Evaluation employed two protocols: prompt-based zero-shot segmentation benchmarked on three diverse public datasets, and downstream fine-tuning

of an UltraSam-initialized Vision Transformer. Downstream tasks spanned both interactive segmentation of anatomical structures and image-level classification.

Universal Vision Models

DINOv3: DINOv3 [SVS⁺25] was developed by Siméoni et al. at Meta AI Research, and released as a technical report in August 2025. The framework extends DINOv2’s self-supervised student-teacher distillation approach to significantly larger model scales. A main contribution is Gram anchoring, a training technique that regularizes cosine similarity between student and teacher patch features using a Gram-matrix objective, reducing the degradation of dense feature maps that emerges in larger models trained for long schedules. The initial 7B model is distilled into a family of smaller variants (ViT-Small, ViT-Base, ViT-Large, and ConvNeXt-based architectures). Pretraining is conducted on a large curated mixture of web images combined with ImageNet-1k. DINOv3 is included in our evaluation as a strong general-purpose vision foundation model baseline, assessed via the standard probing and fine-tuning protocol used for the other models.

7.3 Investigating “Saturation” in Downstream Tasks

A benchmark task is saturated when best-reported performance approaches a practical ceiling, such that further improvements are either statistically marginal or clinically meaningless. In the context of ultrasound foundation model evaluation, saturation is a notable challenge. Tasks that any competent model can “solve” cannot distinguish a strong FM from a weak baseline, undermining the benchmark’s purpose.

First, we aim to characterize the extent of saturation across existing ultrasound benchmarks, by systematically reviewing reported performance for datasets across the six collections described in Section 7.2. Table 7.2 summarizes tasks for which performance at or above 0.90 (on the primary metric) has been reported by at least one benchmark. Based on previous

literature, we identify six classification tasks and seven segmentation tasks showing clear evidence of saturation, spanning a wide range of anatomical regions and task types.

Table 7.2: Ultrasound datasets from existing benchmarks with reported performance at or above 0.90 on the primary metric. Tasks approaching a performance ceiling provide limited discriminative signal for comparing foundation models. Performance values are the best reported across benchmarks; [†][MJL⁺25], [‡][TG], [§][JFR⁺25], [¶][KA24b].

Dataset	Domain	Task	Task Type	Best Reported	Metric
<i>Classification</i>					
UF1990	Uterine	Normal / fibroid classification	Binary	>0.90	Accuracy [†]
STMUS	Skeletal muscle	Normal / pathological classification	Binary	>0.90	Accuracy [†]
Fetal Planes	Fetal	Fetal plane classification (6-class)	Multiclass	>0.95	mAUROC [§]
POCUS	Lung	COVID-19 / pneumonia classification	Multiclass	>0.95	mAUROC [§]
GBUSV	Gallbladder	Gallbladder cancer classification	Binary	>0.95	AUROC [§]
Butterfly	Multi-organ	Organ classification (9-class)	Multiclass	>0.90	mAP [‡]
<i>Segmentation</i>					
AUL	Liver	Liver segmentation	Binary	>0.90	Dice [‡]
Luminous	Lumbar muscle	Lumbar multifidus muscle segmentation	Binary	>0.90	Dice [†]
Open Kidney	Kidney	Kidney capsule segmentation	Binary	>0.90	Dice [†]
CAMUS	Cardiac	LV cavity / myocardium / LA segmentation	Multiclass	>0.90	Dice [†]
BUS-BRA	Breast	Breast tumor segmentation	Binary	>0.90	Dice [†]
JNU-IFM/PSFHS	Pelvis / Fetal	Pubic symphysis & fetal head segmentation	Binary	0.974	Dice [†]
CT2US	Kidney	Kidney segmentation (synthetic)	Binary	0.981	Dice [¶]

We note some key patterns from this analysis. Many saturated classification tasks involve coarse-grained binary distinctions on small, low-diversity datasets (e.g., UF1990, STMUS), where even generic ImageNet-pretrained encoders achieve high accuracy [MJL⁺25]. Among segmentation tasks, simple binary structures on high-contrast images are most susceptible: the JNU-IFM pubic symphysis task reaches Dice = 0.974 and the synthetic CT2US kidney task reaches Dice = 0.981. Some tasks may inflate reported performance through dataset quality issues rather than genuine model capability. POCUS, for instance, consists of scraped and extracted frames from heterogeneous sources, and its reported mAUROC > 0.95 [JFR⁺25] could reflect low task difficulty and poor data curation, rather than model strength.

These findings directly inform dataset selection and experimental design in UltraBench 2.0. Tasks with unambiguous saturation across benchmarks are avoided where possible. However, there are cases that support the retention of a potentially saturated task, such as CAMUS cardiac segmentation or Butterfly organ classification, due to their clinical importance or

the task diversity they offer. In such instances, we recommend that full-dataset evaluation is augmented with a label efficiency experimental setting that constrains training data and restores discriminative difficulty. The task suite described next reflects these criteria directly.

7.4 Tasks

UltraBench 2.0 includes 10 semantic segmentation and 11 image classification tasks to assess model performance across a wide range of clinical applications, ultrasound probe types, and anatomical regions.

Cancer screening. Liver mass classification and segmentation (AUL) [XZL⁺23], ovarian tumor classification and segmentation (MMOTU-2D) [ZLB⁺23], thyroid nodule classification and segmentation (TN3K) [gon26], gallbladder cancer classification (GBCU) [BGR⁺22], breast cancer classification and segmentation (BUS-BRA) [GFGCCdAP24].

Prenatal care. Fetal planes and fetal brain planes classification (FETAL-PLANES) [BACGVA⁺20], fetal head and pubic symphysis segmentation (PSFHS) [CBO⁺24].

Anatomical region segmentation. Cardiac region segmentation (CAMUS) [LSP⁺19], musculoskeletal ultrasound segmentation (STMUS) [Mei21], kidney capsule and regions segmentation (OpenKidney) [SRH⁺22], liver segmentation (AUL) [XZL⁺23].

Disease Diagnosis. COVID-19 detection (COVID-BLUES) [WBR⁺25]. MASLD/NAFLD (MASLD) [BSS⁺18].

Non-clinical applications. Image quality assessment of cardiac ultrasound images (CAMUS) [LSP⁺19].

7.5 The UltraBench 2.0 Framework

7.5.1 Design Principles

UltraBench 2.0 is organized around four design principles:

- Provide a standard interface for using each dataset.
- Assess model performance with limited/reasonable training data.
- The benchmark is designed to be challenging to ensure room for improvement.
- Ease of evaluation.

7.5.2 Dataset Curation and Processing

We began by starting from the datasets and tasks included in UltraBench, and considered other datasets used in other benchmarks/collections to expand the coverage. When considering datasets for inclusion, we assessed the quality, availability, licensing, size, and difficulty of the task. After considering and filtering available datasets, UltraBench 2 includes 11 classification and 10 segmentation tasks across 12 datasets.

For the selected datasets, the images are divided into training and test splits using stratified splitting where applicable to ensure similar label distributions between the two splits. For datasets with multiple images per patient, the dataset is split by patient. As some datasets are published as RGB, RGBA, greyscale, or mixed images, we convert all images to RGB by either stacking the single greyscale channel or removing the alpha channel from RGBA images. Other than standardizing the number of channels, no further pre-processing (e.g., resizing) is applied to the training images to provide maximum freedom for new submissions to innovate on both data augmentation and preprocessing as well as model architecture and training algorithms. For the test set, two versions are published. One version with original-sized images (with the same minimal preprocessing as the training set) and another copy with

the images resized to 256×256 pixels using “letterbox” resizing (i.e., images are resized so that the longest edge measures 256 pixels and then padded to the final size). This ensures, particularly for semantic segmentation tasks, that performance comparisons between models are fair.

7.5.3 Data Setup

UltraBench 2 is built on top of the Hugging Face Datasets library to maximize the ease of use and interoperability with different deep learning frameworks. Each dataset is preprocessed using a similar pipeline and stored as a Hugging Face Dataset, either ahead of time and published to the Hugging Face Hub, if the license permits, or when the user first uses the dataset, for datasets where only the raw dataset may be redistributed, or for datasets that are gated behind research use agreements that the user must sign and download from the original source.

The licensing terms of each dataset dictate whether or not we can redistribute modified or unmodified versions of the data. The licenses divide the datasets into three groups: pre-processed, raw, and local. We host pre-processed datasets (e.g., AUL) and raw datasets (e.g., COVID-BLUES) on the Hugging Face Hub. For raw datasets, the preprocessing is applied at first runtime by the user. For local datasets (e.g., Open Kidney) the user must download the data from the original source and then the preprocessing is applied at first runtime. This tiered structure allows UltraBench 2.0 to include clinically important datasets with non-permissive licenses while maintaining a standardized access and loading experience for users.

7.5.4 Model Evaluation

Performance on classification and segmentation tasks is measured using the F1 and positive Dice scores, respectively. The scores for each task are aggregated into overall scores for each of the five clinical application areas, overall classification and segmentation scores, and overall

scores using a simple average of the individual task scores s_t .

To ensure ease-of-use and reproducibility we provide evaluation scripts as part of the benchmark. These scripts take in predictions in CSV format and generate the required metrics for the benchmark. Using CSV files as the framework-agnostic interface frees researchers to develop their model with their chosen tools and frameworks (e.g., PyTorch or TensorFlow). It also ensures that the metrics are implemented and applied consistently for all models.

7.5.5 Leaderboard

To track the progress of ultrasound foundation models, we provide a leaderboard for open source models and encourage users to evaluate their models on different tasks. Following the principles of benchmarks like GEO-Bench-2 [LLR⁺], submissions are restricted to reproducible open-source models to ensure that evaluations are performed in good faith, as the datasets are public and there are no hidden/private test sets that could be used to prevent bad actors from training on the test data. To ensure consistency and streamline submissions we provide framework-independent evaluation scripts that calculate the benchmark metrics based on prediction files (i.e., per-image labels for classification tasks and run-length encoded (RLE) strings for segmentation tasks), similar to Kaggle and medical image analysis challenges hosted on Grand Challenge.

7.6 Evaluations of Existing Models

To populate the benchmark we perform a retrospective analysis of existing ultrasound foundation models that have been published to date. We also provide implementations of these models and the classification and segmentation decoders to further improve the ease-of-use and ability for researchers to submit their results and perform comparisons.

7.6.1 Evaluation Setup

Each model described in Section 7.2.2 is evaluated in full fine-tuning and probing/head-only settings. In both cases, the models are trained for a maximum of 10,000 steps using a batch size of 32 and a Cosine annealing learning rate scheduler. They are evaluated every 100 steps on the validation set and early stopping is triggered if there is no improvement after 10 validation checks (1,000 training steps). For full fine-tuning, the backbone is frozen for the first 1000 steps and then is unfrozen and the learning rate is warmed up from 10% of the learning rate of the head to match that of the head by a factor of 1.001 every step. All reported results are the mean and standard deviation over three seeds.

7.6.2 Encoder Output Types

The foundation models evaluated in UltraBench 2.0 expose encoder features in three distinct formats, which determine how classification descriptors are formed and which segmentation decoder is used. In each case, we follow the feature extraction convention specified in the model’s original paper and codebase.

Isotropic ViT models. The models under this category include USFM, URFM and USF-MAE. They produce a flat sequence of N patch tokens. For classification, a global descriptor is obtained by mean-pooling across all token positions. This approach is modified for USF-MAE, as they make distinct use of the CLS token for classification. For segmentation, the token sequence is reshaped into a 2D spatial grid and passed to the decoder.

Hierarchical multi-scale model. EchoCare’s Swin Transformer encoder produces multi-scale spatial feature maps at several resolutions. For classification, the coarsest-scale map is spatially pooled to form a global descriptor. For segmentation, the full set of feature maps is passed directly to the SegFormer decoder, which projects and fuses them across scales.

Spatial map models. SAMUS and UltraSAM are SAM-based architectures whose encoders return a single-scale 2D spatial feature map rather than a token sequence. For classification, global average pooling over the spatial dimensions produces a fixed-size descriptor. For segmentation, the feature map is passed directly to the upsampler decoder without any reshaping.

7.6.3 Probing

Probing evaluates the transferability of frozen encoder representations by training lightweight, task-specific heads while keeping the backbone frozen. Features are extracted from the final encoder layer without gradient computation. All probes are optimized with AdamW; hyperparameters are summarized in Table 7.3.

Setting	Classification (Linear Probe)	Segmentation (Upsampler / SegFormer)
Optimizer	AdamW	AdamW
Learning rate	1×10^{-3}	1×10^{-3}
Weight decay	1×10^{-2}	1×10^{-2}
Batch size	32	32
Epochs	100	100
Loss function	Cross-entropy	Dice + CE ($\lambda=0.5$ each)
Image resolution	256×256	256×256

Table 7.3: Training hyperparameters shared across all probing experiments.

Linear probing for classification. For each classification task we attach a single linear layer $\mathbf{W} \in \mathbb{R}^{d \times C}$, where d is the encoder feature dimension and C is the number of output classes. The encoder feature output is extracted using a model-specific protocol. For example, a mean-pool operator is applied to the patch tokens for USFM while we use the CLS token for USF-MAE. Cross-entropy loss is used throughout.

Decoder probing for segmentation. Models are evaluated with a lightweight upsampler decoder. Isotropic ViT models produce patch token sequences $[B, N, D]$, which are spatially re-

organized into 2D grids $[B, D, \sqrt{N}, \sqrt{N}]$ before decoding; CNN/SAM-based models (SAMUS, UltraSAM) pass their spatial feature maps $[B, C, H, W]$ directly. The decoder applies: (1) a 1×1 convolution projecting to 256 channels, (2) four progressive transposed-convolution upsampling stages ($256 \rightarrow 128 \rightarrow 64 \rightarrow 32 \rightarrow 16$ channels, stride 2) each followed by a residual double-convolution block with GELU activations, and (3) a 1×1 classification convolution whose output is bilinearly interpolated to the target resolution. EchoCare, whose hierarchical Swin encoder produces multi-scale feature maps, is paired with a SegFormer-style MLP decoder [XWY⁺21]: per-scale features are linearly projected to 256 dimensions, bilinearly upsampled to the finest available resolution, concatenated, and fused via a 1×1 convolution with batch normalization and ReLU before a final bilinear upsampling to the target resolution. DiceCE loss (equal Dice and cross-entropy weights) is used throughout.

7.6.4 Fine-tuning

Full fine-tuning jointly optimizes the encoder backbone and a task-specific head, and is performed with data augmentation. We adopt a TrivialAugment-style strategy [TG] that randomly applies one of: identity, rotation (up to 30°), translation, random resized crop (scale 0.6–1.0), Gaussian noise, or brightness/contrast jitter. The two-phase backbone warmup protocol described in Section 7.6.1 applies to all fine-tuning experiments.

Classification. The same single linear layer used for linear probing is attached to the encoder and jointly fine-tuned with the backbone. Cross-entropy loss is used throughout.

Segmentation. The segmentation decoder is selected based on each model’s encoder output type. DiceCE loss (equal weights) is used for all segmentation fine-tuning experiments.

7.7 Results

We evaluate fifteen model configurations spanning ten foundation model families across all 21 benchmark tasks, in both head-only (probing) and full fine-tuning regimes. Complete per-task results are reported in Tables 7.4, 7.5, 7.6, and 7.7.

7.7.1 Overall Performance

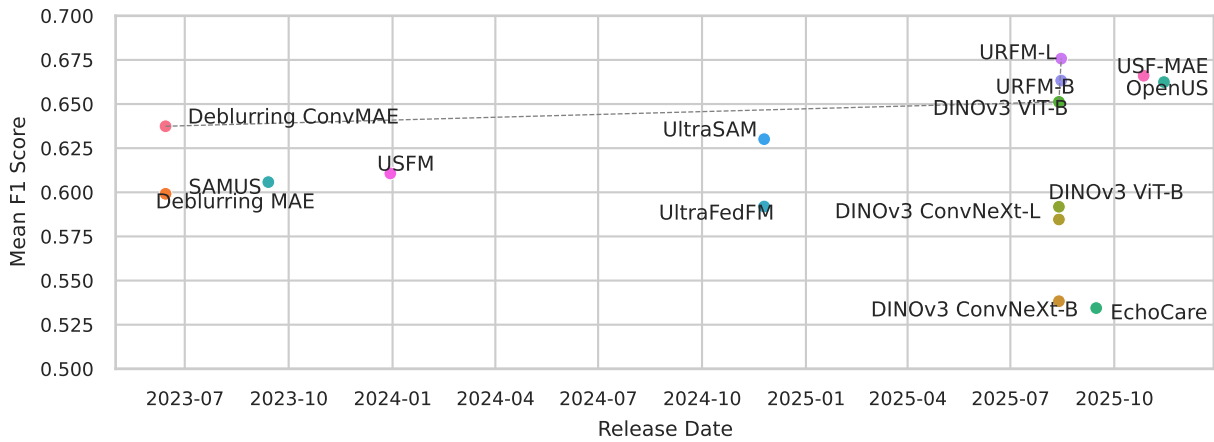


Figure 7.1: Mean F1 score over all classification tasks for each model. The dashed line shows the evolution of the best model over time.

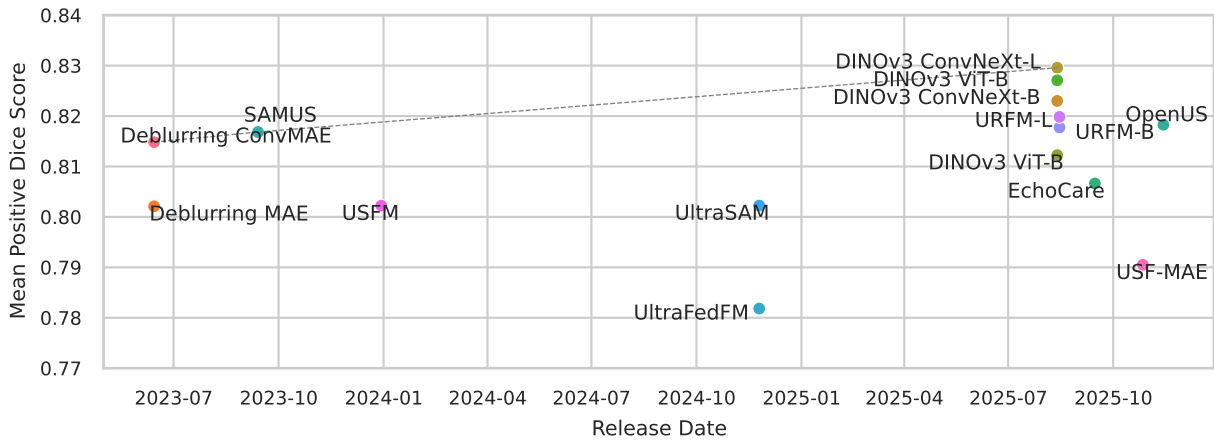


Figure 7.2: Mean positive Dice score over all segmentation tasks for each model. The dashed line shows the evolution of the best model over time.

