THE UNIVERSITY OF CHICAGO

STUDY OF MACHINE LEARNING ALGORITHMS FOR DIAGNOSIS OF INDETERMINATE THYROID NODULES ON ULTRASOUND

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

INTERDISCIPLINARY SCIENTIST TRAINING PROGRAM:
MEDICAL PHYSICS

BY

JOSEPH L COZZI

CHICAGO, ILLINOIS

AUGUST 2025

To my parents, Darlene and Phil, for their boundless support and lifelong love of learning.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

As I reflect on the past few years, I am so grateful for and incredibly indebted to the many people who have made this dissertation possible.

First and foremost, I'd like to thank my thesis advisor Maryellen Giger. She not only served as a wonderful research mentor pushing me to develop the skills required for independent investigation, but also demonstrated daily the kindness, understanding, and collaborative spirit required to sustain a successful and happy career. It has been one of the greatest privileges of my life to be a member of Giger Lab. As such, I want to thank the many members of the lab who generously offered their the time and expertise to contribute to the development of this work: Hui Li, Karen Drukker, Jordan Furhman, Li Lan, Heather Whitney, Madeleine Torcasso, Sasha Edwards, John Papaioannou, and Chun Wai Chan.

Thank you to Xavier Keutgen whose clinical expertise and passion for patient care made this work possible. From building inter-institutional collaborations to data analysis he demonstrated the incredible impact a physician-scientist can make. His mentorship and contributions to this work were invaluable.

Thank you to the members of my thesis committee members, Zheng Feng Lu, Sam Armato, Hui Li. Your guidance and advice has been priceless. Thank you for your continual support of both my research and professional development. I could not be more grateful to have had this group of phenomenal mentors. Thank you as well to my UChicago research collaborators, Julian Conn Busch, Jelani Williams, and Josh Genender, who made substantial research contributions to the work presented in this thesis. In addition, I'd like to thank our thyroid project collaborators at Weill Cornell Medical Center and University of Illinois

Health.

Thank you to the MSTP and GPMP administration without whom the entirety of my time at UChicago would not be possible. I'd especially like to thank Marcus Clark and Alison Anastasio for constantly putting the success and well-being of students first. I'd also like to thank my co-students in both programs for taking on graduate school with me. I'd especially like to thank Joel Toledo, Jason Meier, Geneva Anderberg, Nikki Maheshwari, Randy Melanson, Kevin Chang, Evan Neczypor, and Peter Hahn.

Last but certainly not least, I'd like to thank my friends and family who have provided endless support over shared meals and late nights. To my girlfriend Samantha, thanks for listening to daily research updates and always being there for me. To my parents, Phil and Darlene, and my siblings, Gabriella, Francesca, and Phillip, thank you for all your love and encouragement. I could not have done this without you all.

ABSTRACT

Thyroid nodules are common and seen in over half the US population on autopsy. Though most thyroid nodules are benign, some do represent thyroid cancer and may be treated through thyroidectomy. Nodule risk-assessment begins first with radiologic evaluation on ultrasound. If indicated, nodule diagnosis continues with fine-needle aspiration (FNA) biopsy through which nodules can be grouped as benign, malignant, or indeterminate. Indeterminate thyroid nodules (ITNs), those neither conclusively benign nor malignant on FNA, make up roughly 30% of the nodules biopsied. Even with additional molecular testing performed, nearly half of all ITNs surgically excised are determined to have been benign on post-surgical pathology. There is a clinical need for additional diagnostic tools which aid in the classification of ITNs and reduce the number of 'false-positive' thyroid surgeries.

This dissertation presents the development of artificial intelligence (AI) algorithms for the characterization and classification of indeterminate thyroid nodules on ultrasound. AI algorithms have the potential to segment and classify lesions on ultrasound; however, previous work in the field has focused primarily on FNA-determinate nodules. Therefore, the segmentation and classification of FNA-indeterminate nodules still present a large area of need within the current diagnosis paradigm. Further, as the role of thyroid nodule molecular testing platforms has expanded, there is an increased need for study into the integration of image-based AIs with molecular tests. This dissertation is separated into three primary studies conducted on a multi-institutional dataset.

Segmentation AI for Thyroid Nodules on Ultrasound: An attention-enhanced U-Net algorithm was trained and independently tested for nodule segmentation performance on

ultrasound. Additionally, a fuzzy c-means unsupervised algorithm was validated for internal segmentation of mixed thyroid nodules into solid and cystic components. Results of this study demonstrated high performance of the deep learning segmentation algorithm across all nodule pathologies.

Classification AI for Indeterminate Thyroid Nodules on Ultrasound: Both radiomic and deep transfer learning classification algorithms for ITNs on US were developed. Multiple pre-training and feature selection approaches were investigated to improve algorithm performance. The impact of our automatic segmentation algorithms on radiomic algorithm performance was investigated. Finally, the integration of radiomic and deep learning transfer algorithms was explored for use in tandem with FNA biopsy and molecular testing. Results of this investigation showed a meaningful increase in surgical pathology classification performance when image-based AIs are merged with pre-surgical clinical pathology tests.

Reference-Standard-Free Lesion Segmentation Performance Metric: The quality of nodule segmentation can impact feature calculation and downstream classification. Current segmentation performance metrics require the use of a reference-standard; however, this limits their use once an algorithm has been deployed. In this study, a reference-standard-free segmentation metric, lesion segmentation-sureness (LeSS) score, was developed and independently tested. LeSS was applied to AttU-Net of the first study and the radiomic model of the second study as a test case for its application. The results of this study showed high correlation between LeSS and traditional segmentation performance metrics. It also demonstrated that LeSS analysis could be used to improve segmentation performance and the calculation of two radiomic features.

CHAPTER 1

INTRODUCTION

1.1 Thyroid Cancer Prevalence and Patient Impact

Thyroid nodules are common and may represent a risk of cancer.[1, 2] Thyroid nodules have been reported in >50% of the American population and are most often found incidentally on ultrasound when still asymptomatic.[3, 4] While more than half the population has a thyroid nodule, the American Thyroid Association (ATA) estimates that only 5% of detected nodules represent a possibility of thyroid cancer. These cases are recommended for further diagnostic workup through ultrasound, biopsy, and potentially molecular testing. Thyroid cancer is the 12th most prevalent cancer diagnosis in the USA and has been one of the fastest growing cancers worldwide.[5] Thyroid cancer is 2.9-times more common in women than men and tends to occur earlier in life than most other forms of cancer.[6] The average age of diagnosis is 51 years.[7]

In 2025, the National Cancer Institute projects 44,020 new cases of thyroid cancer in the US with 2,290 associated deaths.[5, 8] In 2020, roughly 1.5 billion USD was spent on the treatment and continued care of well-differentiated thyroid cancer in the US and that is projected to increase to 3.5 billion USD by 2030.[9]

Surgical resection of the entire thyroid (thyroidectomy) or exclusively the affected lobe (lobectomy) are both highly successful treatment options with overall 5-year survival rates of >95%.[10] While thyroid cancer recurrence is possible after lobectomy, there are post-surgical quality of life factors to consider.[10] All patients who undergo total thyroidectomy

require lifelong thyroid hormone supplementation. In contrast, roughly a quarter of patients who undergo lobectomy do not require additional thyroid hormone supplementation due to the functioning of the remaining lobe.[11]

The comparatively early onset of thyroid cancer and need for continuous care after surgery can lead to extended financial and psychological burden. In one study, 46.1% of thyroid cancer survivors reported a psychological-financial burden while only 24.0% of non-thyroid cancer survivors reported the same.[9] A false-positive thyroid cancer case occurs when a nodule is deemed suspicious for cancer, surgically removed, and then re-classified as benign on post-surgical pathology. Of all nodules surgically removed, more than half are false-positive cases.[12, 13] While cancer-free, these patients, still endured the psychological trauma of receiving a cancer diagnosis and maintain the financial burdens of thyroid surgery and hormone replacement therapy. The overall goal of this thesis is the development of machine learning tools for the automatic segmentation and classification of indeterminate thyroid nodules, specifically aimed at reducing the number of benign thyroid nodules sent for surgical excision.

1.2 Clinical Diagnosis of Thyroid Cancer

1.2.1 Ultrasound Evaluation of Nodules - TIRADS

The guidelines diagnosis and management of thyroid nodules are set by the American Thyroid Association (ATA).[14] The first step in the clinical diagnosis of thyroid nodules is a thyroid ultrasound (US) exam.[15, 16] Because thyroid nodules are so prevalent and

the majority are not of clinical concern, ultrasound is well suited for fast and low-cost initial assessment. For these reasons and its lack of ionizing radiation, B-mode ultrasound examinations remain most used despite the higher image quality available with thyroid CT and MRI.[15] Both transverse and longitudinal views of the nodule are taken during thyroid evaluation US exams.

The Thyroid Imaging, Reporting, and Data System (TIRADS), developed by the American College of Radiology (ACR) and first published in 2017, is the current gold standard for nodule risk-stratification on ultrasound.[17] While other ultrasound risk stratification systems are used internationally, such as K-TIRADS and EU-TIRADS, ACR's TIRADS is used in the USA.[18, 19] To calculate a TIRADS score, a radiologist grades the nodule according to 5 diagnostic categories: Composition, Echogenicity, Shape, Margin, and Echogenic Foci. Points assigned across each category are then summed. The TIRADS score and nodule size determine if a nodule ought to undergo biopsy. A summary of the TIRADS score calculation is provided in Figure 1.1.

Overall, larger solid nodules with irregular borders are the most suspicious according to TIRADS. It is important to note that the taller-than-wide designation in the Shape category must be calculated in the transverse view. A high TIRADS score does not explicitly inform nodule pathology, but exclusively determines if nodule biopsy through fine needle aspiration (FNA) is recommended.

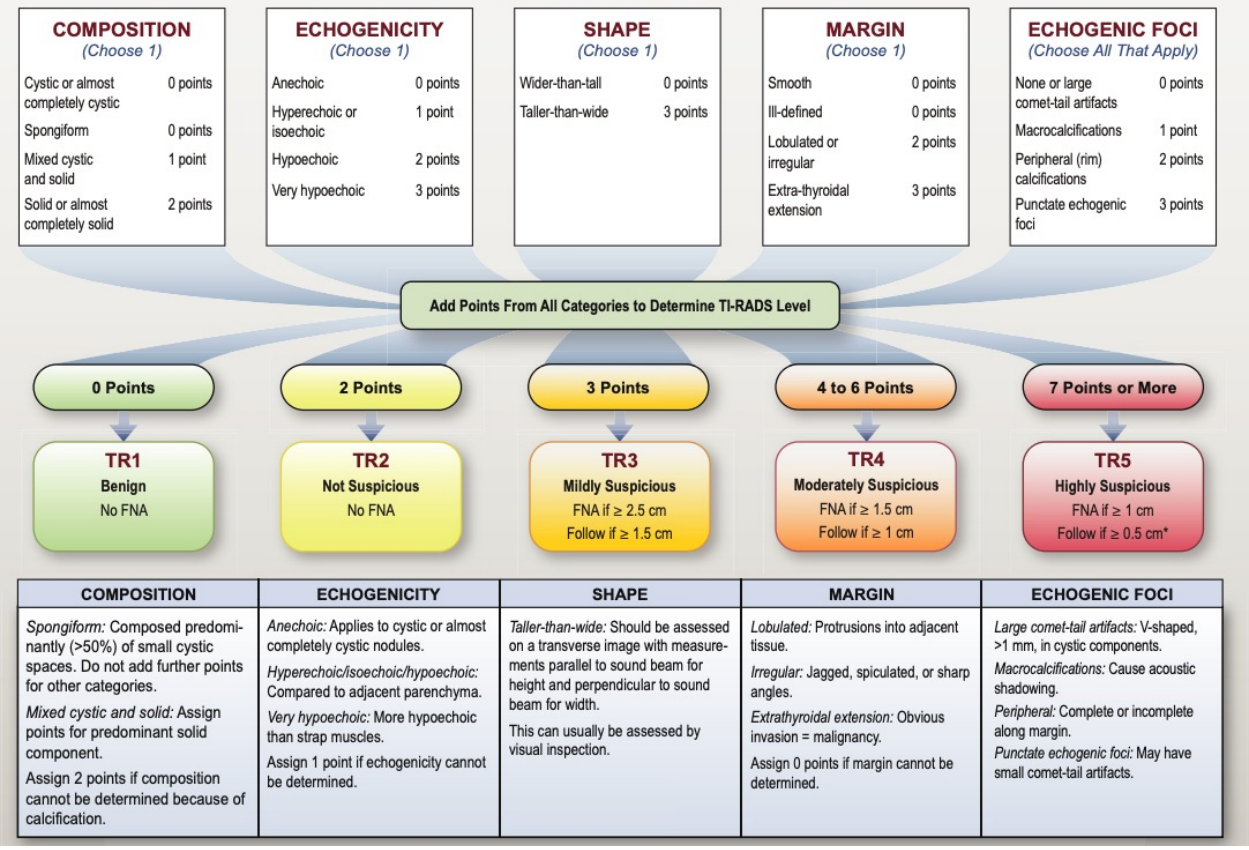

Figure 1.1: TIRADS scoring criteria published in the ACR committee's White Paper.[17]

1.2.2 FNA Biopsy - Bethesda Classification System

If a nodule is recommended by TIRADS for FNA biopsy, specimen collection and assessment can be performed bedside.[14] There are an estimated 600,000 thyroid FNAs each year in the USA.[20, 21] The results of the FNA biopsy are grouped according to the Bethesda System for Reporting Thyroid Cytopathology (Bethesda System).[22] While the Bethesda System was first formalized in 2010 by the National Cancer Institute (NCI), it has undergone updates in 2017 and 2023.[23, 24] According to the Bethesda System, nodules are grouped in six categories with Category I representing non-diagnostic samples, and Categories II, III, IV, V and VI representing increasing malignancy risk. The Bethesda Categories and their

expected malignancy rates are shown in Table 1.1. Nodules grouped in Bethesda Categories III, IV, and V are deemed indeterminate, neither confidently benign nor confidently malignant, and account for roughly a third of all nodules biopsied.[25] Because indeterminate cases represent a risk of cancer, they can be recommended for surgical excision or additional molecular testing. Surgical pathology determines nearly 75% of indeterminate cases sent for excision, without molecular testing, are benign, i.e., false-positives.[13] Therefore, the limitations of our current diagnostic procedures result a low positive predicative value (PPV) for indeterminate nodules.

Bethesda Category	Cytopathologic Description	Frequency	Rate of Malignancy	Clinical Course of Action
I	Non-diagnostic	5-11%	1-4%	Repeat FNA
II	Benign	55-74%	0-3%	Follow
III	Atypia or follicular lesion of undetermined significance	5-15%	5-15%	Molecular Test or Surgical Excision
IV	Suspicious for /or follicular neoplasm	2-25%	15-30%	Molecular Test or Surgical Excision
V	Suspicious for malignancy	1-6%	60-75%	Molecular Test or Surgical Excision
VI	Malignant	2-5%	97-99%	Surgical Excision

Table 1.1: Review of Bethesda categories, their frequency and rate of malignancy, and the clinical action recommended. Table adapted from Kargi et al. [13] Indeterminate categories are shaded.

1.2.3 Molecular Testing

Indeterminate thyroid nodules may also be sent for molecular testing for further evaluation. Thyroid nodule molecular tests are primarily conducted using three platforms: Afirma, ThyroSeq, and ThyGeNext.[26] Afirma, ThyroSeq, and ThyGeNext, provide the physician with a list of genetic markers found, as well as a specialized rate of malignancy (ROM) for that set markers. Note, each molecular testing platform looks at different and confidential sets of molecular and gene mutation panels. While thyroid molecular tests are effective at ruling out cancer, the majority of nodules sent for molecular testing are still deemed suspicious by the molecular testing platforms.[26, 27] Studies report low real-world positive predictive values (PPVs) of 48%, 53%, 67% for Afirma, ThyroSeq V2, and ThyGeNext, respectively.[28, 29, 30] Due to the low PPV and additional financial cost, patients may elect to forgo molecular testing and directly undergo thyroid surgery.

1.2.4 Overview of Diagnostic Pathway

The pathway for the clinical diagnosis of thyroid nodules is presented includes ultrasound evaluation of nodules, fine-needle aspiration (FNA) biopsy of nodules, and the potential use of molecular testing. A schematic of the clinical diagnosis pathway and management of thyroid nodules is shown in Figure 1.2.

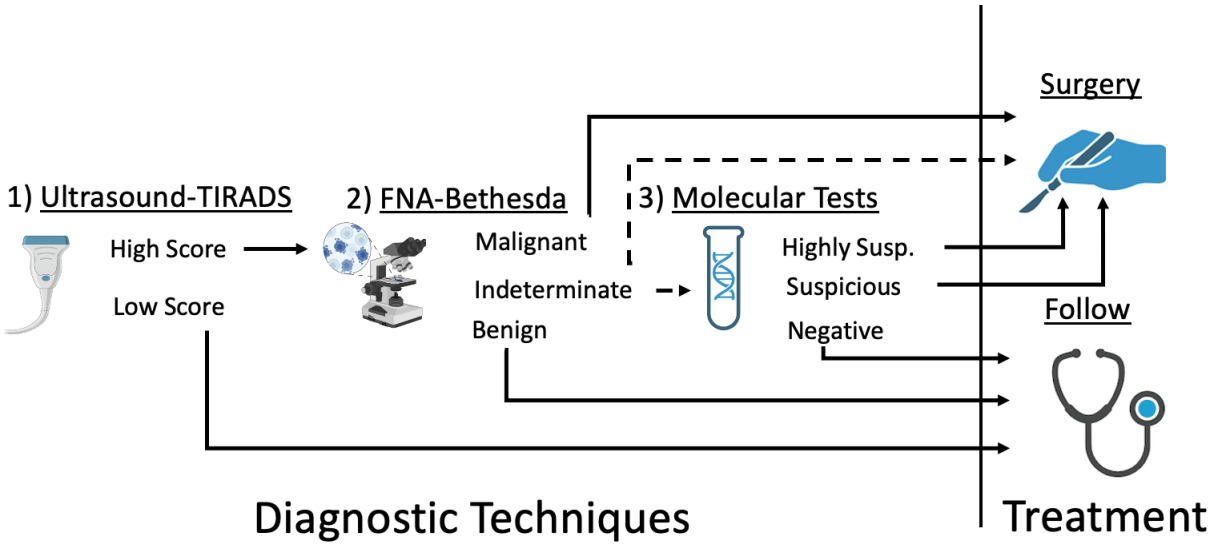

Figure 1.2: Clinical diagnosis and management of thyroid nodules. Dashed arrows indicate indeterminate nodules may either be sent for molecular testing or sent to surgery without further diagnostic workup.

1.3 Literature Review of Machine Learning in Differentiating Benign from Malignant Indeterminate Thyroid Nodules

1.3.1 Review Introduction

Computer-aided diagnosis (CADx) systems utilizing machine learning (ML) have become a focus of research due to their potential to enhance the interpretation and diagnosis of various diseases.[31, 32, 33, 34, 35, 36, 37] Artificial intelligence (AI)-based CADx systems continue to be developed in an attempt to increase diagnostic accuracy with a particular emphasis on the diagnosis of various cancers. Two main subcategories of machine learning used in CADx include radiomic machine learning, based on human-engineered features identified in a region of interest (ROI) or segmented lesion, and deep learning (DL), which uses neural networks to learn and optimize image features across large datasets.[37] AI-based

CADx has the potential to enhance thyroid nodule diagnosis and possibly lead to decreasing the number of invasive tests and unnecessary surgeries.

Many thyroid cancer artificial intelligence articles have been published, including original research, commentaries, and reviews. Systematic reviews have primarily focused on presenting machine learning techniques to a medical audience and describing the use of specific commercially available CADx systems.[37, 38, 39, 40] Among these, there seems to be a consensus that further data, via application of ML to real-world scenarios, need to be gathered in order to draw definitive conclusions about the role of CADx in differentiating benign from malignant thyroid nodules. However, there has not, to our knowledge, been a previous summarization of the current state of ML research evaluating specifically the diagnosis of indeterminate thyroid nodules.

A systematic review was conducted in order to provide a renewed commentary on the field of ML and thyroid cancer diagnosis, specifically analyzing the utility of ML in improving the accuracy of indeterminate thyroid nodules diagnosis. This review was published as "Role of machine learning in differentiating benign from malignant indeterminate thyroid nodules: A literature review" in Health Sciences Review.[41] Additionally, this review contributed to the motivation of the three research studies detailed in this dissertation.

1.3.2 Methods

1.3.2.1 Search Strategy and Selection Criteria

Presented here is our systematic review of published literature on the use of machine learning for the diagnosis of indeterminate thyroid nodules, and we used PRISMA guidelines in the reporting of results.[42] PubMed was used in a literature search on June 21, 2022.

Published articles on the use of ML models for diagnosis of thyroid nodules were selected using the following search query: 1# thyroid cancer OR thyroid nodule [All Fields] 2# deep learning OR machine learning OR radiomics OR artificial intelligence OR computer-aided diagnosis [All Fields]. After screening with title and abstract, remaining articles were assessed individually based on full-text review.

Inclusion criteria were as follows: (1) publication of the study as a full-text manuscript in a peer-reviewed journal, (2) original work (e.g., not a review). Studies were excluded if they (1) did not perform FNA biopsy, (2) excluded all indeterminate lesions, (3) focused on a non-cancer thyroid pathology, (4) described only non-diagnostic uses of their model(s) (e.g. calcification quantification), or (5) used an imaging modality other than grayscale ultrasound (e.g. color Doppler US). We did not place language or date restrictions on the search.

1.3.2.2 Quality Assessment and Data Extraction

Overall quality of the studies included in this review was critically appraised based on the 10-item Joanna Briggs Institute (JBI) Appraisal Checklist for Diagnostic Test Accuracy Studies to assess the quality of included studies.[43, 44] The JBI checklist examines two primary domains: patient selection and index test. Data were extracted and recorded using Excel 2016 (Microsoft Corporation, Redmond, CA). Information collected includes study description (e.g., title, author, year of publication, country of origin), study design (prospective or retrospective), methods characteristics (e.g., type of model used, method of nodule diagnosis), and basic information about the results of the study (e.g., sensitivity, specificity).

1.3.3 Results

1.3.3.1 Literature Search and Selection of Studies

Initial literature search yielded 1215 total articles. We excluded duplicates, unrelated topics, unoriginal work (e.g., reviews), non-full text publications (e.g., correspondences), which left 51 articles for consideration. Of these 51, 6 studies met eligibility criteria and were included in the systematic review.[45, 46, 47, 48, 49, 50] A summary of the search results can be found in Figure 1.3.

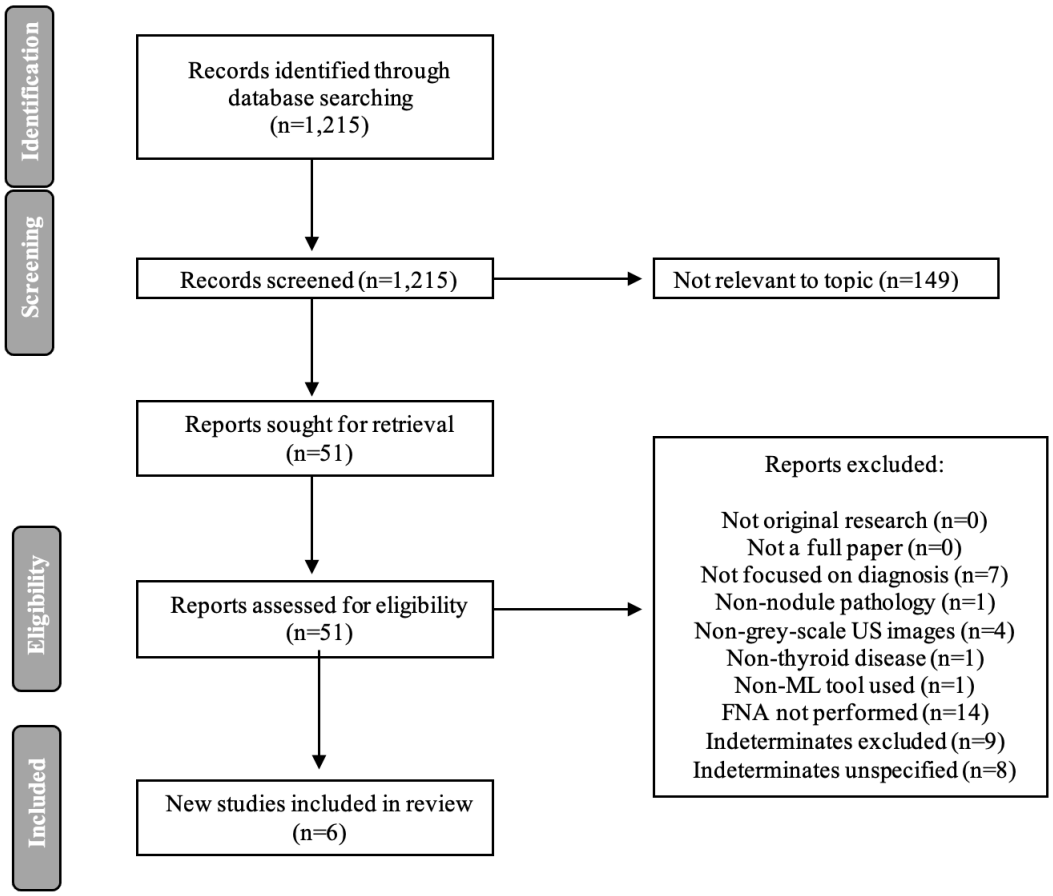

Figure 1.3: Results of article selection procedure for review.[41]

1.3.3.2 Study Attributes

All of the articles eventually selected for review were published after 2019. Studies were

performed in China, Korea, Italy, and USA. All studies were performed at academic centers and were retrospective in design. One study used the commercial machine learning system S-Detect (Samsung Medison Co., Seoul, South Korea).[45] One study used a texture-analysis radiomic model.[50] Four studies compared their algorithms to individual physicians, primarily radiologists.[45, 47, 48, 49] One study compared algorithm performance to a consensus diagnosis made by three physicians together.[46] Final pathology of thyroid nodules was determined by FNA, surgical pathology, or a combination.

1.3.3.3 Study Quality and Bias Assessment

Two studies, Ouyang et al. and Gitto et al. did not use consistent reference standards for all patients.[46, 45] In these studies, researchers used a compilation of FNA and/or surgical pathology for final diagnosis. Indeterminate nodules without surgical pathology were excluded, while benign nodules may be confirmed by observation of lack of growth and/or decrease in size within several months. Malignant nodules were confirmed with surgical pathology. A third study, while including indeterminate nodules, did not specify the number of Bethesda III, IV, and V cases included.[49] Therefore, only three studies could be properly reviewed and evaluated for their uses of machine learning to classify indeterminate nodules. A summary of the cytopathology designations for each study is included in Table 1.2.

1.3.4 Discussion

This systematic review suggests an increase in the number of articles in recent years, likely reflecting both increasing availability of machine learning technology and increased interest in using that technology for diagnosis of thyroid cancer, however only a few articles addressed the more clinically significant issue of indeterminate thyroid lesions. This review

	I	II	III	IV	V	VI	Total
Gitto et al	4	44	12	1	1	0	62
Ouyang et al		92	344	119	624		1169
Zhu et al			128	43	101	195	467
Youn et al			202				202
Zhao et al	Unspecified	Unspecified	Unspecified	Unspecified	Unspecified	Unspecified	849
Keutgen et al		90	77	29	27	70	302
Total	4	136	686	163	726	195	3051

Table 1.2: Distribution of Bethesda categories in articles reviewed.

has identified several trends in the current literature that can help guide the field forward, including the use of commercially available CADx systems and the studies describing the practicality of ML-assisted approaches. However, there are significant shortcomings that warrant further investigation, including inconsistent grouping of indeterminate nodules and a lack of studies including indeterminate nodules.

Our search resulted in only three articles specifically studying the classification of indeterminate nodules, reflecting a significant gap in the current body of thyroid cancer-ML literature.[47, 48, 50] Although Zhu et al. focused on indeterminate nodules, the goal of their study was to demonstrate that a ML model could differentiate between Bethesda Category III nodules and a combined Category IV/V/VI, rather than determine the surgical pathology of indeterminate thyroid nodules.[47] They grouped ultrasound images of Bethesda Category III (AUS/FLUS) nodules ("indeterminate") and Category IV/V/VI ("malignant"), then compared the ability of five separate machine learning models (logistic regression, gradient boost, support vector machine (SVM), random forest, and deep neural network (DNN)) to differentiate between these two groups. They found that DNN had the highest accuracy and specificity in determining whether ultrasound images were of Bethesda III or Bethesda IV/V/VI nodules. However, this approach is distant from the clinical classification of nod-

ules and the decision-making behind surgical procedures, which decreases its clinical utility.

Youn et al. described a study comparing the ability of a deep convolutional neural network (DCNN), a type of deep learning, with that of 8 physicians (4 radiologists, 4 endocrinologists) to determine the true pathology AUS/FLUS nodules.[48] Their study revealed that the diagnostic performance of the DCNN and physicians were not statistically significantly different. Though previous studies have demonstrated the usefulness of ultrasound evaluation in differentiating AUS/FLUS nodules into benign or malignant designations, this study applied machine learning techniques and compared them to the abilities of a handful of physicians.[51, 52] Youn et al. suggested the possibility of the application of previously held knowledge of ultrasound images to the emerging technology of ML. However, this study was limited to Bethesda Category III nodules, not the totality of indeterminate nodules (III/IV/V). We found the designation of Bethesda Category III as indeterminate and Category IV/V/VI as malignant to be inconsistent with the most recent guidelines from the ATA5, which designates Category III/IV/V as indeterminate and VI as malignant.[14] Zhu et al.'s and Youn et al.'s unique assignment of cytopathic subtypes limits these studies use in the USA where the ATA's guidelines are the standard of practice. Keutgen et al. used a radiomics ML model to examine indeterminate thyroid nodules, using a combined Category III/IV/V.[50] This is consistent with the most current ATA definitions, and thus is more clinically relevant. They also demonstrated that their model has similar diagnostic ability to molecular testing, providing evidence that ML may be a comparable and cost-effective alternative to expensive molecular testing. However, they did not compare their model to either physician or physician-assisted approaches, which would provide more directly clini-

cally significant data on their model's practical use.

