A Multimodal Deep Learning Approach Integrated with Clinically Constrained Counterfactual Explainable AI (CXAI) for Shared Decision-Makings

Author

Abe Cit

Abstract

Thyroid nodule malignancy risk stratification remains a significant clinical challenge, with ultrasound-guided fine-needle aspiration (FNA) biopsies frequently yielding indeterminate cytology results (20% of cases) that complicate treatment decisions . While deep learning has shown promise in medical image classification, existing approaches predominantly rely on single-modality analysis, limiting diagnostic accuracy . Furthermore, the "black box" nature of AI systems impedes clinical adoption and shared decision-making between clinicians and patients . This study proposes and validates a multimodal deep learning framework that integrates B-mode ultrasound, strain elastography, and clinical text reports through a bidirectional cross-modal attention mechanism, enhanced by a novel Clinically Constrained Counterfactual Explainable AI (CXAI) module. The CXAI module generates clinically meaningful counterfactual explanations by systematically modifying key imaging and clinical features to demonstrate how diagnostic decisions would change under alternative evidence conditions . The framework was developed on 1,472 cases from Xi'an International Medical Center Hospital and externally validated on 4,530 cases across two clinical centers and public datasets (DDTI, TN3K). Our approach achieved an AUC of 0.937 (95% CI: 0.914–0.960) on internal validation and 0.896 (95% CI: 0.887–0.905) on external validation . The multimodal model significantly outperformed single-modal baseline models including ResNet-50 (AUC:

0.841), DenseNet-121 (AUC: 0.848), and Vision Transformer (AUC: 0.835) (all $p < 0.001$). The CXAI module generated explanations that correlated with clinically recognized biomarkers, enabling radiologists to verify diagnostic decisions and facilitating meaningful shared decision-making. This framework addresses critical barriers to AI adoption in thyroid care by combining superior predictive performance with transparent, clinically meaningful explanations.

Keywords: Thyroid Nodule Malignancy, Multimodal Deep Learning, Counterfactual Explainable AI, Shared Decision-Making, Ultrasound Image Analysis

1. Introduction

1.1 Background

Thyroid nodules are a prevalent clinical finding, with ultrasound detection rates reaching as high as 46.8% in general populations . While the majority of these nodules are benign, approximately 5–15% are malignant , and thyroid cancer has become the fastest-growing malignancy globally . The primary diagnostic pathway for suspicious thyroid nodules involves ultrasound-guided fine-needle aspiration (FNA) biopsy, which serves as the minimally invasive gold standard for cytological evaluation . However, FNA cytology yields indeterminate results—classified as atypia of undetermined significance (AUS) or follicular lesion of undetermined significance (FLUS)—in approximately 20% of cases . This diagnostic gray zone presents a substantial clinical challenge, as patients with indeterminate cytology face difficult decisions between repeat biopsy, molecular testing, diagnostic surgery, or surveillance .

The Thyroid Imaging Reporting and Data System (TI-RADS), developed by the American College of Radiology, provides a standardized framework for interpreting ultrasound features and stratifying malignancy risk . However, subjective interpretation and inter-observer variability continue to limit diagnostic accuracy, particularly for intermediate-risk nodules classified as TI-RADS 4 . The diagnostic challenge is particularly acute for nodules with high-risk features that do not clearly fall into benign or malignant categories, leading to potential overtreatment through unnecessary thyroidectomies or undertreatment due to missed malignancies .

1.2 Problem Statement

Recent advances in artificial intelligence (AI), particularly deep learning, have demonstrated substantial potential in medical image analysis, with some studies reporting diagnostic performance comparable to or exceeding that of human experts in thyroid ultrasound interpretation . Deep learning models trained on large-scale ultrasound datasets have shown promise in automated nodule detection, feature extraction, and malignancy classification . However, several critical limitations impede clinical adoption and integration into shared decision-making processes.

First, the majority of current deep learning approaches for thyroid nodule analysis rely on single-modality inputs, typically B-mode ultrasound images alone . This approach fails to leverage complementary information available from multiple clinical data sources, including strain elastography (which measures tissue stiffness), Doppler ultrasound (which assesses vascularity), and clinical text reports containing patient history and laboratory findings . Recent evidence demonstrates that multimodal integration significantly improves diagnostic accuracy compared to single-modality approaches .