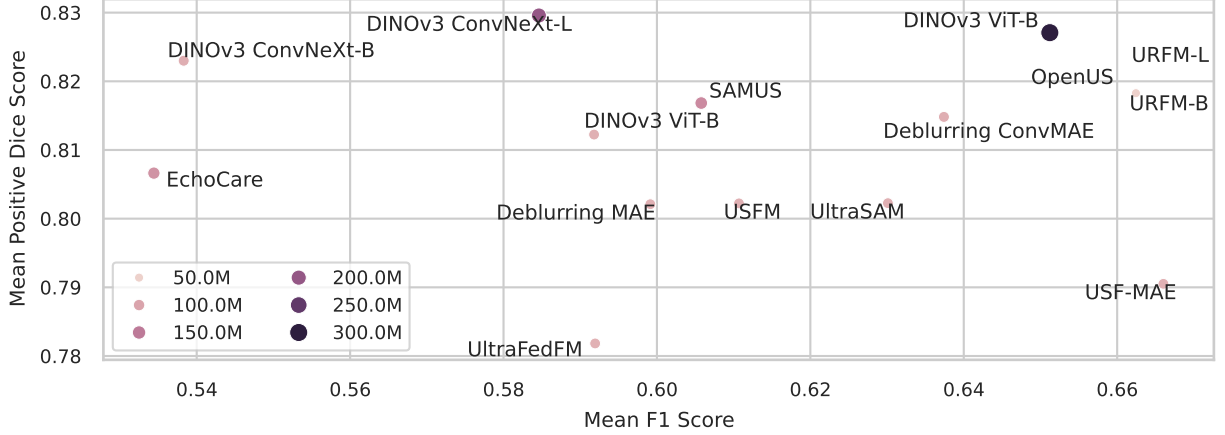


Figure 7.3: Overall classification and segmentation performance per model. The size of each point represents the number of trainable parameters.

Aggregating across the full task suite (Figures 7.1 and 7.2), no single model dominates the benchmark. Under full fine-tuning, URFM-L attains the highest mean classification F1 (0.644), followed closely by USF-MAE (0.628) and OpenUS (0.623); on segmentation, DINOv3 ViT-L leads with a mean positive Dice of 0.827, ahead of DINOv3 ConvNeXt-L (0.821) and URFM-L (0.815). The ordering changes substantially in the head-only regime, where OpenUS achieves the best mean classification F1 (0.601) and URFM-L the best mean segmentation Dice (0.777). Already we see an interesting observation, where a model’s ranking under one evaluation protocol is not necessarily a predictor of its ranking under the other, which is exactly what a standardized benchmark can expose.

Model scale is also a weak predictor of downstream quality (Figure 7.3). Within the DINOv3 family, the Large variants improve on their Base counterparts for segmentation, but the largest ultrasound-specific pretraining datasets do not translate into correspondingly large gains: EchoCare, pretrained on 4.5 million images, ranks last or near-last on several classification tasks, while OpenUS, pretrained on roughly 300,000 exclusively public images, is consistently competitive.

7.7.2 Classification Tasks

Linear Probing

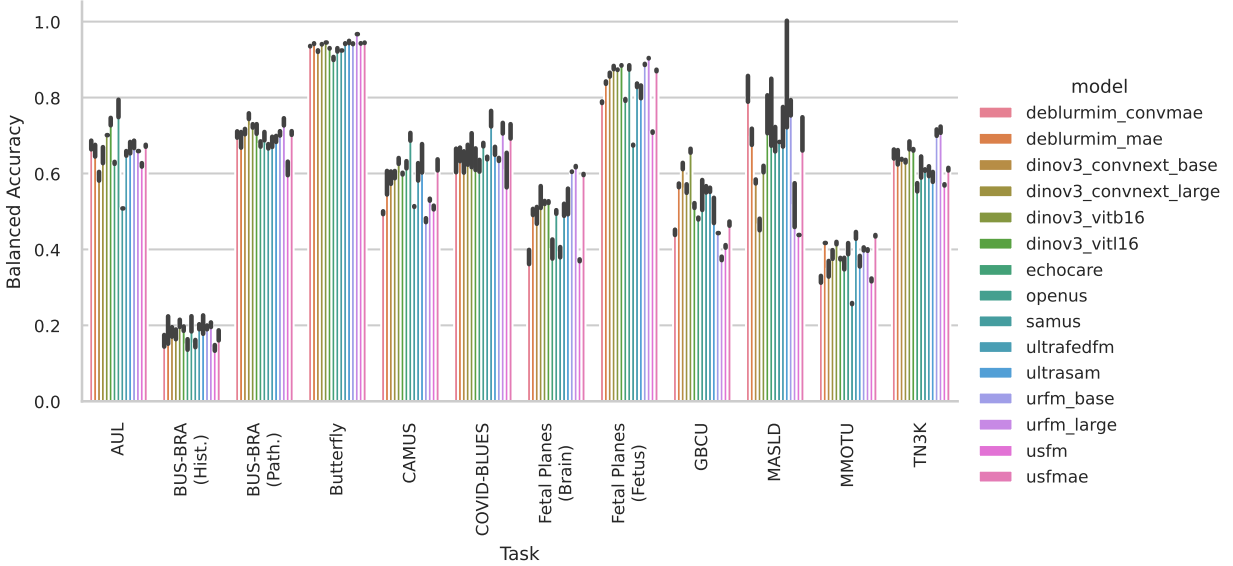


Figure 7.4: Per-task results for each classification task using linear probing.

Linear probing results are reported in Table 7.4 and Figure 7.4. OpenUS achieves the strongest mean performance (0.601), with DINOv3 ViT-L (0.596), DINOv3 ViT-B (0.590) and USF-MAE (0.590) closely behind, while USFM is the weakest model overall (0.468) and is the worst performer on five of the eleven tasks. URFM-L is the most task-dependent model in this regime: it leads on four tasks (BUS-BRA histology, BUS-BRA pathology, fetal brain, and fetal plane) yet collapses to 0.266 on GBCU, the lowest score any model records on that task.

The tasks separate cleanly by difficulty. BUS-BRA histology is the hardest task in the benchmark by a wide margin, with every model scoring between 0.123 and 0.193 macro F1, suggesting that eight-class histological subtyping is not recoverable from a frozen representation with a linear head. MASLD shows the widest inter-model spread of any classification task (0.389–0.858). In contrast, fetal plane classification is the least discriminative task, with most models clustered between 0.77 and 0.89.

Table 7.4: The mean and standard deviation of the macro F1 score over three seeds for each classification task after head-only training. The best result for each task is shown in bold.

	Pathological								Non-Pathological		
	AUL Mass	BUS-BRA Hist.	BUS-BRA Path.	COVID-BLUES	GBCU	MASLD	MMOTU	TN3K	CAMUS	Fetal Brain	Fetal Plane
Deblurring ConvMAE	0.596 $\pm$ 0.015	0.141 $\pm$ 0.020	0.690 $\pm$ 0.005	0.626 $\pm$ 0.038	0.401 $\pm$ 0.009	0.778 $\pm$ 0.008	0.323 $\pm$ 0.013	0.623 $\pm$ 0.006	0.503 $\pm$ 0.006	0.436 $\pm$ 0.021	0.733 $\pm$ 0.005
Deblurring MAE	0.629 $\pm$ 0.013	0.158 $\pm$ 0.013	0.693 $\pm$ 0.002	0.645 $\pm$ 0.023	0.549 $\pm$ 0.009	0.697 $\pm$ 0.031	0.423 $\pm$ 0.001	0.624 $\pm$ 0.023	0.591 $\pm$ 0.039	0.453 $\pm$ 0.027	0.772 $\pm$ 0.009
DINOv3 ConvNeXt-B	0.595 $\pm$ 0.010	0.140 $\pm$ 0.008	0.732 $\pm$ 0.010	0.619 $\pm$ 0.043	0.529 $\pm$ 0.016	0.581 $\pm$ 0.009	0.342 $\pm$ 0.028	0.607 $\pm$ 0.024	0.590 $\pm$ 0.017	0.502 $\pm$ 0.003	0.797 $\pm$ 0.006
DINOv3 ConvNeXt-L	0.647 $\pm$ 0.017	0.142 $\pm$ 0.005	0.741 $\pm$ 0.006	0.640 $\pm$ 0.038	0.559 $\pm$ 0.010	0.437 $\pm$ 0.004	0.389 $\pm$ 0.017	0.604 $\pm$ 0.014	0.600 $\pm$ 0.012	0.536 $\pm$ 0.001	0.830 $\pm$ 0.008
DINOv3 ViT-B	0.707 $\pm$ 0.000	0.182 $\pm$ 0.008	0.687 $\pm$ 0.005	0.651 $\pm$ 0.054	0.602 $\pm$ 0.005	0.620 $\pm$ 0.012	0.420 $\pm$ 0.003	0.657 $\pm$ 0.006	0.645 $\pm$ 0.008	0.525 $\pm$ 0.001	0.799 $\pm$ 0.012
DINOv3 ViT-L	0.719 $\pm$ 0.006	0.149 $\pm$ 0.011	0.725 $\pm$ 0.004	0.623 $\pm$ 0.041	0.582 $\pm$ 0.005	0.769 $\pm$ 0.045	0.364 $\pm$ 0.003	0.638 $\pm$ 0.005	0.602 $\pm$ 0.003	0.551 $\pm$ 0.001	0.836 $\pm$ 0.009
EchoCare	0.628 $\pm$ 0.007	0.129 $\pm$ 0.008	0.657 $\pm$ 0.033	0.618 $\pm$ 0.013	0.472 $\pm$ 0.028	0.762 $\pm$ 0.084	0.350 $\pm$ 0.021	0.576 $\pm$ 0.014	0.619 $\pm$ 0.004	0.415 $\pm$ 0.028	0.683 $\pm$ 0.021
OpenUS	0.750 $\pm$ 0.017	0.166 $\pm$ 0.016	0.706 $\pm$ 0.002	0.668 $\pm$ 0.008	0.582 $\pm$ 0.015	0.715 $\pm$ 0.041	0.400 $\pm$ 0.020	0.607 $\pm$ 0.018	0.706 $\pm$ 0.008	0.506 $\pm$ 0.006	0.807 $\pm$ 0.014
SAMUS	0.527 $\pm$ 0.000	0.125 $\pm$ 0.001	0.628 $\pm$ 0.001	0.640 $\pm$ 0.003	0.536 $\pm$ 0.005	0.715 $\pm$ 0.000	0.242 $\pm$ 0.001	0.595 $\pm$ 0.006	0.496 $\pm$ 0.001	0.382 $\pm$ 0.002	0.680 $\pm$ 0.002
UltraFedFM	0.659 $\pm$ 0.004	0.172 $\pm$ 0.012	0.687 $\pm$ 0.008	0.743 $\pm$ 0.023	0.543 $\pm$ 0.010	0.749 $\pm$ 0.067	0.452 $\pm$ 0.010	0.581 $\pm$ 0.009	0.621 $\pm$ 0.025	0.451 $\pm$ 0.006	0.772 $\pm$ 0.009
UltraSAM	0.656 $\pm$ 0.026	0.162 $\pm$ 0.012	0.681 $\pm$ 0.009	0.655 $\pm$ 0.014	0.516 $\pm$ 0.023	0.858 $\pm$ 0.109	0.363 $\pm$ 0.024	0.576 $\pm$ 0.007	0.652 $\pm$ 0.036	0.506 $\pm$ 0.008	0.771 $\pm$ 0.012
URFM-B	0.658 $\pm$ 0.007	0.182 $\pm$ 0.013	0.750 $\pm$ 0.002	0.625 $\pm$ 0.009	0.423 $\pm$ 0.008	0.664 $\pm$ 0.027	0.406 $\pm$ 0.004	0.669 $\pm$ 0.006	0.470 $\pm$ 0.009	0.588 $\pm$ 0.004	0.857 $\pm$ 0.001
URFM-L	0.610 $\pm$ 0.000	0.193 $\pm$ 0.002	0.769 $\pm$ 0.001	0.716 $\pm$ 0.017	0.266 $\pm$ 0.000	0.495 $\pm$ 0.086	0.390 $\pm$ 0.002	0.668 $\pm$ 0.004	0.540 $\pm$ 0.004	0.611 $\pm$ 0.002	0.885 $\pm$ 0.001
USFM	0.617 $\pm$ 0.006	0.123 $\pm$ 0.006	0.639 $\pm$ 0.002	0.604 $\pm$ 0.051	0.382 $\pm$ 0.000	0.389 $\pm$ 0.000	0.279 $\pm$ 0.006	0.568 $\pm$ 0.003	0.471 $\pm$ 0.005	0.345 $\pm$ 0.004	0.727 $\pm$ 0.001
USF-MAE	0.609 $\pm$ 0.004	0.168 $\pm$ 0.002	0.694 $\pm$ 0.001	0.709 $\pm$ 0.022	0.444 $\pm$ 0.000	0.712 $\pm$ 0.040	0.448 $\pm$ 0.002	0.620 $\pm$ 0.004	0.639 $\pm$ 0.014	0.597 $\pm$ 0.001	0.849 $\pm$ 0.000

Full Fine-tuning

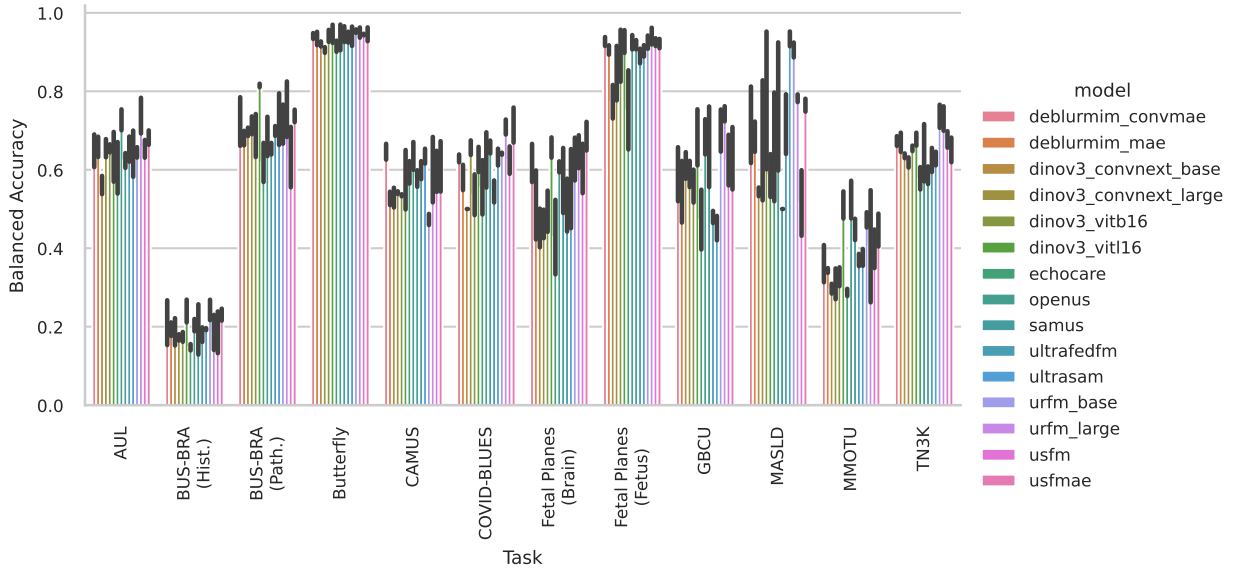


Figure 7.5: Per-task results for each classification task using full fine-tuning.

Full fine-tuning results are reported in Table 7.5 and Figure 7.5. URFM-L attains the highest mean F1 (0.644) and leads on four tasks (AUL Mass 0.720, BUS-BRA pathology 0.774, TN3K 0.695, and fetal plane 0.936). No other model leads more than two tasks, and six of the fifteen configurations top at least one task, indicating that classification performance on this benchmark is distributed rather than concentrated in a single model.

Unfreezing the backbone does not uniformly help. URFM-L gains substantially, moving from mid-tier under probing to the top of the classification ranking, and recovers on GBCU

from 0.266 to 0.706. BUS-BRA histology remains unsolved under fine-tuning (best 0.175, USF-MAE), confirming that its difficulty reflects the task rather than the evaluation protocol.

Table 7.5: The mean and standard deviation of the macro F1 score over three seeds for each classification task after full fine-tuning. The best result for each task is shown in bold.