Prior research has shown that ML-assisted approaches can improve physicians' diagnostic accuracy of thyroid cancer. However, there has not been thorough discussion on how the diagnosis of indeterminate nodules might be impacted by an ML-assisted approach. Zhao et al. is the only study in this review to examine how ML improves physicians' ability to discriminate benign from malignant indeterminate thyroid nodules by utilizing an ML-assisted approach.[49] They found that an ML-assisted US visual approach led to a significantly greater AUC than using an ACR TI-RADS only approach to categorize US images. The ML-assisted approach had identical sensitivity as the ACR TIRADS approach, though specificity of the ML-assisted approach was significantly greater. While TIRADS can provide a point of comparison, it is important to consider TIRADS itself is designed to recommend biopsy and not for direct classification. This study, however, suggests that although ML models are unlikely to replace physicians in the clinical setting, they have tremendous potential to improve physicians' diagnostic ability via an ML-assisted approach.

This review's strength was the breadth of the initial search due to the broadness of our search terms and inclusion criteria. However, there are significant limitations that must be addressed. Firstly, our use of PubMed as the only database may present a bottleneck in information acquisition. Additionally, there is a risk of publication bias when analyzing published studies, as authors often omit non-statistically significant results. We have presented an analysis of study quality and bias assessment to help combat this.

The review has highlighted the limited research conducted into ML for specifically indeterminate nodules. Further research is warranted to explore how CADx can be used as

assistive tools to aid physicians in determining the true nature of indeterminate thyroid nodules, and thus potentially avoid tens of thousands of diagnostic surgeries per year.

1.4 Summary and Dissertation Outline

In Chapter 1, the prevalence of thyroid cancer and impact on patient life were discussed. Next, a summary of the current clinical diagnostic systems for thyroid nodules was given. The over-diagnosis of thyroid nodules as suspicious was shown to result in a large number of benign nodules sent for surgical excision. Finally, a review of previous studies in which machine learning was used to improve the classification of indeterminate thyroid nodules was presented. This review highlighted the gap in literature of machine learning models specifically applied to indeterminate nodules and, in junction with the large clinical need, provided the motivation for the completion of this dissertation.

A summary of the dataset used throughout the dissertation can be found in Chapter 2. This dissertation was constructed with three primary studies. First, the development of segmentation AI for thyroid nodules on ultrasound is presented in Chapter 3. Second, the investigation of classification AIs for thyroid nodules on ultrasound is detailed in Chapter 4. Third, development and testing of a reference-standard-free lesion segmentation metric is presented in Chapter 5. Finally, the conclusions of this dissertation and future directions for the continuation of this work are given in Chapter 6.

CHAPTER 2

DATASET

2.1 Summary of Multi-institutional Cases Collected

2.1.1 Inclusion Criteria

All cases for this dissertation work were collected retrospectively under IRB approval at three institutions: UChicago Medicine (UCM), Weill Cornell Medical Center (WCMC), and University of Illinois Health (UIC). Dataset collection was first initiated at UChicago Medicine. Collaborators at WCMC have contributed cases since 2018 and those at UIC most recently since 2024. Retrospective collection spanned back to 2007, however, 2012 was the first year with multiple indeterminate patients included. All patients included in this study underwent thyroid specific-US and biopsy. Any cases which did not undergo these initial diagnostic exams were excluded. Inclusion criteria for classification tasks further required that all nodules had undergone surgical excision through thyroidectomy or hemithyroidectomy. Surgical pathology was collected and served as reference truth for classification algorithm training and evaluation. If molecular testing was preformed using Afirma, ThyroSeq, or ThyGeNext, the results of that testing were also collected.

2.1.2 Pathology

Pathology designations were named according to the following convention biopsy-surgical pathology i.e., pathology results are noted as cytopathology (FNA) then surgical pathology. This led to four pathology groups: benign-benign (BB), indeterminate-benign (IB), indeterminate-malignant (IM), and malignant-malignant (MM). Initial collection at UCM included not only indeterminate (IB and IM) cases, but also biopsy-determinate cases (BB and MM), which still underwent US, FNA, and surgical excision. While cases with a benign FNA biopsy (Bethesda Category II) may not represent the risks of a cancerous lesion, excision may still be recommended due to their size. A large nodule may impede swallowing, cause discomfort, or be removed for aesthetic reasons. In BB cases, ideally only a lobectomy would have been performed. The BB and MM cases were used in segmentation tasks and as a point of comparison in the classification tasks.

A small number of indeterminate cases, which were sent for molecular testing but did not have a surgical pathology, were also identified. Because these did not meet our classification inclusion criteria, they were exclusively used for segmentation model training, as described in Chapter 3. Only indeterminate cases were collected at WCMC and UIC. As is shown in Figure 2.1, our IB-IM dataset is skewed towards the early 2020s as our collection protocol was refined. Pathological distribution of our dataset is provided in Table 2.1.

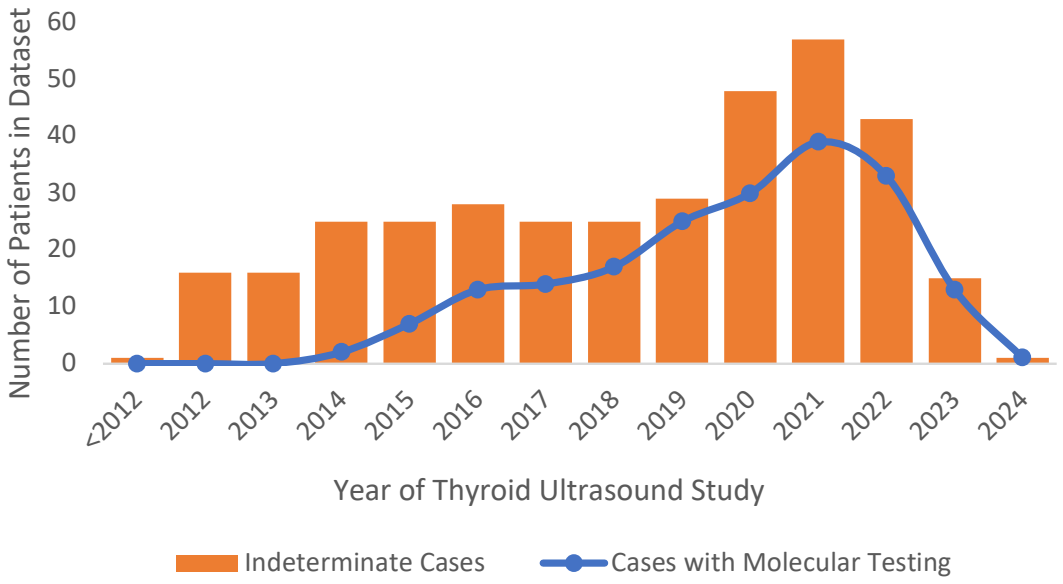

Figure 2.1: Year of treatment for patients with indeterminate nodules included in dataset. The number of those cases which underwent molecular testing is overlaid.

FNA Pathology – Surgical Pathology	UCM			WCMC			UIC		
	Patients	Nodules	Images	Patients	Nodules	Images	Patients	Nodules	Images
Benign-Benign (BB)	111	126	397	0	0	0	0	0	0
Indeterminate-Benign (IB)	131	131	362	65	65	170	36	36	72
Indeterminate-Malignant (IM)	98	100	294	56	56	125	27	27	54
Malignant-Malignant (MM)	99	102	303	0	0	0	0	0	0

Table 2.1: Summary of the pathology of cases collected from each institution.

2.1.3 Image Acquisition

The type and quality of image on which AI systems are trained greatly impacts their performance. As such, ultrasound image acquisition information was extracted. Most nodules were imaged in monochrome (grayscale). Those taken in three-channel color format (RGB) were converted to grayscale using a weighted averaging of each channel (Equation 2.1). [53, 54]

$$Grayscale = 0.298936021293775 * R + 0.587043074451121 * G + 0.114020904255103 * B \quad (2.1)$$

Any image taken with US doppler mode was excluded, and only those deemed diagnostic quality by a team of endocrine surgery physicians were included. Up to 7 images were chosen per nodule. When possible, at least one transverse view image and one longitudinal view image were selected. Over half of the indeterminate ultrasound studies were performed on GE systems, including all UIC cases. Ultrasound pixel size was also extracted for calculation of size features. The distribution ultrasound machine manufacturer for indeterminate images included in this study is shown in Figure 2.2.

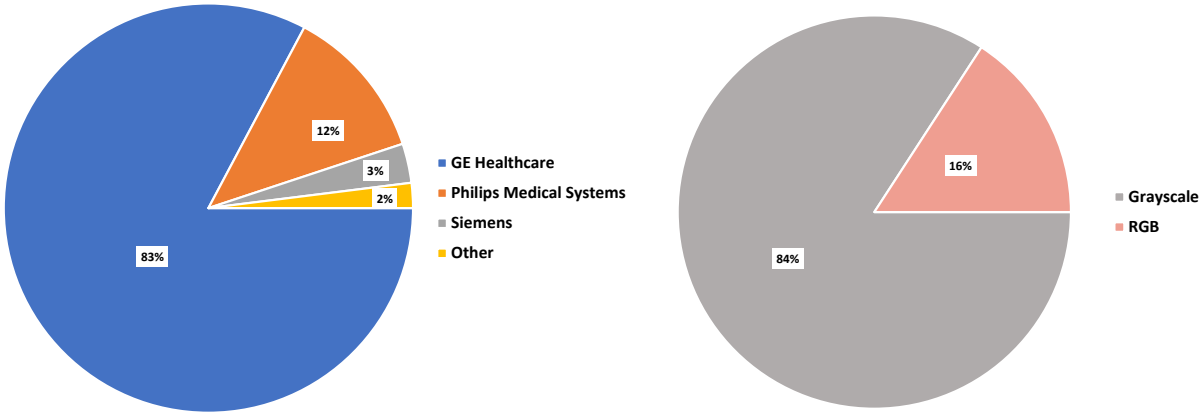

Figure 2.2: Distribution of US image acquisition parameters in dataset: manufacturer (left) and color (right).

2.1.4 Molecular Testing

Because each medical institution offers a different set of molecular testing options, the three-institution collaboration increased the variety of molecular testing platforms. UCM primarily offers ThyGeNext, WCMC primarily uses ThyroSeq, and UIC primarily provides Afirma. The distribution of cases collected with molecular testing is shown in Figure 2.1. In recent years, a higher percentage of cases with molecular testing were included. This is in part because the clinical use of molecular testing has increased. [27] It also is representative of a narrowing of this projects research focus towards the classification of indeterminate thyroid nodules with suspicious molecular testing results (ITN-MS). A summary of the ITN-MS in this data is provided in Table 2.2.

2.2 Data Standardization across Institutions and Molecular Tests

New data collection and integration occurred alongside the development of the AI algorithms (Chapters 3, 4, and 5) of this dissertation. With each institution providing new cases when available, merging information and maintaining a cohesive dataset required establishing a common language framework. Because the Bethesda System 2017 language includes multiple descriptors for Category III (atypia of undetermined significance or follicular lesion of undetermined significance) and Category IV (follicular neoplasm or suspicious for follicular neoplasm), it was particularly imperative that documentation be consistent.[23] When possible, the Bethesda Category number only was shared, and pathology designations (IB

and IM) were validated upon integration.

Further complicating the data collection and use, was that each molecular platform tests for different markers and provides different output statistics as described in Chapter 1. In order to standardize results across the three molecular platforms and cases were split into three categories: Negative, Suspicious, and Highly Suspicious. Inclusion criteria for each category was set in consultation with a UCM endocrine surgeon. Individual designations were made through a combination of mutations identified by the molecular tests and the rates of malignancy (ROM) provided on molecular testing reports. If the ROM was reported as a range, the midpoint was used. For example, a range of malignancy 45%-75% was assigned a 60% ROM. Negative cases (ITN-MN) were those was ROMs $<$10% and/or no mutations. Suspicious cases (ITN-MS) were those with ROMs between 10%-90% and almost all mutations found. Highly suspicious (ITN-MH) cases were those with provided ROMs $>$90% and/or with BRAFV600E mutations, a marker of highly aggressive thyroid cancer, identified.[55] These categories were used in the stratified sampling selection of training and testing cases in Chapter 4.4. The molecular testing cases of most import were those deemed molecular suspicious. Each of the molecular tests had high negative predictive values, however, they still suffered from low positive predictive value (PPV) of the suspicious designation. The dataset's molecular suspicious subset is summarized in Table 2.2 with the PPVs calculated for the molecular testing results of nodules in our dataset.

	Total Dataset	ITN-Molecular Suspicious Subset		
		Afirma	ThyGeNext	ThyroSeq
UCM Cases (Images)	459 (1356)	12 (22)	38 (75)	5 (10)
WCMC Cases (Images)	121 (295)	25 (49)	13 (26)	40 (77)
UIC Cases (Images)	63 (126)	63 (126)	0 (0)	0 (0)
Total Cases (Images)	653 (1777)	100 (197)	51 (101)	45 (87)
PPV of Molecular Tests		0.37	0.55	0.58

Table 2.2: Summary of the Molecular Suspicious Subset. Note the low PPV of the molecular suspicious designation in our dataset.

CHAPTER 3

DEVELOPMENT OF SEGMENTATION AI FOR THYROID NODULES ON ULTRASOUND

3.1 Rationale for Thyroid Nodule Segmentation AI

3.1.1 Introduction

As discussed in Chapter 1, The American College of Radiology (ACR) Thyroid Imaging Reporting and Data System (TIRADS) is commonly used in the United States for evaluating a thyroid nodule's risk of malignancy.[17, 56] It does so based on radiologist review of the nodule characteristics on ultrasound and is used for recommending thyroid nodules to biopsy. Nodules are scored in five categories: composition, echogenicity, shape, margin, and echogenic foci. The decision to perform an FNA is then made by a combination of the summed TIRADS score and the nodule diameter. Accurate whole nodule contouring is needed for consistent TIRADS scoring especially in the categories of shape and margin, and for the calculation of nodule diameter. An automatic segmentation of AI has the potential to aid radiologists in the completion of these diagnostic steps, reduce variability between medical staff, save clinician time, or aid in the development of radiomic machine learning diagnostic systems such as that described in Chapter 4.

The role automatic segmentation could play in the current clinical diagnosis procedure for thyroid cancer is shown in Figure. 3.1.[14, 17, 24, 57] As described in Chapter 2, pathology results are noted as cytopathology (FNA) then surgical pathology, i.e., benign-benign (BB), indeterminate-benign (IB), indeterminate-malignant (IM), malignant-malignant (MM).

An accurate contour of the thyroid nodule on ultrasound is a key component of the clinical diagnosis procedure (Fig. 3.1). Thus, an automatic segmentation of thyroid nodules could aid radiologists in the completion of these diagnostic steps, reduce variability between medical staff, save clinician time, or aid in the development of machine learning diagnostic systems that calculate features based on nodule contours.[50, 41, 59]

This section reports on the development and evaluation of an attention-enhanced automatic segmentation method for thyroid nodules on ultrasound images.

3.2 Attention-enhanced Deep Learning Segmentation of Thyroid Nodules on Ultrasound

The methods and results of Chapter 3.2 were published as "Multi-institutional development and testing of attention-enhanced deep learning segmentation of thyroid nodules on ultrasound" in International Journal of Computer Assisted Radiology and Surgery.[58]

3.2.1 Methods

3.2.1.1 Dataset

The dataset for this study was retrospectively collected and de-identified under Institutional Review Board approval at the University of Chicago Medicine (UCM) and the Weill Cornell Medical Center (WCMC) as described in Chapter 2. A total of 520 patients with thyroid nodules who underwent thyroid ultrasound and FNA at UCM or WCMC were included in this study. Inclusion requirements ensured that the algorithm was trained and tested on images that represent the breadth of thyroid cancer cases of clinical relevance.

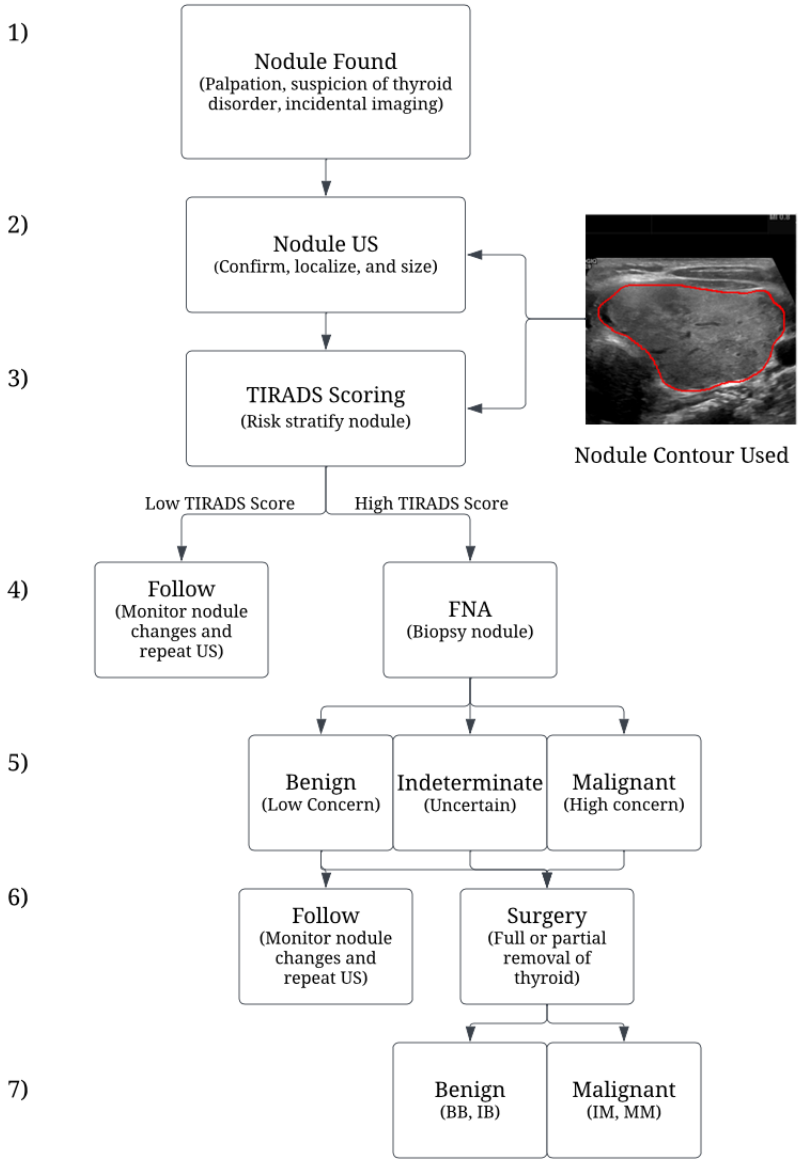

Figure 3.1: Schematic of clinical workflow for diagnosis and treatment of thyroid nodules without molecular testing. A nodule contour is used during steps 2 and 3 to determine the size of nodule as well as calculate a TIRADS score. As used in this dissertation, pathology subgroups written as FNA-surgical pathology include benign-benign (BB), indeterminate-benign (IB), indeterminate-malignant (IM), and malignant-malignant (MM).[58]

All nodules were ≥ 0.6 cm in diameter.[17] UCM and WCMC physicians then manually contoured each image to provide the region of interest (ROI) and establish contour ground truth (reference standard) based on consensus for use in training and testing the segmentation algorithm. In the small subset of ultrasound images in which multiple nodules were visible, only the primary nodule was included for this study. While most ultrasound images were originally acquired as grayscale, a portion of training data contained ultrasound images acquired in RGB color format. Those images were converted to grayscale using weighted averaging via MATLAB's rgb2gray function (Eq. 2.1) in the Image Processing Toolbox.[53, 54]

A total of 1595 images were included for this study and three datasets were constructed, splitting by patient (and not image). The summary of each dataset is presented in Table 3.1. Dataset TV, used for algorithm training and 5-fold cross validation, contained patients from both UCM and WCMC. Datasets T-UCM and T-WCMC, exclusively used for independent algorithm testing, contained only UCM and WCMC patients, respectively.

Name	Patients	Nodules	Images	Institution
Dataset TV	379	398	1334	UCM & WCMC
Dataset T-UCM	84	84	152	UCM
Dataset T-WCMC	57	58	109	WCMC

Table 3.1: Summary table of study datasets: number of patients, nodules, images, purpose, and collection institution. TV designates training and validation dataset. T-UCM and T-WCMC correspond to independent test sets from UCM and WCMC respectively.

3.2.1.2 Image Preprocessing

To explore how image preprocessing affects segmentation performance, algorithms were trained and validated on Dataset TV at three ROI preprocessing levels: un-cropped (UN), cropped and padded (CP), and cropped and resized (CR). An example image demonstrating the three image preprocessing levels is presented in Figure 3.2. Due to the shape of filters in our AttU-Net's hidden layers, image dimensions were required to be multiples of 16; images which did not meet this requirement were 'padded' with a border of zero-value pixels until the dimensions were a multiple of 16.[60]

UN images (example Fig. 3.2A) underwent the least amount of preprocessing only having padding added when necessary. CP images (example Fig. 3.2B) were first cropped to a defined region of interest (ROI) then padded when necessary to meet image dimension requirements. To determine the ROI, the smallest square encompassing the full nodule was calculated using the physician's contour; then, an additional 32 pixels of image were included on all sides of the bounding square. In cases where the nodule neared the image edge, <32 additional pixels were included outside the bounding square. This technique mimics the click and drag ROI selection user-interface envisioned for clinical application. After padding, CP images had varying square dimensions. CR images (example Fig. 3.2C) underwent the most preprocessing. They were first cropped to a defined ROI and then all images were resized to a uniform 256x256 shape. The ROI was defined using the same bounding box and 32-pixel buffer as described for CR images. Resizing was achieved through pixel averaging with python's cv2 library.[61] All CR images were set to the same square 256x256 dimensions.

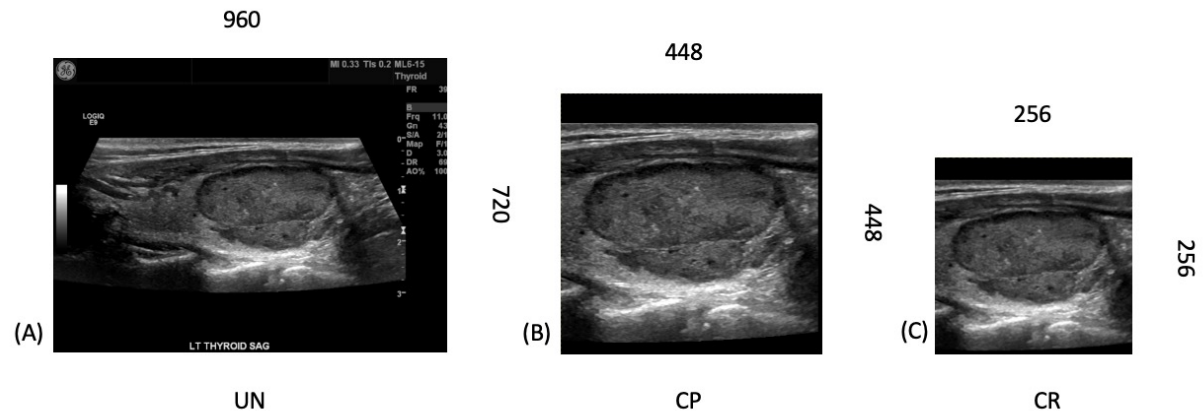

Figure 3.2: Example image at tested preprocessing levels un-cropped (A), cropped with padding (B), and cropped with resizing (C).[58]

3.2.1.3 AttU-Net Architecture

The U-Net architecture has demonstrated high performance in varied medical imaging segmentation tasks.[60, 62] Since being introduced in 2015, attention has been used to improve deep learning algorithms' performance in many applications.[63, 64, 65, 66] An attention-enhanced U-Net model was chosen due to its use of feedforward weighting of spatial information. [62, 67, 68] The design of the attention-enhanced U-Net (AttU-Net) used in this study was initially proposed by Schlemper et al.[67] The attention gates span the two halves of U-Net and allow spatial weights to be passed between corresponding layers of the encoding and decoding phases. For algorithm validation with UN and CP images, the dimensions of the input layer remained variable to accommodate changing image sizes but were fixed at 256x256 for validation with CR images. The hidden layers contained 12 filters for all validation and testing. The trained AttU-Net initially outputs a grayscale mapping which is then binarized by filling holes and selecting for the single largest contour. A schematic of the AttU-Net algorithm is presented in Fig. 3.3.[69]

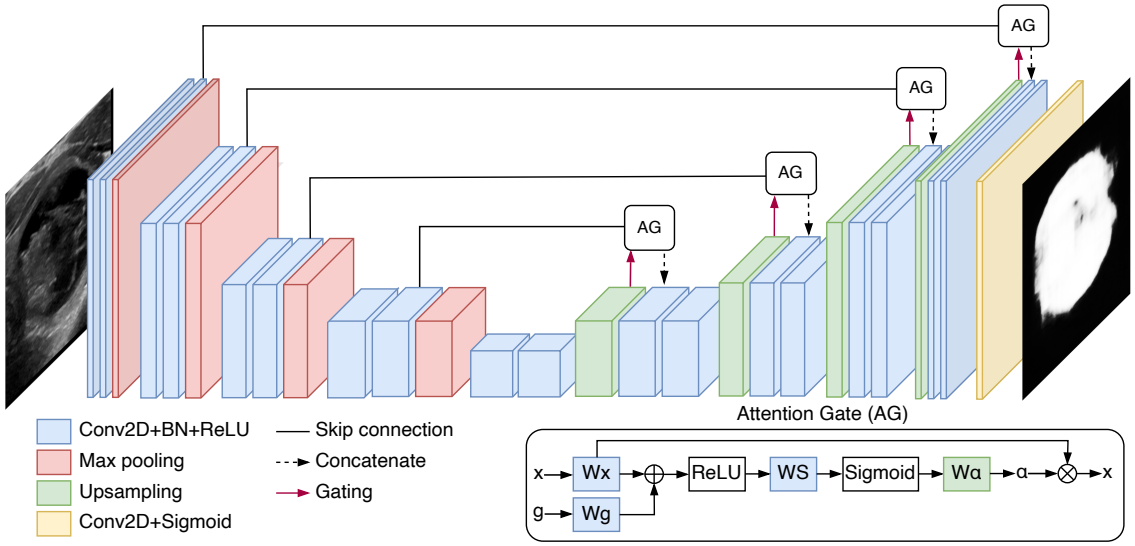

Figure 3.3: Architecture of enhanced U-Net with attention gates (AG) used for automatic thyroid nodule segmentation.[58]

3.2.1.4 Training and Statistical Analysis

3.2.1.4.1 Algorithm Training and Validation

Algorithm validation was conducted with Dataset TV using 5-fold cross validation by actual nodule for each preprocessing level, UN, CP, and CR. The physician contour was used as ground truth for algorithm performance evaluation. Splitting of Dataset TV into the folds during the validation was stratified by cytology at the patient level, ensuring that images of the same thyroid and nodule were selected together, thus preventing leakage between training and validation. The dice similarly coefficient (DSC), a measure of the algorithm's output and ground truth segmentation overlap, served as the performance metric for evaluation of the segmentation performance.[70] The Hausdorff distance (HD), the maximum distance between the algorithm's output and ground truth contours, was also used in algorithm evaluation.[71] While 95% HD and average HD are other commonly used metrics,

the maximum HD was used in paper as it provides the strictest measure by highlighting the largest errors in segmentation performance. For comparison of algorithm performance across this dataset which contains nodules of varying sizes, the HD is reported as a percent of the nodule's effective diameter: %HD. DSC and %HD metrics were calculated at the nodule level to avoid biasing results due to varying number of images per nodule. Note, the "average %HDs" reported represent the nodule level %HDs averaged across the validation fold or test set.

3.2.1.4.2 Independent Multi-Institutional Testing and Analysis

The preprocessing level which yielded the best results during the validation stage was used in the final segmentation algorithm. This final algorithm was trained on the entirety of Dataset TV and tested on institutional independent test sets from UCM and WCMC, Datasets T-UCM and T-WCMC, respectively. Overall algorithm performance was assessed using the average DSC and %HD. Test sets were further subdivided by nodule FNA and surgical pathology results: BB, IB, IM, and MM as shown in Figure 3.1.

3.2.1.4.3 Statistical Evaluation

3.2.1.4.3.1 Statistical Evaluation: ROI Preprocessing Determination

To determine which ROI preprocessing approach yielded the best results, two-tailed t-testing for dependent means was conducted between the 5-fold cross validation results. Both DSC and %HD comparisons were made. Significance was defined at p-value ≤ 0.025, after multiple testing Bonferroni corrections were applied.[72]

3.2.1.4.3.2 Statistical Evaluation: Independent Testing

Similarly testing of the segmentation algorithm's overall performance on the UCM and

WCMC test sets was conducted using the 'a posteriori' bootstrap procedure discussed by Ahn et al.[73] That is, the results of each test set were sampled 2,000 times at a 67% sampling rate using 'a posteriori' bootstrapping. From each of those samples, the differences in T-UCM vs T-WCMC average DSC (ΔDSC) and average %HD (Δ%HD) were calculated. Empiric 95% confidence intervals were determined from the ΔDSC and Δ%HD distributions, because currently there is not strong field agreement on non-inferiority margins for radiology performance metrics, empiric confidence intervals were used to establish equivalence ranges for the algorithm.[73]

To compare algorithm performance on pathology subsets (BB, IB, IM, MM), two-tailed t-testing for independent means was conducted between subset results within each test. Only subsets within the same test set were compared. Significance in this study was defined at p-value ≤ 0.05 for a single comparison; however, multiple testing Bonferroni corrections were applied when appropriate.[72]

3.2.2 Results

3.2.2.1 Algorithm Training and Validation Results

On 5-fold cross-validation conducted with Dataset TV, the CR ROI yielded the highest average fold DSC (std. deviation) 0.916 (0.01) and the lowest average fold %HD (std. deviation) 14.4% (0.7). The CP and UN ROIs yielded average fold DSCs (std. deviation) and average fold %HDs (std. deviation) of 0.815 (0.01), 23.9% (1.0) and 0.518 (0.04), 91.2% (8.1) respectively. Preprocessing with cropping and resizing resulted in a statistically significant (after correction for multiple comparisons, $p<0.025$) higher average DSC and average %HD than the other preprocessing methods. The results of the 5-fold validation and preprocessing

selection process are summarized in Table 3.2. Thus, the CR ROIs were used in the subsequent independent analyses.