Second, the "black box" nature of deep learning models presents a substantial barrier to clinical trust and adoption . Clinicians require interpretable explanations for AI-generated recommendations to verify diagnostic decisions and effectively communicate risks to patients. While post-hoc explainability techniques such as SHAP and LIME have been applied to thyroid disease classification , these methods often fail to provide clinically meaningful explanations that align with established medical knowledge .

Third, the integration of AI decision support into shared decision-making between clinicians and patients remains underdeveloped . Effective shared decision-making requires transparent communication of diagnostic uncertainty and the reasoning behind clinical recommendations. Existing AI systems lack mechanisms for demonstrating how diagnostic conclusions would change under alternative clinical scenarios or feature presentations.

1.3 Objectives of the Study

General Objective:
To develop and validate a multimodal deep learning framework for thyroid nodule malignancy stratification that integrates ultrasound imaging modalities with clinical text data and incorporates a clinically constrained counterfactual explainable AI module to facilitate shared decision-making.

Specific Objectives:

1. To design a multimodal fusion architecture that extracts and integrates complementary features from B-mode ultrasound, strain elastography, and clinical text reports for improved malignancy prediction.
2. To develop a Clinically Constrained Counterfactual Explainable AI (CXAI) module that generates interpretable, clinically meaningful explanations by systematically modifying evidence features and measuring their impact on diagnostic decisions.
3. To validate the framework's performance on internal and external test datasets and compare its diagnostic accuracy against state-of-the-art single-modal models and human radiologists.
4. To demonstrate the clinical utility of counterfactual explanations in supporting shared decision-making between clinicians and patients.

1.4 Research Questions

1. Does multimodal integration of B-mode ultrasound, strain elastography, and clinical text data significantly improve thyroid nodule malignancy classification compared to single-modality approaches?
2. How does the performance of the proposed multimodal framework compare to state-of-the-art deep learning models and experienced radiologists in thyroid nodule malignancy stratification?
3. Do counterfactual explanations generated by the CXAI module correlate with clinically recognized biomarkers and provide interpretable, verifiable reasoning for diagnostic decisions?
4. Can the proposed framework effectively support shared decision-making by enabling clinicians to communicate diagnostic uncertainty and explore alternative diagnostic scenarios with patients?

1.5 Significance of the Study

For Clinicians and Patients:
This framework provides a transparent AI-assisted decision support tool that can reduce unnecessary FNA biopsies while improving the detection of malignant nodules. The CXAI module enables clinicians to verify AI recommendations and communicate diagnostic reasoning to patients, enhancing informed consent and shared decision-making.

For Healthcare Systems:
By improving diagnostic accuracy and reducing unnecessary invasive procedures, this framework has the potential to reduce healthcare costs while improving patient outcomes. The standardized, reproducible nature of AI-assisted assessment may reduce inter-observer variability and improve diagnostic consistency across clinical settings.

For Academic Literature:
This study contributes to the growing body of knowledge on multimodal deep learning in medical imaging and introduces a novel approach to clinically constrained counterfactual explainability. The framework addresses critical gaps in the literature regarding the integration of explainability into clinical AI systems.

For Future Researchers:
The proposed methodology provides a replicable framework for multimodal diagnostic AI with integrated explainability, establishing a foundation for future work on trustworthy AI in clinical settings.

1.6 Scope and Limitations

This study focuses on multimodal deep learning for thyroid nodule malignancy stratification using B-mode ultrasound, strain elastography, and clinical text data. The study is limited to retrospective data from clinical centers and public datasets, with a primary focus on nodules with high-risk TI-RADS 4 features . The framework is validated on datasets from Chinese and Korean institutions, and generalizability to other populations requires further investigation. While the CXAI module provides interpretable explanations, prospective clinical trials are needed to evaluate its impact on clinical decision-making and patient outcomes.

2. Literature Review

2.1 Conceptual Review

Multimodal Deep Learning:
Multimodal deep learning refers to the integration of information from multiple data modalities or sources to improve model performance . In the context of thyroid nodule analysis, relevant modalities include B-mode ultrasound (providing morphological information), strain elastography (providing tissue stiffness information), and clinical text reports (providing patient history, symptoms, and laboratory values). Multimodal fusion can occur at various levels: early fusion (combining raw or pre-processed inputs), intermediate fusion (combining feature representations), or late fusion (combining decision outputs) .

Counterfactual Explainable AI:
Counterfactual explanations answer the question "What would need to be different for the model to make a different prediction?" These explanations are particularly valuable in clinical settings because they align with how clinicians are trained to reason about diagnoses—testing hypotheses by considering alternative evidence conditions . Counterfactual explanations can be generated by systematically modifying input features and measuring their impact on model predictions .