	Pathological								Non-Pathological		
	AUL Mass	BUS-BRA Hist.	BUS-BRA Path.	COVID-BLUES	GBCU	MASLD	MMOTU	TN3K	CAMUS	Fetal Brain	Fetal Plane
Deblurring ConvMAE	0.660 $\pm$ 0.038	0.139 $\pm$ 0.008	0.672 $\pm$ 0.033	0.611 $\pm$ 0.021	0.651 $\pm$ 0.020	0.717 $\pm$ 0.093	0.339 $\pm$ 0.063	0.653 $\pm$ 0.015	0.663 $\pm$ 0.025	0.570 $\pm$ 0.069	0.729 $\pm$ 0.035
Deblurring MAE	0.638 $\pm$ 0.043	0.156 $\pm$ 0.017	0.676 $\pm$ 0.007	0.554 $\pm$ 0.051	0.569 $\pm$ 0.119	0.693 $\pm$ 0.037	0.325 $\pm$ 0.008	0.652 $\pm$ 0.011	0.524 $\pm$ 0.017	0.394 $\pm$ 0.010	0.760 $\pm$ 0.063
DINOv3 ConvNeXt-B	0.561 $\pm$ 0.009	0.142 $\pm$ 0.015	0.712 $\pm$ 0.005	0.323 $\pm$ 0.000	0.604 $\pm$ 0.053	0.544 $\pm$ 0.011	0.273 $\pm$ 0.025	0.606 $\pm$ 0.013	0.511 $\pm$ 0.046	0.464 $\pm$ 0.032	0.755 $\pm$ 0.049
DINOv3 ConvNeXt-L	0.647 $\pm$ 0.013	0.151 $\pm$ 0.007	0.724 $\pm$ 0.006	0.644 $\pm$ 0.020	0.591 $\pm$ 0.019	0.656 $\pm$ 0.159	0.283 $\pm$ 0.040	0.599 $\pm$ 0.052	0.523 $\pm$ 0.026	0.455 $\pm$ 0.017	0.771 $\pm$ 0.032
DINOv3 ViT-B	0.649 $\pm$ 0.020	0.149 $\pm$ 0.010	0.686 $\pm$ 0.013	0.479 $\pm$ 0.075	0.608 $\pm$ 0.040	0.783 $\pm$ 0.174	0.301 $\pm$ 0.035	0.674 $\pm$ 0.009	0.498 $\pm$ 0.004	0.415 $\pm$ 0.018	0.755 $\pm$ 0.073
DINOv3 ViT-L	0.618 $\pm$ 0.044	0.136 $\pm$ 0.003	0.732 $\pm$ 0.024	0.611 $\pm$ 0.044	0.745 $\pm$ 0.029	0.579 $\pm$ 0.077	0.513 $\pm$ 0.040	0.674 $\pm$ 0.027	0.555 $\pm$ 0.097	0.690 $\pm$ 0.023	0.812 $\pm$ 0.094
EchoCare	0.608 $\pm$ 0.055	0.126 $\pm$ 0.007	0.643 $\pm$ 0.011	0.471 $\pm$ 0.154	0.488 $\pm$ 0.141	0.643 $\pm$ 0.192	0.246 $\pm$ 0.027	0.570 $\pm$ 0.005	0.607 $\pm$ 0.030	0.370 $\pm$ 0.046	0.742 $\pm$ 0.037
OpenUS	0.709 $\pm$ 0.029	0.144 $\pm$ 0.014	0.682 $\pm$ 0.019	0.611 $\pm$ 0.087	0.717 $\pm$ 0.046	0.784 $\pm$ 0.156	0.529 $\pm$ 0.038	0.621 $\pm$ 0.005	0.643 $\pm$ 0.035	0.599 $\pm$ 0.034	0.817 $\pm$ 0.015
SAMUS	0.637 $\pm$ 0.012	0.140 $\pm$ 0.003	0.693 $\pm$ 0.051	0.657 $\pm$ 0.017	0.547 $\pm$ 0.024	0.421 $\pm$ 0.000	0.449 $\pm$ 0.034	0.584 $\pm$ 0.017	0.594 $\pm$ 0.019	0.455 $\pm$ 0.023	0.856 $\pm$ 0.017
UltraFedFM	0.653 $\pm$ 0.027	0.165 $\pm$ 0.009	0.696 $\pm$ 0.006	0.436 $\pm$ 0.069	0.556 $\pm$ 0.026	0.727 $\pm$ 0.072	0.368 $\pm$ 0.022	0.587 $\pm$ 0.013	0.614 $\pm$ 0.023	0.477 $\pm$ 0.041	0.776 $\pm$ 0.016
UltraSAM	0.616 $\pm$ 0.045	0.170 $\pm$ 0.012	0.681 $\pm$ 0.015	0.630 $\pm$ 0.025	0.657 $\pm$ 0.055	0.889 $\pm$ 0.029	0.372 $\pm$ 0.028	0.576 $\pm$ 0.018	0.650 $\pm$ 0.018	0.591 $\pm$ 0.091	0.773 $\pm$ 0.013
URFM-B	0.630 $\pm$ 0.030	0.144 $\pm$ 0.003	0.702 $\pm$ 0.004	0.633 $\pm$ 0.003	0.706 $\pm$ 0.007	0.870 $\pm$ 0.008	0.469 $\pm$ 0.027	0.679 $\pm$ 0.005	0.435 $\pm$ 0.015	0.670 $\pm$ 0.050	0.872 $\pm$ 0.035
URFM-L	0.720 $\pm$ 0.068	0.169 $\pm$ 0.025	0.774 $\pm$ 0.028	0.703 $\pm$ 0.024	0.706 $\pm$ 0.017	0.780 $\pm$ 0.008	0.380 $\pm$ 0.167	0.695 $\pm$ 0.007	0.599 $\pm$ 0.109	0.624 $\pm$ 0.075	0.936 $\pm$ 0.003
USFM	0.654 $\pm$ 0.021	0.158 $\pm$ 0.035	0.577 $\pm$ 0.074	0.623 $\pm$ 0.034	0.587 $\pm$ 0.011	0.497 $\pm$ 0.112	0.388 $\pm$ 0.061	0.599 $\pm$ 0.060	0.606 $\pm$ 0.046	0.380 $\pm$ 0.177	0.829 $\pm$ 0.018
USF-MAE	0.688 $\pm$ 0.011	0.175 $\pm$ 0.023	0.689 $\pm$ 0.012	0.710 $\pm$ 0.052	0.678 $\pm$ 0.015	0.762 $\pm$ 0.014	0.445 $\pm$ 0.052	0.625 $\pm$ 0.007	0.611 $\pm$ 0.076	0.636 $\pm$ 0.043	0.885 $\pm$ 0.017

7.7.3 Segmentation Tasks

Decoder Probing

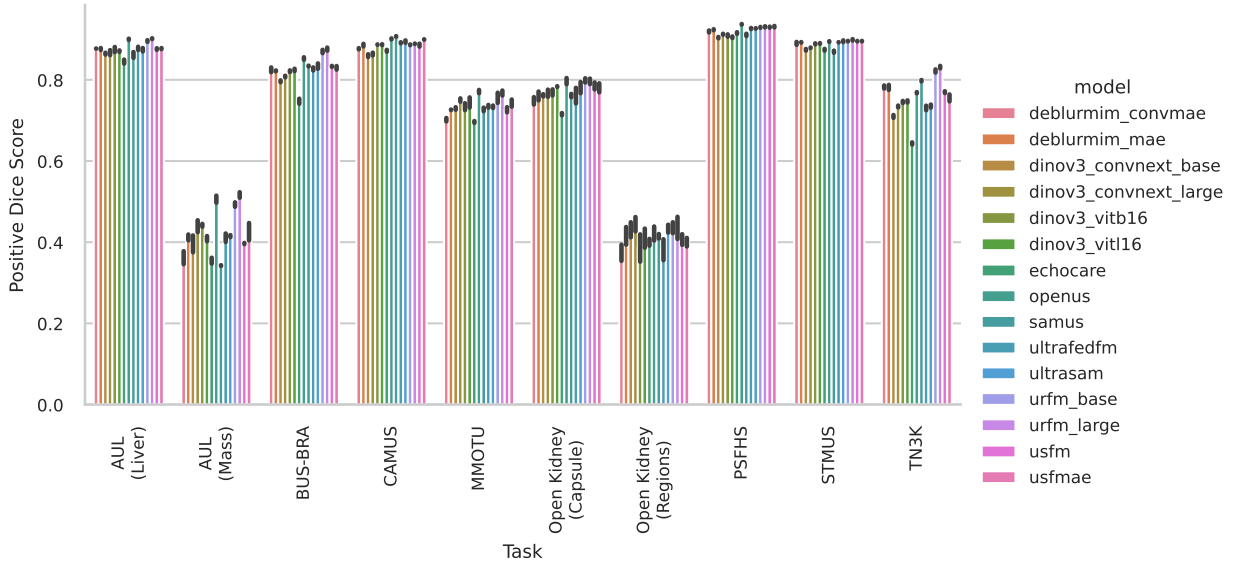


Figure 7.6: Per-task results for each segmentation task using decoder-only training.

Decoder probing results are reported in Table 7.6 and Figure 7.6. Ultrasound-specific pretraining is clearly advantageous in this regime: URFM-L leads with a mean Dice of 0.777,

followed by URFM-B (0.770) and OpenUS (0.767), and URFM-L is the best model on four of the ten tasks. The best DINOv3 variant reaches only 0.737, and a DINOv3 model is top on just one task (OK Regions). EchoCare is the weakest model overall (0.692) and is last on five of the ten tasks, despite having the largest pretraining corpus of any model evaluated.

Task difficulty spans a wide range. OK Regions is by far the hardest segmentation task, with a best Dice of 0.444, reflecting the difficulty of multi-region kidney parcellation. At the opposite end, PSFHS (0.900–0.938), CAMUS (0.841–0.905) and AUL Liver (0.841–0.901) are compressed near ceiling even with a frozen encoder, consistent with the saturation patterns identified in Table 7.2. AUL Mass and TN3K are the most discriminative tasks, with inter-model performance ranges of 0.173 and 0.219 respectively.

Table 7.6: The mean and standard deviation of the positive Dice score over three seeds for each segmentation task after head-only training. The best result for each task is shown in bold.

	Pathological				Anatomical					
	AUL Mass	BUS-BRA	MMOTU	TN3K	AUL Liver	CAMUS	OK Capsule	OK Regions	PSFHS	STMUS
Deblurring ConvMAE	0.369 $\pm$ 0.022	0.804 $\pm$ 0.007	0.702 $\pm$ 0.004	0.748 $\pm$ 0.003	0.874 $\pm$ 0.004	0.841 $\pm$ 0.006	0.748 $\pm$ 0.009	0.374 $\pm$ 0.020	0.920 $\pm$ 0.001	0.853 $\pm$ 0.009
Deblurring MAE	0.411 $\pm$ 0.006	0.801 $\pm$ 0.007	0.726 $\pm$ 0.001	0.736 $\pm$ 0.011	0.875 $\pm$ 0.002	0.859 $\pm$ 0.005	0.760 $\pm$ 0.011	0.416 $\pm$ 0.022	0.922 $\pm$ 0.003	0.864 $\pm$ 0.004
DINOv3 ConvNeXt-B	0.405 $\pm$ 0.020	0.786 $\pm$ 0.008	0.729 $\pm$ 0.003	0.661 $\pm$ 0.009	0.865 $\pm$ 0.002	0.847 $\pm$ 0.002	0.762 $\pm$ 0.003	0.431 $\pm$ 0.019	0.900 $\pm$ 0.006	0.848 $\pm$ 0.005
DINOv3 ConvNeXt-L	0.441 $\pm$ 0.017	0.805 $\pm$ 0.003	0.750 $\pm$ 0.004	0.683 $\pm$ 0.003	0.872 $\pm$ 0.004	0.848 $\pm$ 0.002	0.767 $\pm$ 0.009	0.444 $\pm$ 0.018	0.911 $\pm$ 0.001	0.848 $\pm$ 0.003
DINOv3 ViT-B	0.445 $\pm$ 0.007	0.801 $\pm$ 0.001	0.734 $\pm$ 0.008	0.683 $\pm$ 0.012	0.868 $\pm$ 0.002	0.868 $\pm$ 0.006	0.769 $\pm$ 0.008	0.386 $\pm$ 0.034	0.910 $\pm$ 0.002	0.848 $\pm$ 0.002
DINOv3 ViT-L	0.415 $\pm$ 0.018	0.789 $\pm$ 0.001	0.744 $\pm$ 0.012	0.693 $\pm$ 0.006	0.870 $\pm$ 0.007	0.867 $\pm$ 0.003	0.783 $\pm$ 0.001	0.410 $\pm$ 0.024	0.907 $\pm$ 0.001	0.847 $\pm$ 0.008
EchoCare	0.357 $\pm$ 0.005	0.708 $\pm$ 0.011	0.696 $\pm$ 0.002	0.583 $\pm$ 0.009	0.841 $\pm$ 0.004	0.857 $\pm$ 0.004	0.715 $\pm$ 0.003	0.400 $\pm$ 0.008	0.917 $\pm$ 0.003	0.842 $\pm$ 0.002
OpenUS	0.498 $\pm$ 0.008	0.835 $\pm$ 0.004	0.771 $\pm$ 0.004	0.740 $\pm$ 0.004	0.900 $\pm$ 0.001	0.898 $\pm$ 0.001	0.797 $\pm$ 0.008	0.421 $\pm$ 0.017	0.938 $\pm$ 0.001	0.874 $\pm$ 0.002
SAMUS	0.338 $\pm$ 0.008	0.826 $\pm$ 0.002	0.727 $\pm$ 0.004	0.791 $\pm$ 0.004	0.863 $\pm$ 0.002	0.905 $\pm$ 0.001	0.761 $\pm$ 0.004	0.415 $\pm$ 0.006	0.909 $\pm$ 0.000	0.839 $\pm$ 0.001
UltraFedFM	0.421 $\pm$ 0.016	0.804 $\pm$ 0.003	0.735 $\pm$ 0.003	0.694 $\pm$ 0.003	0.878 $\pm$ 0.004	0.882 $\pm$ 0.003	0.761 $\pm$ 0.019	0.381 $\pm$ 0.026	0.925 $\pm$ 0.002	0.859 $\pm$ 0.004
UltraSAM	0.404 $\pm$ 0.001	0.814 $\pm$ 0.004	0.734 $\pm$ 0.003	0.688 $\pm$ 0.004	0.874 $\pm$ 0.004	0.887 $\pm$ 0.002	0.781 $\pm$ 0.014	0.433 $\pm$ 0.010	0.927 $\pm$ 0.002	0.866 $\pm$ 0.002
URFM-B	0.486 $\pm$ 0.017	0.863 $\pm$ 0.004	0.756 $\pm$ 0.013	0.793 $\pm$ 0.002	0.896 $\pm$ 0.002	0.873 $\pm$ 0.004	0.799 $\pm$ 0.005	0.436 $\pm$ 0.014	0.927 $\pm$ 0.002	0.873 $\pm$ 0.004
URFM-L	0.511 $\pm$ 0.014	0.868 $\pm$ 0.002	0.767 $\pm$ 0.006	0.802 $\pm$ 0.008	0.901 $\pm$ 0.001	0.883 $\pm$ 0.003	0.796 $\pm$ 0.007	0.435 $\pm$ 0.027	0.929 $\pm$ 0.003	0.873 $\pm$ 0.002
USFM	0.396 $\pm$ 0.009	0.825 $\pm$ 0.005	0.726 $\pm$ 0.006	0.724 $\pm$ 0.010	0.877 $\pm$ 0.003	0.873 $\pm$ 0.003	0.785 $\pm$ 0.009	0.407 $\pm$ 0.013	0.929 $\pm$ 0.001	0.873 $\pm$ 0.001
USF-MAE	0.429 $\pm$ 0.008	0.806 $\pm$ 0.012	0.743 $\pm$ 0.009	0.713 $\pm$ 0.008	0.877 $\pm$ 0.002	0.892 $\pm$ 0.001	0.780 $\pm$ 0.011	0.400 $\pm$ 0.011	0.930 $\pm$ 0.002	0.868 $\pm$ 0.003

Full Fine-tuning

Full fine-tuning results are reported in Table 7.7 and Figure 7.7, and they invert the probing picture. DINOv3 ViT-L achieves the best mean Dice (0.827) and a DINOv3 variant is the top performer on seven of the ten segmentation tasks, including AUL Mass (0.678), MMOTU (0.853), AUL Liver (0.921), STMUS (0.889), OK Capsule (0.882, ConvNeXt-B) and OK Regions (0.518, ConvNeXt-L); on BUS-BRA it ties URFM-L at 0.895. The remaining tasks are led by ultrasound-specific models: URFM-L on TN3K (0.825), and SAMUS on CAMUS

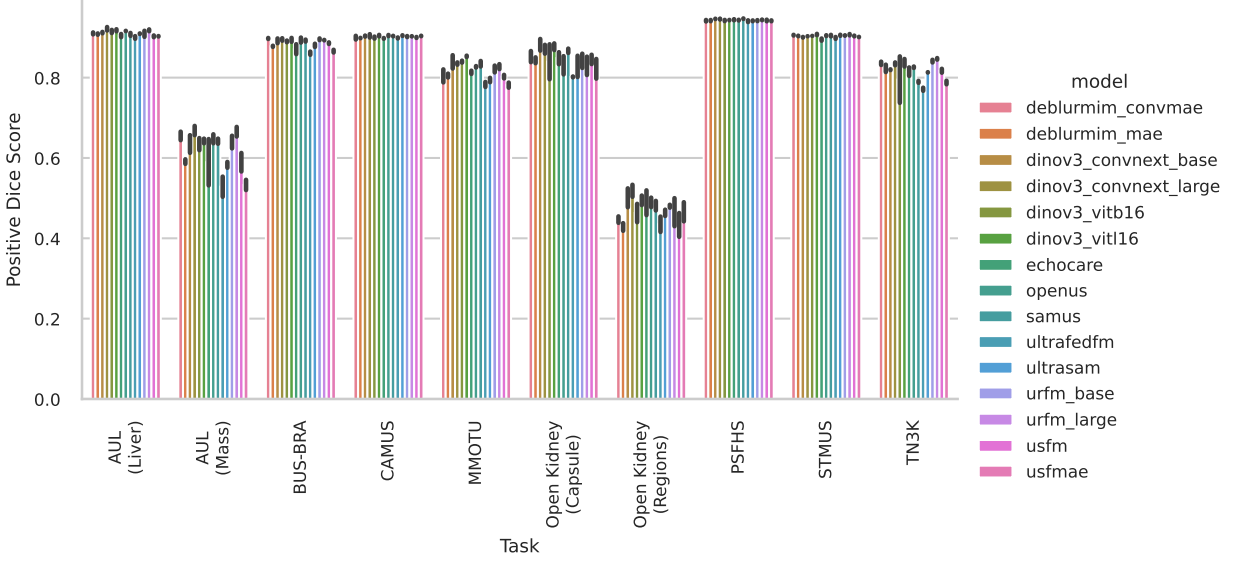


Figure 7.7: Per-task results for each segmentation task using full fine-tuning.

(0.903) and PSFHS (0.946).

Here we observe that a general-purpose vision model pretrained on natural images, given no ultrasound-specific pretraining at all, is the strongest segmentation backbone on this suite once the encoder is allowed to adapt, while the same model family is distinctly mid-tier when the encoder is frozen. The implication may be that much of what ultrasound-specific pretraining currently provides is a representation that transfers well without adaptation, rather than a better starting point for adaptation.

Fine-tuning also narrows the performance differences. Eight of the ten tasks have an inter-model spread below 0.09, and four (PSFHS 0.006, CAMUS 0.011, STMUS 0.018, AUL Liver 0.021) are effectively saturated, with every model within two points of the best. Only AUL Mass (0.165) and OK Regions (0.090) retain meaningful separation, which makes those two tasks the primary source of discriminative signal for segmentation in the current benchmark.