Model Name	Preprocessing Level	Nodules	Images	Avg. DSC of 5 Folds (SD)	Avg. %HD of 5 Folds (SD)
UN Model	Un-Cropped + Padded	398	1334	0.518 (0.04)	91.2% (8.1)
CP Model	Cropped + Padded	398	1334	0.815 (0.01)	23.9% (1.0)
CR Model	Cropped + Resized	398	1334	0.916 (0.01)	14.4% (0.7)

$p = 3x10^{-5}$ $p = 6x10^{-5}$ $p = 3x10^{-5}$ $p = 5x10^{-5}$

Table 3.2: 5-fold cross validation results with dataset TV at different preprocessing levels. P-values of statistical comparisons are labelled with brackets; significance is defined at $p < 0.025$ after Bonferroni multiple comparison correction.

3.2.2.2 Independent Multi-Institutional Testing

Finally, the AttU-Net algorithm was trained on all 1334 cropped and resized images of Dataset TV. The images in the independent test sets from UCM and WCMC were cropped and resized according to the CR ROI preprocessing procedure optimized through the prior 5-fold validation results. AttU-Net produced an average DSC (std. deviation) of 0.915 (.04) and average %HD (std. deviation) of 12.9% (4.6) on the UCM independent test set and an average DSC (std. deviation) 0.922 (.03) and average %HD (std. deviation) of 13.4% (6.3) on the WCMC independent test set (Table 3.3). Sample cases are presented in Fig 3.4. Statistical testing failed to show a difference between the T-UCM test subsets: BB, IB, IM, MM (Table 3.4) ($p>0.0166$) and failed to show a difference between the T-WCMC test subsets: IB, IM ($p>0.05$). Similarity testing established equivalence of the segmentation algorithm's overall performance across the two independent institutional test sets up to margins of $\Delta\text{DSC} \leq 0.013$ and $\Delta\%\text{HD} \leq 1.73\%$ (Fig. 3.5).

Model	Preprocessing	Training Set	Testing Set	Avg. DSC (SD)	Avg. %HD (SD)
AttU-Net	Cropped + Resized	Dataset TV Nodules: 398 Images: 1334	Dataset T-UCM Nodules: 84 Images: 152	0.915 (0.04)	12.9 (4.6)
AttU-Net	Cropped + Resized	Dataset TV Nodules: 398 Images: 1334	Dataset T-WCMC Nodules: 58 Images: 109	0.922 (0.03)	13.4 (6.3)

Table 3.3: Multi-institutional independent testing results of attention-enhanced segmentation system. Results given are averaged across nodules.

Pathology	T-UCM				T-WCMC			
	Nodules	Images	Avg. DSC (SD)	Avg. %HD (SD)	Nodules	Images	Avg. DSC (SD)	Avg. %HD (SD)
BB	23	41	0.921 (0.03)	13.3 (5.1)	--	--	--	--
IB	16	28	0.917 (0.05)	11.5 (4.3)	22	40	0.921 (0.03)	14.0 (6.4)
IM	15	30	0.923 (0.05)	11.5 (3.9)	36	69	9.22 (0.04)	13.0 (6.2)
MM	30	53	0.906 (0.04)	14.1 (4.5)	--	--	--	--

Table 3.4: Multi-institutional testing results of the AttU-Net of the four pathological subsets, BB, IB, IM, and MM. Results are presented as averages at the nodule level for each subset. T-WCMC contained only indeterminate, IB and IM, cases.

3.2.3 Discussion

Based on the results of the multi-institutional testing, the AttU-Net algorithm presented in this work appeared to have the potential for use in clinical assessments via TIRADS and in development of AI algorithms for predicting thyroid nodule malignancy [17, 50, 59]. The multi-institutional independent test average DSCs of 0.915 (0.04) and 0.922 (0.03) for

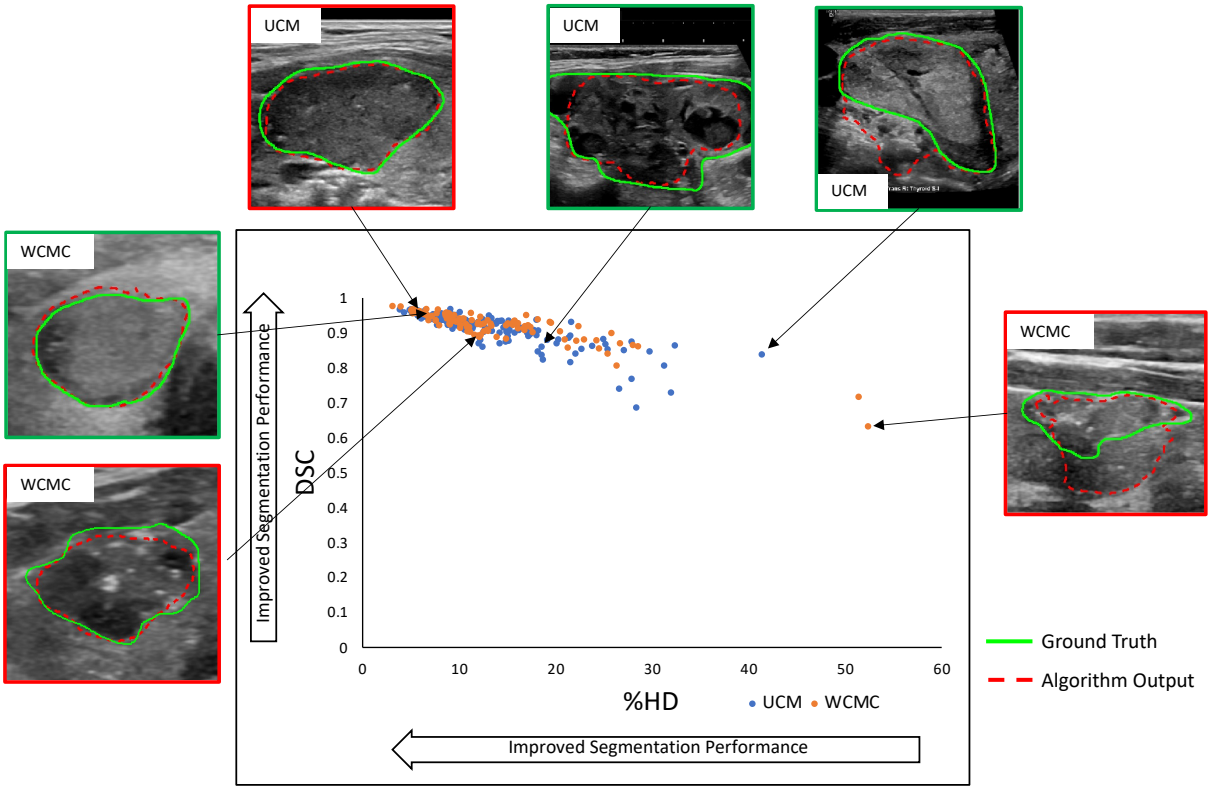

Figure 3.4: Graph of independent testing metrics; DSC versus %HD for independent testing images. Individual images are highlighted to provide examples throughout the full range of results. Images of malignant and benign nodules have borders of red and green, respectively. Inside each image, the green solid line shows ground truth while the red dashed line delineates the algorithm output.[58]

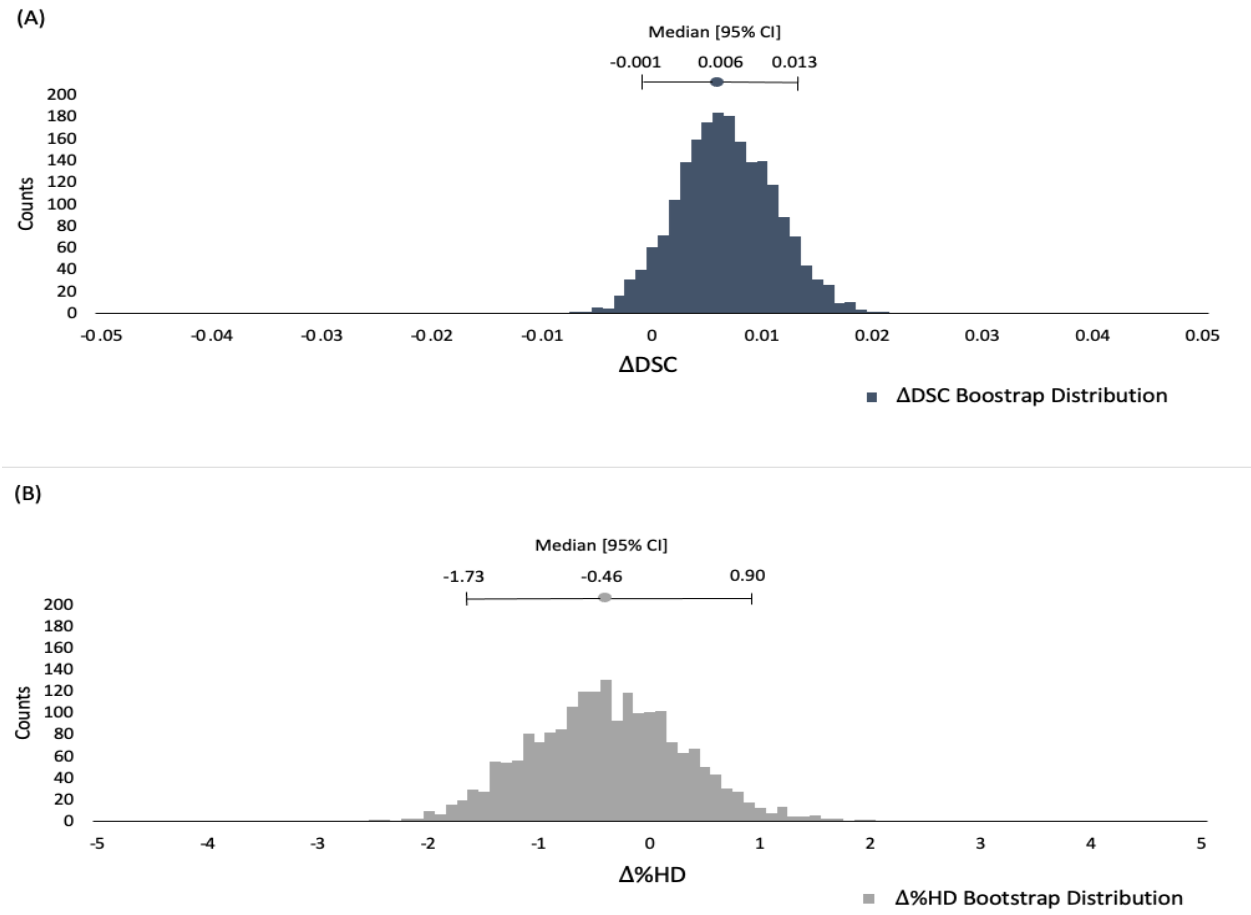

Figure 3.5: Similarity testing of our AI algorithm's performance between two independent institutional test sets (UCM and WCMC). Median ΔDSC and Δ%HD and their respective empirical 95% CIs from the 'a posteriori' bootstrap procedure is shown. (A) Because the ΔDSC confidence interval crosses zero, equivalence is established up to a margin of ΔDSC ≤ 0.013. (B) Similarity testing using the %HD metrics confirms the results seen with DSC and establishes equivalence up to a margin of $\Delta\%HD \leq 1.73\%$.[58]

segmentation of thyroid nodules on ultrasound were within the ranges for automatic contouring of other lesions such as breast lesions on ultrasound and %HDs 12.9% (4.6) and 13.4% (6.3) were within the ranges of reported clinician intra-observer and inter-observer variability when determining thyroid lesions dimensions on ultrasound.[74, 75] This study additionally represented a unique advancement as it was the first, to the best of the authors knowledge to assess a thyroid nodule segmentation algorithm's performance across combined cytological and surgical pathology determinations including indeterminate-benign and indeterminate-malignant nodules.

An automatically generated ultrasound thyroid nodule contour could be used in objectively delineating the nodule for TIRADS calculation. Such calculations are often performed on subjectively manually drawn contours and therefore are suspect to high clinician variability.[75, 76] Additionally, automatic contouring of ultrasound thyroid nodules could improve the efficiency of expanding training datasets for radiomic thyroid nodule classification systems by reducing the labor of manual contouring. It could also expedite the clinical application of radiomic systems by allowing a thyroid ultrasound to be contoured and then classified immediately after image acquisition. The algorithm's performance across two independent test sets from different institutions which yielded small equivalence margins highlighted the potential robustness and generalizability of this segmentation method. The algorithm's potential robustness was further demonstrated by high performance on indeterminate and malignant cases, which tend to be more difficult due to irregular borders. The algorithm's success especially on indeterminate thyroid nodules (Table 3.4) additionally suggested this algorithm could be used in machine learning systems focused on classifying thyroid lesions

outside the capability of FNA.[50, 59]

Future work could explore how the algorithm's performance is impacted by differences in ROI selection due to physician and ultrasound technician variability in a clinical setting. One limitation of this study is the training dataset size used in developing this algorithm. Collection of more training images may improve the performance of the algorithm. Future research efforts will include expansion of the dataset with the end goal of high-performance segmentation of bedside ultrasound. This algorithm has potential to impact not only nodule risk stratification but also bedside FNA biopsy acquisition.

This study presented the development and independent testing of a deep learning algorithm for segmenting thyroid nodules on ultrasound. The AttU-Net algorithm's sustained high performance was demonstrated on two independent test sets from separate medical institutions, UCM and WCMC. This study merits further exploration for integration into clinical and research workflows as an accurate segmentation tool.

3.3 Unsupervised AI for Internal Segmentation of Thyroid Nodules on Ultrasound

3.3.1 Motivation

Within TIRADS, thyroid nodules are grouped into four composition categories on US. In increasing order of nodule suspicion, they are: 1) cystic or almost completely cystic, 2) spongiform, 3) mixed cystic and solid (mixed) 4) solid or almost completely solid (solid).[17]

Because all nodules in our dataset were recommended for biopsy by TIRADS, they were primarily characterized as the one of the final two composition categories, mixed or solid. Upon discussion with a UCM endocrine surgeon, it was noted that physicians consider how much of the nodule is cystic and how much is solid, when evaluating mixed nodules. The calculation of the fraction of the nodule which is solid (SF) is defined according to Equation 3.1.

$$SF = (SolidArea)/(TotalNoduleArea) \tag{3.1}$$

To perform this calculation, the internal segmentation of mixed thyroid nodules into their cystic and solid components was required. While whole thyroid nodule segmentation was achieved through the training of deep learning model, this study will use an unsupervised fuzzy-c means algorithm.[77, 78, 79] Works by Chen et. al and Whitney et. al have demonstrated the success of fuzzy-c means segmentation algorithms for the segmentation of internal lesion components, including contouring cystic components on ultrasound.[80, 81] This study presents the application of fuzzy-c means internal component segmentation in the task of segmenting mixed composition thyroid nodules on ultrasound.

3.3.2 Validation Analyses

3.3.2.1 Dataset and Preprocessing

Nodule internal component segmentation, for the calculation of SF, was validated on a dataset of 108 mixed composition nodules which included 140 images from IB nodules and 202 images from IM nodules. Image cropping and resizing was performed according to the

methodologies optimized in Chapter 3.2.

3.3.2.2 Image Filtering for Component Segmentation

Cystic components, as fluid-filled sacs, are expected to be anechoic and nearly homogeneous. In contrast, solid components which are composed of tissue may have more variation and noise. Including a measure of local randomness or variation, entropy and standard deviation [STD], respectively, could aid in fuzzy-c means pixel assignment.[82] Local entropy and standard deviation filters were applied to each grayscale image using a 3x3 pixel neighborhood. These filters were applied in Matlab using the Image Processing Toolbox.[54] Composite images of grayscale values with local entropy values and local STD layered were constructed. An example of each filter applied to a nodule image is shown in Figure 3.6.

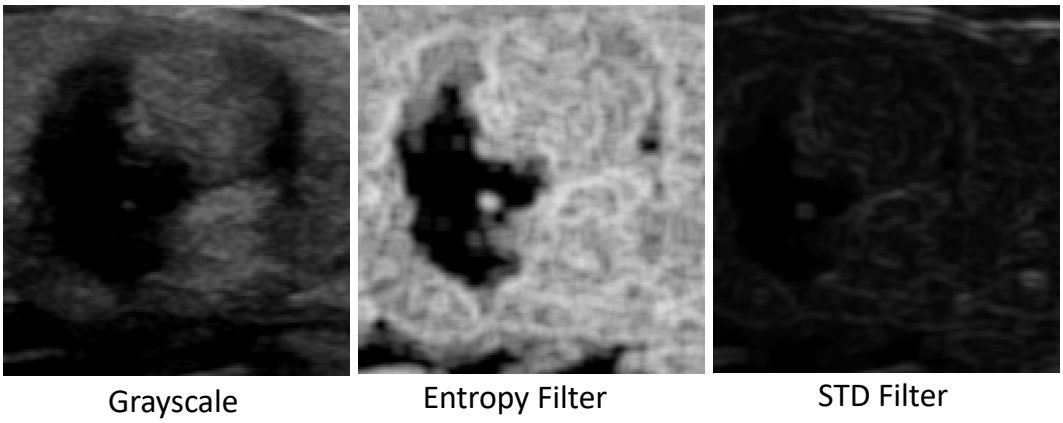

Figure 3.6: Figure 3.6 Example of original grayscale image nodule and entropy and standard deviation (STD) filters applied.

3.3.2.3 Segmentation with Fuzzy-C Means

Internal component segmentation was performed using a fuzzy-c means algorithm.[77, 78] Partition of pixels into two clusters was chosen to best capture the two internal compo-

nents present in mixed thyroid nodules. Final cluster assignment for each pixel was given at 50% inclusion probability. Because cystic components are fluid-filled, they are anechoic and appear very dark on ultrasound. In contrast, solid components, made of thyroid tissue, are echoic and may appear brighter on ultrasound. Therefore, whichever cluster had a lower centroid was considered representative of the nodule's cystic components and that which had a higher centroid considered representative of the nodule's solid components. The Xie-Beni index (XBI), ratio of intra-cluster compactness to inter-cluster distance, was used to evaluate fuzzy-c means clustering results.[83] The portion of the nodule deemed solid (SF) with and without filtered images included was calculated. A t-test for paired samples was used for statistical comparison of XBI and SF changes, with Bonferroni multiple corrections applied (alpha$\leq$0.025).[83, 72]

3.2.3 Validation Results

Inclusion of the entropy filtered image with the grayscale image resulted in a higher average XBI [STD] of 0.47 [0.18]. This was statistically significantly higher ($p<0.025$) the XBI [STD] of the grayscale image alone, 0.21 [0.06]. Additionally, inclusion of filtered images increased the portion of the nodule categorized as solid and correspondingly decreased the portion of nodule categorized as cystic. With the entropy filter added, there was an average change in SF of +6%. This difference was found to be statistically significant ($p < 0.025$). An example cystic component segmentation is provided in Figure 3.7.

3.3.4 Discussion

The inclusion of additional filters to the grayscale image resulted in overall less cluster agreement and an increase in the calculated SF of mixed nodules. In this dissertation, sub-

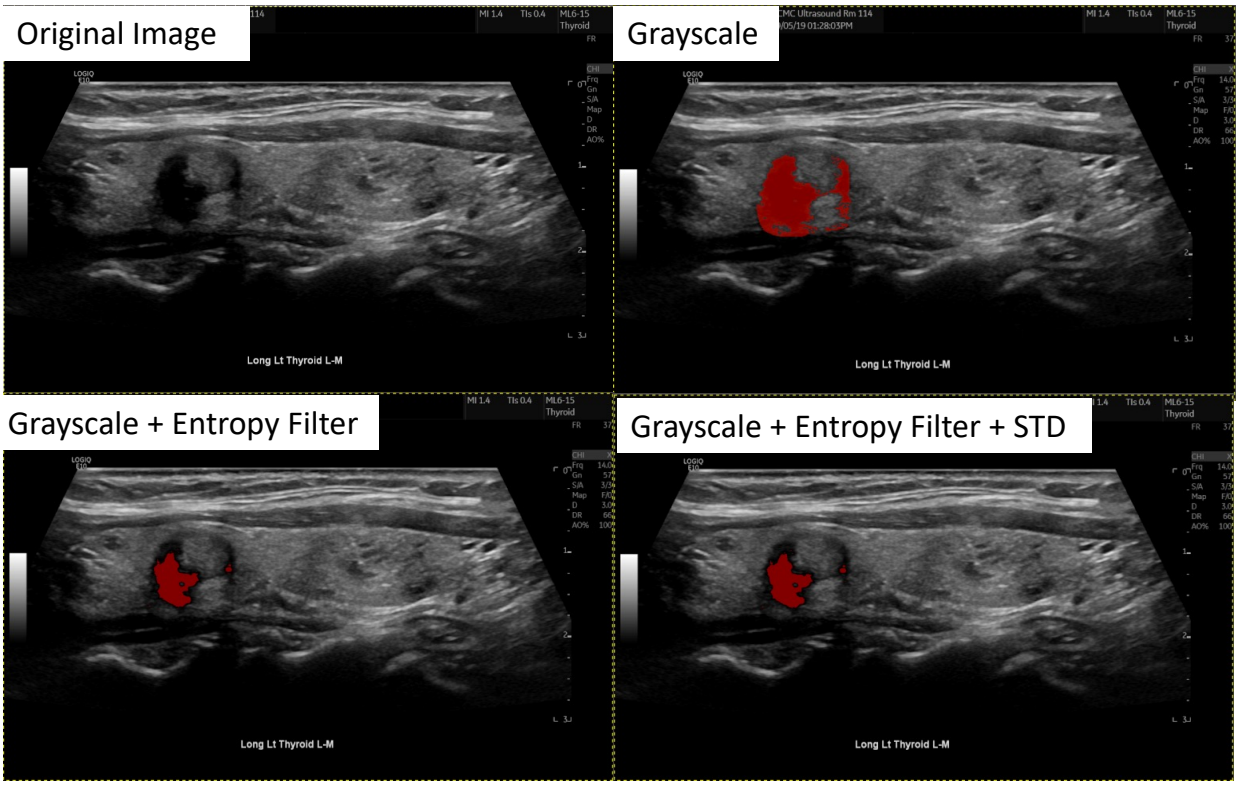

Figure 3.7: Figure 3.7 Example fuzzy-c means segmentation of cystic component with additional entropy and standard deviation filters added as channels. The area of the cystic component was decreased with additional filtered channels.

sequent fuzzy-c means internal solid and cystic component segmentation will be performed using the grayscale image alone.

3.4 Summary

This chapter presented the development and independent testing of a deep learning algorithm for segmenting thyroid nodules on ultrasound. The AttU-Net algorithm's sustained high performance was demonstrated on two independent test sets from separate medical institutions, UCM and WCMC. Additionally, the validation of internal segmentation fuzzy-

c methodology demonstrated adding an additional entropy filter led to an increase in the calculated size of the mixed nodule's solid component. The impact of automatic whole nodule segmentation and internal cystic and solid component segmentation on radiomic feature calculation and nodule classification is detailed in the next chapter.

CHAPTER 4

INVESTIGATION OF CLASSIFICATION AI FOR INDETERMINATE THYROID NODULES ON ULTRASOUND

4.1 Rationale for Indeterminate Thyroid Nodule Classification AI

Improving the classification of indeterminate thyroid nodules before surgical excision could help avoid unnecessary surgeries, as described in Chapter 1. Because an ultrasound (US) study is already standard of care for thyroid nodule risk stratification, a thyroid nodule classification algorithm AI for US could be directly integrated into current diagnostic procedures. [14, 17] The classification AIs presented throughout this chapter could aid nodule diagnosis at multiple stages: at the time of ultrasound, after FNA, and after molecular testing. A deeper understanding of how best to integrate FNA-biopsy results and molecular testing results with classification AI could result in improved clinical utility. Radiomic classification algorithms (Chapter 4.2.1), deep transfer learning classification algorithms (Chapter 4.2.2), and finally, ensemble classification algorithms (Chapters 4.2.3 and 4.3) were investigated for this classification task.

4.2 Radiomic Classification of Indeterminate Thyroid Nodules on Ultrasound

4.2.1 Development and Validation of Radiomic Algorithm

4.2.1.1 Motivation

Radiomic algorithms have been shown to be able to accurately classify lesions on ultrasound.[84, 59] Radiomic algorithms may be preferred over deep learning algorithms in medical applications because all features are human-engineered and therefore directly explainable.[85] As a result, radiomic algorithms can often be more easily understood by end-user, medical professionals, while still maintaining high classification performance.

4.2.1.2 Methods

4.2.1.2.1 Dataset

For initial radiomic algorithm validation, a subset of the indeterminate cases from the dataset detailed in Chapter 2 was selected. Only images originally taken in grayscale format at UCM were included, as these were the first cases available for training and validation purposes. Physician reference standard contours were used to localize the lesion. Only pixels within the lesion were considered for calculation of radiomic features. A summary of the cases included for algorithm validation is given in Table 4.1

4.2.1.2.2 Radiomic Features

For this radiomic algorithm, twenty-five features were extracted from each US image using the physician reference contours. Fourteen texture features were calculated using the gray-level co-occurrence matrix (GLCM). Six histogram features were determined based off

Pathology	Images	Nodules
IB	217	70
IM	218	67

Table 4.1: Summary of cases used for radiomic algorithm validation.

the distribution of pixel values inside the lesion. Four shape features and one size feature were calculated from the nodule contour. Because these final five features were exclusively determined by the nodule contour, they are directly impacted by changes in nodule segmentation. A list of features included in our radiomic model and a description of each is provided in Table 4.2. This group of features has been used for many other radiomic classification tasks, and most recently in Keutgen et al., 2022, which was discussed in Chapter 1.2.[86, 87, 50]

4.2.1.2.3 Standardization of Radiomic Features Calculation

In order to standardize GLCM texture and histogram feature calculation, pixel values of all images were set between 0 and 255. This was particularly important for the images originally taken in RGB format.[53, 54] Histogram equalization within images was not applied, as it failed to demonstrate a benefit during initial model validation. Pixel bin widths of 32 were used for histogram feature calculation. Additionally, a batch normalization of all features was parameterized on the training set and applied to validation and test sets.

4.2.1.2.4 Radiomic Classifier Selection

Three classifiers were studied in initial validation of the radiomic algorithm: linear discriminate analysis (LDA), support vector machine with a linear kernel (SVM-L), and support

Feature	Description	Origin
Contrast	Measure of local image variations	GLCM
Correlation	Measure of image linearity	GLCM
Difference entropy (DiffEntropy)	Measure of randomness of the difference of neighboring pixels' grayscale values	GLCM
Difference variance	Measure of variations of difference of grayscale value between pixel-pairs	GLCM
Energy	Measure of image uniformity	GLCM
Entropy	Measure of randomness of the grayscale values	GLCM
Homogeneity	Measure of image uniformity	GLCM
Information measure of correlation (ICM) 1	Measure of nonlinear grayscale value dependence	GLCM
ICM 2	Measure of nonlinear grayscale value dependence	GLCM
Maximum correlation coefficient	Measure of nonlinear grayscale value dependence	GLCM
Sum average	Measure of the overall image brightness	GLCM
Sum entropy	Measure of randomness of the sum of gray-levels of neighboring pixels	GLCM
Sum variance	Measure of spread in the sum of the gray-levels of pixel-pairs distribution	GLCM
Sum of squares (variance)	Measure of spread in gray-level distribution	GLCM
Maximum grayscale pixel value	Maximum grayscale pixel value of the nodule	Pixel Value Histogram
Minimum grayscale pixel value	Minimum grayscale pixel value of the nodule	Pixel Value Histogram
Average grayscale pixel value	Average grayscale pixel value of the nodule	Pixel Value Histogram
Standard deviation	Measure of the amount of variation in grayscale pixel values	Pixel Value Histogram
Skewness	Measure of the asymmetry of the grayscale pixel value distribution	Pixel Value Histogram
Kurtosis	Measure of the tailedness of the grayscale pixel value distribution	Pixel Value Histogram
Irregularity	Deviation of the nodule perimeter from the perimeter of a circle	Contour - Shape
Circularity	Similarity of the nodule shape to a circle	Contour - Shape
Compactness	Ratio of the nodule size to the area of a circle with the same perimeter	Contour - Shape
Depth-to-Width Ratio	Ratio of nodule depth to nodule width	Contour - Shape
Effective Diameter	Diameter of a circle with the same area as the nodule	Contour - Size

Table 4.2: Radiomic features name, a description of the feature, and from where the feature is calculated. This table was adapted from Keutgen et al.[50]

vector machine with a radial basis function kernel (SVM-RBF).[88, 89, 90] Both discriminate analysis and support vector machines classify through the creation of hyperplanes. SVMs are often better able to handle higher dimensionality classification problems as they make less stringent assumptions regarding covariance.[89] Whichever classifier yielded the highest receiver operating characteristic curve (ROC) area under the curve (AUC) on 5-fold cross validation was used in subsequent studies.[91] Initial algorithm validation was conducted with features calculated using physician reference contours.