Table 7.7: The mean and standard deviation of the positive Dice score over three seeds for each segmentation task after full fine-tuning. The best result for each task is shown in bold.

	Pathological				Anatomical					
	AUL Mass	BUS-BRA	MMOTU	TN3K	AUL Liver	CAMUS	OK Capsule	OK Regions	PSFHS	STMUS
Deblurring ConvMAE	0.647 $\pm$ 0.003	0.885 $\pm$ 0.010	0.804 $\pm$ 0.016	0.820 $\pm$ 0.005	0.909 $\pm$ 0.004	0.896 $\pm$ 0.001	0.852 $\pm$ 0.014	0.447 $\pm$ 0.008	0.943 $\pm$ 0.001	0.886 $\pm$ 0.002
Deblurring MAE	0.582 $\pm$ 0.010	0.872 $\pm$ 0.002	0.805 $\pm$ 0.005	0.803 $\pm$ 0.007	0.901 $\pm$ 0.002	0.895 $\pm$ 0.001	0.843 $\pm$ 0.007	0.428 $\pm$ 0.009	0.943 $\pm$ 0.001	0.885 $\pm$ 0.001
DINOv3 ConvNeXt-B	0.625 $\pm$ 0.032	0.888 $\pm$ 0.004	0.839 $\pm$ 0.017	0.796 $\pm$ 0.011	0.905 $\pm$ 0.006	0.899 $\pm$ 0.001	0.882 $\pm$ 0.015	0.501 $\pm$ 0.024	0.945 $\pm$ 0.001	0.882 $\pm$ 0.004
DINOv3 ConvNeXt-L	0.637 $\pm$ 0.023	0.892 $\pm$ 0.001	0.835 $\pm$ 0.003	0.811 $\pm$ 0.011	0.919 $\pm$ 0.005	0.900 $\pm$ 0.002	0.871 $\pm$ 0.011	0.518 $\pm$ 0.015	0.944 $\pm$ 0.001	0.878 $\pm$ 0.009
DINOv3 ViT-B	0.513 $\pm$ 0.092	0.884 $\pm$ 0.002	0.840 $\pm$ 0.002	0.812 $\pm$ 0.008	0.910 $\pm$ 0.005	0.894 $\pm$ 0.001	0.840 $\pm$ 0.044	0.463 $\pm$ 0.023	0.942 $\pm$ 0.001	0.887 $\pm$ 0.002
DINOv3 ViT-L	0.678 $\pm$ 0.023	0.895 $\pm$ 0.002	0.853 $\pm$ 0.002	0.821 $\pm$ 0.003	0.921 $\pm$ 0.003	0.898 $\pm$ 0.004	0.877 $\pm$ 0.008	0.495 $\pm$ 0.012	0.943 $\pm$ 0.001	0.889 $\pm$ 0.003
EchoCare	0.586 $\pm$ 0.060	0.866 $\pm$ 0.004	0.813 $\pm$ 0.004	0.783 $\pm$ 0.003	0.903 $\pm$ 0.007	0.892 $\pm$ 0.001	0.848 $\pm$ 0.015	0.489 $\pm$ 0.031	0.943 $\pm$ 0.002	0.871 $\pm$ 0.004
OpenUS	0.652 $\pm$ 0.015	0.888 $\pm$ 0.003	0.827 $\pm$ 0.002	0.807 $\pm$ 0.011	0.914 $\pm$ 0.007	0.900 $\pm$ 0.003	0.831 $\pm$ 0.023	0.489 $\pm$ 0.012	0.943 $\pm$ 0.001	0.878 $\pm$ 0.008
SAMUS	0.636 $\pm$ 0.020	0.892 $\pm$ 0.003	0.835 $\pm$ 0.007	0.791 $\pm$ 0.002	0.911 $\pm$ 0.004	0.903 $\pm$ 0.002	0.867 $\pm$ 0.005	0.481 $\pm$ 0.012	0.946 $\pm$ 0.001	0.885 $\pm$ 0.005
UltraFedFM	0.532 $\pm$ 0.007	0.848 $\pm$ 0.003	0.783 $\pm$ 0.006	0.747 $\pm$ 0.006	0.900 $\pm$ 0.001	0.892 $\pm$ 0.003	0.802 $\pm$ 0.001	0.435 $\pm$ 0.019	0.940 $\pm$ 0.001	0.874 $\pm$ 0.004
UltraSAM	0.575 $\pm$ 0.006	0.875 $\pm$ 0.001	0.794 $\pm$ 0.005	0.790 $\pm$ 0.005	0.903 $\pm$ 0.004	0.902 $\pm$ 0.001	0.828 $\pm$ 0.027	0.463 $\pm$ 0.008	0.943 $\pm$ 0.001	0.881 $\pm$ 0.005
URFM-B	0.629 $\pm$ 0.019	0.892 $\pm$ 0.003	0.822 $\pm$ 0.009	0.824 $\pm$ 0.003	0.917 $\pm$ 0.002	0.896 $\pm$ 0.001	0.842 $\pm$ 0.018	0.480 $\pm$ 0.004	0.940 $\pm$ 0.000	0.886 $\pm$ 0.002
URFM-L	0.662 $\pm$ 0.011	0.895 $\pm$ 0.001	0.827 $\pm$ 0.006	0.825 $\pm$ 0.007	0.919 $\pm$ 0.002	0.898 $\pm$ 0.001	0.829 $\pm$ 0.023	0.465 $\pm$ 0.035	0.942 $\pm$ 0.001	0.885 $\pm$ 0.003
USFM	0.582 $\pm$ 0.006	0.876 $\pm$ 0.007	0.802 $\pm$ 0.004	0.805 $\pm$ 0.005	0.905 $\pm$ 0.001	0.897 $\pm$ 0.002	0.845 $\pm$ 0.012	0.433 $\pm$ 0.030	0.943 $\pm$ 0.001	0.886 $\pm$ 0.004
USF-MAE	0.518 $\pm$ 0.007	0.850 $\pm$ 0.000	0.781 $\pm$ 0.006	0.757 $\pm$ 0.003	0.906 $\pm$ 0.003	0.901 $\pm$ 0.001	0.822 $\pm$ 0.025	0.466 $\pm$ 0.024	0.941 $\pm$ 0.001	0.879 $\pm$ 0.002

7.8 Future Directions

7.8.1 Few-Shot Scaling

Few-shot scaling experiments measure the label efficiency of each foundation model by varying the fraction of training data available during probe training. We define n -shot as n training samples per class and sweep a range of shot fractions from a small percentage up to the full training set.

This evaluation directly extends the saturation analysis in Section 7.3. For tasks where full-data performance is near ceiling across all models, the few-shot regime restores discriminative difficulty: models with stronger representations require substantially less labeled data to reach plateau performance, and inter-model gaps are widest at the lowest data regimes. Conversely, tasks that remain undifferentiated even at minimal supervision provide another lens into performance saturation, and could also be candidates for replacement in future benchmark iterations.

7.8.2 Expanding Task Coverage

UltraBench 2.0 primarily covers 2D frame classification and segmentation only. Other clinically significant ultrasound tasks, such as object detection via bounding boxes, landmark

localization, cardiac metric regression and video-based temporal analysis, are not represented. As we expand the capacity of this benchmark, we will look to close this gap by adding support for tasks that assess spatial localization beyond segmentation masks and any form of temporal reasoning.

7.9 Discussion

Saturation in practice. The results provide an empirical characterization of task difficulty and inter-model discriminability within UltraBench 2.0, largely confirming the saturation predictions from Section 7.3. PSFHS, CAMUS, AUL Liver and STMUS segmentation show minimal inter-model spread, and under full fine-tuning all four fall within roughly two points across every model evaluated, offering little signal for ranking foundation models. These patterns validate the core concern motivating the few-shot scaling experiments: for tasks where all competent models converge near ceiling performance, restricting training data may restore the discriminative difficulty needed to meaningfully separate them. In contrast, BUS-BRA histology classification (best F1: 0.193), OK Regions segmentation (best Dice: 0.518), MMOTU classification (best F1: 0.529) and AUL Mass segmentation identify themselves as the key tasks in the current benchmark for distinguishing foundation model quality. A notable finding is the severity of USFM’s underperformance under probing (weakest mean classification F1 at 0.468, and lowest score on five of eleven tasks), suggesting that pretraining on a large dataset does not automatically yield strong performance and raises another research question of understanding how pretraining data diversity affects learned representations. EchoCare reinforces the same point: the largest pretraining corpus in the comparison, at 4.5 million images, yields the weakest segmentation probing performance of any model.

Domain-specific versus general-purpose pretraining. The clearest cross-cutting result is that the value of ultrasound-specific pretraining depends entirely on the evaluation protocol. With a frozen encoder, ultrasound-pretrained models dominate: URFM-L, URFM-B and

OpenUS occupy the top three segmentation positions, and the best DINOv3 variant trails them. Once the encoder is unfrozen, that ordering inverts and DINOv3 becomes the strongest segmentation backbone on seven of ten tasks. There are two interpretations here. Either ultrasound pretraining is mostly creating a good frozen representation rather than a better optimization starting point, or the fine-tuning protocol advantages the more heavily-regularized general-purpose models. Distinguishing these interpretations is a future goal, but it is a question that would not have come to light without a good benchmark to support experimentation.

Limitations. A direct consequence of architectural heterogeneity across the evaluated models is that the probing heads are not homogenized in terms of parameters. For classification, the linear probe maps from the encoder’s embedding dimension to the number of output classes. Since embedding dimensions vary across models, the parameter count of the classification head differs per model, and also between tasks. For segmentation, we mitigate this where possible by using a shared upsampler decoder, but nuances remain. First, EchoCare’s hierarchical Swin backbone is designed to produce multi-scale feature maps, and collapsing it to a single-scale output would disregard the intent of the model. Thus, we pair it with a SegFormer decoder that accepts all the feature scale outputs, which ends up having more trainable parameters than the upsampler. Second, even within the upsampler, the input dimensionality differs between isotropic ViT models, which produce patch token sequences that are spatially rearranged before decoding, and spatial map models (SAMUS, UltraSAM), which pass feature maps directly, yielding different channel depths at the decoder input. As a result, probing comparisons across models may partially conflate representation quality with probe head capacity.

The leaderboard is a second limitation. Because all datasets are publicly accessible and there are no private held-out test sets, the benchmark cannot technically prevent training on test data, and instead relies on the requirement that submissions be reproducible open-source

models.

Takeaways. UltraBench 2.0 establishes a foundation for systematic, reproducible comparisons of ultrasound foundation models. This resource comes at a moment when the field is producing new architectures faster than evaluation methodology can keep pace. Our results already reveal meaningful insights: not all models generalize uniformly across task types, pretraining corpus scale does not necessarily correlate to improved downstream performance, model rankings are not stable across adaptation regimes, and a general-purpose vision model can outperform every ultrasound-specific model on segmentation once fine-tuned. The difficulty gradient across tasks remains wide enough to support meaningful ranking for most models today, though the compression observed under full fine-tuning suggests this will not hold indefinitely without the few-shot protocol described in Section 7.8.1. Lastly, as the leaderboard grows with community submissions, UltraBench 2.0 will evolve from a static readout of current performance standards into a real-time tracker of whether the field is making genuine progress in understanding clinical ultrasound.

CHAPTER 8

Conclusions

The central question motivating this dissertation was whether deep learning could make a difference in how clinical ultrasound is used for diagnostic tasks, and if so, what the right way to build and evaluate such systems looks like. We arrive at an answer through a progression of works that reflect the technical advancements in AI methodology happening in parallel. Task-specific supervised models demonstrate attention-worthy performance, with limits that turn out to be informative. These limits, regarding external generalization and annotation cost, are not just coincidental, but a direct result of focusing representation learning on labeled data. This observation is what motivates the shift to self-supervised pretraining pursued in the second half of this dissertation.

8.1 Summary

The dissertation traces a progression from task-specific deep learning for thyroid nodule diagnostics to a broader perspective on how to build and evaluate foundation models for ultrasound imaging more generally. Part 1 (Chapters 3–5) established what is possible with supervised, task-specific approaches and also where those approaches break down. Part 2 (Chapters 6–7) was motivated directly by those limitations, asking whether the pretraining objective and the representation learning setup could avoid tunnel visioning on better architectures within the supervised paradigm.

Chapter 3 provides a systematic review of deep learning methods for thyroid nodule

ultrasound, reporting architectures, datasets, clinical tasks, and reported performance claims across the literature. The most consistent finding was that most models were trained and evaluated on single retrospective datasets, and prospective use of a model in a new environment required a non-trivial amount of re-validation to ensure expected performance.

Chapter 4 introduces a joint detection and segmentation framework that couples image-level nodule presence detection with boundary delineation in a single model. This setup better reflects how a radiologist simultaneously decides whether a nodule is present and where it is. Adding the anomaly detection module proved to be a helpful addition, leading to improved image-wide detection across all tested architectures. Despite the correlated drop in segmentation quality of small nodules, we were able to establish that these are typically clinically irrelevant. Lastly, external validation confirmed a meaningful generalization gap, illustrating that task-specific architectural improvements alone are insufficient to address data distribution shifts.

Chapter 5 introduces ThyGraph, a patient-level graph construction approach for thyroid nodule malignancy classification. This method treats the full ultrasound study as a graph, with edges weighted by anatomical proximity derived from OCR-extracted spatial annotations. The core finding was that encoding inter-image relationships within graph structure consistently outperformed approaches that treat each image independently under the MIL framework. We also found that patch-level features carried more discriminative information for malignancy than frame-level embeddings. Lastly, the self-attention graph pooling mechanism enabled clinically interpretable visualizations, with the model correctly focusing on the anatomical lobe containing the biopsied nodule in examined cases.

Chapter 6 proposes US-JEPA, a JEPA-based ultrasound foundation model pretrained on the largest publicly available multi-anatomy ultrasound corpus assembled to date. The model achieved state-of-the-art or competitive performance across the majority of classification tasks evaluated. Controlled comparisons with matched pretraining data confirmed that the improvements over standard JEPA and MAE baselines hold independently of data scale,

attributing observed gains to the SALT pretraining objective and the USrc region conditioning. In low-label regimes, US-JEPA substantially outperformed existing ultrasound foundation models, requiring fewer downstream labels to achieve strong performance. Robustness to image corruptions, including blur, contrast depletion, and speckle noise, was consistently stronger than reconstruction-based baselines, providing empirical support to the claim that latent prediction avoids allocating capacity to uninformative pixel-level variance.

Chapter 7 addresses a structural gap in the ultrasound foundation model literature: the absence of a standardized evaluation framework that would allow meaningful comparison across models. Prior work compared against limited or non-standardized baselines, making it impossible to objectively assess progress. UltraBench 2.0 provides a consolidated set of downstream tasks and a tool for the community to use to unify evaluation of existing foundation models. Without a standardized evaluation substrate, the field cannot distinguish genuine progress from favorable benchmark selection, and this benchmark is designed to provide a reproducible reference point for future work.

8.2 Future Work

The first, and most intuitive extension is to bring the findings of Part 2 full-circle to address the limitations observed in Part 1. The task-specific models in Chapters 4 and 5 relied, respectively, on task-specific annotated datasets and ImageNet-pretrained CNN backbones as feature extractors. US-JEPA representations were shown to be substantially informative and transferable; thus, using them for training initialization or as fixed feature extractors would be a natural next step. This comparison would also provide a concrete downstream evaluation of US-JEPA on the initial clinical tasks and datasets that motivated this dissertation, closing the loop on some unresolved questions.

A second direction concerns the anatomical density problem identified in Chapter 6. US-JEPA’s pretraining corpus, despite being the largest publicly available ultrasound dataset

assembled to date, is not uniformly distributed across anatomies. Distribution shift at the organ- or pathology-level in the pretraining data can correlate with downstream task performance. The UCLA thyroid ultrasound dataset curated for Chapters 4 and 5 is already a starting point for 1) a larger thyroid-specific pretraining approach or 2) fine-tuning of general-purpose models like US-JEPA for thyroid diagnostics.

A third direction relates to model portability. Ultrasound’s primary clinical advantage is in its accessibility: it is bedside, point-of-care, and deployed widely in resource-limited settings. Large ViT-B and ViT-L models may not integrate with on-device hardware restrictions or be practical in these environments. Knowledge distillation approaches from US-JEPA into smaller encoders, or SSL pretraining of lightweight architectures, could provide high-quality ultrasound representations for constrained settings. The impact would be invaluable in low-resource environments where ultrasound is often the only available imaging modality and where standardized AI-assisted diagnostics could have the most transformative impact.

And lastly, UltraBench 2.0 provides coverage for classification and segmentation, two fundamental tasks in ultrasound AI. However, the task space for ultrasound, especially with regards to clinically meaningful predictions, is much broader. Expanding the benchmark to include video-based and 3D volume tasks, regression-focused endpoints, and clinically curated scenarios would provide a more complete picture of what current models can and cannot do. More importantly, incorporating real-world distribution shift evaluation across scanner manufacturers, geographic regions, and patient populations would move beyond synthetic corruptions toward practically meaningful robustness measurement.

As we rapidly make concurrent advancements in deep learning research and applied medical AI, it becomes progressively more important that we build the tools necessary to understand how and why we are observing improvement. Research in reproducibility, interpretability and benchmarking will continue to be driving forces behind the discoveries in this field. This dissertation lays the groundwork for how such methods may be used to improve AI-based diagnostics in ultrasound.