4.2.1.3 Radiomic Algorithm Validation Results

The SVM with a linear kernel had the highest AUC for classification of thyroid nodules on US with all radiomic features. A summary of the initial validation results using different types of classifiers is shown in Table 4.3

Classifier	Avg. AUC (10itr.)	Standard Dev.
LDA	0.71	0.04
SVM-L	**0.72**	**0.04**
SVM-RGB	0.60	0.05

Table 4.3: Algorithm performance validation with three classifiers: linear discriminate analysis (LDA), support vector machine with a linear kernel (SVM-L) and support vector machines with a radial basis function kernel (SVM-RBF).

4.2.2 Impact of Nodule Segmentation AI on Radiomic Characterization and Classification

4.2.2.1 Impact of Whole Nodule Segmentation on Radiomic Performance

4.2.2.1.1 Motivation

Segmentation of nodules on ultrasound is a key step in the clinical risk stratification system TIRADS and for development of machine learning nodule classification models.[17] In ACRs TIRADs, nodule diameter, nodule border irregularity, and depth-to-width ratio are important features taken from the whole nodule contour and used to determine if FNA is required.[17] Pixel-based segmentation metrics including Dice similarity coefficient (DSC) and Hausdorff distance (HD) are commonly used to assess segmentation model performance.[70, 71] While DSC and HD remain the gold standard of segmentation model performance evaluation, additional testing may be required to determine how model performance impacts classification tasks. This work presents the use of radiomic features and radiomic classification modeling as a method of evaluation complementary to DSC and HD for an attention-enhanced U-Net thyroid nodule segmentation model (AttU-Net).[60, 67] In doing so, this project aims to gain a greater understanding of how DSC and HD are related to common radiomic features and to gain further confidence in the applicability of a segmentation model when used in classification tasks.

4.2.2.1.2 Methods

A subset of the dataset detailed in Chapter 2 was used for this study. The dataset was divided into a training and testing set at the patient level. A summary of the dataset used for

this study is presented in Table 4.2. Nodules were contoured by two UCM endocrine surgeons with a consensus contour generated. The physician contours (PC) served as segmentation reference standard and were used to generate reference standard radiomic features. This was the same methodology as Chapter 3.2

The attention-enhanced U-Net (AttU-Net) segment mentation model was constructed

Surgical Pathology	Training Set		Testing Set	
	Images	Nodules	Images	Nodules
Malignant	482	135	89	49
Benign	702	198	64	36
Unknown*	150	46	0	0
Totals:	1334	379	153	85

Table 4.4: Summary of training and testing datasets. *Nodules with unknown surgical pathology were exclusively used in segmentation algorithm training and excluded from classification related work.

by added attention gates to the max pooling layers of a base U-Net architecture.[67] The AttU-Net model architecture is the same as described in Chapter 3.2. Before being inputted into the AttUNet, images were cropped to a square ROI, determined as the smallest square enclosing the full nodule, and then resized to 256x256, allowing for a fixed input-layer size.[61] The AttU-Net was trained on the 1334 images of the training set and tested on the 153 images of the test set. DSCs and HDs were calculated for model output contours (MOCs).[70, 71]

For use in the radiomic classification model, 25 radiomic features were selected: 14 gray-level co-occurrence matrix (GLCM) features, 6 histogram-based features, 4 shape features, and 1 size feature.[86, 87, 50] These features and their descriptions are listed in Table 4.1.

For each test image, two radiomic features sets were extracted, one using the physician contour (PC features) and one using the AttU-Net model output contour (MOC features). The percent difference between PC generated and MOC generated features was calculated. Pearson correlation coefficients were calculated between percent feature differences and the pixel-based metrics (DSC and HD). Bonferroni multiple testing correction was applied.[72]

As optimized by the validation study in Chapter 4.2.1, a linear SVM to classify thyroid nodule pathology was trained on the radiomic features of training set images calculated using PCs.[90] This model was only trained on images with surgical malignant or benign designations. This classification model was then tested on the test set radiomic MOC features and PC features. Posteriori bootstrapping was performed on the SVM output scores for MOC features and PC features.[73, 92] 2000 samples were extracted, each of which contained 2/3 of the full test set (i.e., 57 of 85 nodules). It is important to note that sampling was conducted ensuring all images from the same nodule were selected together. A nodule malignancy score generated by averaging the scores of each image of that nodule, and ROC analysis was performed at the nodule level.[91] The AUC of the ROC curve served as the figure of merit for classification of nodules based on MOC features or PC features. Confidence intervals [CI] were determined empirically through the bootstrapping method.[73, 92] Difference testing of the radiomic classification performance using each the two test feature sets (PC vs. MOC) was completed using the 95% CIs. Equivalence testing was then conducted using the same 95% CIs to establish equivalence margins as described by Ahn et al.[73]

4.2.2.1.3 Results

The AttU-Net yielded an average DSC [STD] of 0.916 [0.05] and an average HD [STD]

of 2.8mm [2.2mm]. Skewness had the largest average percent difference, 149%, when calculating test set features using the MOC versus the PC. DSC and HD were both found to be most strongly correlated with effective diameter ($p<0.001$). Interestingly, DSC was found to be statistically significantly correlated with 16 out of the 25 features. Comparatively, HD was found to be statistically significantly correlated with only 8 out of the 25 features. Table 4.5 lists the correlation between the two pixel-based metrics, DSC and HD, and the change in radiomic features from ground truth PC to model segmentation MOC. Note, a higher DSC represents improved model segmentation and decreased percent feature difference (i.e., a negative Pearson correlation coefficient). Additionally, it is interesting that while skewness had the highest percent difference, its change was not statistically correlated with DSC or HD.

The linear SVM model yielded a median AUC [95% CI.] of 0.723 [0.655, 0.825] when tested with radiomic features extracted using the AttU-Net output contours (MOCs). The classification model yielded an AUC [empirical 95% CI] of 0.736 [0.654, 0.825] when tested with ground truth radiomic features extracted using the PCs. These results are shown in Table 4.6, and median ROC are curves displayed in Figure 4.1.

The median difference ΔAUC [95% CI] between model performance was -0.0136 [-0.062, 0.042]. Note, a negative difference signifies that the PC AUC was larger than the MOC AUC. We did not observe a statistically significant difference ($p>0.05$) between model performance using PC features and model performance using MOC features. Additionally, the lower bound of the 95% CI demonstrated that the radiomic classification performances using features calculated from physician reference contours and features calculated AttU-Net

Feature Name	DSC-Feature Difference Corr. Coefficient (r) (p<0.001)	HD-Feature Difference Corr. Coefficient (r) (p<0.001)
Difference Entropy	-0.29	--
Energy	-0.45	--
Entropy	-0.42	0.34
ICM2	-0.40	0.32
SUM Average	-0.36	--
Sum Entropy	-0.43	0.36
Sum Variance	-0.26	--
Variance	-0.26	--
Kurtosis	-0.35	--
Mean Gray Value	-0.51	0.36
Standard Deviation	-0.38	--
Effective Diameter	-0.82	0.47
Irregularity	-0.54	0.27
Circularity	-0.56	0.37
Compactness	-0.57	0.31
Depth to Width Ratio	-0.53	--

Table 4.5: Statistically significant Pearson correlation coefficients ($p<0.001$) between pixel-based segmentation metrics, DSC and HD, and percent difference in radiomic feature values calculated from physician contours vs. model output contours.

	PC AUCs	MOC AUCs	ΔAUC
Median Value	0.736	0.723	-0.0136
Empirical 95% CI	[0.654, 0.825]	[0.655, 0.825]	*[-0.062, 0.042]

Table 4.6: Summary of classification results tested through bootstrapping *As the ΔAUC 95% CI spans zero ($p>0.05$), statistical difference was failed to be shown between these sets of results.

output contours was equivalent up to a margin of 0.062. A graph of the equivalence testing margins is presented in Figure 4.2.

4.2.2.1.4 Discussion

The AttU-Net segmentation model not only demonstrated high performance as measured

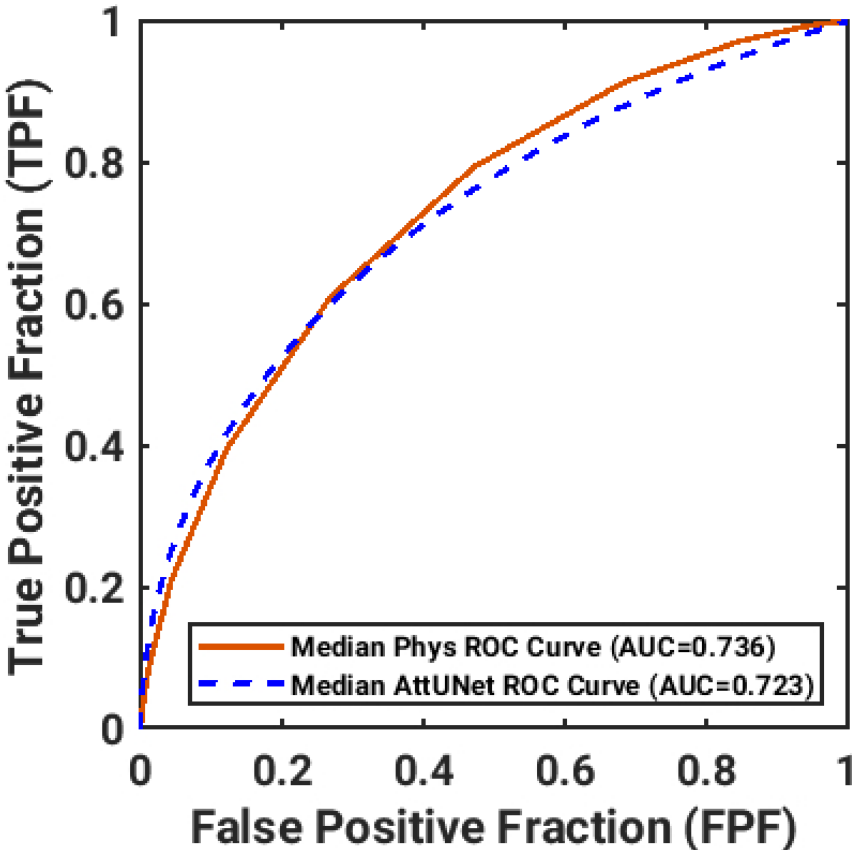

Figure 4.1: Median ROC curves classifying indeterminate nodule pathology using radiomics features extracted from physician coutours (Phys) and those extracted from out attention U-Net algorithm output contour(AttU-Net).

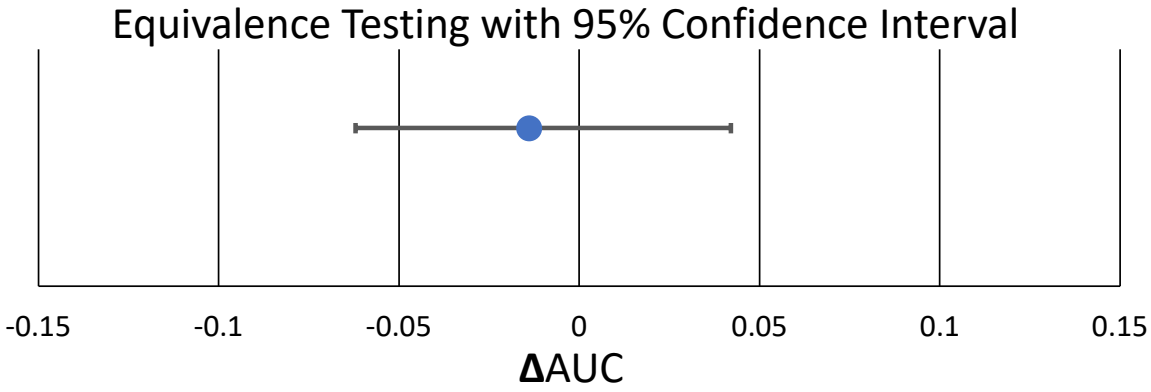

Figure 4.2: Equivalence testing results. Equivalence between using classification performance using radiomic features calculated from physician contours vs. AttU-Net contours is established up to a margin of $\Delta\text{AUC} < 0.062$.

by pixel-based segmentation metrics, DSC and HD, but also yielded a nodule classification AUC similar to that generated using ground truth segmentation. This work particularly highlighted that DSC and HD, while sensitive to radiomic shape and size differences, are less sensitive to segmentation changes which alter kurtosis, contrast, and correlation radiomic features. Overall, this work demonstrated how calculation of radiomic features can be useful in understanding and evaluating segmentation model performance and its role as part of a classification system.

4.2.2.2 Impact of Internal Component Segmentation on Radiomic Performance

4.2.2.2.1 Motivation

To calculate a TIRADS Score, as described in Chapter 1.2.1, radiologists must evaluate to what degree a nodule is cystic or solid.[17] For mixed composition nodules, the fraction of the nodule which is solid (SF) could be used to inform risk assessment as a new feature in our radiomic model. Additionally, texture features calculated from mixed nodules may be skewed due to the presence of major cystic components. Therefore, calculation of GLCM texture features specifically from a mixed nodule's solid component may lead to more informative features.

4.2.2.2.2 Methods

Indeterminate nodules in our dataset were labeled either mixed composition (with both solid and cystic components present) or solid composition. This designation was performed following the ACR guidelines for TIRADS scoring.[17] A summary of the cases used for this study, a subset of the dataset detailed in Chapter 2, is provided in Table 4.7.

Using an unsupervised fuzzy-c means algorithm, nodules with mixed composition des-

Composition	IB Images	IM Images	Total Images (Nodules)
Mixed	140	202	342 (108)
Solid	211	78	289 (93)

Table 4.7: Summary of indeterminate thyroid nodule dataset used for this study by composition: mixed cystic and solid (mixed) or fully solid (solid).

ignations were segmented into cystic and solid components according to the methodology described in Chapter 3.3.[77, 80, 81] The solid fraction (SF) was calculated according to equation 3.1. Nodules with solid composition designations were assigned an SF of 1.

Two systems for incorporating internal segmentation results into the radiomic algorithm were studied. First, the SF was added as a feature alongside the original 25 features of the radiomic algorithm input to the SVM-L.[90] This integrated algorithm, containing 26 total features, is referred to as radiomics+SF. Second, the GLCM texture features for mixed composition nodules were extracted exclusively from the largest solid component. All other features in mixed composition nodules and all features, including GLCM texture features, in solid composition nodules were extracted from the whole nodule. This integration strategy, containing 25 features, is referred to as radiomics-solid-GLCM. The average ROC AUC from 5-fold cross validation over 10 iterations was used as the statistic of merit for algorithm validation.[91] Classification and ROC analysis results were calculated at the nodule level.

4.2.2.2.3 Results

The validation results of including fuzzy-c means internal segmentation are presented in Table 4.8. Both integration methods radiomics+SF and radiomics-solid-GLCM yielded

lower average AUCs over the 10 iterations of 5-fold cross validation.

4.2.2.2.4 Discussion

Radiomics and Internal Components Algorithm	Avg. AUC	Avg. Standard Deviation
Radiomics	0.70	0.04
Radiomics + SF	0.68	0.04
Radiomics-Solid-GLCM	0.68	0.04

Table 4.8: Summary of indeterminate thyroid nodule dataset used for this study by composition: mixed cystic and solid (mixed) or fully solid (solid).

The two systems for integrating fuzzy-c means internal component segmentation outputs with our radiomic model did not result in higher algorithm performance. As such, fuzzy-c means SF was not included as a feature in future radiomic models. Additionally, GLCM features were extracted from whole nodules segmentations for both mixed and solid nodules in all subsequent radiomic algorithms.

4.2.3 Integration of FNA Cytopathology Results with Radiomic Classification AI

4.2.3.1 Motivation

The integration of information from multiple diagnostic modalities can have a synergistic effect in medical classification tasks. In this study, we explore the impact of combining FNA cytopathology Bethesda scores with our ultrasound radiomic model (Chapter 4.2.1) in

the task of classifying indeterminate thyroid nodules.[24] The integrated radiomic and FNA pathway is shown in Figure 4.3.

4.2.3.2 Methods

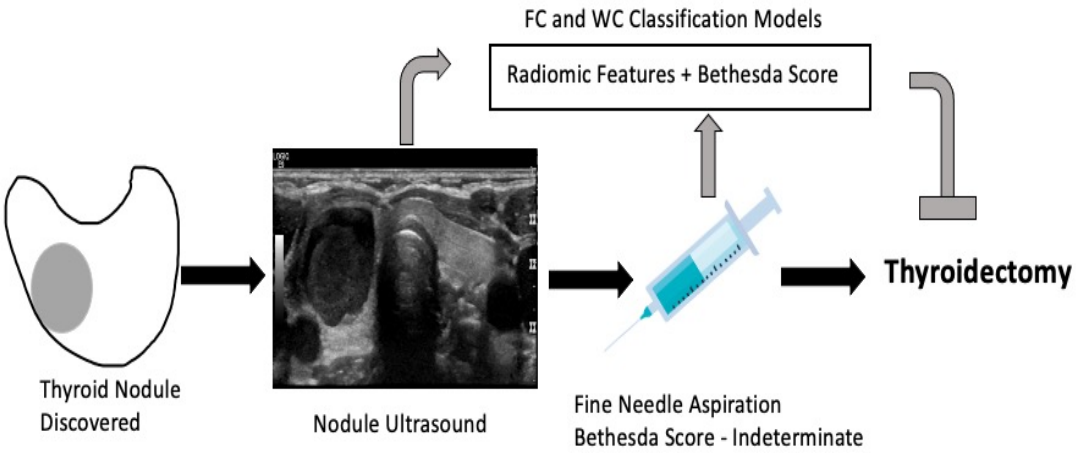

Figure 4.3: Schematic of the clinical role for the combination of our radiomic algorithm and Bethesda Score for indeterminate thyroid nodule classification.

4.2.3.2.1 Dataset

The dataset, a subset of that described in Chapter 2, included 449 grayscale ultrasound images from 141 thyroid nodules (N=139 patients). All nodules had been clinically designated indeterminate by fine needle aspiration (FNA) cytopathology with Bethesda scores of III, IV, and V, and due to their indeterminate status, each was surgically removed.[24] Subsequent surgical pathology classified 67 nodules as malignant and 74 nodules as benign. The distributions of Bethesda scores and nodule pathologies are shown in Figure. 4.4. From the ultrasound images, 25 human-engineered radiomic features (14 texture, 4 shape, 1 size, and 6 histogram-based features) were extracted and used to train an SVM classifier (RA).[86, 87, 12] This is the same algorithm as described in Chapter 4.2.1.

4.2.3.2.2 Radiomic and FNA Score Combination Methods

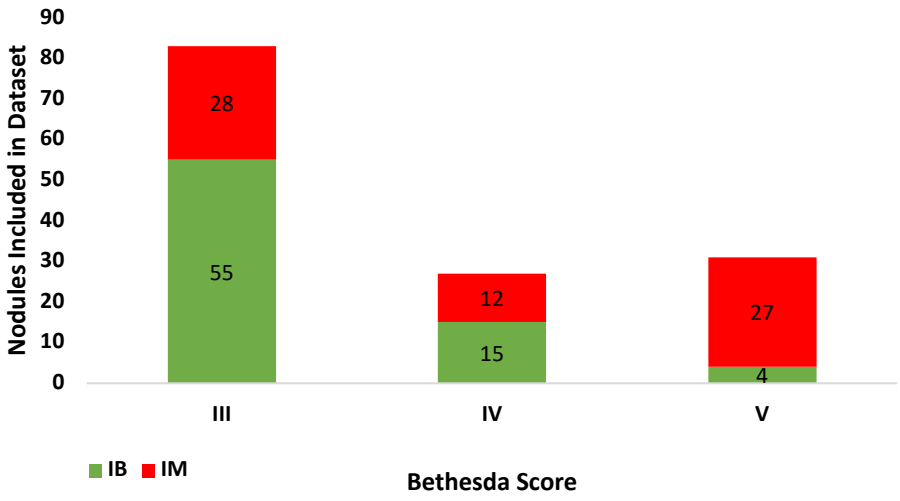

Figure 4.4: Bethesda scores and pathology of indeterminate study dataset.

Two systems were tested for combining the cytopathology score with radiomic features: feature combination (FC) and weighted combination (WC). In FC, the Bethesda score was included as a categorical feature input, along with the radiomic features, to the SVM classifier. In WC, the nodule's radiomics classifier (RA) score was multiplied by the expected malignancy rate of the nodule's Bethesda category (III: 0.20, IV: 0.33, V: 0.63) to generate a Bethesda-weighted malignancy score.[13, 93]

4.2.3.2.3 Algorithm Evaluation

Classification systems were evaluated at the nodule level using 100 iterations of 5-fold cross validation stratified across both Bethesda score and surgical pathology. Area under the receiver operating characteristic (ROC) curve (AUC) served as the figure of merit.[91] Model mean AUCs were compared using two-tailed t-testing with Bonferroni correction for multiple comparisons (alpha-corrected = 0.025).[72]

4.2.3.3 Results

The radiomics classifier (RA) yielded a mean AUC [STD] of 0.72 [0.04]. FC and WC approaches yielded mean AUCs [STD] of 0.80 [0.4] and 0.79 [0.04], respectively. Both combination approaches performed statistically significantly ($p<0.025$) better than the radiomic features alone. The average ROC curves for each algorithm are graphed in Figure 4.5

4.2.4.4 Discussion

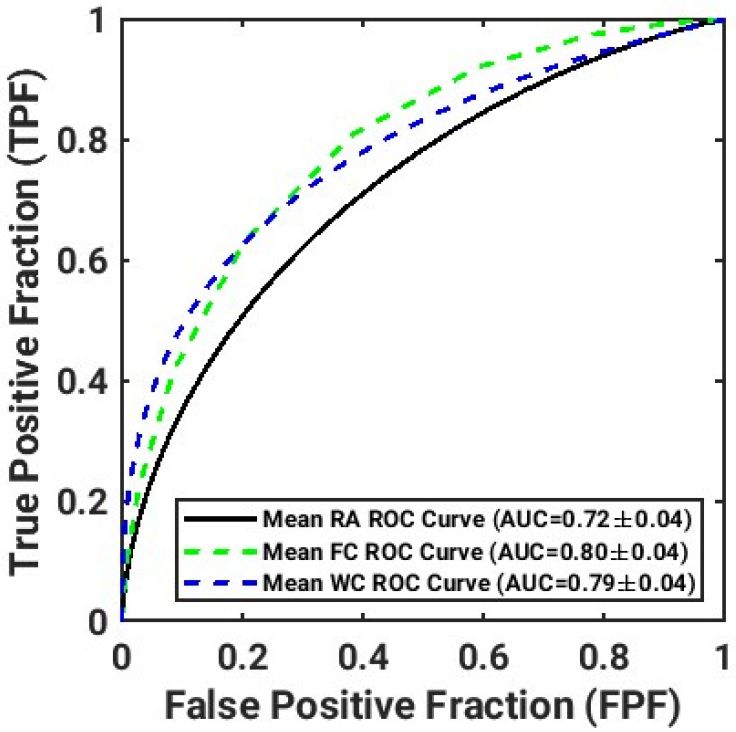

Figure 4.5: Comparison of mean classification results of radiomics only (RA), feature combination (FC) and weighted combination (WC) algorithms. Both radiomic and Bethesda combination models (FC and WC) had classification performances with AUCs statistically significantly ($p<0.025$) better than that of the radiomic model.

This study assessed two systems for merging cytopathological scores with radiomic features and demonstrated the merit of using multi-modal information in the task of classifying indeterminate thyroid nodules. Both combination algorithms resulted in improved classifi-

cation. In the FC, the highest performing algorithm, the Bethesda Score was selected 100% of the time, highlighting its importance to the combined model's success.

4.3 Deep Transfer Learning Classification of Indeterminate Thyroid Nodules on Ultrasound

4.3.1 Development and Validation of Deep Transfer Learning Algorithms

4.3.1.1 Motivation

Deep learning (DL) algorithms, convolutional neural networks with multiple hidden layers, have been shown to handle complex image classification tasks with high performance.[94] DL algorithms have the potential to outperform radiomic algorithms because they are not constrained to features which are mathematically designed by humans.[85] To train the millions of parameters contained in modern DL algorithms without overfitting, a large dataset is required. As is the case for many medical tasks, the dataset for this project (1500 total images) was too small to train a deep learning algorithm from scratch without severe overfitting. To compensate for dataset size, a transfer learning approach was utilized for algorithm training. In deep transfer learning (DTL), a DL algorithm is 'pre-trained' on a large dataset of general images in order to learn gross classification DL features, also known as embeddings.[95, 96, 97, 98] In deep transfer learning feature extraction (DTL-FE), DL features optimized by pre-training are calculated for the task-specific images and input into

a new external classifier.[96] In deep transfer learning fine-tuning (DTL-FT), the final layers, including the classification layer of the pre-trained DL algorithm are re-trained with task-specific images.[99] By keeping the majority of the algorithm frozen, the number of parameters available for training better matches the size of the task-specific dataset. As presented in the following sections, both transfer learning with feature extraction and transfer learning with fine-tuning were explored in the development of our deep learning classification algorithm.

4.3.1.2 Methods

4.3.1.2.1 Dataset

The pre-training dataset used for DTL algorithm development and validation was ImageNet.[100] ImageNet contains over 14 million images (e.g.. lion, wolf, boat) with nearly 22,000 associated labels (e.g. feline, canine, vessel). For DTL algorithm validation, a set of indeterminate cases from the dataset described in Chapter 2 was selected. Validation of the DTL-FE algorithm was conducted on dataset with 70 IB nodules and 69 IM nodules (Table 4.1). Because validation of DTL-FT algorithm required the training of convolution layers, a larger validation set was selected with 142 IB nodules and 123 IM nodules. All images were cropped to a region of interest using a physician reference standard contour with the same protocol as discussed in Chapter 3.2. Images were then resized to 224x244 to meet the dimensionality requirements of our deep learning algorithms.[61]

4.3.1.2.2 Deep Transfer Learning Feature Extraction (DTL-FE) Algorithm

The DTL-FE algorithm used a VGG19 network architecture (Figure 4.6).[101] Features, or embeddings, from each the VGG19 max pooling layers were extracted, 1472 in total.[97,

98, 64, 102] Three feature selection methods were explored: low variance filtering, high correlation filtering, and the Mann-Whitney U test (MWU-T) filtering.[103, 104] Low variance filtering (LVF) removed features with overall variances less than 1x10E-6. This was used to remove any constant features. High correlation filtering (HCF) aimed at removing redundancy in the DL features. For any pair of features with correlation coefficient greater than 0.80, only one feature was retained. In Mann-Whitney U feature selection, a Mann-Whitney U non-parametric test was performed on each feature in the training set to select for features with meaningful differences between IB and IM classification labels.[104] Those with different classification distributions $p < 0.025$ were retained. After feature selection, data reduction through principal component analysis (PCA) was implemented in order to reduce dimensionality and combat overfitting.[105, 103] Deep learning features were then input into an SVM with a linear kernel (SVM-L).[90] This was similar to the radiomic model and later enabled the merging of the radiomic algorithm and DTL-FE algorithm, as detailed in Chapter 4.4 and Chapter 4.5. Algorithm design was validated through 5-fold cross validation with ROC AUC serving as the statistic of merit. The algorithm with the highest AUC was used in subsequent studies.

4.3.1.2.3 Deep Transfer Learning Fine-Tuning (DTL-FT) Algorithm

The DTL-FT algorithm used the same VGG19 network architecture pre-trained on ImageNet.[101] All layers before layer 18 (Figure 4.6) were kept frozen, while layer 18 and those after were continued to be trained. InceptionV3 and DenseNet121 were also considered for the DTL-FT protocol.[106, 107] For comparison of DTL-FT architectures, as close to the same number of parameters as algorithm architecture allowed were left trainable. For each

architecture a classification head consisting of 6 fully connected layers was added; it included an average-pooling layer, two dense rectified linear unit (ReLU) layers, two dropout layers, and a final dense sigmoid classification layer. In addition to the dropout layers within the classification head, an early stopping training protocol was used to mitigate overfitting. DTL-FT architectures were compared with a validation set using the average ROC AUC as the statistic of merit.[91] The algorithm with the highest validation AUC was used in subsequent studies.

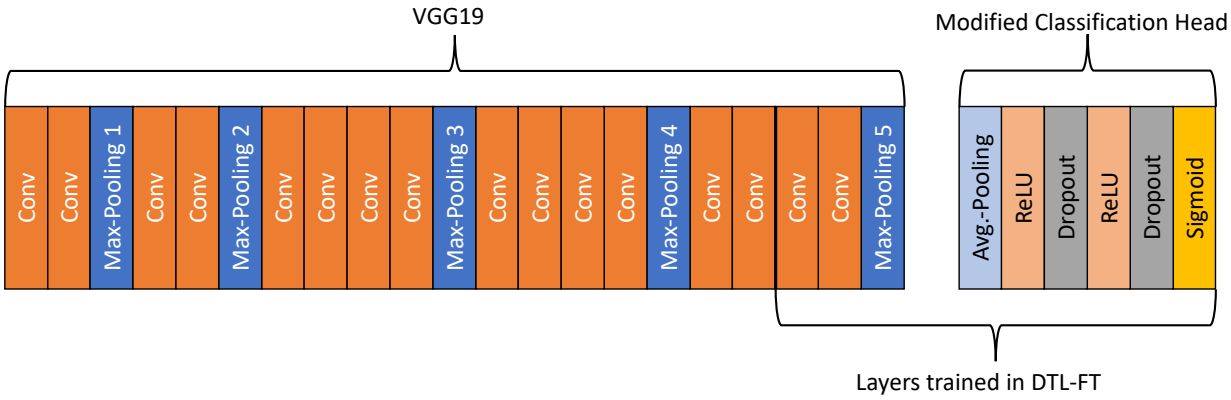

Figure 4.6: VGG19 and classification head architecture used in deep transfer learning algorithms. The max-pooling layers (dark blue) from which features, also known as embeddings, were extracted in DTL-FE are highlighted. The layers which remained unfrozen for continued training in DTL-FT are bracketed and labeled.