Bibliography

- [AAMMAA⁺23] Ali Abbasian Ardakani, Afshin Mohammadi, Mohammad Mirza-Aghazadeh-Attari, Fariborz Faeghi, Thomas J. Vogl, and U. Rajendra Acharya. Diagnosis of metastatic lymph nodes in patients with papillary thyroid cancer: A comparative multi-center study of semantic features and deep learning-based models. *Journal of Ultrasound in Medicine*, 42(6):1211–1221, 2023.
- [AAMMAAA23] Ali Abbasian Ardakani, Afshin Mohammadi, Mohammad Mirza-Aghazadeh-Attari, and U Rajendra Acharya. An open-access breast lesion ultrasound image database: Applicable in artificial intelligence studies. *Computers in Biology and Medicine*, 152:106438, January 2023.
- [AB18] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, November 2018. arXiv:1610.01644 [stat].
- [ABCP⁺23] Syed Z. Ali, Zubair W. Baloch, Beatrix Cochand-Priollet, Fernando C. Schmitt, Philippe Vielh, and Paul A. VanderLaan. The 2023 Bethesda System for Reporting Thyroid Cytopathology. *Thyroid*, 33(9):1039–1044, September 2023.
- [AC22] Erik K. Alexander and Edmund S. Cibas. Diagnosis of thyroid nodules. *The Lancet Diabetes & Endocrinology*, 10(7):533–539, 2022.
- [ACB⁺19] Zeynettin Akkus, Jason Cai, Arunnit Boonrod, Atefeh Zeinoddini, Alexander D. Weston, Kenneth A. Philbrick, and Bradley J. Erickson. A Survey of

- Deep-Learning Applications in Ultrasound: Artificial Intelligence–Powered Ultrasound for Improving Clinical Workflow. *Journal of the American College of Radiology*, 16(9, Part B):1318–1328, September 2019.
- [ADGKF19] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 28:104863, November 2019.
- [ADM⁺23] Mahmoud Assran, Quentin Duval, Ishan Misra, et al. Self-supervised learning from images with a joint-embedding predictive architecture. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023.
- [Aga22] Agata Momot. Common Carotid Artery Ultrasound Images, November 2022.
- [AIB⁺25] Hugues Van Assel, Mark Ibrahim, Tommaso Biancalani, Aviv Regev, and Randall Balestrieri. Joint Embedding vs Reconstruction: Provable Benefits of Latent Space Prediction for Self Supervised Learning, October 2025. arXiv:2505.12477 [cs.LG].
- [AML⁺25] Jakob Ambsdorf, Asbjørn Munk, Sebastian Llambias, Anders Nymark Christensen, Kamil Mikolaj, Randall Balestrieri, Martin Tolsgaard, Aasa Feragen, and Mads Nielsen. General Methods Make Great Domain-specific Foundation Models: A Case-study on Fetal Ultrasound, June 2025. arXiv:2506.19552 [cs].
- [AvdL] Habib Ahmad and Aad van der Lugt. The Radiology Assistant : TI-RADS - Thyroid Imaging Reporting and Data System.
- [AYK⁺21] Deepak Anand, Kumar Yashashwi, Neeraj Kumar, Swapnil Rane, Peter H. Gann, and Ameya Sethi. Weakly supervised learning on unannotated H&E-

- stained slides predicts BRAF mutation in thyroid cancer with high accuracy. *Journal of Pathology*, 255(3):232–242, 2021.
- [B⁺21] Rishi Bommasani et al. On the opportunities and risks of foundation models, 2021.
- [BACGVA⁺20] Xavier P. Burgos-Artizzu, David Coronado-Gutiérrez, Brenda Valenzuela-Alcaraz, Elisenda Bonet-Carne, Elisenda Eixarch, Fatima Crispi, and Eduard Gratacós. Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes. *Scientific Reports*, 10(1):10200, June 2020.
- [Bao19] Han Bao. Investigations of the influences of a CNN’s receptive field on segmentation of subnuclei of bilateral amygdalae, 2019.
- [BCY⁺20] Zhantao Bai, Lei Chang, Ruiguo Yu, Xiaoli Li, Xun Wei, Ming Yu, Zhikang Liu, Jie Gao, Jiacheng Zhu, Yating Zhang, Shuo Wang, and Zhizhong Zhang. Thyroid nodules risk stratification through deep learning based on ultrasound images. *Medical Physics*, 47(12):6355–6365, 2020.
- [BDPW22] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022.
- [BGP⁺24] Adrien Bardes, Quentin Garrido, Jean Ponce, et al. V-JEPA: Latent video prediction for visual representation learning, 2024.
- [BGR⁺22] Soumen Basu, Mayank Gupta, Pratyaksha Rana, Pankaj Gupta, and Chetan Arora. Surpassing the human accuracy: Detecting gallbladder cancer from usg images with curriculum learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20854–20864. IEEE, 6 2022.

- [BHL⁺23] Shruthi Bannur, Stephanie Hyland, Qianchu Liu, et al. Learning to exploit temporal structure for biomedical vision-language processing, 2023.
- [BKK⁺20] Sherry Bhalla, Harpreet Kaur, Rupinder Kaur, Saloni Sharma, and G. P. S. Raghava. Expression based biomarkers and models to classify early and late-stage samples of papillary thyroid carcinoma. *PLOS ONE*, 15(4):e0231629, 2020.
- [BL24] Randall Balestriero and Yann LeCun. Learning by Reconstruction Produces Uninformative Features For Perception, February 2024. arXiv:2402.11337 [cs.CV].
- [BMR⁺20] Tom B. Brown, Benjamin Mann, Nick Ryder, et al. Language models are few-shot learners, 2020.
- [BPA⁺21] Fabiano Bini, Andrea Pica, Luca Azzimonti, Alessandro Giusti, Lorenzo Ruinelli, Franco Marinozzi, and Pierpaolo Trimboli. Artificial intelligence in thyroid field—A comprehensive review. *Cancers*, 13(19):4740, 2021.
- [BSBW⁺20] Marcin Barczyński, Małgorzata Stopa-Barczyńska, Bartłomiej Wojtczak, Agnieszka Czarniecka, and Aleksander Konturek. Clinical validation of S-DetectTM mode in semi-automated ultrasound classification of thyroid lesions in surgical office. *Gland Surgery*, 9(S2):S77–S85, 2020.
- [BSM⁺23] Zachary M. C. Baum, Shaheer U. Saeed, Zhe Min, Yipeng Hu, and Dean C. Barratt. MR to Ultrasound Registration for Prostate Challenge - Dataset, June 2023.
- [BSS⁺18] Michał Byra, Grzegorz Styczynski, Cezary Szmigielski, Piotr Kalinowski, Łukasz Michałowski, Rafał Paluszkiewicz, Bogna Ziarkiewicz-Wróblewska, Krzysztof Zieniewicz, Piotr Sobieraj, and Andrzej Nowicki. Transfer learning

- with deep convolutional neural network for liver steatosis assessment in ultrasound images. *International Journal of Computer Assisted Radiology and Surgery*, 13(12):1895–1903, 12 2018.
- [BSS⁺25] Mahdi-Reza Borna, Hanan Saadat, Mohammad Mehdi Sepehri, Hossein Torkashvand, Leila Torkashvand, and Shamim Pilehvari. AI-powered diagnosis of ovarian conditions: insights from a newly introduced ultrasound dataset. *Frontiers in Physiology*, 16, July 2025.
- [BTS⁺21] Moritz Böhlend, Lars Tharun, Tim Scherr, Ralf Mikut, Veit Hagenmeyer, Lester D. R. Thompson, Sven Perner, and Markus Reischl. Machine learning methods for automated classification of tumors with papillary thyroid carcinoma-like nuclei: A quantitative analysis. *PLOS ONE*, 16(9):e0257635, 2021.
- [But18] Butterfly Network. Butterflynetwork/mitgrandhack2018. <https://github.com/ButterflyNetwork/MITGrandHack2018>, 2018. GitHub repository.
- [BWC⁺21] Jannis Born, Nina Wiedemann, Manuel Cossio, Charlotte Buhre, Gabriel Brändle, Konstantin Leidermann, Julie Goulet, Avinash Aujayeb, Michael Moor, Bastian Rieck, and Karsten Borgwardt. Accelerating detection of lung pathologies with explainable ultrasound image analysis. *Applied Sciences*, 11(2):672, 1 2021.
- [C⁺24] Richard J. Chen et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 2024.
- [CA17] Edmund S. Cibas and Syed Z. Ali. The 2017 bethesda system for reporting thyroid cytopathology. *Thyroid*, 27(11):1341–1346, 2017.
- [ÇAL⁺16] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: Learning dense volumetric segmentation from

- sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, pages 424–432. Springer, 2016.
- [CBO⁺24] Gaowen Chen, Jieyun Bai, Zhanhong Ou, Yaosheng Lu, and Huijin Wang. PSFHS: Intrapartum ultrasound image dataset for AI-based segmentation of pubic symphysis and fetal head. *Scientific Data*, 11(1):436, May 2024.
- [CBP⁺17] Young Jun Choi, Jung Hwan Baek, Hyo Sung Park, Won Hyuk Shim, Tae Yong Kim, Young Kee Shong, and Jeong Hyun Lee. A computer-aided diagnosis system using artificial intelligence for the diagnosis and characterization of thyroid nodules on ultrasound: Initial clinical assessment. *Thyroid*, 27(4):546–552, 2017.
- [CDL⁺19] Eric Z. Chen, Xu Dong, Xiaoxiao Li, Hongda Jiang, Ruichen Rong, and Junyan Wu. Lesion attributes segmentation for melanoma detection with multi-task U-Net. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 485–488, 2019.
- [CFS⁺23] Yanling Cui, Caijian Fu, Changfeng Si, Jinjian Li, Yue Kang, Yanping Huang, and Kefeng Cui. Analysis and comparison of the malignant thyroid nodules not recommended for biopsy in ACR TIRADS and AI TIRADS with a large sample of surgical series. *Journal of Ultrasound in Medicine*, 42(6):1225–1233, 2023.
- [CK] Anagha Choudhari and Aishwarya Korde. PCOS detection using ultrasound images.
- [CKK⁺09] Ji Soo Choi, Jinna Kim, Jin Young Kwak, Min Jung Kim, Hang Seok Chang, and Eun-Kyung Kim. Preoperative staging of papillary thyroid carcinoma: Comparison of ultrasound imaging and CT. *American Journal of Roentgenology*, 193(3):871–878, 2009.

- [CKK⁺10] Seon Hyeong Choi, Eun-Kyung Kim, Jin Young Kwak, Min Jung Kim, and Eun Ju Son. Interobserver and intraobserver variations in ultrasound assessment of thyroid nodules. *Thyroid*, 20(2):167–172, 2010.
- [CKNH20a] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020.
- [CKNH20b] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations, July 2020. arXiv:2002.05709 [cs].
- [CTM⁺21a] Mathilde Caron, Hugo Touvron, Ishan Misra, et al. Emerging properties in self-supervised vision transformers, 2021.
- [CTM⁺21b] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers, May 2021. arXiv:2104.14294 [cs].
- [CWB⁺17] Jianning Chi, Ekta Walia, Paul Babyn, Jimmy Wang, Gary Groot, and Mark Eramian. Thyroid nodule classification in ultrasound images by fine-tuning deep convolutional neural network. *Journal of Digital Imaging*, 30(4):477–486, 2017.
- [CYZM21] Wenzhuo Chen, Yao Yao, Peng Zheng, and Joyce Malywanga. Development of a set of DNA methylation markers in the diagnosis and prognosis of papillary thyroid carcinoma by machine learning. In *2021 7th Annual International Conference on Network and Information Systems for Computers (ICNISC)*, pages 635–639. IEEE, 2021.
- [DBK⁺21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

- [DCL⁺15] Cosimo Durante, Giuseppe Costante, Giuseppe Lucisano, Rocco Bruno, Domenico Meringolo, Alessandra Paciaroni, Efisio Puxeddu, Massimo Torlontano, Salvatore Tumino, Marco Attard, Livia Lamartina, Antonio Nicolucci, and Sebastiano Filetti. The natural history of benign thyroid nodules. *JAMA*, 313(9):926–935, 2015.
- [DCLT19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, 2019.
- [DCY⁺22] Grant Duffy, Paul P. Cheng, Neal Yuan, Bryan He, Alan C. Kwan, Matthew J. Shun-Shin, Kevin M. Alexander, Joseph Ebinger, Matthew P. Lungren, Florian Rader, David H. Liang, Ingela Schnittger, Euan A. Ashley, James Y. Zou, Jignesh Patel, Ronald Witteles, Susan Cheng, and David Ouyang. High-Throughput Precision Phenotyping of Left Ventricular Hypertrophy With Cardiovascular Deep Learning. *JAMA Cardiology*, 7(4):386–395, April 2022.
- [DDJ⁺20] Xinyun Duan, Shukai Duan, Ping Jiang, Ruijuan Li, Yan Zhang, Juan Ma, Hua Zhao, and Hongbing Dai. An ensemble deep learning architecture for multilabel classification on TI-RADS. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 576–582. IEEE, 2020.
- [DDS⁺09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
- [DGL⁺18] Cosimo Durante, Giorgio Grani, Livia Lamartina, Sebastiano Filetti, Su-

- san J. Mandel, and David S. Cooper. The diagnosis and management of thyroid nodules: A review. *JAMA*, 2018.
- [DHSN19] Jianrui Ding, Zichen Huang, Mengdie Shi, and Chunping Ning. Automatic thyroid ultrasound image segmentation based on U-shaped network. In *2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–5, 2019.
- [DTL⁺21] James M. Dolezal, Aleksandra Trzcinska, Chih-Yi Liao, Sara Kochanny, Elizabeth Blair, Nishant Agrawal, Xavier M. Keutgen, Peter Angelos, Nicole A. Cipriani, and Alexander T. Pearson. Deep learning prediction of BRAF-RAS gene expression signature identifies noninvasive follicular thyroid neoplasms with papillary-like nuclear features. *Modern Pathology*, 34(5):862–874, 2021.
- [DYW⁺22] Yi Ding, Qiqi Yang, Yiqian Wang, Dajiang Chen, Zhiguang Qin, and Jian Zhang. MallesNet: A multi-object assistance based network for brachial plexus segmentation in ultrasound images. *Medical Image Analysis*, 80:102511, August 2022.
- [E⁺94] Shereen Ezzat et al. Thyroid incidentalomas: Prevalence by palpation and ultrasonography. *Archives of Internal Medicine*, 154(16):1838–1840, 1994.
- [EKN⁺17] Andre Esteva, Brett Kuprel, Roberto A. Novoa, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.
- [ELW21] John Eisenbrey, Andrej Lyshchik, and Corinne Wessner. Ultrasound data of a variety of liver masses, 2021.
- [EMB⁺25] Edward Ellis, Robert Mendel, Andrew Bulpitt, Nasim Parsa, Michael F. Byrne, and Sharib Ali. Self-Supervised Ultrasound-Video Segmentation

with Feature Prediction and 3D Localised Loss, July 2025. arXiv:2507.18424 [cs].

- [EXM⁺22] Ashkan Ebadi, Pengcheng Xi, Alexander MacLean, Adrian Florea, Stéphane Tremblay, Sonny Kohli, and Alexander Wong. COVIDx-US: An Open-Access Benchmark Dataset of Ultrasound Imaging Data for AI-Driven COVID-19 Analytics. *Frontiers in Bioscience*, 27(7):198, June 2022.
- [fda18] FDA 510(k): AmCAD-UT detection 2.2. https://www.accessdata.fda.gov/cdrh_docs/pdf18/K180006.pdf, 2018.
- [fda20] FDA 510(k): MEDO thyroid. https://www.accessdata.fda.gov/cdrh_docs/pdf20/K203502.pdf, 2020.
- [fda21] FDA 510(k): Koios DS. https://www.accessdata.fda.gov/cdrh_docs/pdf21/K212616.pdf, 2021.
- [FW16] James A. Fagin and Samuel A. Wells. Biologic and clinical perspectives on thyroid cancer. *New England Journal of Medicine*, 375(11):1054–1067, 2016.
- [GDWG21] Yanjun Guo, Xingguang Duan, Chengyi Wang, and Huiqin Guo. Segmentation and recognition of breast ultrasound images based on an expanded U-Net. *PLoS ONE*, 16(6):e0253202, June 2021.
- [GFGCCdAP24] Wilfrido Gómez-Flores, Maria Julia Gregorio-Calas, and Wagner Coelho de Albuquerque Pereira. BUS-BRA: A breast ultrasound dataset for assessing computer-aided diagnosis systems. *Medical Physics*, 51(4):3110–3123, 4 2024.
- [GLJ⁺18] Liang Gao, Rui Liu, Yan Jiang, Wei Song, Yi Wang, Jiachen Liu, Jing Wang, Di Wu, Shanshan Li, Aimin Hao, and Bo Zhang. Computer-aided

- system for diagnosing thyroid nodules on ultrasound: A comparison with radiologist-based clinical assessments. *Head & Neck*, 40(4):778–783, 2018.
- [GMC⁺24] Yimeng Geng, Gaofeng Meng, Mingcong Chen, Guanglin Cao, Mingyang Zhao, Jianbo Zhao, and Hongbin Liu. Force Sensing Guided Artery-Vein Segmentation via Sequential Ultrasound Images. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 656–666, Cham, 2024. Springer Nature Switzerland.
- [gon26] haifan gong. [haifangong/TRFE-Net-for-thyroid-nodule-segmentation](#), February 2026. original-date: 2021-02-11T08:33:39Z.
- [GPC⁺16] Varun Gulshan, Lily Peng, Marc Coram, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22):2402–2410, 2016.
- [GSA⁺20] Jean-Bastien Grill, Florian Strub, Florent Altché, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, 2020.
- [GSFD24] Giorgio Grani, Marialuisa Sponziello, Sebastiano Filetti, and Cosimo Durante. Thyroid nodules: diagnosis and management. *Nature Reviews Endocrinology*, 20(12):715–728, 2024.
- [H⁺24] Ibrahim Ethem Hamamci et al. A foundation model utilizing chest CT volumes and radiology reports for supervised-level zero-shot detection of abnormalities, 2024.
- [HAB⁺16] Bryan R. Haugen, Erik K. Alexander, Keith C. Bible, Gerard M. Doherty, Susan J. Mandel, Yuri E. Nikiforov, Furio Pacini, Gregory W. Randolph, Anna M. Sawka, Martin Schlumberger, Kathryn G. Schuff, Steven I. Sherman, Julie Ann Sosa, David L. Steward, R. Michael Tuttle, and Leonard

- Wartofsky. 2015 american thyroid association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer. *Thyroid*, 26(1):1–133, 2016.
- [HBH⁺17] Alexander Hann, Lucas Bettac, Mark M. Haenle, Tilmann Graeter, Andreas W. Berger, Jens Dreyhaupt, Dieter Schmalstieg, Wolfram G. Zoller, and Jan Egger. Algorithm guided outlining of 105 pancreatic cancer liver metastases in Ultrasound. *Scientific Reports*, 7(1):12779, October 2017.
- [HBL⁺23] Qi He, Sophia Bano, Jing Liu, Wentian Liu, Danail Stoyanov, and Siyang Zuo. Query2²: Query over queries for improving gastrointestinal stromal tumour detection in an endoscopic ultrasound. *Computers in Biology and Medicine*, 152:106424, January 2023.
- [HCM⁺25] Vlad Hondru, Florinel Alin Croitoru, Shervin Minaee, Radu Tudor Ionescu, and Nicu Sebe. Masked Image Modeling: A Survey, July 2025. arXiv:2408.06687 [cs].
- [HCS⁺22] Xiao Han, Lei Chang, Kun Song, Li Cheng, Mingyi Li, and Xun Wei. Multitask network for thyroid nodule diagnosis based on TI-RADS. *Medical Physics*, 49(8):5064–5080, 2022.
- [HCX⁺21] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021.
- [HDJ⁺22] Han Huang, Yijie Dong, Xiaohong Jia, Jianqiao Zhou, Dong Ni, Jun Cheng, and Ruobing Huang. Personalized Diagnostic Tool for Thyroid Cancer Classification Using Multi-view Ultrasound. In Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, Lecture

Notes in Computer Science, pages 665–674, Cham, 2022. Springer Nature Switzerland.

- [HFW⁺20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.
- [HGM⁺82] B. Hamberger, H. Gharib, L. J. Melton, J. R. Goellner, and A. R. Zinsmeister. Fine-needle aspiration biopsy of thyroid nodules. impact on thyroid practice and cost of care. *American Journal of Medicine*, 73(3):381–384, 1982.
- [HHP21] Minkyung Han, Eun Ju Ha, and Ji Hyun Park. Computer-aided diagnostic system for thyroid nodules on ultrasonography: Diagnostic performance based on the thyroid imaging reporting and data system classification and dichotomous outcomes. *American Journal of Neuroradiology*, 42(3):559–565, 2021.
- [HHZ⁺24] Xiaowen Hou, Menglei Hua, Wei Zhang, Jianxin Ji, Xuan Zhang, Huiru Jiang, Mengyun Li, Xiaoxiao Wu, Wenwen Zhao, Shuxin Sun, Lei Cao, and Liuying Wang. An ultrasonography of thyroid nodules dataset with pathological diagnosis annotation for deep learning. *Scientific Data*, 11(1):1272, November 2024.
- [HLM⁺15] Jenny K. Hoang, Jill E. Langer, William D. Middleton, Carol C. Wu, Lynwood W. Hammers, John J. Cronan, Franklin N. Tessler, Edward G. Grant, and Lincoln L. Berland. Managing incidental thyroid nodules detected on imaging: white paper of the ACR incidental thyroid findings committee. *Journal of the American College of Radiology*, 12(2):143–150, 2015.

- [HW24] Dan Hendrycks and Thomas Woodside. Devising ML Metrics, February 2024.
- [HZL⁺22] Ping Huang, Bing Zheng, Mingyi Li, Lin Xu, Shahid Rabbani, Ahmed Majet Mayet, Chuanyi Chen, Bin Zhan, and Huang Jun. The diagnostic value of artificial intelligence ultrasound S-Detect technology for thyroid nodules. *Computational Intelligence and Neuroscience*, 2022:1–7, 2022.
- [HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.
- [HZZ⁺23] Jin Huang, Jingwen Zhang, Yimin Zhang, Xiaoxiao Li, Xiao Ma, Jingwen Deng, Hui Shen, Du Wang, liye mei, and Cheng Lei. BUSI_whu: Breast Cancer Ultrasound Image Dataset, October 2023.
- [IPR⁺26] Vedrana Ivezić, Mara Pleasure, Ashwath Radhachandran, Saarang Panchavati, Shreeram Athreya, Vivek Sant, Benjamin Emert, Gregory Fishbein, Corey Arnold, and William Speier. Pretext Matters: An Empirical Study of SSL Methods in Medical Imaging, March 2026.
- [Iqb23a] Ahmed Iqbal. BUSC Dataset. 1, March 2023. Mendeley Data.
- [Iqb23b] Ahmed Iqbal. BUS_UC. 1, October 2023. Mendeley Data.
- [IRK⁺19] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghighi, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison, January 2019. arXiv: 1901.07031.

- [ITW18] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International Conference on Machine Learning*, 2018.
- [JDK⁺23] Parikshit Juvekar, Reuben Dorent, Fryderyk Kögl, Erickson Torio, Colton Barr, Laura Rigolo, Colin Galvin, Nick Jowkar, et al. The Brain Resection Multimodal Imaging Database (ReMIND), 2023.
- [JFR⁺25] Yuncheng Jiang, Chun-Mei Feng, Jinke Ren, Jun Wei, Zixun Zhang, Yiwen Hu, Yunbi Liu, Rui Sun, Xuemei Tang, Juan Du, Xiang Wan, Yong Xu, Bo Du, Xin Gao, Guangyu Wang, Shaohua Zhou, Shuguang Cui, and Zhen Li. From pretraining to privacy: federated ultrasound foundation model with self-supervised learning. *npj Digital Medicine*, 8(1):714, November 2025.
- [JKH⁺19] Eun Young Jeong, Hee Lim Kim, Eun Ju Ha, Sae Yeon Park, Young Jun Cho, and Minkyung Han. Computer-aided diagnosis system for thyroid nodules on ultrasonography: diagnostic performance and reproducibility based on the experience level of operators. *European Radiology*, 29(4):1978–1985, 2019.
- [JMK⁺22] Xiaoxiao Jia, Zhe Ma, Deguang Kong, Yunxiang Li, Hao Hu, Liyun Guan, Jia Yan, Ruihao Zhang, Yun Gu, Xin Chen, Li Shi, Xin Luo, Qian Li, Bo Bai, Xiangdong Ye, Haifa Zhai, Hui Zhang, Yuhao Dong, Lei Xu, Jianqiao Zhou, and CAAU. Novel human artificial intelligence hybrid framework pinpoints thyroid nodule malignancy and identifies overlooked second-order ultrasonographic features. *Cancers*, 14(18):4440, 2022.
- [JPO⁺22] Zhenhua Jin, Songwang Pei, Lanling Ouyang, Lingling Zhang, Xiaochun Mo, Qijia Chen, Jiaying You, Liang Chen, Bin Zhang, and Shu Zhang.

- Thy-Wise: An interpretable machine learning model for the evaluation of thyroid nodules. *International Journal of Cancer*, 151(12):2229–2243, 2022.
- [JRAE⁺24] Yuliana Jimenez, Maria Jose Rodriguez-Alvarez, Lider Escudero, Carlos Sandoval, and Vasudevan Lakshminarayanan. Ultrasound Breast images denoising using Generative Adversarial Networks (GANs), 2024.
- [JSO⁺21] Samir Jain, Ayan Seal, Aparajita Ojha, Anis Yazidi, Jan Bures, Ilja Tacheci, and Ondrej Krejcar. A deep CNN model for anomaly detection and localization in wireless capsule endoscopy images. *Computers in Biology and Medicine*, 137:104789, 2021.
- [JZL⁺24] Jing Jiao, Jin Zhou, Xiaokang Li, Menghua Xia, Yi Huang, Lihong Huang, Na Wang, Xiaofan Zhang, Shichong Zhou, Yuanyuan Wang, and Yi Guo. USFM: A universal ultrasound foundation model generalized to tasks and organs towards label efficient image analysis. *Medical Image Analysis*, 96:103202, August 2024.
- [KA24a] Kim-Ann. ftsvd/USAnotAI, June 2024.
- [KA24b] Zeki Kuş and Musa Aydin. MedSegBench: A comprehensive benchmark for medical image segmentation in diverse data modalities. *Scientific Data*, 11(1):1283, November 2024.
- [KCH⁺22] Young-Joong Kim, Youseok Choi, Soo-Jong Hur, Kyong-Soo Park, Hyun-Jin Kim, Mirinae Seo, Min Kyoung Lee, So-Lyung Jung, and Chan Kwon Jung. Deep convolutional neural network for classification of thyroid nodules on ultrasound: Comparison of the diagnostic performance with that of radiologists. *European Journal of Radiology*, 152:110335, 2022.
- [KED⁺22] Markus Krönke, Christine Eilers, Desislava Dimova, Melanie Köhler, Gabriel Buschner, Lilit Schweiger, Lemonia Konstantinidou, Marcus Makowski,

- James Nagarajah, Nassir Navab, Wolfgang Weber, and Thomas Wendler. Tracked 3D ultrasound and deep neural network-based thyroid segmentation reduce interobserver variability in thyroid volumetry. *PLOS ONE*, 17(7):e0268550, July 2022.
- [KGZ⁺] Qingbo Kang, Jun Gao, Hongkai Zhao, Zhu He, Kang Li, and Qicheng Lao. D2MAE: Diffusional Deblurring MAE for Ultrasound Image Pre-training.
- [KHH19] Hee Lim Kim, Eun Ju Ha, and Minkyung Han. Real-world performance of computer-aided diagnosis system for thyroid nodules using ultrasonography. *Ultrasound in Medicine & Biology*, 45(10):2672–2678, 2019.
- [KJS⁺25] Sekeun Kim, Pengfei Jin, Sifan Song, Cheng Chen, Yiwei Li, Hui Ren, Xiang Li, Tianming Liu, and Quanzheng Li. EchoFM: Foundation Model for Generalizable Echocardiogram Analysis. *IEEE Transactions on Medical Imaging*, 44(10):4049–4062, October 2025.
- [KLG⁺25] Qingbo Kang, Qicheng Lao, Jun Gao, Wuyongga Bao, Zhu He, Chenlin Du, Qiang Lu, and Kang Li. URFM: A general Ultrasound Representation Foundation Model for advancing ultrasound image diagnosis, August 2025.
- [KLH⁺20] Jieun Koh, Eunjung Lee, Kyunghwa Han, Eun-Kyung Kim, Eun Ju Son, Yu-Mee Sohn, Mirinae Seo, Mihye Kwon, Jung Hyun Yoon, Jong Hyun Lee, Yoo Mee Park, Sumi Kim, Jung Hee Shin, and Jin Young Kwak. Diagnosis of thyroid nodules on ultrasonography by a deep convolutional neural network. *Scientific Reports*, 10(1):15245, 2020.
- [KLHZ17] Guixia Kang, Kui Liu, Beibei Hou, and Ningbo Zhang. 3D multi-view convolutional neural networks for lung nodule classification. *PLOS ONE*, 12(11):e0188290, November 2017.

- [KLM⁺22] Xavier M. Keutgen, Hua Li, Kelechi Memeh, Jeanne Conn Busch, Jeffrey Williams, Liu Lan, David Sarne, Brendan Finnerty, Peter Angelos, Thomas J. Fahey, and Maryellen L. Giger. A machine-learning algorithm for distinguishing malignant from benign indeterminate thyroid nodules using ultrasound radiomic features. *Journal of Medical Imaging*, 9(03), 2022.
- [KMR⁺23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, et al. Segment anything. In *IEEE International Conference on Computer Vision*, 2023.
- [koaia] Koios DS, koios medical, inc. <https://grand-challenge.org/aiforradiology/product/koios-ds/>. accessed January 29, 2024.
- [koib] Koios DSTM thyroid. <https://koiosmedical.com/products/koios-ds-thyroid/>. Koios Medical, accessed January 29, 2024.
- [KSS20] Cari M. Kitahara, Julie A. Sosa, and Meredith S. Shiels. Influence of Nomenclature Changes on Trends in Papillary Thyroid Cancer Incidence in the United States, 2000 to 2017. *The Journal of Clinical Endocrinology and Metabolism*, 105(12):e4823–4830, December 2020.
- [KT86] Frederick W. Kremkau and Kenneth J. Taylor. Artifacts in ultrasound imaging. *Journal of Ultrasound in Medicine*, 5(4):227–237, 1986.
- [KW17] Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks, February 2017.
- [KWZ18] E. J. Kuo, J. X. Wu, and K. A. Zanoocco. Cost effectiveness of immediate biopsy versus surveillance of intermediate-suspicion thyroid nodules. *Surgery*, 164(6):1330–1335, 2018.

- [KZ20] Tahsin M. Khan and Martha A. Zeiger. Thyroid nodule molecular testing: Is It Ready for Prime Time? *Frontiers in Endocrinology*, 11:590128, 2020.
- [LCA⁺17] Chang Liu, Yu Cao, Marlon Alcantara, Benyuan Liu, Maria Brunette, Jesus Peinado, and Walter Curioso. TX-CNN: Detecting tuberculosis in chest X-ray images using convolutional neural network. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 2314–2318, Beijing, China, September 2017. IEEE Press.
- [LCHL14] Xiaodan Liang, Qingxing Cao, Rui Huang, and Liang Lin. Recognizing focal liver lesions in contrast-enhanced ultrasound with discriminatively trained spatio-temporal model. In *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, pages 1184–1187, Beijing, China, April 2014. IEEE.
- [LCJ⁺21] Hyunna Lee, Young Jun Chai, Hyeonwoo Joo, Kyoungjune Lee, Ji Young Hwang, Seok-Mo Kim, Kwangsoo Kim, Inn-Chul Nam, Ji Young Choi, Hyeong Won Yu, Myung-Chul Lee, Hiroo Masuoka, Akira Miyauchi, Kyu Eun Lee, Sumi Kim, and Hyoun-Joong Kong. Federated learning for thyroid ultrasound image analysis to protect personal information: Validation study in a real health care environment. *JMIR Medical Informatics*, 9(5):e25869, 2021.
- [LCT⁺23] Jun Li, Junyu Chen, Yucheng Tang, Ce Wang, Bennett A. Landman, and S. Kevin Zhou. Transforming medical imaging with Transformers? A comparative review of key properties, current progresses, and future perspectives. *Medical Image Analysis*, 85:102762, April 2023.
- [LDY⁺22] Zixian Liu, Lewis Deyer, Ao Yang, Sijing Liu, Jing Gong, Yilin Yang, Mingqian Huang, Anish Doshi, Maoxue Lu, Denise Lee, Timothy Deyer,

- and Xueyang Mei. Automated machine learning-based radiomics analysis versus deep learning-based classification for thyroid nodule on ultrasound images: a multi-center study. In *2022 IEEE 22nd International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 23–28. IEEE, 2022.
- [LHC⁺21] Xiaowen Liang, Yong Huang, Yuming Cai, Jiahui Liao, and Zhengrong Chen. A computer-aided diagnosis system and thyroid imaging reporting and data system for dual validation of ultrasound-guided fine-needle aspiration of indeterminate thyroid nodules. *Frontiers in Oncology*, 11:611436, 2021.
- [LHL⁺25] Xianhang Li, Chen Huang, Chun-Liang Li, Eran Malach, Josh Susskind, Vimal Thilak, and Etai Littwin. Rethinking JEPA: Compute-Efficient Video SSL with Frozen Teachers, September 2025. arXiv:2509.24317 [cs].
- [LLK19] Junhyun Lee, Inyeop Lee, and Jaewoo Kang. Self-Attention Graph Pooling, April 2019.
- [LLM⁺23] Michał Ludwig, Bartłomiej Ludwig, Agnieszka Mikuła, Sławomir Biernat, Jerzy Rudnicki, and Krzysztof Kaliszewski. The use of artificial intelligence in the diagnosis and classification of thyroid nodules: An update. *Cancers*, 15(3):708, 2023.
- [LLR⁺] Alexandre Lacoste, Nils Lehmann, Pau Rodriguez, Evan David Sherwin, Hannah Kerner, Bjorn Lutjens, Jeremy Irvin, David Dao, Hamed Alemohammad, Alexandre Drouin, Mehmet Gunturkun, Gabriel Huang, David Vazquez, Dava Newman, Yoshua Bengio, Stefano Ermon, and Xiao Xi-ang Zhu. GEO-Bench: Toward Foundation Models for Earth Monitoring. *Advances in Neural Information Processing Systems*.
- [LLZ⁺22] Zhi Lin, Junhao Lin, Lei Zhu, Huazhu Fu, Jing Qin, and Liansheng Wang.

A New Dataset and A Baseline Model for Breast Lesion Detection in Ultrasound Videos, July 2022. arXiv:2207.00141 [eess].

[LSA⁺24] Etai Littwin, Omid Saremi, Madhu Advani, Vimal Thilak, Preetum Nakkiran, Chen Huang, and Joshua Susskind. How JEPA Avoids Noisy Features: The Implicit Bias of Deep Linear Self Distillation Networks, July 2024. arXiv:2407.03475 [cs.LG].

[LSP⁺19] Sarah Leclerc, Erik Smistad, João Pedrosa, Andreas Østvik, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Pierre-Marc Jodoin, Thomas Grenier, Carole Lartizien, Jan D’hooge, Lasse Lovstakken, and Olivier Bernard. Deep Learning for Segmentation Using an Open Large-Scale Dataset in 2D Echocardiography. *IEEE Transactions on Medical Imaging*, 38(9):2198–2210, September 2019.

[LXC⁺25] Gongning Luo, Mingwang Xu, Hongyu Chen, et al. Tumor Detection, Segmentation and Classification Challenge on Automated 3D Breast Ultrasound: The TDSC-ABUS Challenge, January 2025. arXiv:2501.15588 [eess].

[LXY⁺24] Xian Lin, Yangyang Xiang, Li Yu, and Zengqiang Yan. Beyond Adapting SAM: Towards End-to-End Ultrasound Image Segmentation via Auto Prompting. In Marius George Linguraru et al., editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15008, pages 24–34. Springer Nature Switzerland, Cham, 2024.

[LYLC20] Xiao Liang, Jing Yu, Jiahui Liao, and Zhengrong Chen. Convolutional neural network for breast and thyroid nodules diagnosis in ultrasound imaging. *BioMed Research International*, 2020:1–9, 2020.