4.3.1.3 Validation Results

In initial validation, MWU-T filtering included the fewest features, 330; HCF and LVF included 1127 features and 1403 features, respectively. As shown in Figure 4.7, the DTL-FE algorithm performed best when MWU-T filtering was implemented and data was reduced to the first 50 features from PCA. The validation results of DTL-FE are presented in Figure

4.7.

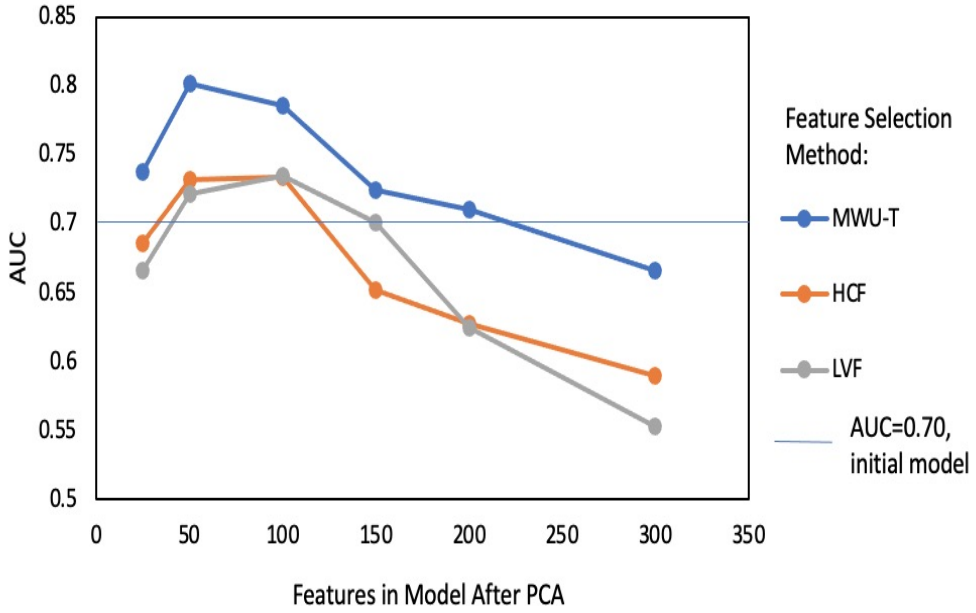

Figure 4.7: Validation results of deep transfer learning - feature extraction (DTL-FE) algorithm performance with feature selection: Mann-Whitney U filtering (MWU-T), high correlation filtering (HCF), or low variance filtering (LVF) and dimensionality reduction through principle component analysis. Without feature selection and dimensionality reduction, the DTL-FE model had an AUC of 0.70.

The deep transfer learning with fine-tuning (DTL-FT) algorithm with a VGG19 architecture had the highest validation AUC: 0.75.[101] The average AUC and the number of trainable layers and trainable parameters are listed in Table 4.9.

4.3.1.3 Discussion

Classification with MWU-T feature selection yielded the highest average AUC in this validation study. Because MWU-T filtering selects for features with separate distributions across pathology, it makes sense that this methodology was most effective for our classification task.[104] It will be used in subsequent applications of the DTL-FE algorithm. Since the VGG19 architecture yielded the highest average AUC on 5-Fold cross validation, it will

Algorithm	Total Parameters	Total Layers	Fine-Tuned Parameters	Fine-Tuned Layers	Avg. AUC
VGG19	20,221,761	28	4,916,993	9	0.75
InceptionV3	22,393,377	317	4,965,377	33	0.67
DenseNet121	7,365,953	433	4,905,857	215	0.68

Table 4.9: Comparison of deep transfer learning fine-tuned algorithm validation performance with VGG19, DenseNet121, and InceptionV3 architectures.

be used in subsequent applications of the DTL-FE algorithm.[101] While InceptionV3 and DenseNet121 are newer architectures, the linearity of VGG19 may be better suited to our limited fine-tuning dataset.[101, 106, 107]

4.3.2 Comparison of Transfer Learning Classification AI Performance on Cytopathologic Determinate vs. Indeterminate Nodules

4.3.2.1 Motivation

FNA biopsy Bethesda II (Benign) and Bethesda IV (Malignant) designations are highly accurate classifications of thyroid nodules, as determined by subsequent surgical excision.[24] These FNA-determinate nodules, labeled BB and MM in our dataset, represent classic benign and cancerous presentations, respectively. Many papers have published classification performances classifying exclusively or primarily FNA-determinate cases.[41] As described in

Chapter 1.2, 9 papers published fully excluded indeterminate nodules and 8 failed to specify if indeterminate nodules were included. Finally, 2 of the 6 papers, which specifically included indeterminate nodules, grouped indeterminate and determinate together for algorithm testing.[41] The goal of this study was to compare the performance of our deep transfer learning algorithms when trained and validated on determinate vs. indeterminate nodules.

4.3.2.2 Methods

For this validation study, a subset of determinate and indeterminate nodules was included from the dataset described in Chapter 2. The percent of FNA-indeterminate (IB/IM) vs FNA-determinate (BB/MM) cases included in the training set was varied through a pathology-stratified random selection of nodules. Stratification occurred at the nodule level. The training set compositions evaluated were 100% IB/IM, 75% IB/IM and 25% BB/MM, 50% IB/IM and 50% BB/MM, 25% IB/IM and 75% BB/MM, 100% BB/MM. Additionally, two validation sets were generated: one exclusively of FNA-determinate (BB/MM) cases and one of FNA-indeterminate (IB/IM) cases. These validation sets were unchanged as the training set varied. A summary of the cases included are listed in Table 4.10. Both the DTL-FE and DTL-FT algorithms of Chapter 4.3.1 were included in this study. Algorithm classification performance on the validation set was evaluated through ROC analysis with AUC as the figure of merit.

4.3.2.3 Results

Both the DTL-FE algorithm and the DTL-FT algorithm performed best in the task of classifying determinate nodules (BB/MM), with AUCs [STD] of 0.90 [0.04] and 0.93 [0.04] respectively, when trained fully on determinate cases. In contrast, the best classification

	FNA Indeterminate		FNA – Determinate	
	IM – Images (Nodules)	IB – Images (Nodules	MM – Images (Nodules)	BB – Images (Nodules)
Training	91 (273)	102 (315)	90 (315)	70 (245)
Validation	15 (30)	16 (29)	30 (54)	24 (44)

Table 4.10: Summary of dataset used for comparison of DTL-FE and DTL-FT algorithm performance in classifying determinate vs indeterminate nodules on ultrasound.

performance of indeterminate nodules occurred when the algorithm was trained on a mix of determinate and indeterminate cases. The DTL-FE and DTL-FT yeilded highest AUCs [STD] of only 0.67 [0.06] and 0.69 [0.06] respectively, in the task of classifying indeterminate nodules. The results of this study for both DTL-FE and DTL-FT are presented in Figure 4.8

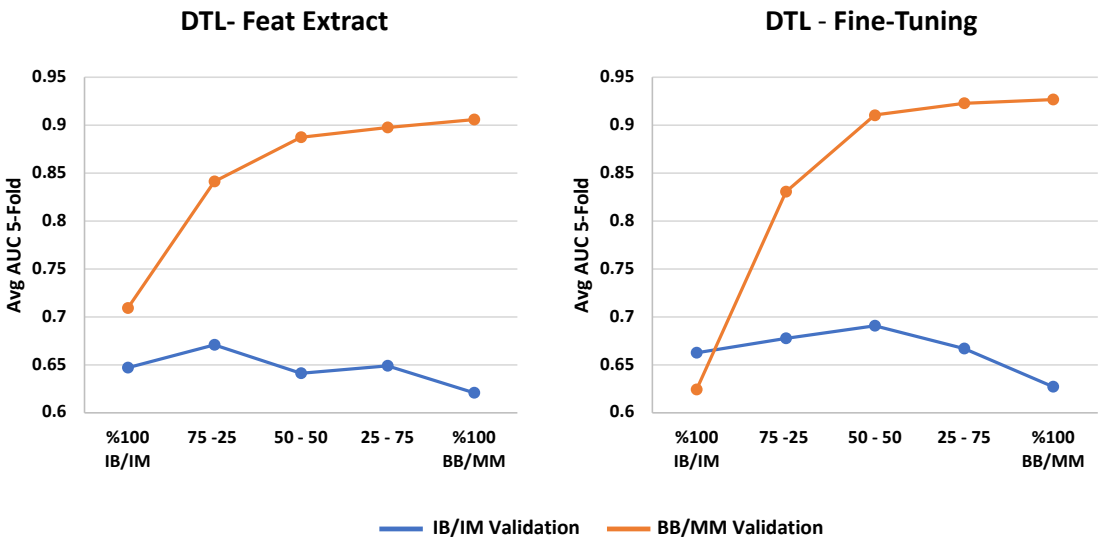

Figure 4.8: Comparison of DTL algorithm performance on determinate vs indeterminate thyroid nodules when changing training pathology distribution.

4.3.2.4 Discussion

Both the DTL-FE and DTL-FT algorithms performed best when classifying determinate cases (BB/MM). Each algorithm had its highest classification of indeterminate cases (IB/IM) when trained on a combination of IB/IM and BB/MM and lowest when trained on exclusively BB/MM cases. These results demonstrated that algorithms specifically aimed at classifying indeterminate nodules may benefit from the inclusion FNA-determinate training cases but only up to an extent. Algorithms trained primarily on FNA-determinate cases can classify determinate nodules well but in turn may be much worse at classifying indeterminate nodules.

Both DTL-FE and DTL-FT algorithms high performances, AUC$>$0.90, in classifying determinate (BB/MM) nodules were particularly meaningful as these results were similar to those of papers reviewed in Chapter 1, which included determinate cases.[41] In agreement with the results presented here, papers reviewed (Chapter 1) whose algorithms were validated on primarily indeterminate cases yielded lower classification performances.[41] This study was the first, to the best of authors knowledge, to have directly compared the same algorithm's performance across determinate and indeterminate thyroid nodule test sets.

This validation study provided support for our DTL methodologies. Both DTL-FE and DTL-FT algorithm methodologies were shown to be highly effective in the task of classifying determinate thyroid nodules. Conversely, indeterminate nodules were demonstrated to be more difficult to classify on US and may have a lower classification performance possible.

4.3.3 Comparison of ImageNet vs. RadImageNet Pre-Training on Deep Transfer Learning Classification Performance

4.3.3.1 Motivation

When applying transfer learning, the more similar the pre-training tasks are to the end classification task, the more task-relevant features your algorithm has the opportunity to learn.[95, 96] ImageNet has been widely used and demonstrated success with many classification tasks; however, the pre-training images included in ImageNet are not medically related.[100, 108] To fill this gap within the field of medical transfer learning algorithms, RadImageNet (RIN), a pretraining dataset specifically containing magnetic resonance imaging, computed tomography, and US images, was constructed.[109] RadImageNet contains >1.4 million images with associated anatomical annotations. The goal of this study was to validate if pre-training with a smaller set of medical images, using RadImageNet, allowed for improved transfer learning performance compared to pre-training with the larger non-medical, ImageNet.

4.3.3.2 Methods

The ultrasound subset of RadImageNet (RIN-US) containing 389,885 US images was used for pre-training in this study. [109] Within RIN-US, the subset of 92,599 thyroid images was also selected (RIN-ThyUS) for pre-training. A summary of the US images, as well as labels in RadImageNet, is provided in Table 4.11.

Four pre-training protocols were established: pre-training with ImageNet, pre-training with RIN-US, pre-training with RIN-ThyUS, and finally beginning with ImageNet pre-

Location	Anatomic Label	Number of Images
Abdomen/Pelvis	Aorta	18105
Abdomen/Pelvis	Common biliary duct	3646
Abdomen/Pelvis	Gallbladder	39743
Abdomen/Pelvis	Kidney	111252
Abdomen/Pelvis	Ovary	3595
Abdomen/Pelvis	Portal vein	769
Abdomen/Pelvis	Uterus	13312
Abdomen/Pelvis	Bladder	758
Abdomen/Pelvis	Fibroid	2095
Abdomen/Pelvis	Inferior vena cava	957
Abdomen/Pelvis	Liver	74889
Abdomen/Pelvis	Pancreas	21645
Abdomen/Pelvis	Spleen	6520
Neck	Thyroid General	57513
Neck	Thyroid Nodule	35086

Table 4.11: Summary of RadImageNet's 389,885 ultrasound images and 15 labels used for pre-training (RIN-US). The RadImageNet thyroid subset (RIN-ThyUS) is boxed. [109]

trained weights then continuing pre-training with RIN-ThyUS. The DTL-Feature Extraction algorithm developed in 4.2.2 was used for this study, and 10 iterations of 5-fold cross validation was employed. When pre-training with RIN subsets, the same training protocol was employed as in the original publication and validation of RadImageNet by Mei et al., 2022.[109] A training and validation set containing 106 IM nodules and 118 IB nodules was used for this study. As in Chapter 4.3.1., all images underwent cropping to an ROI and

resizing to 224x224 dimensions.[61] The average algorithm ROC AUC across 10 iterations was used as the statistic of merit.

4.3.3.3 Results

Pre-training with RIN-ThyUS alone yielded the highest average AUC [STD] of 0.75 [0.04]. The full results are shown in Table 4.12

Pre-Training	Number of Extracted Feats	Feature Selection	PCA Features	Avg AUC 5-Fold 10-iterations (STD)
ImageNet	1472	Mann-Whitney U	50	0.71 (0.04)
RIN-US	1472	Mann-Whitney U	50	0.70 (0.04)
RIN-ThyUS	1472	Mann-Whitney U	50	0.75 (0.04)
ImageNet -> RIN-ThyUS	1472	Mann-Whitney U	50	0.69 (0.04)

Table 4.12: Comparison of performance of deep transfer learning – feature extraction (DTL-FE) algorithm when pre-trained with ImageNet, the ultrasound subset of RadImageNet (RIN-US), the thyroid ultrasound subset of RadImageNet (RIN-ThyUS), and starting from pre-trained ImageNet weights then continuing pre-training with RIN-ThyUS. Pre-training with ThyUS yielded the highest average AUC and is highlighted.

4.3.3.4 Discussion

This validation study demonstrates that pre-training with images closer to the end task may improve algorithm performance. Interestingly, the combination of ImageNet and then RIN-ThyUS had a lower AUC on average than RIN-ThyUS from random weights and performed similarly to pre-training with only ImageNet. This may be due to a bias towards the starting weights, if local minimums had already been reached with ImageNet.

4.3.4 Interpretability-Driven Integration of Radiomic and Deep Transfer Learning Algorithms

4.3.4.1 Motivation

Radiomics and Deep Learning have the ability to capture unique diagnostic features from medical images. It is hypothesized that the merging of these unique features may generate higher overall classification performance.[85, 110] Additionally, because radiomic features are human-engineered with clinical correlates, they can be interpretable. In combining radiomic features with deep learning, we aim to not only improve classification performance, but also to increase algorithm interpretability. In this research, a novel method is presented for integrating radiomics and deep transfer learning in the task of classifying cytopathologically indeterminate thyroid nodules (ITN) on ultrasound as malignant or benign.

4.3.4.2.1 Dataset and Image Preprocessing

The dataset for this project was a subset of that described in Chapter 2, and only included indeterminate nodules. The dataset included 222 images from 69 unique ITNs with a final pathology of malignant (IM) and 254 images from 82 unique ITNs with a final pathology of benign (IB). Each nodule was represented by up to 7 images; both longitudinal and transverse ultrasound views were included when possible. The dataset for this study, a subset of that described in Chapter 2, is summarized in Table 4.13.

Nodule reference standard contours form UCM physicians were used for radiomic feature extraction. Nodule contours were also used in selecting a region of interest (ROI) for deep transfer learning feature extraction (DTL-FE). Ultrasound images were cropped to and ROI

Cytopathology	Surgical Pathology	Number of US images	Number of unique nodules
Indeterminant	Malignant	222	69
Indeterminant	Benign	254	82
Totals:		476	151

Table 4.13: Summary of cases used to study the integration of radiomics and deep transfer learning

and resized to 224x224 before being input into the deep learning transfer model (DTLM).[61] This was the same preprocessing performed in Chapter 4.3.1. Image preprocessing is demonstrated in Figure 4.9.

4.3.4.2.2 Deep Learning and Radiomic Features

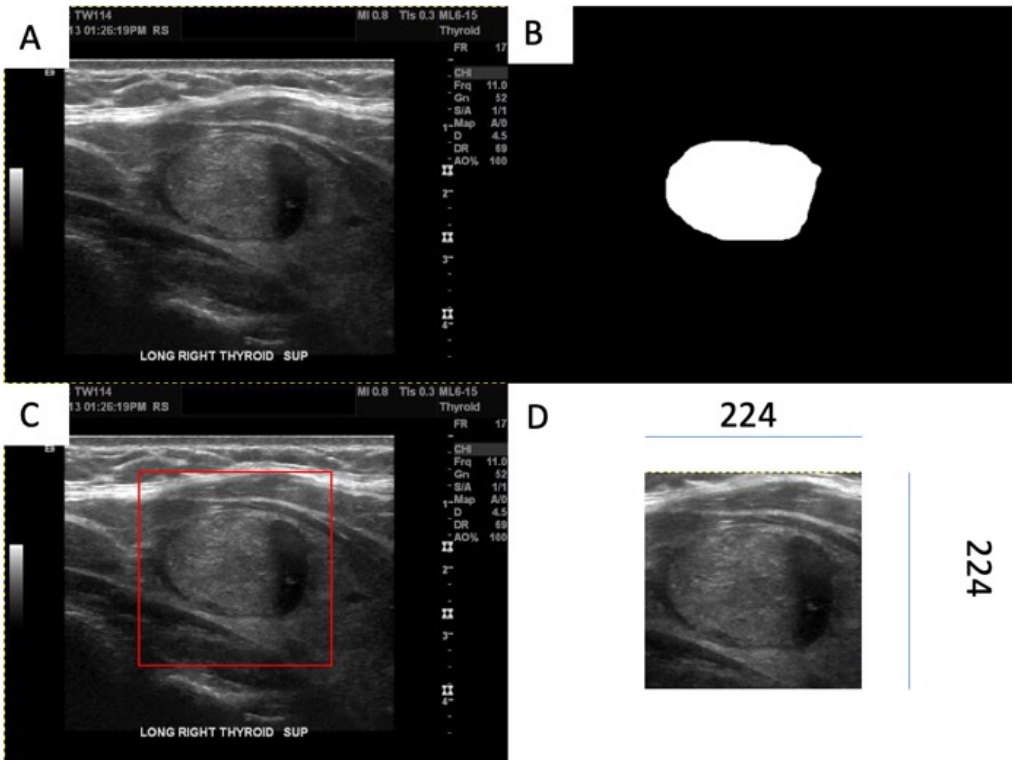

Figure 4.9: Ultrasound image preprocessing: initial image (A) and contour (B) input to RM, designation of ROI from contour (C) and cropped and resized input for DTLM (D).

The deep learning transfer features for this study were extracted from VGG19 architecture

pretrained on ImageNet, a publicly available database of more than 14 million images.[101, 100, 97, 111, 102] For each ultrasound image, 1472 features were extracted. Feature selection was performed with Mann-Whitney U filtering (MWU-F), as described in Chapter 4.3.1.[104] A MWU-F inclusion criteria of p-value < 0.35 was used utilized for this study. Principle component analysis then reduced MWU-F selected feature dimensionality leading to 50 deep learning components for each image. This process was optimized in Chapter 4.3.1. A total of 25 human-engineered features were utilized: 14 GlCM texture, 4 shape, 1 size, and 6 histogram-based features.[50, 80, 112, 113, 114] A list of these radiomic features is provided in Table 4.2.

4.3.4.2.3 Classification Schemes

The DTLM used all 50 deep learning features with a single linear support vector machine (SVM) classifier to score each image as benign or malignant. Similarly, the RM used a single linear SVM trained on all 25 human-engineered features. Two combination predictive models were built: a simple classifier combination model (SCM) and an interpretability-driven combination model (ICM).[110] The SCM merged the 25 radiomic features and the 50 DTLM features by feeding them all into a singular linear SVM classifier. In comparison, the ICM began first by sorting radiomic features into three categories: Echogenicity-Related, Composition-Related, and Shape/Margin-Related. The radiomic features designations can be found in Table 4.14. These radiomic designations were made in consultation with the endocrinologist who outlined nodules in this dataset. The TIRADS scoring system for used for categorical definitions.[17]

The 50 DTLM features of training images were correlated with each of the radiomic

Clinical Descriptor	Echogenicity-Related	Composition-Related	Shape/Margin-Related
Features	Contrast, Correlation, Energy, Homogeneity, Entropy, Difference Variance, Difference Entropy, IMC1, IMC2, Max Correlation Coefficient, Maximum, Minimum	Variance, Sum Average, Sum Variance, Sum Entropy, Skewness, Kurtosis, Average, Standard Deviation	Effective Diameter, Irregularity, Circularity, Compactness, Depth to Width Ratio

Table 4.14: Categorization of radiomic features for Interpretability-driven combination model

features of those same images and grouped in the category with which they had the highest absolute value average correlation. These designations were reset during each fold of the validation scheme. The RM and DTLM features were used in training three separate linear SVMs generating Echogenicity-Related, Composition-Related, and Shape/Margin-Related malignancy scores for each image, which were then averaged to provide the three nodule level scores. Finally, the three clinical descriptor related scores were averaged to provide an overall nodule malignancy score per lesion. A comparison of architectures of the combination models is shown in Figure 4.10.

Model performance evaluation was conducted at the nodule level. When there were multiple images of a single nodule, the individual image malignancy scores were averaged to generate a nodule malignancy score. Five-fold cross-validation by nodule over 100 iterations was used to assess model performance. Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) served as the figure of merit. Model mean AUCs were compared using two sample t-testing with Bonferroni correction for multiple comparisons.

4.3.4.3 Results

The RM and DTLM yielded mean AUCs with 95% confidence intervals of 0.75[.67,.83] and 0.70 [.72,.78], respectively, in the task of identifying thyroid nodule final pathology on

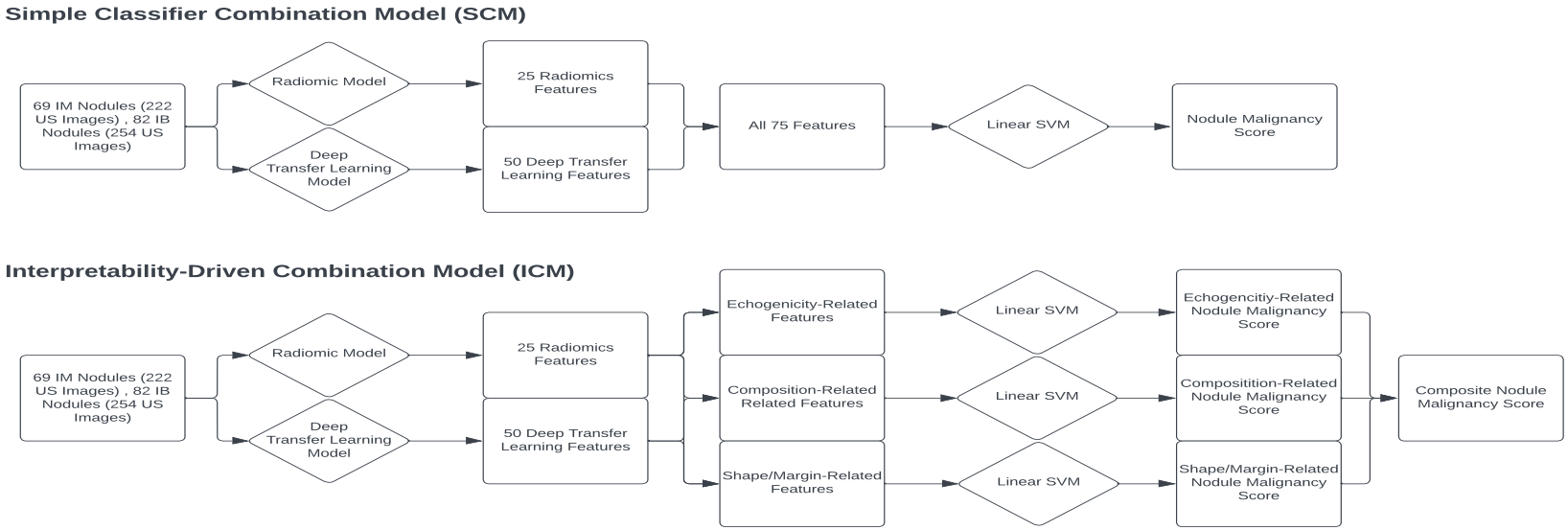

Figure 4.10: Comparison of simple combination model (SCM) and interpretability-driven combination model (ICM) workflow architectures

ultrasound. The malignancy scores of each nodule from the RM are plotted against those from the DTLM in Figure 4.11. The malignancy scores of the RM and DTLM were correlated with a Pearson correlation coefficient of 0.51. While the results of these models are positively correlated, their variability implies that the deep learning features and radiomic features could be quantifying different aspects of the lesions on US.

The SCM and ICM yielded mean nodule AUCs [95% CI] of 0.77 [0.70,0.84] and 0.76

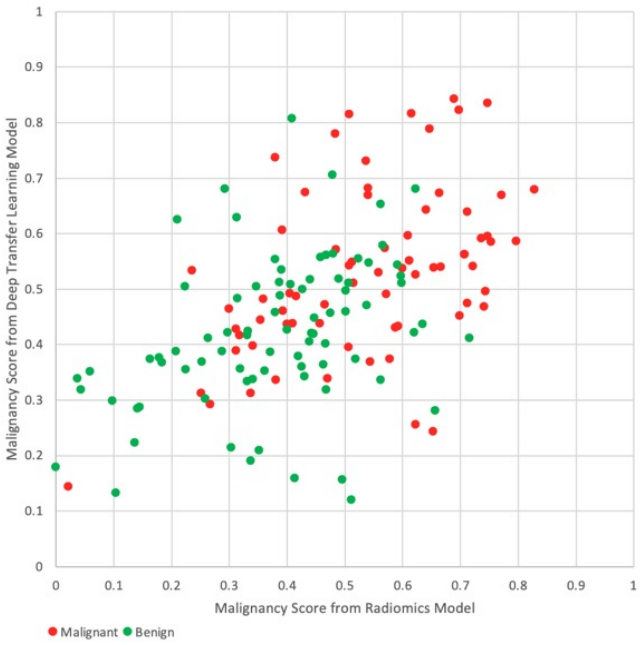

Figure 4.11: Comparison of RM and DTLM Malignancy Scoring. Color denotes surgical pathology. RM and DTLM have a Pearson correlation coefficient of 0.51. Each data point represents a single actual lesion (N=151).

[.69,.84], respectively, in the task of identifying thyroid nodule final pathology on ultrasound. The Composition-Related score had the highest individual characteristic-related mean AUC across the 100 iterations. While the SCM and ICM average AUCs were higher than the RM and DTLM, we failed to show this improvement was statistically significant ($p>0.01$). A

comparison of the DTLM, RM, SCM, and ICM ROC Curves is presented in Figure 4.12.

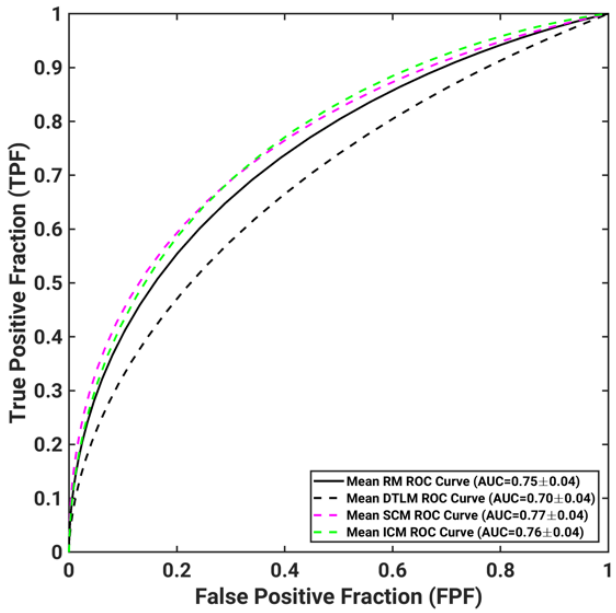

Figure 4.12: ROC Curves of Deep Transfer Learning Model (DTLM), Radiomic Model (RM), Simple Combination Model (SCM) and Interpretability-driven Combination Model (ICM) with AUC±Standard Error in the task of determining thyroid nodule final pathology on ultrasound.

4.3.4.4 Discussion

This work aimed to combine radiomics with deep learning to harness the interpretability of human engineered features with the power of deep learning models. While we failed to show a statistical difference between the DTLM and RM AUCs, the Pearson correlation coefficient of their outputs was only 0.51. Therefore, while the DTLM and RM aggregate model performances are similar, the variability on individual malignancy scores suggests that the radiomic and deep transfer features may be incorporating different aspects of the image into the classification decision. Merging features from both the DTLM and RM, the combination models failed to show a statically significant improvement; however, it did result

in the highest average AUCs over these 100 runs. While the ICM failed to show a statistical improvement in average AUC, the multi-tiered classification system of the ICM may be viewed as an improvement in model interpretability. By combining correlated features from both the RM and DTLM and increasing the total number of model features as compared to RM and DTLM, the ICM was able to train three separate classifiers. By providing Echogenicity-Related, Composition-Related, and Shape/Margin-Related malignancy scores in the terminology utilized by the current clinical visual thyroid nodule US risk stratification system (TIRADS), the ICM may be able to provide clinicians increased direction as to which aspects of the nodule are most contributing to the model classification. Additional reader studies with clinical experts are needed to assess the effect such a system would have on practical clinician decision making. Because thyroid cancer is highly curable with surgery, an emphasis on confidence in a machine learning informed decision would be important for decreasing the rates of unnecessary surgery.[115, 116] This sort of interpretable informed decision making will be key for clinical integration of AI driven systems.[116] Further work is required to validate our findings and to extend our methodology beyond correlation to better provide machine learning interpretability in clinical practice.

This study assessed two models for integrating radiomics and deep learning in the task of classifying indeterminate thyroid nodules on ultrasound. The first was a simple combination that established the potential for maintaining model performance while merging deep learning and radiomic features. The second demonstrated the efficacy of a clinically-driven combination model with means for potential interpretability of AI output in the task of predicting pathology status of ITN from US images. Further work is warranted to explore

ways in which combination modeling might be utilized alongside physician decision making to confidently identify thyroid cancer and avoid unnecessary thyroidectomies.

4.4 Integration of Image-Based AIs with Molecular Testing for Classification of Indeterminate Thyroid Nodules

4.4.1 Stratified Selection of Training and Testing Sets

4.4.1.1 Motivation

Selection of appropriate training and testing sets is an important step in machine learning model evaluation.[117, 118] As multi-year studies progress, training and testing sets are often separated temporally, with the cases collected first used in training and the most recent cases used for testing. This method, however, can lead to training and testing sets with biases in feature distributions. This study presents a comparison between temporal training/testing splits with that of a stratification procedure.

4.4.1.2 Methods

A dataset of indeterminate thyroid nodule ultrasound (US) studies was retrospectively collected from two institutions: University of Chicago Medicine (UCM) and Weill Cornell Medical Center (WCMC). In total, 315 cases from 2012-2023 were included. The US study features collected: study location (UCM or WCMC), US machine manufacturer (GE, Phillips, or Siemens), and US intensity representation (grayscale or RBG). Case pathology features collected: cytopathology Bethesda score (III, IV, or V), molecular testing platform (Afirma, ThyGeNext, ThyroSeq, or none), molecular testing result (benign, suspicious, or

highly suspicious), and surgical pathology (benign or malignant).

Temporal training/testing splits were simulated by using cases from the last two years, three years, or four years of data collection as the test set (Temporal – 2yr, Temporal – 3yr, and Temporal – 4yr,). A stratified sampling (SS) training/testing split was generated for comparison using the MIDRC SS Algorithm.[119, 120] The SS matched dataset characteristics across training and testing. The Jensen-Shannon divergence (JSD), a measure of distribution similarity, was calculated using the MIDRC REACT Tool and used to compare training and testing feature distributions.[121, 120, 118] A lower JSD indicates more similar distributions.[121] A paired t-test (alpha = 0.05) was used to test the difference in average JSD across the training/testing separation methods.

4.4.1.3 Results

The average JSDs [STD] of training vs. testing case feature distributions were 0.039 [0.028], 0.148 [0.096], 0.148 [0.087], and 0.152 [0.086] for SS, Temporal –2yr, Temporal – 3yr, and Temporal – 4yr, respectively. From the paired t-test, the average JSD of the SS was found to be statistically significantly lower than that of the Temporal – 3yr (p=0.007) and Temporal – 4yr (p=0.007). The largest differences in JSD between the SS and temporal splits were found in molecular testing platform and result distributions. The JSD values of temporal and stratified sampling methods is shown in Figure 4.13.

4.4.1.4 Discussion

Two of the temporal training/testing splits had statistically significantly difference training/testing feature distributions than stratified sampling in a thyroid US study dataset. The distribution of indeterminate nodule surgical pathology has remained relatively con-

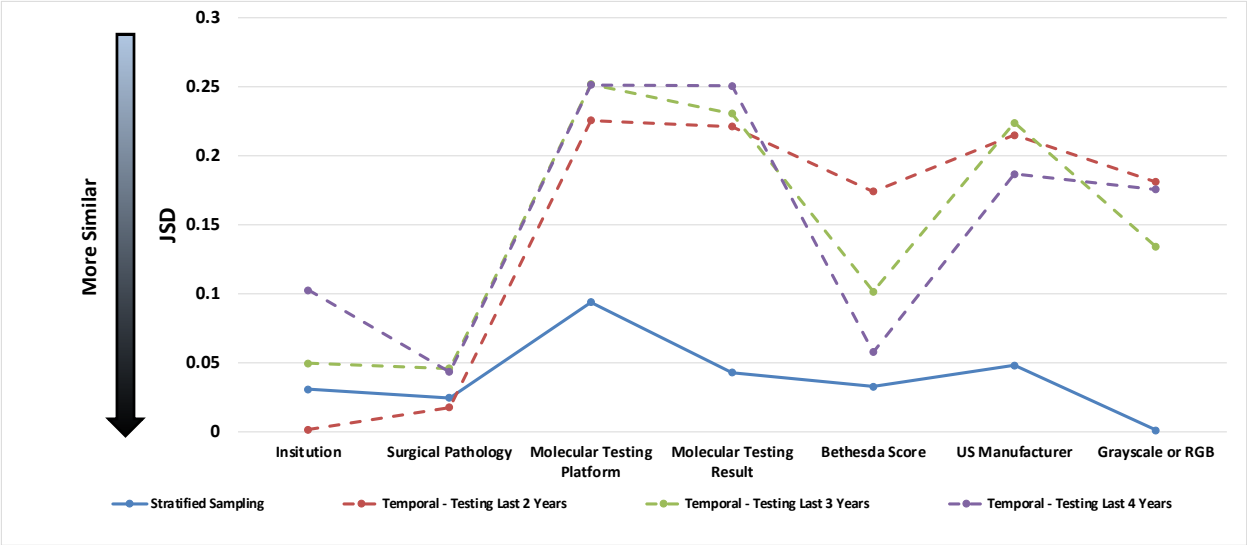

Figure 4.13: Jensen-Shannon divergence across dataset features when selecting training and testing sets through stratified vs. temporal sampling.

stant throughout the last many years, as has the consistency of UCM and WCMC regularly contributing cases to the dataset. This is reflected by all train-test split methods, stratified or temporal, yielding low JSDs. On the other hand, use of molecular tests has increased over the time of dataset collection, disproportionally placing them in temporal testing sets. This resulted in large molecular testing platform and molecular testing result JSDs. Because the stratified sampling yielded consistently the lowest JSDs pertaining to molecular testing, it will be used to generate test sets throughout the following study aimed at integrating our radiomic and deep transfer learning algorithms with commercially available molecular testing platforms.

4.4.2 Ensemble Classification of Indeterminate Thyroid Nodules

4.4.2.1 Motivation

The use of molecular testing has greatly expanded over the last decade.[27] Increases in the number of mutations tested and the availability of microRNA panels has led to performance improvements. With the use of molecular tests, the false positive surgical rate was approximately 25% lower than without. However, as discussed in Chapter 1.2.3 and reflected in our dataset (Table 2.2), the positive predictive value of molecular testing is still at ∼0.5.[28, 29] Integration of image-based AIs, such as those developed earlier in this chapter, with molecular testing could lead to improved classification performance and a further reduction in false positive thyroid surgeries.

4.4.2.2 Methods

4.4.2.2.1 Dataset

All indeterminate nodules in the dataset detailed in Chapter 2 were included for classification this study. Based on the results of Chapter 4.4.1, the stratified selection procedure was used to generate the training/validation and testing sets in a ratio of 70/30, respectively.[119, 120] Stratification was optimized across the following case features: institution, Bethesda score, molecular testing platform, molecular testing result, US manufacturer, and US image color. Molecular testing results were categorized as negative, suspicious, and highly suspicious according to the categorization criteria for molecular testing results were detailed in Chapter 2.2. A summary of the molecular testing results compared to final pathology is given in Table 4.15.

Dataset	Full Dataset		Molecular Testing – Subset	
Pathology	IM	IB	IM	IB
Training/Validation	103	140	Negative: 2 Suspicious: 29 Highly Suspicious: 17	Negative: 23 Suspicious: 69 Highly Suspicious: 2
Independent Testing	51	59	Negative: 2 Suspicious: 16 Highly Suspicious: 10	Negative: 9 Suspicious: 22 Highly Suspicious: 0

Table 4.15: Summary of dataset with molecular testing subset used to study the integration of our ensemble AI with molecular testing results.

4.4.2.2.2 Image-Based Ensemble AI

By combining radiomics and deep transfer learning algorithms, classification performance may be improved.[85, 110, 122] For this ensemble classifier, the best performing algorithms from Chapters 4.2 and 4.3 were combined. The radiomic and Bethesda score classifier (RBC), employing the feature combination algorithm (FC) of Chapter 4.2.3, was included.[50] Additionally, two deep transfer learning feature extraction algorithms from Chapter 4.3.3 were included, one pre-trained on ImageNet (DTL-FE-IGN) and the other on RadImageNet (DTL-FE-RIN).[100, 109] Classifier outputs were aggregated through the weighted average given in Eq. 4.1. This weighting of classifiers was optimized using the training/validation set through a pair-wise combinatorial averaging of classifier scores. The ensemble classifier yielded the highest validation results when the RBC algorithm score was weighted more heavily than those of the two DTL-FE algorithms. This was expected as the RBC algorithm had previously demonstrated the highest individual classification performance.

$$Ensemble = (0.25 * DTL-FE-IGN) + (0.25 * DTL-FE-RIN) + (0.5 * RBC) \quad (4.1)$$

4.4.2.2.2 Integration of Molecular Testing with Ensemble Image-Based AI

If a nodule underwent molecular testing, its molecular testing results were merged with the ensemble nodule score through a rule-based integration. Nodules with negative molecular testing results (ITN-MN) were assigned a score of 0 and those with highly suspicious results (ITN-MH) a score of 1. Nodules with molecular testing suspicious results (ITN-MS) were assigned the image-based ensemble classifier score. A schematic of the molecular testing and ensemble AI combination procedure is given in Figure 4.14.

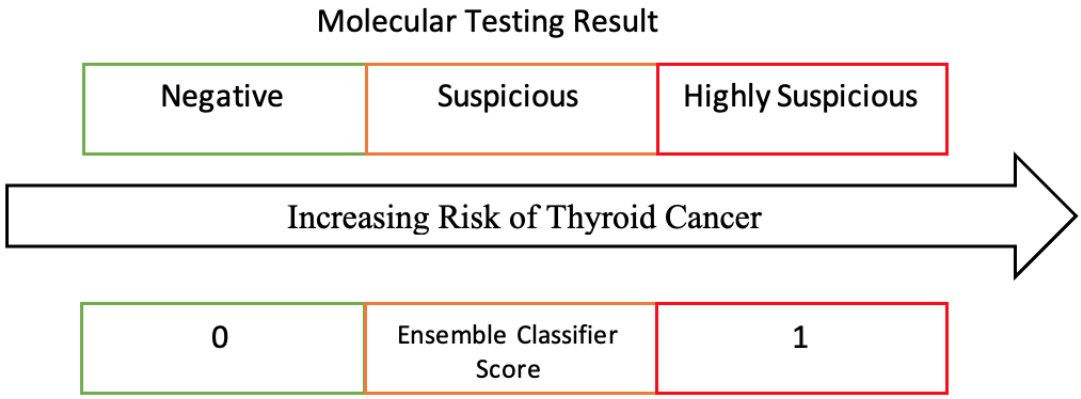

Figure 4.14: Schematic of the rule-based merging of molecular testing results and our image-based AI output.

4.4.2.2.3 Estimation of False Positive Surgeries Avoided

A primary goal for this chapter was the development of classification AIs to reduce the number of false positive surgeries on benign thyroid nodules. To estimate the additional surgeries avoided by the addition of our ensemble AI into the clinical workflow, exclusively molecular testing suspicious nodules were considered. ITN-MS represent the subset of nodules which are most often misclassified and result in the largest number of false positive

surgeries. To calculate how many false positive surgeries could be avoided while maintaining 100% sensitivity, the decision threshold was set to include the lowest scored surgically malignant nodule. All ITN-MS benign nodules, which had ensemble classifier scores below the threshold, were considered false positive surgeries avoided. An example of setting this decision threshold is provided in Figure 4.15.

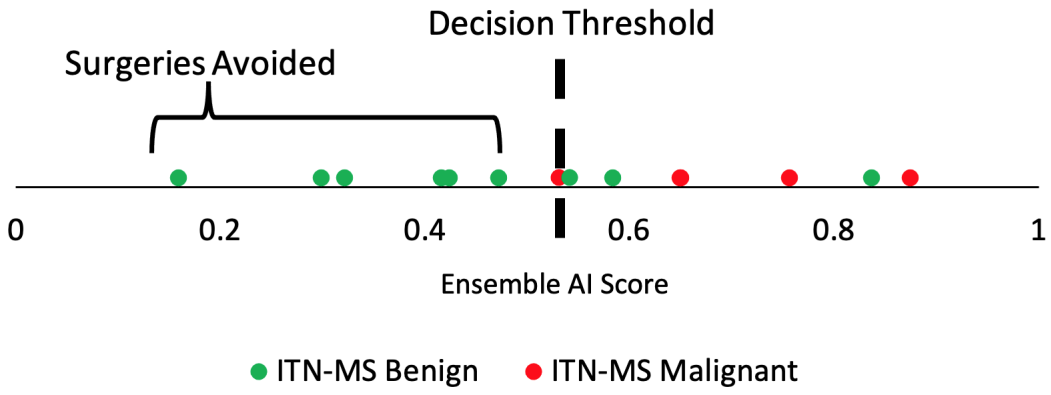

Figure 4.15: Example setting of decision threshold to maintain 100% sensitivity and determine false positive surgeries avoided with use of ensemble AI.

4.4.2.2.4 Evaluation of Image-Based Ensemble AI with Molecular Testing Classifiers

The performance of the image-based ensemble classifier and the joint image-based AI with molecular testing was first evaluated using the 5-fold cross validation on the training/validation dataset. Stratified 5-fold cross validation was conducted over 25 iterations and the average ROC AUC of the ensemble classifier was calculated.[91] The same features used to stratify the selected training/validation and independent testing set were applied to throughout 5-fold validation splitting.[119, 120] The ensemble AI algorithm alone and the

merged ensemble AI with molecular testing algorithm performances were calculated on the independent testing set.[91]

The percent of false positive surgeries avoided was calculated through each of the 5-fold cross validation folds across 25 iterations as well as on the independent test set. A 95% CI interval was calculated to evaluate if there was a statistically significant reduction in false positive surgeries. When estimating false positive surgeries avoided in the independent test set, a different threshold was set for each molecular testing platform to account for platform specific differences in suspicious designation, i.e., what Afirma identifies as suspicious, may be different than what ThyroSeq and ThyGeNext identify.[26]

4.4.2.3 Results

In 5-fold cross validation averaged over 25 iterations, the ensemble classifier alone yielded an average AUC [Std. Err.] of 0.73 [0.04], and after the integration with molecular testing, an average AUC [Std. Err.] of 0.83 [0.04]. On the independent test set, the ensemble classifier yielded an AUC [STD] of 0.70 [0.04] and with integration of molecular testing an AUC [STD] of 0.75 [0.05]. The ensemble algorithm classification results are summarized in Table 4.16.

Across the 25 5-fold iterations, on average [95% CI] 14.3% [7.3%,21.3%] of false positive surgeries were avoided, while maintaining 100% sensitivity. Applying the ensemble AI to the independent test set ThyroSeq suspicious nodules resulted in 2 of 5 false positive surgeries avoided. When applied to ThyGeNext suspicious nodules in the independent test set, the ensemble classifier led to 1 of 4 false positive surgeries avoided. Finally, when the ensemble classifier was applied to Afirma suspicious cases, no false positive surgeries were avoided.

Algorithm	5-Fold Cross Validation (25itr)		Independent Testing	
	Avg. AUC (Std Err.)	n	AUC (Std. Dev.)	n
All Indeterminate Cases – Ensemble AI	0.73 (0.04)	243	0.70 (0.04)	110
Cases with Molecular Testing– Ensemble AI + MT	0.83 (0.04)	136	0.75 (0.05)	58

Table 4.16: Summary of ensemble AI algorithm classification results with integration of molecular testing results. The number of nodules (n) included for each classification task is listed.

4.4.2.4 Discussion

The merging of the ensemble AI classifier and molecular testing results yielded the highest AUC for the classification of indeterminate thyroid nodules in both cross validation and independent testing. When applied to molecular suspicious cases, the ensemble AI classifier was shown, through cross validation, to statistically significantly reduce the number of surgeries performed on benign nodules. The independent testing molecular testing suspicious subsets were small; however, 40% of ThyroSeq and 25% of ThyGeNext suspicious benign lesions in the testing set could have been avoided with application of our ensemble classification AI.

4.5 Summary

In this chapter multiple radiomic, deep transfer learning, and ensemble classifiers were investigated. Integration of the radiomic algorithm with cytopathology Bethesda scores showed

statistically significant classification improvement.[24] The equivalence of radiomic classification performance with features calculated from physician contours and those calculated from our whole nodule segmentation AI was demonstrated up to a reasonable margin.[123] Pre-training deep transfer learning algorithms on RadImageNet task specific subsets was shown to have the potential to improve algorithm performance.[109] Finally, our ensemble radiomic and deep transfer learning AI algorithm yielded high classification performance on indeterminate nodules and showed the potential to reduce the number of false positive thyroid nodule surgeries.

CHAPTER 5

REFERENCE-STANDARD-FREE LESION SEGMENTATION PERFORMANCE METRIC (LESION SEGMENTATION-SURENESS SCORE)

5.1 Rationale for Sureness Metrics in AI Lesion Segmentation

Machine learning has demonstrated the ability to accurately segment abnormalities of interest on medical imaging. Machine learning enabled segmentation algorithms are already deployed in clinical settings for automatic contouring, lesion detection, and the calculation of diagnosis-relevant radiomic features.[124, 125, 126] Automatic segmentation can also be used in research settings to accelerate the time-intensive work required for manual segmentation.

Conventional metrics for evaluating segmentation algorithm performance are the Sorenson-Dice coefficient (DSC), a measure of the overlap of two shapes, and the average Hausdorff distance (aHD), the average distance from the edge of one shape to any point on the edge of the other shape.[70, 71] Both of these evaluation metrics, often used in independent testing of segmentation algorithm performance, require a reference standard (such as an expert's delineation of the contour). When deployed, segmentation algorithms are expected to perform near the testing average in aggregate; however, on any individual case, performance may vary greatly. Therefore, a case-by-case measure of algorithm segmentation performance, without the need for a reference standard, could be useful for evaluating algorithm output segmentations in real-time. This study proposes a reference-standard-free lesion segmenta-

tion performance metric: lesion segmentation-sureness (LeSS) score.

By developing and validating the LeSS score, this study aims to expand repeatability analysis to the challenges and applications of computerized segmentation.[127] Sureness metrics using repeatability analysis have already been developed and validated for machine learning classification tasks.[128, 129] Classification-sureness measures consistency in classification probabilities in order to weight the integration of an algorithm's output into clinical diagnosis making. The visuospatial nature of segmentation tasks presents the unique opportunity for segmentation-sureness analysis to convey information from individual image regions, as well as to provide a score for the image on the whole. In practice, if the algorithm is highly inconsistent on only a small region in the image, clinicians may only need to review that portion.

LeSS may also be useful in segmentation algorithm development. Knowledge of algorithm consistency on types of images or regions within images may allow researchers to better understand the strengths and weaknesses of their algorithm. It may also allow for algorithm improvement through up-sampling of low LeSS score types of images in training or by implementing rule-based segmentation within low LeSS score cases. Automatic segmentations can be used to calculate radiomic features of size, shape, and texture for lesion characterization and classification.[126, 123] LeSS-based algorithm adjustments could impact calculation of lesion radiomic features and potentially downstream lesion classification. A schematic of potential clinical and model development uses of LeSS analysis in segmentation tasks is presented in Figure 5.1.

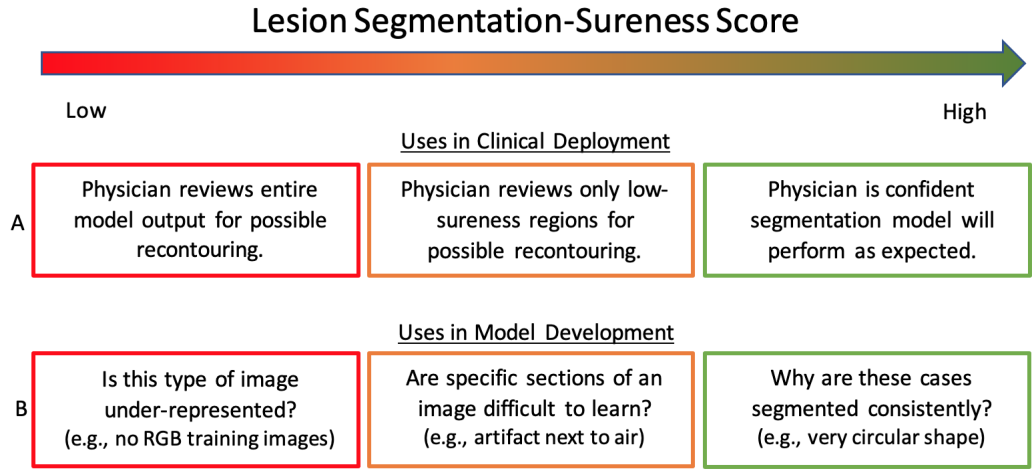

Figure 5.1: A) A use case for lesion segmentation-sureness score once algorithm has been deployed in a clinical setting. B) A use case for lesion segmentation-sureness analysis while the segmentation algorithm is under development in a research setting.

The segmentation of thyroid nodules on ultrasound is an important step in calculating a risk stratification score through the American College of Radiology's Thyroid Imaging Reporting and Data Systems (TIRADS).[17] This study will utilize the segmentation AI algorithm developed in Chapter 3.2 and radiomic classification AI developed in Chapter 4.2 for development and validation of lesion segmentation-sureness analysis.[58, 50]

5.2 Extraction of AI Lesion Segmentation-Sureness from Imaging Data

5.2.1 Methods

5.2.1.1 Dataset

For this study, a subset of the full dataset detailed in Chapter 2 was selected, including

cases from the University of Chicago Medicine (UCM) and the Weill Cornell Medical Center (WCMC). In total, 492 thyroid nodules were included in this study. As in Chapter 3, nodules were manually contoured by a group UCM physicians, and a consensus contour was applied when necessary. Physician contours were used as the reference standard in segmentation model training and evaluation. The dataset was split at the patient level into a training set containing cases from both UCM and WCMC and two independent testing sets, one from UCM (Test-UCM) and one from WCMC (Test-WCMC). A summary of the study datasets is presented in Table 5.1

Dataset	Malignant		Benign		Totals	
	Images	Nodules	Images	Nodules	Images	Nodules
Training Set	482	143	702	207	1184	350
Test-UCM	83	45	69	39	152	84
Test-WCMC	69	36	40	22	109	58

Table 5.1: Summary of training and testing datasets used for segmentation-sureness validation and testing

5.2.1.2 Segmentation Algorithm

The segmentation algorithm used for this study was an attention-enhanced U-Net (AttU-Net).[60, 67] Image preprocessing included the conversion of any RGB formatted images to grayscale by weighted averaging using matlabs rgb2gray function.[54] All images were cropped to a square region of interest and resized to 256x256.[61] Each segmentation algorithm used the same architecture with 12 internal filters. Segmentation prediction contours

were determined by finding the largest continuous border in the binarized segmentation prediction maps. The contour was then filled, such that there were no holes inside the segmentation output. This post-processing of the algorithm output ensured a single continuous segmentation mask. The preprocessing methodology and algorithm training schema used were validated and independently tested across two institutions in Cozzi et al., 2025, as presented in Chapter 3.2.[58]

5.2.1.3 Lesion Segmentation-Sureness Analysis and Score

5.2.1.3.1 Segmentation Repeatability Profile

To explore algorithm consistency, small changes in the training set were introduced and their effects on outputs were measured. Using a 0.632 bootstrapping procedure on the training set (TS), 100 sub-training sets (S-TS) were generated.[92] Each S-TS was then used to train an individual segmentation algorithm (SAttU-Net), resulting in 100 algorithms with the variation in each attributed to the variations in training set data. Using the SAttU-Net algorithms, 100 segmentation predictions were generated for each test image.[128, 129] The methodology performed to generate multiple segmentation predictions for each image is presented in Figure 5.2A.

The spatial discrepancies across the multiple segmentation outputs served as a measure of algorithm uncertainty at each point. Because thyroid nodules have gross circular or ovular shapes, a polar coordinate system allowed visualization along the axes of greatest variation (Fig. 5.2B). Each predicted contour was converted into polar coordinates using the average center of mass of all outputs as a standard center point. Interpolation along the contour every 0.07 degrees ensured all predicted contours had standardized angles for comparison.

The median algorithm output contour was determined at each polar angle. For every point on the median output contour, the empiric 95% confidence interval (95CI), the radial distance between the empiric 2.5% and 97.5% contours (Fig 5.2B), was calculated. An example segmentation repeatability profile is shown in polar coordinates as well as overlaid on the example image in Figure 5.2B.

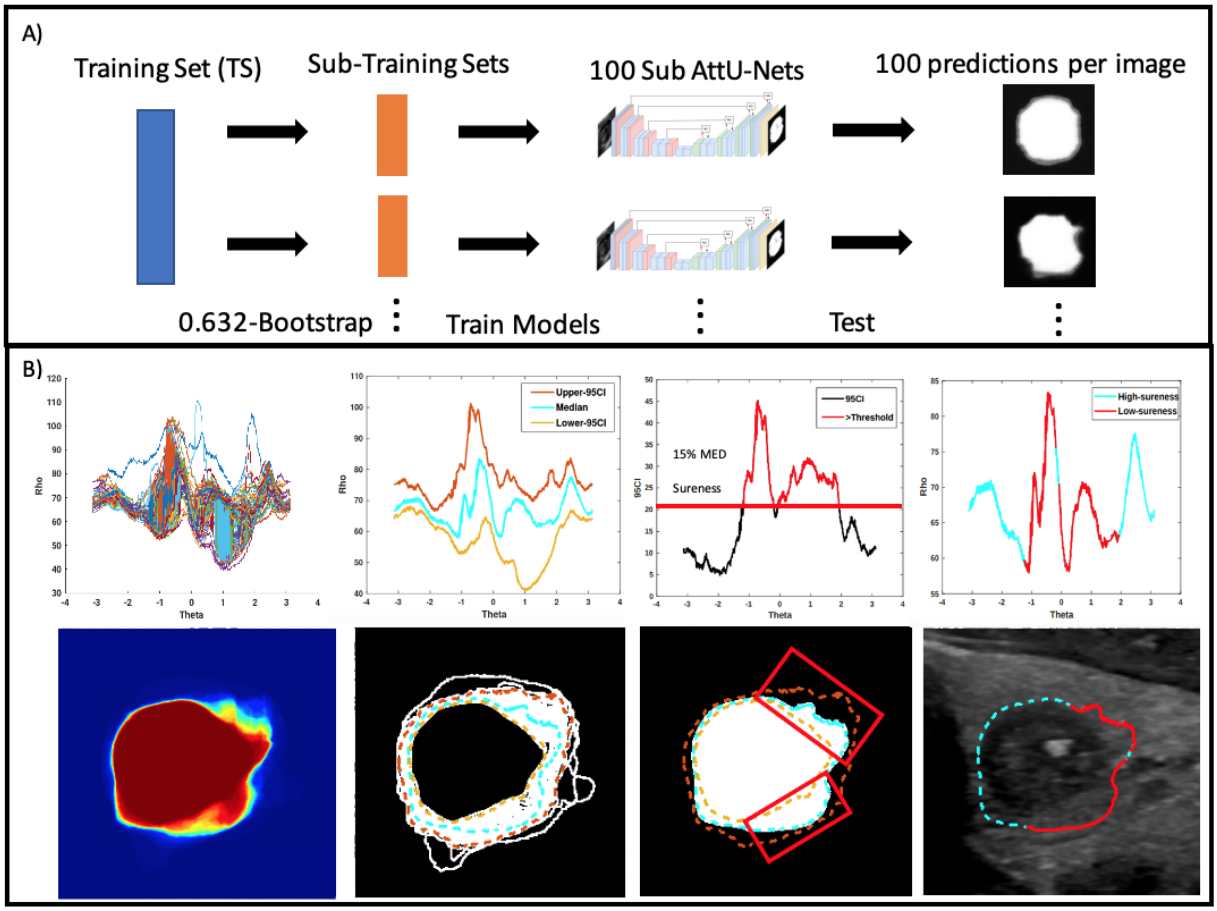

Figure 5.2: A) Methodology of building a segmentation repeatability distribution through bootstrapping. B) Example separation of high and low sureness regions and the calculation of lesion segmentation-sureness (LeSS) score shown in polar coordinates (top) and on the image (bottom). [Left to Right]: Graph of all 100 sub-prediction contours, determination of median contour and 95CI, separation of median contour into high and low sureness regions by 15% MED sureness margin, LeSS score equal to 0.54.

5.2.1.3.2 Calculation of Lesion Segmentation-Sureness Score (LeSS)

The median algorithm output contour was separated into high and low sureness regions based on algorithm variability quantified by each point's empiric 95CI. A sureness margin separated sections of the median contour into high-sureness (empiric 95CIs within the sureness margin) and low-sureness (empiric 95CIs outside the margin). In order to account for the range of nodule sizes present in our data set, this margin was calculated as a percent of the effective diameter of the algorithm's median output contour (MED). This strategy, in keeping with our methodology goals, does not require a reference truth. A margin of 15% MED was used for the testing of sureness analysis in this study of thyroid nodules. Margins of 10% MED and 20% MED were also explored during initial methodology validation. We note the sureness margin would be adjusted when used for other segmentation applications to best suit the specific end goals of those tasks. Finally, the lesion segmentation-sureness (LeSS) score was defined as the fraction of total contour categorized as high sureness, empiric 95CI within the sureness margin. The setting of a sureness margin and calculation of LeSS is demonstrated in Figure 5.2B.