- [LZW⁺23] Jiajia Li, Pingping Zhang, Teng Wang, Kaixuan Wang, and Bin Sheng. LEPset, June 2023.
- [LZZ⁺19] Xin Li, Shu Zhang, Qi Zhang, Xun Wei, Yixin Pan, Jun Zhao, Xin Xin, Chunlai Qin, Xiaoya Wang, Jie Li, Fengjiao Yang, Yanfang Zhao, Meng Yang, Qiang Wang, Zhiyun Zheng, Xueying Zheng, Xiaowei Yang, Christopher T. Whitlow, Metin N. Gurcan, Long Zhang, Xiaolong Wang, Boris C. Pasche, Ming Gao, Wei Zhang, and Ke Chen. Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. *Lancet Oncology*, 20(2):193–201, 2019.
- [MC11] Christopher L. Moore and Joshua A. Copel. Point-of-care ultrasonography. *New England Journal of Medicine*, 364(8):749–757, 2011.
- [MDE⁺25] Youssef Megahed, Robin Ducharme, Aylin Erman, Mark Walker, Steven Hawken, and Adrian D. C. Chan. USF-MAE: Ultrasound Self-Supervised Foundation Model with Masked Autoencoding, November 2025. arXiv:2510.22990 [eess].
- [Mei21] Kristen Meiburger. DATASET for "Deep learning segmentation of transverse musculoskeletal ultrasound images for neuromuscular disease assessment", July 2021.
- [Mei22] Kristen Meiburger. DATASET for: "Carotid Ultrasound Boundary Study (CUBS)", March 2022.
- [MFS⁺26] Alif Munim, Adibvafa Fallahpour, Teodora Szasz, Ahmadreza Attarpour, River Jiang, Brana Sooriyakanthan, Maala Sooriyakanthan, et al. Echo-JEPA: A Latent Predictive Foundation Model for Echocardiography, 2026.

- [MJL⁺25] Chen Ma, Jing Jiao, Shuyu Liang, Junhu Fu, Qin Wang, Zeju Li, Yuanyuan Wang, and Yi Guo. TinyUSFM: Towards Compact and Efficient Ultrasound Foundation Models, 2025.
- [MLR⁺22] Xueyan Mei, Zelong Liu, Philip M. Robson, Brett Marinelli, Mingqian Huang, Amish Doshi, Adam Jacobi, Chendi Cao, Katherine E. Link, Thomas Yang, Ying Wang, Hayit Greenspan, Timothy Deyer, Zahi A. Fayad, and Yang Yang. RadImageNet: An Open Radiologic Deep Learning Research Dataset for Effective Transfer Learning. *Radiology: Artificial Intelligence*, 4(5):e210315, September 2022.
- [MMZ⁺25] Adrien Meyer, Aditya Murali, Farahdiba Zarin, Didier Mutter, and Nicolas Padoy. Ultrasam: a foundation model for ultrasound using large open-access segmentation datasets. *International Journal of Computer Assisted Radiology and Surgery*, September 2025.
- [MS12] Ara Mohebbati and Ashok R. Shaha. Anatomy of thyroid and parathyroid glands and neurovascular relations. *Clinical Anatomy*, 25(1):19–31, 2012.
- [MSH⁺16] Anna Montoya, DJ Sterling, Hasnin, kaggle446, shirzad, Will Cukierski, and yffud. Ultrasound nerve segmentation. <https://kaggle.com/competitions/ultrasound-nerve-segmentation>, 2016. Kaggle.
- [MWB55] J. D. Mortensen, L. B. Woolner, and W. A. Bennett. Gross and microscopic findings in clinically normal thyroid glands. *Journal of Clinical Endocrinology and Metabolism*, 15(10):1270–1280, 1955.
- [NFT⁺25] William Ndzimbong, Cyril Fourniol, Loic Themyr, Nicolas Thome, Yvonne Keeza, Benoît Sauer, Pierre-Thierry Piéchaud, Arnaud Méjean, Jacques Marescaux, Daniel George, Didier Mutter, Alexandre Hostettler, and Toby Collins. TRUSTED: The Paired 3D Transabdominal Ultrasound and CT

Human Data for Kidney Segmentation and Registration Research. *Scientific Data*, 12(1):615, April 2025.

- [noaa] Common cancer sites — cancer stat facts. <https://seer.cancer.gov/statfacts/html/common.html>. SEER, accessed 2023-01-25.
- [noab] Stanford AIMI shared datasets. <https://stanfordaimi.azurewebsites.net/datasets/a72f2b02-7b53-4c5d-963c-d7253220bfd5>. accessed 2023-01-27.
- [noa11] Definition of thyroid - NCI Dictionary of Cancer Terms - NCI, February 2011. Archive Location: nciglobal,ncicenterprise.
- [NPM⁺20] Shyam Natarajan, Alan Priester, Daniel Margolis, Jiaoti Huang, and Leonard Marks. Prostate MRI and Ultrasound With Pathology and Coordinates of Tracked Biopsy (Prostate-MRI-US-Biopsy), 2020.
- [NZT⁺22] Xueli Nie, Xuguang Zhou, Tong Tong, Xuefang Lin, Lihong Wang, Hongyu Zheng, Jufeng Li, Ershu Xue, Shengwei Chen, Meijuan Zheng, Chunxia Chen, and Meng Du. N-Net: A novel dense fully convolutional neural network for thyroid nodule segmentation. *Frontiers in Neuroscience*, 16:872601, 2022.
- [OCE] OCEBM Levels of Evidence Working Group. The oxford levels of evidence 2. <https://www.cebm.ox.ac.uk/resources/levels-of-evidence/ocebm-levels-of-evidence>. Oxford Centre for Evidence-Based Medicine, accessed August 7, 2023.
- [ODM⁺24] Maxime Oquab, Timothée Darcet, Théo Moutakanni, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

- [OHG⁺] David Ouyang, Bryan He, Amirata Ghorbani, Matt P Lungren, Euan A Ashley, David H Liang, and James Y Zou. EchoNet-Dynamic: a Large New Cardiac Motion Video Data Resource for Medical Machine Learning.
- [OHG⁺20] David Ouyang, Bryan He, Amirata Ghorbani, Neal Yuan, Joseph Ebinger, Curtis P. Langlotz, Paul A. Heidenreich, Robert A. Harrington, David H. Liang, Euan A. Ashley, and James Y. Zou. Video-based AI for beat-to-beat assessment of cardiac function. *Nature*, 580(7802):252–256, April 2020.
- [OSF⁺18] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y. Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention U-Net: Learning where to look for the pancreas, 2018.
- [PĆD⁺24] Anna Pawłowska, Anna Ćwierz-Pieńkowska, Agnieszka Domalik, Dominika Jaguś, Piotr Kasprzak, Rafal Matkowski, Łukasz Fura, Andrzej Nowicki, and Norbert Zolek. A Curated Benchmark Dataset for Ultrasound Based Breast Lesion Analysis (Breast-Lesions-USG), 2024.
- [PHK⁺16] Daniel G. Pankratz, Zhen Hu, Seo Young Kim, Robert J. Monroe, Meng G. Wong, S. Thomas Traweek, Richard T. Kloos, P. Sean Walsh, and Glen C. Kennedy. Analytical performance of a gene expression classifier for medullary thyroid carcinoma. *Thyroid*, 26(11):1573–1580, 2016.
- [PHS⁺19] Vivian Y. Park, Kyunghwa Han, Yeong Kyeong Seong, Min Hye Park, Eun-Kyung Kim, Hee Jung Moon, Jung Hyun Yoon, and Jin Young Kwak. Diagnosis of thyroid nodules: Performance of a deep learning convolutional neural network model vs. radiologists. *Scientific Reports*, 9(1):17843, 2019.
- [PKJ⁺10] Chan Sub Park, Sang Hee Kim, So Lyung Jung, Bong Joo Kang, Ji Young Kim, Jeong Jun Choi, Mi Suk Sung, Hyeon Woo Yim, and Seung Hee

- Jeong. Observer variability in the sonographic evaluation of thyroid nodules. *Journal of Clinical Ultrasound*, 38(6):287–293, 2010.
- [PKOK21] Kee-Soo Park, Sung Hak Kim, Jae Ho Oh, and Sang Youn Kim. Highly accurate diagnosis of papillary thyroid carcinomas based on personalized pathways coupled with machine learning. *Briefings in Bioinformatics*, 22(4):bbaa336, 2021.
- [PLL⁺21] Sui Peng, Yihao Liu, Weiming Lv, Longzhong Liu, Qian Zhou, Hong Yang, Jie Ren, Guangjian Liu, Xiaodong Wang, Xuehua Zhang, Qiang Du, Fangxing Nie, Gao Huang, Yuchen Guo, Jie Li, Jinyu Liang, Hangtong Hu, Han Xiao, Zelong Liu, Fenghua Lai, Qiuyi Zheng, Haibo Wang, Yanbing Li, Erik K Alexander, Wei Wang, and Haipeng Xiao. Deep learning-based artificial intelligence model to assist thyroid nodule diagnosis and management: a multicentre diagnostic study. *The Lancet Digital Health*, 3(4):e250–e259, April 2021.
- [PMB⁺21] Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, Roger Chou, Julie Glanville, Jeremy M. Grimshaw, Asbjørn Hróbjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Luke A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting, and David Moher. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372:n71, 2021.
- [PSW⁺21] Susan C. Pitt, Megan C. Saucke, Elizabeth M. Wendt, David F. Schneider, Jason Orne, Christine L. Macdonald, Nadine P. Connor, and Rebecca S. Sip-

pel. Patients’ reaction to diagnosis with thyroid cancer or an indeterminate thyroid nodule. *Thyroid*, 31(4):580–588, 2021.

- [PVN⁺15] Lina Pedraza, Carlos Vargas, Fabián Narváez, Oscar Durán, Emma Muñoz, and Eduardo Romero. An open access thyroid ultrasound image database. In *10th International Symposium on Medical Information Processing and Analysis*, volume 9287, pages 188–193. SPIE, 2015.
- [PYL⁺20] Kpal N. Patel, Linwah Yip, Carrie C. Lubitz, Elizabeth G. Grubbs, Barbara S. Miller, Wen Shen, Peter Angelos, Herbert Chen, Gerard M. Doherty, Thomas J. III Fahey, Electron Kebebew, Virginia A. Livolsi, Nancy D. Perrier, Jennifer A. Sipos, Julie A. Sosa, David Steward, Ralph P. Tufano, Christopher R. McHenry, and Sally E. Carty. The american association of endocrine surgeons guidelines for the definitive surgical management of thyroid disease in adults. *Annals of Surgery*, 271(3):e21, 2020.
- [PZL21] Huitong Pan, Quan Zhou, and Longin Jan Latecki. SGUNet: Semantic guided UNet for thyroid nodule segmentation. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 630–634, 2021.
- [RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241. Springer, 2015.
- [RKH⁺21] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

- [Ros06] Douglas S. Ross. Predicting thyroid malignancy. *The Journal of Clinical Endocrinology & Metabolism*, 91(11):4253–4255, 2006.
- [RSH⁺22] Gregory W. Randolph, Julie A. Sosa, Yue Hao, Trevor E. Angell, David C. Shonka, Virginia A. LiVolsi, Paul W. Ladenson, Todd C. Blevins, Quan-Yang Duh, Ronald Ghossein, Monica Harrell, Kepal N. Patel, Michael H. Shanik, S. Thomas Traweek, P. Sean Walsh, Michael W. Yeh, Amr H. Abdelhamid Ahmed, Allen S. Ho, Richard J. Wong, Joshua P. Klopfer, Jing Huang, Glen C. Kennedy, Richard T. Kloos, and Peter M. Sadow. Preoperative identification of medullary thyroid carcinoma (MTC): Clinical validation of the afirma MTC RNA-sequencing classifier. *Thyroid*, 32(9):1069–1076, 2022.
- [RSK⁺22] Ekaterina Redekop, Karthik V. Sarma, Adam Kinnaird, Anthony Sisk, Steven S. Raman, Leonard S. Marks, William Speier, and Corey W. Arnold. Attention-guided prostate lesion localization and grade group classification with multiple instance learning. In *Medical Imaging with Deep Learning*, 2022.
- [RWC⁺19] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019.
- [SB24] Wei Shao and Wayne Brisbane. Micro-Ultrasound Prostate Segmentation Dataset, January 2024.
- [SBD⁺21] Yasaman Sharifi, Mohamad Amin Bakhshali, Toktam Dehghani, Morteza DanaiAshgzari, Mahdi Sargolzaei, and Saeid Eslami. Deep learning on ultrasound images of thyroid nodules. *Biocybernetics and Biomedical Engineering*, 41(2):636–655, 2021.

- [SCG⁺24] Kirsten R. Steffner, Matthew Christensen, George Gill, Michael Bowdish, Justin Rhee, Abirami Kumaresan, Bryan He, James Zou, and David Ouyang. Deep learning for transesophageal echocardiography view classification. *Scientific Reports*, 14(1):11, January 2024.
- [SCM⁺19] Jiyong Song, Young Jun Chai, Hiroo Masuoka, Sung-Woo Park, Sumi Kim, Ji Young Choi, Hyoun-Joong Kong, Kyu Eun Lee, Jonghwa Lee, Naksu Kwak, Ka Hee Yi, and Akira Miyauchi. Ultrasound image analysis using deep learning algorithm for the diagnosis of thyroid nodules. *Medicine*, 98(15):e15133, 2019.
- [SLL⁺19] Wenfeng Song, Shuai Li, Ji Liu, Hong Qin, Bo Zhang, Shuyang Zhang, and Aimin Hao. Multitask cascade convolution neural networks for automatic thyroid nodule detection and recognition. *IEEE Journal of Biomedical and Health Informatics*, 23(3):1215–1224, 2019.
- [SLL⁺22a] Adam Stenman, Nina Linder, Mikael Lundin, Caj Haglund, Johanna Arola, and Johan Lundin. A deep learning-based algorithm for tall cell detection in papillary thyroid carcinoma. *PLOS ONE*, 17(8):e0272696, 2022.
- [SLL⁺22b] Jiawei Sun, Chunying Li, Zhengda Lu, Mu He, Tong Zhao, Xiaoqin Li, Liugang Gao, Kai Xie, Tao Lin, Jianfeng Sui, Qianyi Xi, Fan Zhang, and Xinye Ni. TNSNet: Thyroid nodule segmentation in ultrasound imaging using soft shape supervision. *Computer Methods and Programs in Biomedicine*, 215:106600, 2022.
- [SRH⁺22] Rohit Singla, Cailin Ringstrom, Grace Hu, Victoria Lessoway, Janice Reid, Christopher Nguan, and Robert Rohling. The Open Kidney Ultrasound Data Set, December 2022. arXiv:2206.06657 [eess].
- [STN⁺22] Katrine Z. Swan, Johnson Thomas, Viveque E. Nielsen, Maria L. Jespersen,

- and Søren J. Bonnema. External validation of AIBx, an artificial intelligence model for risk stratification, in thyroid nodules. *European Thyroid Journal*, 11(2), 2022.
- [SVS⁺25] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, August 2025. arXiv:2508.10104 [cs].
- [SWZ⁺23] Jiawei Sun, Bobo Wu, Tong Zhao, Liugang Gao, Kai Xie, Tao Lin, Jianfeng Sui, Xiaoqin Li, Xiaojin Wu, and Xinye Ni. Classification for thyroid nodule using ViT with contrastive learning in ultrasound images. *Computers in Biology and Medicine*, 152:106444, January 2023.
- [Sza14] Thomas L. Szabo. *Diagnostic Ultrasound Imaging: Inside Out*. Academic Press, 2nd edition, 2014.
- [SZLC21] Run Su, Deyun Zhang, Jinhui Liu, and Chuandong Cheng. MSU-Net: Multi-scale U-Net for 2d medical image segmentation. *Frontiers in Genetics*, 12, 2021.
- [SZZ⁺20] Rui Song, Long Zhang, Chao Zhu, Ji Liu, Jingfen Yang, and Tao Zhang. Thyroid nodule ultrasound image classification through hybrid feature cropping network. *IEEE Access*, 8:64064–64074, 2020.
- [T⁺17] Franklin N. Tessler et al. 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, 2017.

- [TG] Adam Tupper and Christian Gagné. Revisiting Data Augmentation for Ultrasound Images.
- [TG97] G. H. Tan and H. Gharib. Thyroid incidentalomas: management approaches to nonpalpable nodules discovered incidentally on thyroid imaging. *Annals of Internal Medicine*, 126(3):226–231, 1997.
- [Thi] Thingiverse.com. Kevin’s Thyroid by j0z.
- [Tia23] Tiantian Yang. uterine fibroid ultrasound images, March 2023.
- [TMOB⁺25] Amina Turki, Ahmed Mahdi Obaid, Hatem Bellaaj, Mohamed Ksantini, and Abdulla Altaee. Gallbladder Diseases Dataset, 2025.
- [Top19] Eric J. Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56, 2019.
- [TSTM⁺23] Adam Taha, Bahaa Saad, Stephanie Taha-Mehlitz, Vincent Ochs, Jana El-Awar, Mourad M. Mourad, Katrin Neumann, Christian Glaser, Robert Rosenberg, and Philippe C. Cattin. Analysis of artificial intelligence in thyroid diagnostics and surgery: A scoping review. *American Journal of Surgery*, pages S0002–9610(23)00624–4, 2023.
- [TT23] Franklin N. Tessler and Johnson Thomas. Artificial Intelligence for Evaluation of Thyroid Nodules: A Primer. *Thyroid*, 33(2):150–158, February 2023.
- [TTLTD⁺23] David Toro-Tobon, Rodrigo Loo-Torres, Monica Duran, Jungwei W. Fan, Naykky Singh Ospina, Yonghui Wu, and Juan P. Brito. Artificial intelligence in thyroidology: A narrative review of the current applications, associated challenges, and future directions. *Thyroid*, 33(8):903–917, 2023.

- [V⁺16] Salvatore Vaccarella et al. Worldwide thyroid-cancer epidemic? The increasing impact of overdiagnosis. *New England Journal of Medicine*, 375:614–617, 2016.
- [V⁺24] Eugene Vorontsov et al. A foundation model for clinical-grade computational pathology and biomarker discovery in oncology. *Nature Medicine*, 2024.
- [VBD⁺24] Noelia Vallez, Gloria Bueno, Oscar Deniz, Miguel Angel Rienda, and Carlos Pastor. BUS-UCLM: Breast ultrasound lesion segmentation dataset, February 2024.
- [VCC⁺18] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.
- [VOIL20] Santiago Vitale, José Ignacio Orlando, Emmanuel Iarussi, and Ignacio Larrabide. Improving realism in patient-specific abdominal ultrasound simulation using CycleGANs. *International Journal of Computer Assisted Radiology and Surgery*, 15(2):183–192, February 2020.
- [VSP⁺17] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.
- [Wan23] Mingyu Wang. 1st place solution in MICCAI 2020 TN-SCUI challenge. <https://github.com/WAMAWAMA/TNSCUI2020-Seg-Rank1st>, 2023.
- [WBR⁺25] Nina Wiedemann, Dianne de Korte-de Boer, Matthias Richter, Sjors van de Weijer, Charlotte Buhre, Franz A. M. Eggert, Sophie Aarnoudse, Lotte Grevendonk, Steffen Röber, Carlijn M.E. Remie, Wolfgang Buhre, Ronald Henry, and Jannis Born. COVID-BLUeS - a Prospective Study on the Value of AI in Lung Ultrasound Analysis. *IEEE Journal of Biomedical and Health Informatics*, 29(9):6301–6310, September 2025.