5.2.1.4 Comparison of LeSS Analysis with Traditional Segmentation Metrics

The applicability of the LeSS score, as a reference-standard-free metric, was assessed in comparison to DSC, which required physician reference contour for calculation.[70] A higher DSC reflects better segmentation performance. The correlation between LeSS and DSC of the median algorithm output contour was tested. Segmentation performance differences between the low-sureness regions and high-sureness regions in an image were compared using the aHD.[71] A lower aHD reflects better segmentation performance. The effectiveness of

the sureness-based correction methods was assessed through the changes in the aHD of low-sureness regions before and after correction. Segmentation outputs with LeSS equal to 1 were excluded from this calculation as they had no regions corrected. Wilcoxon signed rank tests (alpha = 0.05) was used for statistical difference testing to account for the non-parametric data.[130] The Bonferroni multiple comparisons correction was applied when appropriate.[72] This testing was completed separately on both independent test sets.

5.2.2 Results

5.2.2.1 LeSS Segmentation-Sureness Analysis Results

At the 15% MED margin, the median LeSS was 0.94 for images in Test-UCM and 0.93 for images in Test-WCMC. In Test-UCM, 39 of 152 images had LeSS scores equal to 1, while 28 of 109 images were in Test-WCMC. All images were used in the testing of the LeSS score as a reference-standard-free lesion segmentation performance metric. Table 5.2 provides a summary of the LeSS results in relation to dataset pathology for Test-UCM and Test-WCMC.

Lesion Segmentation-Sureness Score	Test-UCM		Test-WCMC	
	Total Percent	Malignant (n)/ Benign (n)	Total Percent	Malignant(n)/ Benign (n)
1	26%	58% (23)/ 42% (16)	26%	68% (19)/ 32% (9)
<1	74%	53% (60)/ 47% (53)	74%	62% (50)/ 38% (31)

Table 5.2: Summary of Segmentation Results on Test-UCM and Test-WCMC. Images with segmentation-sureness less than 1 were used in examining sureness-based corrections in Chapter 5.3.

5.2.2.2 Comparison of LeSS Analysis with Traditional Segmentation Metrics

In the following section of independent testing results, values are presented: Test-UCM, Test-WCMC. The average DSC [Standard Deviation] for our segmentation methodology was 0.922 [0.04] on Test-UCM and 0.929 [0.04] on Test-WCMC. A statistically significant correlation between a nodule's LeSS score and its DSC was found in both independent test sets, Test-UCM and Test-WCMC, with Pearson correlation coefficients of 0.69 (p=2.0x10-22) and 0.58 (p=2.4x10-10), respectively. Within a contour, the aHD of low-sureness regions was found to be statistically significantly larger (worse segmentation) than that of high-sureness regions in both independent test sets (p=2.0x10-8, p=7.5x10-7). The median aHD of low-sureness regions was found to be 56% and 50% larger than high sureness regions in Test-UCM and Test-WCMC, respectively.

5.2.3 Discussion

Lesion segmentation-sureness analysis has the potential to augment segmentation algorithm usage by providing performance feedback to the user without the need for a reference standard. The strong correlation between LeSS score and DSC suggests the algorithm's variability can be indicative of segmentation performance. Because the LeSS score is calculated without a reference standard, it can highlight on a case-by-case basis images with which the segmentation algorithm may perform below average after algorithm deployment. In a clinical setting, those cases with low LeSS could be flagged for review by a physician. The low-sureness regions of a contour were shown to perform statistically significantly worse than sections within the sureness margin. Further detail, therefore, can be offered as part of the algorithm-user interface by displaying low-sureness regions visually, as in Figure 5.2B

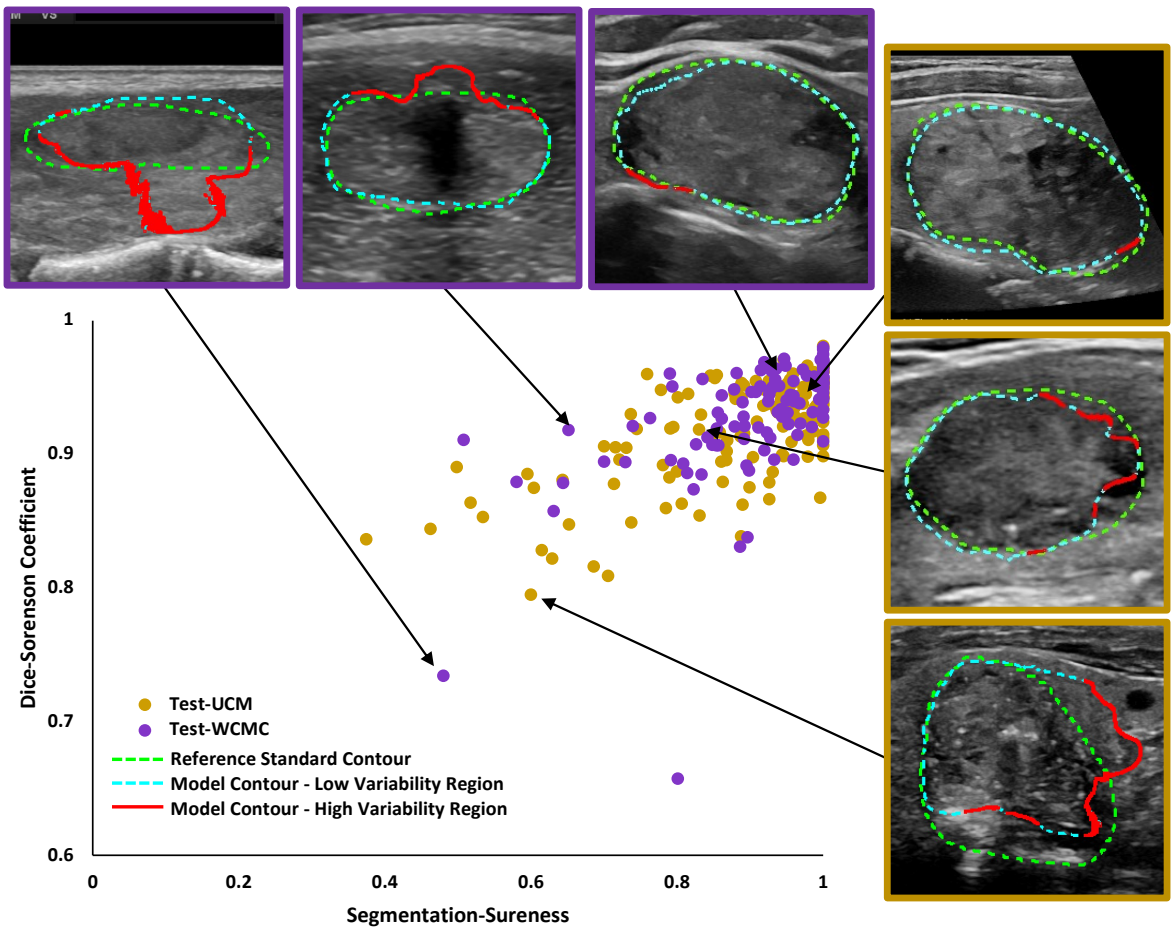

Figure 5.3: Lesion Segmentation-Sureness (LeSS) Score vs. DSC of Independent Test Sets. A statistically significant correlation was found in both test sets. Pearson correlation coefficients were 0.69 (p=2.0x10-22) for Test-UCM and 0.58 (p=2.4x10-10) for Test-WCMC, respectively. A set of representative examples from each test are shown with lesion segmentation-sureness analysis on the median algorithm contour highlighted vs. reference standard contours.

and Figure 5.3. This would provide the clinician not only with a measure of the algorithm's overall performance (LeSS score), but also where specifically within the image the algorithm is most likely to have erred. By setting the sureness margin, one can fine tune the sensitivity of this analysis to the demands of their specific segmentation task. There is a trade-off when setting the sureness margin between identifying poor performing cases and increasing the review workload of the algorithm user.

The lesion segmentation-sureness results provided information related to the development of our AttU-Net algorithm as well.[67, 58] LeSS scores were similarly distributed across institutions (UCM vs WCMC) and nodule pathology (benign vs malignant). This provides support for the sufficient distribution of those categories in training. Upon qualitative review, low-sureness regions were often positioned next to areas of low signal on the ultrasound (e.g., near edges or the tracheal border). This suggests our algorithm might benefit from the inclusion of more training cases with proximal nodules in transverse views.

5.3 Impact of Lesion Sureness-Based Corrections on Thyroid Ultrasound Segmentation Performance and on Radiomic Classification

5.3.1 Methods

5.3.1.1 Dataset and Segmentation Model

The same dataset and segmentation model were used in this study as in Chapter 5.2.[67, 58] A summary of the dataset can be found in Table 5.1. The corrections applied are based off the calculation of lesion segmentation-sureness, as described in Chapter 5.2. Only those images with LeSS scores < 1 were included for sureness-based correction analysis: 113 images in Test-UCM and 81 in Test-WCMC (Table 5.2). No correction could be applied to lesions with LeSS scores equal to 1.

5.3.1.2 Sureness-Based Lesion Segmentation Correction

Because low-sureness regions were where the segmentation algorithm was least consistent, they may benefit from rule-based automatic contouring. Two contour interpolation methods, cartesian and polar, were tested to remove variability in low-sureness regions and potentially improve segmentation results. In the cartesian interpolation correction, low-sureness regions are replaced with a linear interpolation between high-sureness regions within cartesian coordinates. In the polar interpolation correction, regions of low-sureness are replaced with an interpolation of the region within polar coordinates. This resulted a curved final contour approximating the expected gross-ovular shape of lesions. An example of each correction is shown in Figure 5.4.

5.3.1.3 Study of Segmentation-Sureness Analysis' Impact on Radiomic Features and Classification

For each nodule image and associated contour, 25 radiomic features were extracted, including 14 gray-level co-occurrence matrix features, 6 histogram-based features, 4 shape features, and 1 size feature.[86, 87, 50] These features were inputted into a support vector machine with a linear kernel for prediction of surgical nodule pathology, benign or malignant.[90] Malignancy scores of images from the same nodule were averaged to give a nodule-level score. Radiomic algorithm classification performance was measured at the nodule level. This thyroid nodule feature set and classification methodology is the same from Chapter 4.2. Nodule features were calculated from the physician reference standard contours (reference features), median AttU-Net output contours, polar corrected AttU-Net output contours, and cartesian corrected AttU-Net output contours. The percent difference

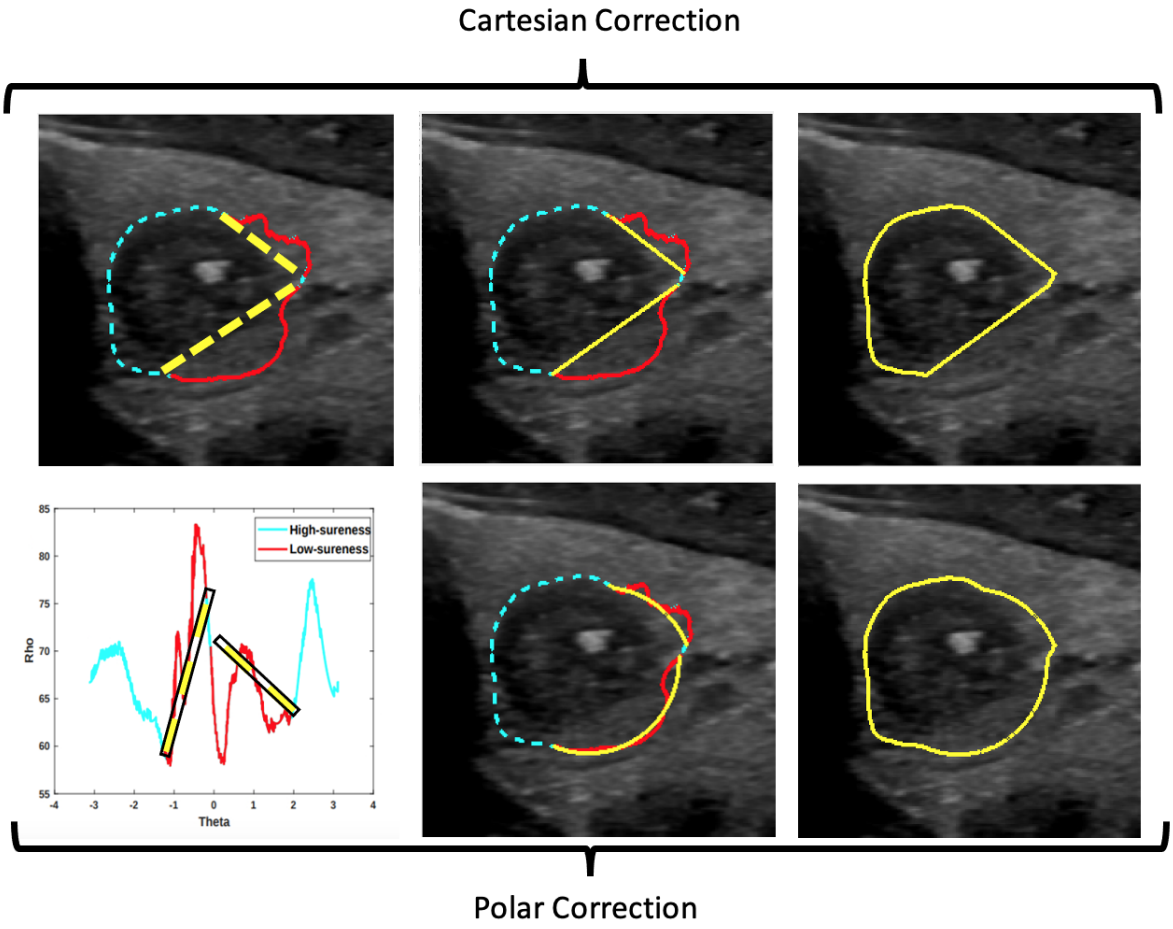

Figure 5.4: An example of the cartesian (top) and polar (bottom) segmentation corrections (yellow) applied to low-sureness regions (red).

between reference features and those calculated using the AttU-Net output contours was determined. The methodology of this comparison is the same as in Chapter 4.2.2.[123] Because the 4 shape features (irregularity, circularity, compactness, and depth-to-width ratio) and 1 size feature (effective diameter) are calculated directly from segmentation contour, statistical testing was performed on this subset. Wilcoxon signed rank test, with a Bonferroni multiple comparison correction (alpha-corrected = 0.005), was used to assess if sureness-based

corrections resulted in statistically significant more accurate or less accurate feature calculation.[130, 72]

The radiomic classifier algorithm was trained on the full training set of reference features. Radiomic classification algorithm performance was evaluated at the nodule level for each set of test features (reference, algorithm median output, algorithm cartesian corrected, algorithm polar corrected). In order to preserve statistical power while evaluating at the nodule level, classification evaluations were performed on all test cases combined (Test-UCM+Test-WCMC). The impact of the sureness-based segmentation corrections on classification performance was measured. Receiver operating characteristic (ROC) analysis was performed, and the area under the curve (AUC) from ROC analysis was used as the statistic of merit for classification performance evaluation. The Delong test (alpha=0.05) was used to test differences in classifier performance with and without sureness-based segmentation corrections applied.[131]

5.3.2 Results

5.3.2.1 Comparison of Sureness-Based Corrections with Traditional Segmentation Metrics

The polar correction yielded a statistically significant decrease (improved segmentation) in aHD of low-sureness regions in Test-UCM (p=2.6x10-6) and Test-WCMC (p=2.4x10-3). The polar correction yielded a median decrease in aHD of 9% (Test-UCM) and 6% (Test-WCMC) when applied. The cartesian correction failed to yield a statistically significant change in the aHD of low-sureness regions in either test set.

5.3.2.2 Study of LeSS Analysis' Impact on Radiomic Features and Classification

Irregularity and compactness features were statistically significantly ($p<0.005$) closer in value to reference features when calculated with the polar and the cartesian corrections applied than without corrections in both test sets. The percent difference from reference truth of radiomic features calculated using the AttU-Net median output without correction, with the cartesian correction applied, and with the polar correction applied are shown in Figure 5.5. The set of points in the upper right corner of each plot corresponds to the calculation of skewness, where even a small change in the contour can result in a large difference in feature value.

On all test cases, the radiomic classification algorithm yielded an AUC [95CI] of 0.712

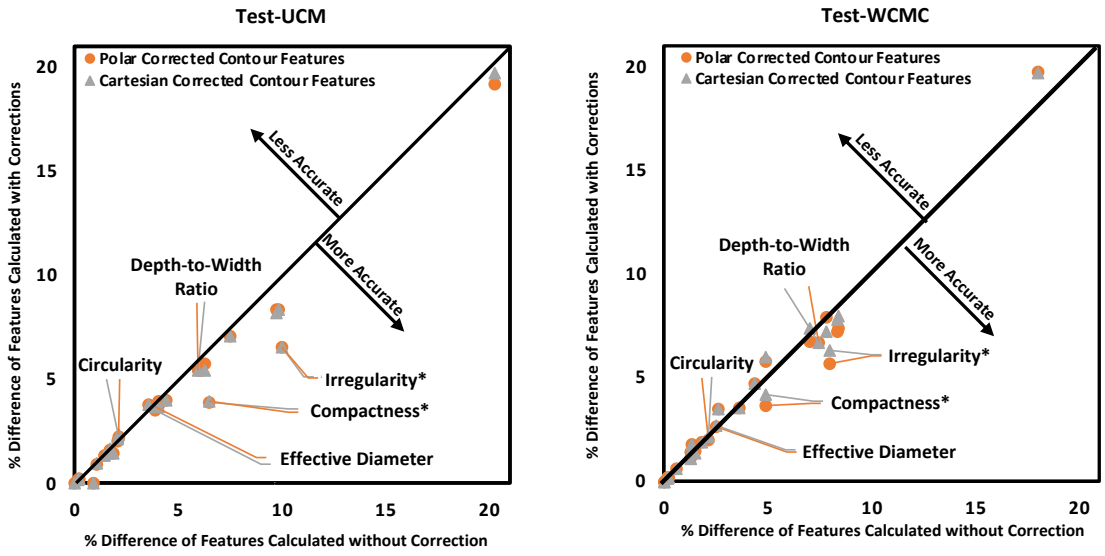

Figure 5.5: Percent difference from reference of radiomic features calculated from the AttU-Net median output contours without correction, with polar correction, and with cartesian correction for the UCM (left) and WCMC (right) test sets. Geometrical shape and size features calculated directly from the contours are labeled: irregularity, circularity, compactness, depth-to-width ratio, and effective diameter. Statistically significant differences determined by the Wilcoxon signed rank test (Bonferroni corrected alpha = 0.005) are shown with *.

[0.626, 0.799] with features extracted from reference contours, and AUC [95CI] of 0.717 [0.631,0.803] with features extracted from AttU-Net median contours. We failed to show statistically significant differences between classification performance with or without sureness-based corrections ($p>0.05$). The results of radiomic classification testing on cases with sureness-based correction applied are shown in Table 5.3.

Testing Set Radiomic Features Extracted from Different Contours	Test-UCM+Test-WCMC (LeSS Score < 1)	
	AUC	95% CI
AttU-Net Median Contour	0.726	[0.634, 0.818]
AttU-Net Median Contour w/ Polar Correction	0.731	[0.641, 0.822]
AttU-Net Median Contour w/ Cartesian Correction	0.731	[0.641, 0.821]

Table 5.3: Performance of radiomic algorithm testing results on features calculated AttU-Net median contours without correction, and with the polar and cartesian corrections applied. Test images with sureness-based correction applied (LeSS Score < 1) were used for this comparison.

5.3.3 Discussion

Sureness-based segmentation corrections have the potential to improve segmentation algorithm performance. In this study, only the polar correction improved the segmentation itself by decreasing the aHD.[71] As thyroid nodules are mostly ovular, it was expected that the polar correction, which resulted in a curved contour, would be more useful for this task. The cartesian correction may have more utility when segmenting objects with long straight edges, such as leg bones. Both corrections, polar and cartesian, made the calculation of two

radiomic shape features (irregularity and compactness) more accurate. [86, 87, 50] Despite improved segmentation and feature calculation, the sureness-based segmentation corrections did not result in a statistically significant increase in classification performance.

5.4 Summary

This chapter presented a framework for applying sureness analysis to segmentation tasks and introduced a new metric: lesion segmentation-sureness (LeSS) score. On the test case of segmenting thyroid nodules on ultrasound, LeSS analysis showed the potential to identify images on which the algorithm's segmentation performance may be below expected levels and highlight regions within a contour which are likely to have high deviations from reference standard. Sureness-based segmentation corrections were demonstrated to improve calculation of two shape features; however, we failed to show a statistically significant increase in classification performance when corrections were applied.

CHAPTER 6

SUMMARY AND FUTURE DIRECTIONS

6.1 Summary

The major contributions of this dissertation were the development of machine learning segmentation and classification algorithms for the pre-surgical classification of indeterminate thyroid nodules. Clinically informed integrations of AI algorithms and clinical tests were presented. Additionally, contributions to the field of algorithm repeatability were made through the novel application of sureness analysis to segmentation tasks and the calculation of a new reference-standard-free segmentation algorithm performance metric.

Chapter 1 established the clinical need for improved pre-surgical classification of indeterminate thyroid nodules. False-positive surgeries, benign nodules excised due to suspicious characteristics, occur because the current clinical diagnosis strategies are not yet sufficiently specific. While numerous advances in clinical nodule diagnostic tools have been made within the last decade, still half of all indeterminate nodules recommended for surgery are determined to be benign on post-surgical pathology. The goal of this dissertation was the development of machine learning algorithms which could aid in and improve indeterminate thyroid nodule classification. This aim was achieved through the development of task specific segmentation AIs, creation of an ensemble classification AI, and the establishment of a framework for segmentation AI 'self-correction'.

The whole nodule segmentation AI developed in Chapter 3 demonstrated high performance across all pathology subgroups. Additionally, the equivalence of classification per-

formance using segmentation AI contours versus physician contours for feature calculation supports the use of our segmentation AI algorithm in conjunction with the radiomic AI of Chapter 4.

The ultrasound-based classification AIs of Chapter 4 were evaluated for incorporation into multiple stages of the diagnostic pathway: 1) nodule evaluation on US, 2) nodule evaluation with FNA-biopsy, and 3) nodule evaluation with molecular testing. To the best of the authors knowledge, this represents the first evaluation of pre-training on the RadImageNet thyroid ultrasound subset. New ensemble classification architectures were evaluated to combine, not only the machine learning algorithms, but also the cytopathologic and molecular testing results. When applied to molecular testing suspicious nodules (ITN-MS), our ensemble AI demonstrated the potential to reduce false positive surgeries by 14.3% on average while maintaining 100% sensitivity. While this validation was performed on small number of ITN-MS cases, a 14.3% reduction would result in an estimated 4,054 benign surgeries avoided in the United States each year. This calculation is shown in Figure 6.1, based on the ∼600,000 FNA tests performed in the USA annually.[21]

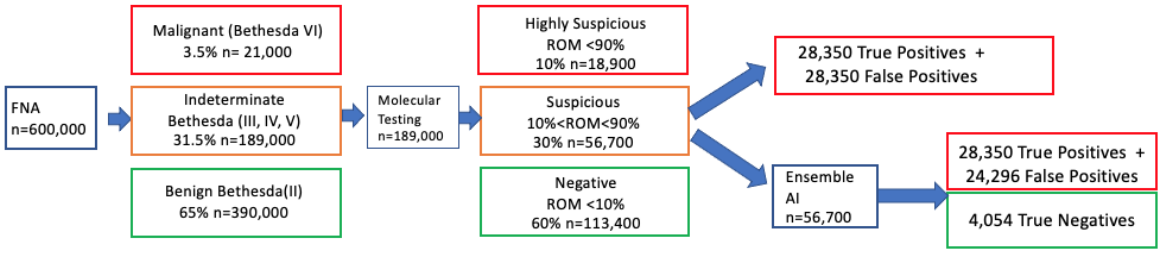

Figure 6.1: Estimate of false positive surgeries which could be avoided yearly with the 14.3% reduction demonstrated by our ensemble AI classification model.

The lesion segmentation-sureness (LeSS) score was developed in Chapter 5 and validated as a reference-standard-free measure of segmentation algorithm performance. In practice, the LeSS score could be used to automatically flag segmentations for review. The segmentation sureness-based corrections were shown to improve both the segmentation algorithm performance and the calculation of two shape features. In addition to their demonstrated use for the improvement of thyroid nodule segmentation, this analysis framework with sureness-based segmentation corrections could be applied for the improvement of segmentation algorithms for different tasks.

6.2 Future Directions

The results presented in this dissertation provide support for continued investigation into the development and integration of machine learning systems for the classification of indeterminate thyroid nodules. A primary limiting factor was the dataset size. Collection of more molecular suspicious cases would allow for a larger testing set and more statistical power. It would also allow for the development of ensemble classification AIs specific to each molecular testing platform. A larger dataset could additionally allow for increased training data when training DTL fine-tuned models. The promising results presented provide support for the potential use of these AI algorithms in a joint classification system: nodules segmented through the whole nodule segmentation AI algorithm (Chapter 3), segmentation sureness-based corrections applied to those segmentation outputs (Chapter 5), and the resulting radiomic features input to the ensemble classification AI (Chapter 4). Eventually, reader

studies with radiologist and endocrine surgeons could be conducted to test if the joint AI classification system has a meaningful impact on clinical decision making.

REFERENCES

[1] H. Sung *et al.*, "Global cancer statistics 2020; GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA Cancer J Clin*, vol. 71, no. 3, pp. 209–249, 2021.

[2] D. Dean and H. Gharib, "Epidemiology of thryroid nodules," *Best Pract Res Clin Encrinol Metab*, vol. 22, pp. 901–911, 2008.

[3] M. C. Frates *et al.*, "Prevalence and distribution of carcinoma in patients with solitary and multiple thyroid nodules on sonography," *J Clin Endocrinol Metab*, vol. 91, no. 9, pp. 3411–3417, 2006.

[4] R. M. Tuttle, D. W. B. D. Ball, and D. Q. Doherty, "Thyroid carinoma," *J Natl Compr Canc Netw*, vol. 8, no. 11, pp. 1228–1274, 2010.

[5] R. L. Siegel, T. B. Kratzer, A. N. Giaquinto, H. Sung, and A. Jemal, "Cancer statistics, 2025," en, *CA Cancer J. Clin.*, vol. 75, no. 1, pp. 10–45, Jan. 2025.

[6] R. Rahbari, L. Zhang, and E. Kebebew, "Thyroid cancer gender disparity," en, *Future Oncol.*, vol. 6, no. 11, pp. 1771–1779, Nov. 2010.

[7] A. C. Society, *Cancer facts & figures 2025*, en, https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/2025-cancer-facts-figures.html, Accessed: 2025-6-8.

[8] N. C. Institute, *Seer cancer stat facts: Thyroid cancer. 2024*, en, https://seer.cancer.gov/statfacts/html/thyro.html, Accessed: 2025-6-8.

[9] C. E. Barrows, J. M. Belle, A. Fleishman, C. C. Lubitz, and B. C. James, "Financial burden of thyroid cancer in the united states: An estimate of economic and psychological hardship among thyroid cancer survivors," en, *Surgery*, vol. 167, no. 2, pp. 378–384, Feb. 2020.

[10] V. Hsiao *et al.*, "Complication rates of total thyroidectomy vs hemithyroidectomy for treatment of papillary thyroid microcarcinoma: A systematic review and meta-analysis," en, *JAMA Otolaryngol. Head Neck Surg.*, vol. 148, no. 6, pp. 531–539, Jun. 2022.

[11] M. Wilson, A. Patel, W. Goldner, J. Baker, Z. Sayed, and A. L. Fingeret, "Postoperative thyroid hormone supplementation rates following thyroid lobectomy," en, *Am. J. Surg.*, vol. 220, no. 5, pp. 1169–1173, Nov. 2020.

[12] X. Keutgen, F. Filicori, and T. Fahey, "Molecular diagnosis for indeterminate thyroid nodules on fine needle aspiration: Advances and limitations," *Expert Rev Mol Diagn*, vol. 6, pp. 613–623, 2013.

[13] A. Y. Kargi, M. P. Bustamante, and S. Gulec, "Genomic profiling of thyroid nodules: Current role for ThyroSeq next-generation sequencing on clinical decision-making," en, *Mol. Imaging Radionucl. Ther.*, vol. 26, no. Suppl 1, pp. 24–35, Feb. 2017.

[14] B. R. Haugen *et al.*, "2015 american thyroid association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: The american thyroid association guidelines task force on thyroid nodules and differentiated thyroid cancer," en, *Thyroid*, vol. 26, no. 1, pp. 1–133, Jan. 2016.

[15] B. Gorman and C. C. Reading, "Ultrasonography, CT, MRI of the thyroid gland," in *Functional and Morphological Imaging of the Endocrine System*, Boston, MA: Springer US, 2000, pp. 73–102.

[16] J. I. Lew and C. C. Solorzano, "Use of ultrasound in the management of thyroid cancer," en, *Oncologist*, vol. 15, no. 3, pp. 253–258, Mar. 2010.

[17] F. N. Tessler *et al.*, "ACR thyroid imaging, reporting and data system (TI-RADS): White paper of the ACR TI-RADS committee," *J. Am. Coll. Radiol.*, vol. 14, no. 5, pp. 587–595, May 2017.

[18] E. J. Ha, D. G. Na, and J. H. Baek, "Korean thyroid imaging reporting and data system: Current status, challenges, and future perspectives," en, *Korean J. Radiol.*, vol. 22, no. 9, pp. 1569–1578, Sep. 2021.

[19] G. Russ, S. J. Bonnema, M. F. Erdogan, C. Durante, R. Ngu, and L. Leenhardt, "European thyroid association guidelines for ultrasound malignancy risk stratification of thyroid nodules in adults: The EU-TIRADS," en, *Eur. Thyroid J.*, vol. 6, no. 5, pp. 225–237, Sep. 2017.

[20] D. S. Dean and H. Gharib, *Fine-Needle Aspiration Biopsy of the Thyroid Gland*, E. South, Ed. Dartmouth, MA: National Library of Medicine, 2023.

[21] J. A. Sosa, J. W. Hanna, K. A. Robinson, and R. B. Lanman, "Increases in thyroid nodule fine-needle aspirations, operations, and diagnoses of thyroid cancer in the united states," en, *Surgery*, vol. 154, no. 6, 1420–6, discussion 1426–7, Dec. 2013.

[22] E. S. Cibas, S. Z. Ali, and NCI Thyroid FNA State of the Science Conference, "The bethesda system for reporting thyroid cytopathology," en, *Am. J. Clin. Pathol.*, vol. 132, no. 5, pp. 658–665, Nov. 2009.

[23] E. S. Cibas and S. Z. Ali, "The 2017 bethesda system for reporting thyroid cytopathology," en, *Thyroid*, vol. 27, no. 11, pp. 1341–1346, Nov. 2017.

[24] S. Z. Ali, Z. W. Baloch, B. Cochand-Priollet, F. C. Schmitt, P. Vielh, and P. A. VanderLaan, "The 2023 bethesda system for reporting thyroid cytopathology," en, *Thyroid*, vol. 33, no. 9, pp. 1039–1044, Sep. 2023.

[25] X. M. Keutgen, F. Filicori, and T. J. Fahey rd, "Molecular diagnosis for indeterminate thyroid nodules on fine needle aspiration: Advances and limitations," *Expert Rev. Mol. Diagn*, vol. 13, no. 6, pp. 613–623, 2013.

[26] M. Zhang and O. Lin, "Molecular testing of thyroid nodules: A review of current available tests for fine-needle aspiration specimens," en, *Arch. Pathol. Lab. Med.*, vol. 140, no. 12, pp. 1338–1344, Dec. 2016.

[27] J. Patel, J. Klopper, and E. E. Cottrill, "Molecular diagnostics in the evaluation of thyroid nodules: Current use and prospective opportunities," en, *Front. Endocrinol. (Lausanne)*, vol. 14, p. 1 101 410, Feb. 2023.

[28] C. E. Nasr *et al.*, "Real-world performance of the afirma genomic sequencing classifier (GSC)-A meta-analysis," en, *J. Clin. Endocrinol. Metab.*, vol. 108, no. 6, pp. 1526–1532, May 2023.

[29] H. Guan *et al.*, "The real-world performance of ThyroSeqV.2 to diagnose thyroid "neoplasm requiring surgery"," *Am J Cancer Res*, vol. 10, no. 11, pp. 3838–3851, 2020.

[30] Z. Yang, K. Ijaz, and M. Esebua, "Molecular testing using ThyGeNEXT/ThyraMIR in thyroid nodules with indeterminate cytology: A single medical institute experience," en, *Diagn. Cytopathol.*, vol. 53, no. 7, pp. 351–356, Jul. 2025.

[31] T. Paulraj and K. S. V. Chellliah, "Computer-aided diagnosis of lung cancer in computed tomography scans: A review," *Curr. Med. Imaging Rev.*, vol. 14, no. 3, pp. 374–388, May 2018.

[32] M. R. Mohebian, H. R. Marateb, M. Mansourian, M. A. Mañanas, and F. Mokarian, "A hybrid computer-aided-diagnosis system for prediction of breast cancer recurrence (HPBCR) using optimized ensemble learning," en, *Comput. Struct. Biotechnol. J.*, vol. 15, pp. 75–85, 2017.

[33] D. Sheth and M. L. Giger, "Artificial intelligence in the interpretation of breast cancer on MRI," en, *J. Magn. Reson. Imaging*, vol. 51, no. 5, pp. 1310–1324, May 2020.

[34] O. F. Ahmad *et al.*, "Artificial intelligence and computer-aided diagnosis in colonoscopy: Current evidence and future directions," en, *Lancet Gastroenterol. Hepatol.*, vol. 4, no. 1, pp. 71–80, Jan. 2019.

[35] P. V. Nayantara, S. Kamath, K. N. Manjunath, and K. V. Rajagopal, "Computer-aided diagnosis of liver lesions using CT images: A systematic review," en, *Comput. Biol. Med.*, vol. 127, no. 104035, Dec. 2020.

[36] N. Razmjooy *et al.*, "Computer-aided diagnosis of skin cancer: A review," en, *Curr. Med. Imaging Rev.*, vol. 16, no. 7, pp. 781–793, Jan. 2020.

[37] L. Xu *et al.*, "Computer-aided diagnosis systems in diagnosing malignant thyroid nodules on ultrasonography: A systematic review and meta-analysis," en, *Eur. Thyroid J.*, vol. 9, no. 4, pp. 186–193, 2020.

[38] F. Bini *et al.*, "Artificial intelligence in thyroid field-a comprehensive review," en, *Cancers (Basel)*, vol. 13, no. 19, p. 4740, Sep. 2021.

[39] D. Zhang *et al.*, "A review of the role of the S-Detect computer-aided diagnostic ultrasound system in the evaluation of benign and malignant breast and thyroid masses," en, *Med. Sci. Monit.*, vol. 27, e931957, Sep. 2021.

[40] E. J. Ha and J. H. Baek, "Applications of machine learning and deep learning to thyroid imaging: Where do we stand?" en, *Ultrasonography*, vol. 40, no. 1, pp. 23–29, Jan. 2021.

[41] J. Conn Busch, J. L. Cozzi, H. Li, L. Lan, M. L. Giger, and X. M. Keutgen, "Role of machine learning in differentiating benign from malignant indeterminate thyroid nodules: A literature review," *Health Sciences Review*, vol. 7, no. 3, pp. 379–423, 2023.

[42] M. J. Page *et al.*, "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," en, *Syst. Rev.*, vol. 10, no. 1, p. 89, Mar. 2021.

[43] P. F. Whiting *et al.*, "QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies," en, *Ann. Intern. Med.*, vol. 155, no. 8, pp. 529–536, Oct. 2011.

[44] J. M. Campbell *et al.*, "Diagnostic test accuracy: Methods for systematic review and meta-analysis," en, *Int. J. Evid. Based Healthc.*, vol. 13, no. 3, pp. 154–162, Sep. 2015.

[45] S. Gitto *et al.*, "A computer-aided diagnosis system for the assessment and characterization of low-to-high suspicion thyroid nodules on ultrasound," en, *Radiol. Med.*, vol. 124, no. 2, pp. 118–125, Feb. 2019.

[46] F.-S. Ouyang *et al.*, "Comparison between linear and nonlinear machine-learning algorithms for the classification of thyroid nodules," en, *Eur. J. Radiol.*, vol. 113, pp. 251–257, Apr. 2019.

[47] Y. Zhu, Q. Sang, S. Jia, Y. Wang, and T. Deyer, "Deep neural networks could differentiate bethesda class III versus class IV/V/VI," en, *Ann. Transl. Med.*, vol. 7, no. 11, p. 231, Jun. 2019.

[48] I. Youn *et al.*, "Diagnosing thyroid nodules with atypia of undetermined significance/follicular lesion of undetermined significance cytology with the deep convolutional neural network," en, *Sci. Rep.*, vol. 11, no. 1, p. 20 048, Oct. 2021.

[49] C.-K. Zhao *et al.*, "A comparative analysis of two machine learning-based diagnostic patterns with thyroid imaging reporting and data system for thyroid nodules: Diagnostic performance and unnecessary biopsy rate," en, *Thyroid*, vol. 31, no. 3, pp. 470–481, Mar. 2021.

[50] X. M. Keutgen *et al.*, "A machine-learning algorithm for distinguishing malignant from benign indeterminate thyroid nodules using ultrasound radiomic features," en, *J. Med. Imaging (Bellingham)*, vol. 9, no. 3, p. 034 501, May 2022.

[51] L.-Y. Gao *et al.*, "Ultrasound is helpful to differentiate bethesda class III thyroid nodules," en, *Medicine (Baltimore)*, vol. 96, no. 16, e6564, Apr. 2017.

[52] J. H. Lee *et al.*, "Risk stratification of thyroid nodules with atypia of undetermined significance/follicular lesion of undetermined significance (AUS/FLUS) cytology using ultrasonography patterns defined by the 2015 ATA guidelines," en, *Ann. Otol. Rhinol. Laryngol.*, vol. 126, no. 9, pp. 625–633, Sep. 2017.

[53] I.-R. Recommendation, *Studio encoding parameters of digital television for standard 4:3 and wide-screen 16:9 aspect ratios*, 2011.

[54] T. M. Inc, *Image processing toolbox version: 11.6 (r2022b)*, Natick, Massachusetts, 2022.

[55] J. A. Silver *et al.*, "BRAF V600E mutation is associated with aggressive features in papillary thyroid carcinomas $\leq$ 1.5 cm," en, *J. Otolaryngol. Head Neck Surg.*, vol. 50, no. 1, p. 63, Nov. 2021.

[56] C. Floridi *et al.*, "Ultrasound imaging classifiications of thyroid nodules for malignancy risk stratification and clinical management: State of the art," *Gland Surg*, no. 3, pp. 233–244, 2019.

[57] I. Renuka, G. Bala, C. Aparna, R. Kumari, and K. Sumalatha, "The bethesda system for reporting thyroid cytopathology: Interpretation and guidlines in surgical treatment," *Indian J Otolaryngol Head Neck Surg*, vol. 64, no. 4, pp. 305–311, 2012.

[58] J. L. Cozzi *et al.*, "Multi-institutional development and testing of attention-enhanced deep learning segmentation of thyroid nodules on ultrasound," en, *Int. J. Comput. Assist. Radiol. Surg.*, vol. 20, no. 2, pp. 259–267, Feb. 2025.

[59] X. Gao, X. Ran, and W. Ding, "The progress of radiomics in thyroid nodules," en, *Front. Oncol.*, vol. 13, p. 1 109 319, Mar. 2023.

[60] T. Falk *et al.*, "U-Net: Deep learning for cell counting, detection, and morphometry," *Nature Methods*, vol. 16, pp. 67–70, 2019.

[61] G. Bradski, "The OpenCV library," *Dr.Dobb's Journal*, vol. 25, no. 11, pp. 120–125, 2000.

[62] N. Siddique, S. Paheding, C. Elkin, and V. Devabhaktuni, "U-Net and its variants for medical image segmentation: A review of theory and applications," *IEEE Access*, vol. 9, pp. 82 031–82 057, 2021.

[63] A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[64] P. Anderson, X. B. C. He, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and vqa," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, Salt Lake City, 2018.

[65] S. Jetley, N. Lord, N. Lee, and P. Torr, "Learn to pay attention," in *6th International Confrence on Learning Prepresentations(ICLR)*, Vancouver, 2018.

[66] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations(ICLR)*, San Diego, 2015.

[67] J. Schlemper *et al.*, "Attention gated networks: Learning to leverage salient regions in medical images," *Medical Image Analysis*, vol. 53, pp. 197–207, 2019.

[68] J. Wu *et al.*, "U-Net combined with multi-scale attention mechanism for liver segmentation in CT images," *BMC Med Inform Decis Mak*, no. 283, 2021.

[69] J. Genender *et al.*, "Attention U-Net segmentation of indeterminate nodules on thyroid ultrasound," in *AAPM 64th Annual Meeting*, Washington, DC, 2022.

[70] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, Jul. 1945.

[71] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing images using the hausdorff distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 9, pp. 850–863, 1993.

[72] O. J. Dunn, "Multiple comparisons among means," *Journal of American Statistical Assocaition*, vol. 56, no. 293, pp. 52–63, 1961.

[73] S. Ahn, S. H. Park, and K. H. Lee, "How to demonstrate similarity by using noninferiority and equivalence statistical testing in radiology research," *Radiology*, vol. 267, no. 2, pp. 328–338, 2013.

[74] M. R. Ferreira *et al.*, "Comparative analysis for current deep learning networks for breast lesion segmentation in ultrasound images," *Annu Int Conf IEEE Med Biol Soc*, pp. 2878–3881, 2022.

[75] H. J. Lee *et al.*, "Intraobserver and interobserver variability in ultraosund measurements of thyroid nodules," *J Ultrasound Med*, vol. 37, no. 1, pp. 173–178, 2018.

[76] J. Alyani *et al.*, "Interobserver variability in ultrasound assessment of thyroid nodules," *Medicine (Baltimore)*, vol. 101, no. 41, 2022.

[77] D. Gustafson and W. Kessel, "Fuzzy clustering with a fuzzy covariance matrix," in *1978 IEEE Conference on Decision and Control including the 17th Symposium on Adaptive Processes*, San Diego, CA, USA: IEEE, Jan. 1978.

[78] J. C. Bezdek, *Pattern recognition with fuzzy objective function algorithms* (Advanced Applications in Pattern Recognition), en. New York, NY: Springer, May 2012.

[79] H. Zhou, G. Schaefer, and C. Shi, "Fuzzy c-means techniques for medical image segmentation," in *Fuzzy Systems in Bioinformatics and Computational Biology*, ser. Studies in fuzziness and soft computing, Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 257–271.

[80] W. Chen, M. L. Giger, U. Bick, and G. M. Newstead, "Automatic identification and classification of characteristic kinetic curves of breast lesions on DCE-MRI," en, *Med. Phys.*, vol. 33, no. 8, pp. 2878–2887, Aug. 2006.

[81] H. M. Whitney *et al.*, "AI-based automated segmentation for ovarian/adnexal masses and their internal components on ultrasound imaging," en, *J. Med. Imaging (Bellingham)*, vol. 11, no. 4, p. 044 505, Jul. 2024.

[82] M. Althouse and C.-I. Chang, "Image segmentation by local entropy methods," in *Proceedings., International Conference on Image Processing*, vol. 3, 1995, 61–64 vol.3.

[83] X. Xie and G. Beni, "A validity measure for fuzzy clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 11, pp. 841–847, 1991.

[84] Y. Wang, "Comparison study of radiomics and deep Learning-Based methods for thyroid nodules classification using ultrasound images," *IEEE Access*, vol. 8, pp. 52 010–52 017, 2020.

[85] V. S. Parekh and M. A. Jacobs, "Deep learning and radiomics in precision medicine," *Expert Rev Precis Med Drug Dev*, vol. 4, no. 2, pp. 59–72, 2019.

[86] M. L. Giger, "Computer-aided diagnosis of breast lesions in medical images," *Comput. Sci. Eng.*, vol. 2, no. 5, pp. 39–45, 2000.

[87] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-3, no. 6, pp. 610–621, 1973. DOI: 10.1109/TSMC.1973.4309314.

[88] E. M. Fisher, "Linear discriminant analysis," *Statistics & Discrete Methods of Data Sciences*, vol. 392, pp. 1–5, 1936.

[89] I. Gokcen and J. Peng, "Comparing linear discriminant analysis and support vector machines," in *Advances in Information Systems*, ser. Lecture notes in computer science, Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 104–113.

[90] C. Cortes and V. Vapnik, "Support-vector networks," en, *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.

[91] C. E. Metz, "Basic principles of ROC analysis," *Semin. Nucl. Med.*, vol. 8, no. 4, pp. 283–298, Oct. 1978.

[92] B. Efron and R. Tibshirani, "Improvements on cross-validation: The .632+ bootstrap method," *J. Am. Stat. Assoc.*, vol. 92, no. 438, p. 548, Jun. 1997.

[93] M. J. Magister, I. Chaikhoutdinov, E. Schaefer, N. Williams, B. Saunders, and D. Goldenberg, "Association of thyroid nodule size and bethesda class with rate of malignant disease," en, *JAMA Otolaryngol. Head Neck Surg.*, vol. 141, no. 12, pp. 1089–1095, Dec. 2015.

[94] C. Chen, N. A. Mat Isa, and X. Liu, "A review of convolutional neural network based methods for medical image classification," en, *Comput. Biol. Med.*, vol. 185, no. 109507, Feb. 2025.

[95] M. Iman, H. R. Arabnia, and K. Rasheed, "A review of deep transfer learning and recent advancements," en, *Technologies (Basel)*, vol. 11, no. 2, p. 40, Mar. 2023.

[96] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, "Transfer learning for medical image classification: A literature review," en, *BMC Med. Imaging*, vol. 22, no. 1, p. 69, Apr. 2022.

[97] N. Antropova, B. Q. Huynh, and M. L. Giger, "A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets," *Medical physics*, vol. 44, pp. 5162–5171, 2017.

[98] B. Q. Huynh, H. Li, and M. L. Giger, "Digital mammographic tumor classification using transfer learning from deep convolutional neural networks," *J. Med. Imaging (Bellingham)*, vol. 3, no. 3, p. 034 501, Jul. 2016.

[99] A. Davila, J. Colan, and Y. Hasegawa, "Comparison of fine-tuning strategies for transfer learning in medical image classification," en, *Image Vis. Comput.*, vol. 146, no. 105012, p. 105 012, Jun. 2024.

[100] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[101] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations (ICLR 2015)*, 2015, pp. 1–14.

[102] K. Mendel, H. Li, D. Sheth, and M. Giger, "Transfer learning from convolutional neural networks for computer-aided diagnosis: A comparison of digital breast ftomosynthesis and full-field digital mammography," *Academic radiology*, vol. 26, no. 6, pp. 735–743, 2019.

[103] M. A. Salam, A. Taher, M. Samy, and K. Mohamed, "The effect of different dimensionality reduction techniques on machine learning overfitting problem," en, *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 4, 2021.

[104] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Ann. Math. Stat.*, vol. 18, no. 1, pp. 50–60, Mar. 1947.

[105] K. Pearson, "LIII. on lines and planes of closest fit to systems of points in space," *Lond. Edinb. Dublin Philos. Mag. J. Sci.*, vol. 2, no. 11, pp. 559–572, Nov. 1901.

[106] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826.

[107] G. Huang, Z. Liu, L. v. d. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI: IEEE, Jul. 2017.

[108] M. A. Morid, A. Borjali, and G. Del Fiol, "A scoping review of transfer learning research on medical image analysis using ImageNet," en, *Comput. Biol. Med.*, vol. 128, no. 104115, p. 104 115, Jan. 2021.

[109] X. Mei *et al.*, "RadImageNet: An open radiologic deep learning research dataset for effective transfer learning," en, *Radiol. Artif. Intell.*, vol. 4, no. 5, e210315, Sep. 2022.

[110] H. M. Whitney, H. Li, Y. Ji, P. Liu, and M. L. Giger, "Comparison of breast MRI tumor classification using Human-Engineered radiomics, transfer learning from deep convolutional neural networks, and fusion methods," *Proceedings of the IEEE*, vol. 108, pp. 163–177, 2020.

[111] R. Anderson, H. Li, Y. Ji, P. Liu, and M. L. Giger, "Evaluating deep learning techniques for dynamic contrast-enhanced MRI in the diagnosis of breast cancer"," *SPIE*, vol. 2019, pp. 26–32, 2019.

[112] W. Chen, M. L. Giger, H. Li, U. Bick, and G. M. Newstead, "Volumetric texture analysis of breast lesions on contrast-enhanced magnetic resonance images," *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 58, no. 3, pp. 562–571, 2007.

[113] N. Bhooshan *et al.*, "Combined use of t2-weighted MRI and t1-weighted dynamic contrast-enhanced MRI in the automated analysis of breast lesions," en, *Magn. Reson. Med.*, vol. 66, no. 2, pp. 555–564, Aug. 2011.

[114] K. Robinson, H. Li, L. Lan, D. Schacht, and M. Giger, "Radiomics robustness assessment and classification evaluation: A two-stage method demonstrated on multivendor FFDM," *Medical physics*, vol. 46, no. 5, pp. 2145–2156, 2019.

[115] L. Davies and H. G. Welch, "Thyroid cancer survival in the united states: Observational data from 1973 to 2005," *Arch Otolaryngol Head Neck Surg*, vol. 136, no. 5, pp. 440–444, 2005.

[116] M. Reyes *et al.*, "On the interpretability of artificial intelligence in radiology: Challenges and opportunities," *Radiol Artif Intell*, vol. 2, no. 3, 2020.

[117] Q. Wei and R. L. Dunbrack Jr, "The role of balanced training and testing data sets for binary classifiers in bioinformatics," en, *PLoS One*, vol. 8, no. 7, e67863, Jul. 2013.

[118] H. M. Whitney *et al.*, "Longitudinal assessment of demographic representativeness in the medical imaging and data resource center open data commons," *J. Med. Imaging (Bellingham)*, vol. 10, no. 6, p. 61 105, Nov. 2023.

[119] N. Baughan *et al.*, "Sequestration of imaging studies in MIDRC: Stratified sampling to balance demographic characteristics of patients in a multi-institutional data commons," en, *J. Med. Imaging (Bellingham)*, vol. 10, no. 6, p. 064 501, Nov. 2023.

[120] MIDRC, *Algorithms*, https://www.midrc.org/algorithms, Accessed: 2025-7-27.

[121] J. Lin, "Divergence measures based on the shannon entropy," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 145–151, 1991. DOI: 10.1109/18.61115.

[122] J. L. Cozzi *et al.*, "Novel integration of radiomics and deep transfer learning for diagnosis of indeterminate thyroid nodules on ultrasound," in *Medical Imaging 2023: Computer-Aided Diagnosis*, K. M. Iftekharuddin and W. Chen, Eds., San Diego, United States: SPIE, Apr. 2023, p. 5.

[123] J. L. Cozzi, H. Li, L. Lan, X. Keutgen, and M. L. Giger, "Evaluation of automatic segmentations through performance of radiomic features in the classification of thyroid nodules on ultrasound," in *AAPM 66th Annual Meeting*, Los Angeles, CA, 2024.

[124] P. J. Doolan *et al.*, "A clinical evaluation of the performance of five commercial artificial intelligence contouring systems for radiotherapy," en, *Front. Oncol.*, vol. 13, p. 1 213 068, Aug. 2023.

[125] K. G. van Leeuwen *et al.*, "Comparison of commercial AI software performance for radiograph lung nodule detection and bone age prediction," en, *Radiology*, vol. 310, no. 1, e230981, Jan. 2024.

[126] R. N. Finnegan *et al.*, "Geometric and dosimetric evaluation of a commercial AI autocontouring tool on multiple anatomical sites in CT scans," en, *J. Appl. Clin. Med. Phys.*, e70067, Mar. 2025.

[127] S. S. Alahmari, D. B. Goldgof, P. R. Mouton, and L. O. Hall, "Challenges for the repeatability of deep learning models," *IEEE Access*, vol. 8, pp. 211 860–211 868, 2020.

[128] K. Drukker, L. Pesce, and M. Giger, "Repeatability in computer-aided diagnosis: Application to breast cancer diagnosis on sonography," en, *Med. Phys.*, vol. 37, no. 6, pp. 2659–2669, Jun. 2010.

[129] H. M. Whitney *et al.*, "Role of sureness in evaluating AI/CADx: Lesion-based repeatability of machine learning classification performance on breast MRI," en, *Med. Phys.*, vol. 51, no. 3, pp. 1812–1821, Mar. 2024.

[130] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics*, vol. 1, pp. 80–83, 1945.

[131] E. R. Delong, D. M. Delong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach," *Biometrics*, vol. 44, no. 3, pp. 837–845, 1988.

[132] S. T. Mufti and R. Molah, "The bethesda system for reporting thyroid cytopathology: A five-year retrospective review of one center experience," en, *Int. J. Health Sci. (Qassim)*, vol. 6, no. 2, pp. 159–173, Jun. 2012.

[133] S. Guth, U. Theune, J. Aberle, A. Galach, and C. M. Bamberger, "Very high prevalence of thyroid nodules detected by high frequency (13 MHz) ultrasound examination," en, *Eur. J. Clin. Invest.*, vol. 39, no. 8, pp. 699–706, Aug. 2009.

[134] G. H. Tan and H. Gharib, "Thyroid incidentalomas: Management approaches to nonpalpable nodules discovered incidentally on thyroid imaging," en, *Ann. Intern. Med.*, vol. 126, no. 3, pp. 226–231, Feb. 1997.

[135] J. P. Brito *et al.*, "The accuracy of thyroid nodule ultrasound to predict thyroid cancer: Systematic review and meta-analysis," en, *J. Clin. Endocrinol. Metab.*, vol. 99, no. 4, pp. 1253–1263, Apr. 2014.

[136] D. C. Malheiros, S. Canberk, D. N. Poller, and F. Schmitt, "Thyroid fnac: Causes of false-positive results," en, *Cytopathology*, vol. 29, no. 5, pp. 407–417, Oct. 2018.

[137] Ş. Canberk, P. Fırat, and F. Schmitt, "Pitfalls in the cytological assessment of thyroid nodules," en, *Turk Patoloji Derg.*, vol. 31 Suppl 1, pp. 18–33, 2015.

[138] Y. E. Nikiforov *et al.*, "Impact of the multi-gene ThyroSeq next-generation sequencing assay on cancer diagnosis in thyroid nodules with atypia of undetermined significance/follicular lesion of undetermined significance cytology," en, *Thyroid*, vol. 25, no. 11, pp. 1217–1223, Nov. 2015.

[139] P. Valderrabano *et al.*, "Evaluation of ThyroSeq v2 performance in thyroid nodules with indeterminate cytology," *Endocr. Relat. Cancer*, vol. 24, no. 3, pp. 127–136, Mar. 2017.

[140] M. Nishino, "Molecular cytopathology for thyroid nodules: A review of methodology and test performance," en, *Cancer Cytopathol.*, vol. 124, no. 1, pp. 14–27, Jan. 2016.

[141] H.-B. Zhao *et al.*, "A comparison between deep learning convolutional neural networks and radiologists in the differentiation of benign and malignant thyroid nodules on CT images," en, *Endokrynol. Pol.*, vol. 72, no. 3, pp. 217–225, Feb. 2021.

[142] S. Guth, U. Theune, J. Aberle, A. Galach, and C. M. Bamberger, "Very high prevalence of thyroid nodules detected by high frequency (13 MHz) ultrasound examination," en, *Eur. J. Clin. Invest.*, vol. 39, no. 8, pp. 699–706, Aug. 2009.

[143] M. Frates *et al.*, "Prevelence and distribution of carinoma in patients with solitary and multiple thyroid nodules on sonography," *J Clin Endocrinol Metab*, vol. 91, no. 9, 2006.

[144] A. Y. Kargi, M. P. Bustamante, and S. Gulec, "Genomic profiling of thyroid nodules: Current role for ThyroSeq Next-Generation sequencing on clinical Decision-Making," *Mol Imaging Radionucl Ther*, vol. 26, no. 1, pp. 24–35, 2017.

[145] R. M. Tuttle and A. S. Alzahrani, "Risk stratification in differentiated thyroid cancer: From detection to final follow-up," *J Clin Endocrinol Metab*, vol. 104, no. 9, pp. 4087–4100, 2019.

[146] X. Goa, X. Ran, and W. Ding, "The progress of radiomics in thyroid nodules," *Front Oncol*, vol. 13, no. 1109319, 2023.

[147] T. M. Inc, *Fuzzy logic toolbox (r2023b)*, Natick, Massachusetts, 2023.

[148] Z. Zhang, "Variable selection with stepwise and best subset approaches," en, *Ann. Transl. Med.*, vol. 4, no. 7, p. 136, Apr. 2016.

[149] K. Leclair, K. Bell, L. Furuya-Kanamori, S. A. Doi, D. O. Francis, and L. Davies, "Evaluation of gender inequity in thyroid cancer diagnosis: Differences by sex in US thyroid cancer incidence compared with a meta-analysis of subclinical thyroid cancer rates at autopsy," *JAMA Intern Med*, vol. 1, no. 10, pp. 1351–1358, 2021.

[150] B. R. Haugen, "American thyroid association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: What is new and what has changed?" *Cancer*, vol. 123, no. 3, pp. 372–381, 2017.

[151] E. J. Ha, "US fine-needle aspiration biopsy for thyroid malignancy: Diagnostic performance of seven society guidelines applied to 2000 thyroid nodules," *Radiology*, vol. 287, no. 3, pp. 893–900, 2018.

[152] S. Sorrenti *et al.*, "Artificial intelligence for thyroid nodule characterization: Where are we standing?" *Cancers (Basel)*, vol. 14, no. 14, 2022.

[153] J. L. Cozzi *et al.*, "Assessing role of feature selection in deep transfer classification of indeterminant thyroid nodules," in *AAPM 64th Annual Meeting*, vol. 49, Washington, DC, 2022, E690–E691.

PEER-REVIEWED PUBLICATIONS

[41] J. Conn Busch, J. L. Cozzi, H. Li, L. Lan, M. L. Giger, and X. M. Keutgen, "Role of machine learning in differentiating benign from malignant indeterminate thyroid nodules: A literature review," *Health Sciences Review*, vol. 7, no. 3, pp. 379–423, 2023.

[58] J. L. Cozzi *et al.*, "Multi-institutional development and testing of attention-enhanced deep learning segmentation of thyroid nodules on ultrasound," en, *Int. J. Comput. Assist. Radiol. Surg.*, vol. 20, no. 2, pp. 259–267, Feb. 2025.

ABSTRACTS AND PRESENTATIONS

[122] J. L. Cozzi *et al.*, "Novel integration of radiomics and deep transfer learning for diagnosis of indeterminate thyroid nodules on ultrasound," in *Medical Imaging 2023: Computer-Aided Diagnosis*, K. M. Iftekharuddin and W. Chen, Eds., San Diego, United States: SPIE, Apr. 2023, p. 5.

[123] J. L. Cozzi, H. Li, L. Lan, X. Keutgen, and M. L. Giger, "Evaluation of automatic segmentations through performance of radiomic features in the classification of thyroid nodules on ultrasound," in *AAPM 66th Annual Meeting*, Los Angeles, CA, 2024.

[153] J. L. Cozzi *et al.*, "Assessing role of feature selection in deep transfer classification of indeterminant thyroid nodules," in *AAPM 64th Annual Meeting*, vol. 49, Washington, DC, 2022, E690–E691.