- [WC97] C. Wang and C. Crapo. The epidemiology of thyroid disease and implications for screening. *Endocrinology and Metabolism Clinics of North America*, 26(1), 1997.
- [WCC⁺20] Yi Wang, Eun Jung Choi, Younhee Choi, Hao Zhang, Gong Yong Jin, and Seok-Bum Ko. Breast Cancer Classification in Automated Breast Ultrasound Using Multiview Convolutional Neural Network with Transfer Learning. *Ultrasound in Medicine & Biology*, 46(5):1119–1132, May 2020.
- [WGY⁺20] Xun Wei, Ming Gao, Ruiguo Yu, Zhikang Liu, Qing Gu, Xiaoli Liu, Zhiyun Zheng, Xueying Zheng, Jiacheng Zhu, and Shu Zhang. Ensemble deep learning model for multicenter classification of thyroid nodules on ultrasound images. *Medical Science Monitor*, 26, 2020.
- [WJB25] Stefano Woerner, Arthur Jaques, and Christian F. Baumgartner. A comprehensive and easy-to-use multi-domain multi-task medical imaging meta-dataset. *Scientific Data*, 12(1):666, April 2025.
- [WLC⁺18] Zhiwei Wang, Chaoyue Liu, Danpeng Cheng, Liang Wang, Xin Yang, and Kwang-Ting Cheng. Automated Detection of Clinically Significant Prostate Cancer in mp-MRI Images Based on an End-to-End Deep Neural Network. *IEEE Transactions on Medical Imaging*, 37(5):1127–1139, May 2018.
- [WLC⁺22] Yuchen Wang, Zhongyu Li, Xiangxiang Cui, Liangliang Zhang, Xiang Luo, Meng Yang, and Shi Chang. Key-frame Guided Network for Thyroid Nodule Recognition Using Ultrasound Videos. In Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, Lecture Notes in Computer Science, pages 238–247, Cham, 2022. Springer Nature Switzerland.

- [WLS⁺18] Jing Wang, Shanshan Li, Wei Song, Hong Qin, Bo Zhang, and Aimin Hao. Learning from weakly-labeled clinical data for automatic thyroid nodule classification in ultrasound images. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 3114–3118. IEEE, 2018.
- [WSSL83] Robert F. Wagner, Stephen W. Smith, James M. Sandrik, and H. Lopez. Statistics of speckle in ultrasound B-scans. *IEEE Transactions on Sonics and Ultrasonics*, 30(3):156–163, 1983.
- [WTTZM22] Benjamin Wildman-Tobriner, Ethan Taghi-Zadeh, and Maciej A. Mazurowski. Artificial intelligence (AI) tools for thyroid nodules on ultrasound, from the AJR special series on AI applications. *American Journal of Roentgenology*, 219(4):547–554, 2022.
- [WWL⁺22] Bo Wang, Zhexin Wan, Congwei Li, Min Zhang, Yan Shi, Xiaoyan Miao, Yi Jian, Yu Luo, Jiahong Yao, and Wenbin Tian. Identification of benign and malignant thyroid nodules based on dynamic AI ultrasound intelligent auxiliary diagnosis system. *Frontiers in Endocrinology*, 13:1018321, 2022.
- [WXL⁺22] Yanli Wang, Lina Xu, Wei Lu, Xin Kong, Ke Shi, Lingling Wang, and Deguang Kong. Clinical evaluation of malignancy diagnosis of rare thyroid carcinomas by an artificial intelligent automatic diagnosis system. *Endocrine*, 80(1):93–99, 2022.
- [WYJ⁺22] Xiantao Wu, Pengyi Yu, Chongwen Jia, Ning Mao, Kun Che, Gang Li, Haicheng Zhang, Yongli Mou, and Xicheng Song. Radiomics analysis of computed tomography for prediction of thyroid capsule invasion in papillary thyroid carcinoma: A multi-classifier and two-center study. *Frontiers in Endocrinology*, 13:849065, 2022.
- [WZG⁺26] Chong Wang, Yabin Zhang, Yunhe Gao, Maya Varma, Clemence Mottez,

- Faidra Patsatzi, Jiaming Liu, Jin Long, Jean-Benoit Delbrouck, Sergios Gatidis, Akshay S. Chaudhari, and Curtis P. Langlotz. A data- and compute-efficient chest X-ray foundation model beyond aggressive scaling, February 2026. arXiv:2602.22843 [cs].
- [WZW⁺20] Qi Wei, Shi-En Zeng, Li-Ping Wang, Yan-Jun Yan, Tao Wang, Jia-Wei Xu, Meng-Yue Zhang, Wen-Zhi Lv, Xin-Wu Cui, and Christoph F. Dietrich. The value of S-Detect in improving the diagnostic performance of radiologists for the differential diagnosis of thyroid nodules. *Medical Ultrasonography*, 22(4):415, 2020.
- [WZZ⁺20] Lituan Wang, Lei Zhang, Minjuan Zhu, Xiaofeng Qi, and Zhang Yi. Automatic diagnosis for thyroid nodules in ultrasound images by deep neural networks. *Medical Image Analysis*, 61:101665, April 2020.
- [XGH⁺24] Yutong Xie, Lin Gu, Tatsuya Harada, Jianpeng Zhang, Yong Xia, and Qi Wu. Rethinking masked image modelling for medical image representation. *Medical Image Analysis*, 98:103304, December 2024.
- [XWW⁺22] Dong Xu, Yanli Wang, Huanhuan Wu, Wei Lu, Weibo Chang, Jiahong Yao, Meiyang Yan, Chenghao Peng, Chen Yang, Lingling Wang, and Lina Xu. An artificial intelligence ultrasound system’s ability to distinguish benign from malignant follicular-patterned lesions. *Frontiers in Endocrinology*, 13:981403, 2022.
- [XWY⁺21] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, volume 34, pages 12077–12090, 2021.
- [XYZ⁺19] Siyue Xia, Jiahui Yao, Wenjun Zhou, Yuli Dong, Suying Xu, Jian Zhou,

- and Weiwei Zhan. A computer-aided diagnosing system in the evaluation of thyroid nodules—experience in a specialized thyroid center. *World Journal of Surgical Oncology*, 17(1):210, 2019.
- [XZL⁺23] Yiming Xu, Bowen Zheng, Xiaohong Liu, Tao Wu, Jinxiu Ju, Shijie Wang, Yufan Lian, Hongjun Zhang, Tong Liang, Ye Sang, Rui Jiang, Guangyu Wang, Jie Ren, and Ting Chen. Improving artificial intelligence pipeline for liver malignancy diagnosis using ultrasound images and video frames. *Briefings in Bioinformatics*, 24(1), 1 2023.
- [YDZ⁺23] Jiewen Yang, Xinpeng Ding, Ziyang Zheng, Xiaowei Xu, and Xiaomeng Li. GraphEcho: Graph-Driven Unsupervised Domain Adaptation for Echocardiogram Video Segmentation, September 2023. arXiv:2309.11145 [cs].
- [YGC⁺22] Qinghan Yang, Chong Geng, Ruyue Chen, Chen Pang, Run Han, Lei Lyu, and Yuang Zhang. DMU-Net: Dual-route mirroring U-Net with mutual learning for malignant thyroid nodule segmentation. *Biomedical Signal Processing and Control*, 77:103805, 2022.
- [YHC⁺18] Young Jong Yoo, Eun Ju Ha, Young Jun Cho, Hee Lim Kim, Minkyung Han, and So Yeon Kang. Computer-aided diagnosis of thyroid nodules via ultrasonography: Initial clinical experience. *Korean Journal of Radiology*, 19(4):665, 2018.
- [YKA⁺22] Rikiya Yamashita, Tara Kapoor, Minhaj Nur Alam, Alfia Galimzianova, Saad Ali Syed, Mete Ugur Akdogan, Emel Alkim, Andrew Louis Wentland, Nikhil Madhuripan, Daniel Goff, Victoria Barbee, Natasha Diba Sheybani, Hersh Sagreiya, Daniel L. Rubin, and Terry S. Desser. Toward Reduction in False-Positive Thyroid Nodule Biopsies with a Deep Learning-based Risk

Stratification System Using US Cine-Clip Images. *Radiology: Artificial Intelligence*, 4(3):e210174, May 2022.

- [YPL⁺19] Shi Yin, Qimmu Peng, Hongming Li, Zhengqiang Zhang, Xinge You, Hangfan Liu, Katherine Fischer, Susan L. Furth, Gregory E. Tasian, and Yong Fan. Multi-instance Deep Learning with Graph Convolutional Neural Networks for Diagnosis of Kidney Diseases Using Ultrasound Imaging. *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging and Clinical Image-based Procedures*, 11840:146–154, October 2019.
- [YPM⁺18] Moi Hoon Yap, Gerard Pons, Joan Martí, Sergi Ganau, Melcior Sentís, Reyer Zwiggelaar, Adrian K. Davison, and Robert Martí. Automated Breast Ultrasound Lesions Detection Using Convolutional Neural Networks. *IEEE Journal of Biomedical and Health Informatics*, 22(4):1218–1226, July 2018.
- [YPW⁺23] Jichen Yang, Lacey C. Page, Lara Wagner, Benjamin Wildman-Tobriner, Lily Bisset, Donald Frush, and Maciej A. Mazurowski. Thyroid nodules on ultrasound in children and young adults: Comparison of diagnostic performance of radiologists’ impressions, ACR TI-RADS, and a deep learning algorithm. *American Journal of Roentgenology*, 220(3):408–417, 2023.
- [YSW⁺21] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. MedMNIST v2 – A large-scale lightweight benchmark for 2D and 3D biomedical image classification, October 2021.
- [YWL⁺22] Pengyi Yu, Xiantao Wu, Junhao Li, Ning Mao, Haicheng Zhang, Guangjia Zheng, Xiao Han, Lin Dong, Kun Che, Qi Wang, Gang Li, Yongli Mou, and Xicheng Song. Extrathyroidal extension prediction of papillary thyroid can-

- cer with computed tomography based radiomics nomogram: A multicenter study. *Frontiers in Endocrinology*, 13:874396, 2022.
- [YYD⁺21] Zaoming Yan, Bailin Yang, Fangfang Dong, Meiying Yan, Dong Xu, Changrui Zhu, Guohua Cheng, and Linyang He. Ultrasound-based thyroid follicular neoplasm classification using fine-tuning and graph convolutional network. In *Twelfth International Conference on Graphics and Image Processing (ICGIP 2020)*, volume 11720, pages 493–499. SPIE, January 2021.
- [YYY⁺20] Bin Yang, Meiying Yan, Zaoming Yan, Changrui Zhu, Dong Xu, and Fangfang Dong. Segmentation and classification of thyroid follicular neoplasm using cascaded convolutional neural network. *Physics in Medicine & Biology*, 65(24):245040, 2020.
- [ZAJ⁺21] Yong-Chen Zhu, Ayman AlZoubi, Sabah Jassim, Quan Jiang, Yi Zhang, Ya-Bing Wang, Xiang-Dong Ye, and Hang Du. A generic deep learning framework to classify thyroid and breast lesions in ultrasound images. *Ultrasonics*, 110:106300, 2021.
- [ZCR⁺25] Xiaoyu Zheng, Xu Chen, Awais Rauf, Qifan Fu, Benedetta Monosi, Felice Rivellese, Myles J. Lewis, Shaogang Gong, and Gregory Slabaugh. OpenUS: A Fully Open-Source Foundation Model for Ultrasound Image Analysis via Self-Adaptive Masked Contrastive Learning, November 2025. arXiv:2511.11510 [cs.CV].
- [ZCY⁺22] Shi-Xuan Zhao, Yang Chen, Kai-Fu Yang, Yan Luo, Bu-Yun Ma, and Yong-Jie Li. A Local and Global Feature Disentangled Network: Toward Classification of Benign-Malignant Thyroid Nodules From Ultrasound Image. *IEEE Transactions on Medical Imaging*, 41(6):1497–1509, June 2022.

- [ZDJ⁺20] Shu Zhang, Hang Du, Zhenhua Jin, Yuchao Zhu, Yi Zhang, Fan Xie, Ming Zhang, Xin Tian, Jie Zhang, and Yu Luo. A novel interpretable computer-aided diagnosis system of thyroid nodules on ultrasound based on clinical experience. *IEEE Access*, 8:53223–53231, 2020.
- [ZDJ⁺22] Yong Zhu, Hang Du, Quan Jiang, Tao Zhang, Xinru Huang, Yi Zhang, Xiaoming Shi, Jiajia Shan, and Ayman AlZoubi. Machine learning assisted doppler features for enhancing thyroid cancer diagnosis: A multi-cohort study. *Journal of Ultrasound in Medicine*, 41(8):1961–1974, 2022.
- [ZF21] Xiongfeng Zhu and Qianjin Feng. MVC-NET: Multi-View Chest Radiograph Classification Network With Deep Fusion. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 554–558, Nice, France, April 2021. IEEE.
- [ZIF⁺24] Luoting Zhuang, Vedrana Ivezic, Jeffrey Feng, Chushu Shen, Ashwath Radhachandran, Vivek Sant, Maitraya Patel, Rinat Masamed, Corey Arnold, and William Speier. Patient-level thyroid cancer classification using attention multiple instance learning on fused multi-scale ultrasound image features. *AMIA Annual Symposium Proceedings*, 2023:1344–1353, January 2024.
- [ZJB⁺21] Yong-Chen Zhu, Peng-Fei Jin, Juan Bao, Quan Jiang, and Xin Wang. Thyroid ultrasound image classification using a convolutional neural network. *Annals of Translational Medicine*, 9(20):1526, 2021.
- [ZJD⁺20] Hao Zhou, Yueming Jin, Lei Dai, Min Zhang, Yao Qiu, Kun Wang, Jie Tian, and Jian Zheng. Differential diagnosis of benign and malignant thyroid nodules using deep learning radiomics of thyroid ultrasound images. *European Journal of Radiology*, 127:108992, 2020.

- [ZLB⁺23] Qi Zhao, Shuchang Lyu, Wenpei Bai, Linghan Cai, Binghao Liu, Guangliang Cheng, Meijing Wu, Xiubo Sang, Min Yang, and Lijiang Chen. Mmotu: A multi-modality ovarian tumor ultrasound image dataset for unsupervised cross-domain semantic segmentation, 2023.
- [ZLH⁺25] Huan Zhang, Qianglin Liu, Xiaolin Han, Lijuan Niu, and Weidong Sun. TN5000: An ultrasound image dataset for thyroid nodule detection and classification. *Scientific Data*, 12(1):1437, 8 2025.
- [ZLR⁺22] Xinyu Zhang, Vincent C. S. Lee, Jia Rong, Feng Liu, and Haoyu Kong. Multi-channel convolutional neural network architectures for thyroid cancer detection. *PLOS ONE*, 17(1):e0262128, 2022.
- [ZSS⁺22] Yue Zou, Yi Shi, Fangyuan Sun, Jianfeng Liu, Yanying Guo, Haijun Zhang, Xin Lu, Yang Gong, and Sheng Xia. Extreme gradient boosting model to assess risk of central cervical lymph node metastasis in patients with papillary thyroid carcinoma: Individual prediction using SHapley Additive exPlanations. *Computer Methods and Programs in Biomedicine*, 225:107038, 2022.
- [ZSTL18] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. UNet++: A nested U-Net architecture for medical image segmentation, 2018.
- [ZWCW20] Yi Zhang, Qi Wu, Yao Chen, and Yi Wang. A clinical assessment of an ultrasound computer-aided diagnosis system in differentiating thyroid nodules with radiologists of different diagnostic experience. *Frontiers in Oncology*, 10:557169, 2020.
- [ZWT20] Hao Zhou, Kun Wang, and Jie Tian. Online transfer learning for differential

- diagnosis of benign and malignant thyroid nodules with ultrasound images. *IEEE Transactions on Biomedical Engineering*, 67(10):2773–2780, 2020.
- [ZWZ⁺25] Hongyuan Zhang, Yuheng Wu, Mingyang Zhao, Zhiwei Chen, Rebecca Li, Fei Zhu, Haohan Zhao, Xiaohua Yuan, Meng Yang, et al. A Fully Open and Generalizable Foundation Model for Ultrasound Clinical Applications, September 2025. arXiv:2509.11752 [cs].
- [ZXU⁺23] Sheng Zhang, Yanbo Xu, Naoto Usuyama, et al. BiomedCLIP: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs, 2023.
- [ZXU⁺25] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, et al. A Multimodal Biomedical Foundation Model Trained from Fifteen Million Image–Text Pairs. *NEJM AI*, 2(1):AIoa2400640, January 2025.
- [ZZL⁺21] Qi Zhang, Shu Zhang, Jie Li, Yixin Pan, Jun Zhao, Yuhao Feng, Yanfang Zhao, Xiaoya Wang, Zhiyun Zheng, Xiaowei Yang, Liwei Liu, Chunlai Qin, Ke Zhao, Xinxin Liu, Chunbao Li, Long Zhang, Chuanqi Yang, Naiqian Zhuo, Haozheng Zhang, Jinling Liu, Jie Gao, Xingyu Di, Fanwei Meng, Weinan Ji, Meng Yang, Xin Xin, Xun Wei, Rong Jin, Lan Zhang, Xiaolong Wang, Fei Song, Xueying Zheng, Ming Gao, Ke Chen, and Xin Li. Improved diagnosis of thyroid cancer aided with deep learning applied to sonographic text reports: a retrospective, multi-cohort, diagnostic study. *Cancer Biology & Medicine*, 19(5):733–741, 2021.
- [ZZY⁺21] Jiacheng Zhu, Shu Zhang, Ruiguo Yu, Zhikang Liu, Hantong Gao, Bo Yue, Xinxin Liu, Xueying Zheng, Ming Gao, and Xun Wei. An efficient deep convolutional neural network model for visual localization and automatic

diagnosis of thyroid nodules on ultrasound images. *Quantitative Imaging in Medicine and Surgery*, 11(4):1368–1380, 2021.