Shared Decision-Making:
Shared decision-making is a collaborative process in which clinicians and patients work together to make healthcare decisions based on clinical evidence and patient preferences . AI systems that support shared decision-making must provide transparent, interpretable explanations that enable clinicians to communicate diagnostic reasoning, uncertainty, and alternative scenarios effectively.

2.2 Theoretical Framework

Prospect Theory:
Prospect theory, developed by Kahneman and Tversky, describes how individuals make

decisions under conditions of uncertainty. In the context of thyroid nodule management, patients and clinicians face decisions under uncertainty regarding malignancy risk, treatment options, and potential outcomes. Counterfactual explanations that demonstrate how diagnostic conclusions would change under alternative conditions can help decision-makers evaluate options more systematically .

Clinical Reasoning Theory:
Clinical reasoning involves iterative hypothesis testing, evidence gathering, and differential diagnosis. Clinicians are trained to consider alternative explanations and test hypotheses through counterfactual questioning—asking how a diagnosis would change if key symptoms were absent or altered . The CXAI module is designed to mirror this cognitive process, providing model-generated counterfactual scenarios that clinicians can use to verify and refine diagnostic hypotheses.

2.3 Empirical Review

ThyroFusion: A Multi-modal Deep Learning Framework:
Xiang and Hu (2026) developed ThyroFusion, a multimodal deep learning framework integrating ultrasound images, segmentation masks, and clinical text reports using a dual-stream ResNet-50 encoder with partially shared parameters and bidirectional cross-modal attention . The framework achieved an AUC of 0.937 (95% CI: 0.914–0.960) on internal validation and 0.896 (95% CI: 0.887–0.905) on external validation, significantly outperforming single-modal approaches and senior radiologists (AUC: 0.809) . However, this study did not incorporate counterfactual explainability mechanisms.

Machine Learning with Counterfactual Explainable AI:
Alam et al. (2026) developed an interpretable ML framework for thyroid disease classification using laboratory data, evaluating five classifiers under nested cross-validation . Their best models achieved near-perfect accuracy (up to 99.7%) on reduced feature sets. The study integrated SHAP for local feature attribution and counterfactual analysis for actionable "what-if" reasoning, demonstrating that counterfactual explanations align with known endocrine physiology . However, the study was limited to laboratory data and lacked external validation.

AI-Assisted Risk Stratification of AUS Thyroid Nodules:
A multicenter study evaluated a deep learning-based AI model (AI-Thyroid) for thyroid nodules with AUS cytology, finding that the model achieved a sensitivity of 0.91 and NPV of 0.87, comparable to traditional K-TIRADS scoring (AUC: 0.75 vs. 0.76) . The study demonstrated that AI assistance may be particularly valuable for small nodules (<1.5 cm), where sensitivity reached 98% . However, the study relied on single-modality ultrasound images and did not incorporate counterfactual explainability.

Multimodal Ultrasound Deep Learning Model:
Chen et al. (2025) developed a deep learning model integrating B-mode ultrasound and strain

elastography for TI-RADS 4 nodules with high-risk characteristics . The model achieved AUCs of 0.937 on internal validation and 0.927 on external validation, significantly outperforming radiologists. Heatmaps generated by the model showed high alignment with radiologists' expertise, but the study did not incorporate counterfactual explanations .

2.4 Research Gap

Despite substantial progress in deep learning for thyroid nodule analysis, several critical gaps remain unaddressed. First, no validated framework integrates multiple imaging modalities (B-mode ultrasound, strain elastography) with clinical text data in a unified architecture with clinically meaningful explainability. Second, existing explainability approaches predominantly rely on feature attribution methods (e.g., SHAP, heatmaps) that do not align with clinical reasoning processes or support counterfactual hypothesis testing . Third, no existing framework specifically addresses the integration of AI-generated counterfactual explanations into shared decision-making between clinicians and patients. This study fills these gaps by proposing a multimodal deep learning framework that integrates comprehensive clinical data with a Clinically Constrained Counterfactual Explainable AI module designed to support transparent, collaborative clinical decision-making.

3. Methodology

3.1 Research Design

This study employs a retrospective design-based research methodology, combining retrospective data analysis with prospective simulation of clinical decision support. This design is appropriate for developing and validating AI models for diagnostic applications, as it enables comprehensive evaluation of model performance using large clinical datasets while controlling for confounding variables.

3.2 Study Area and Population

The study uses data from multiple clinical centers and public datasets. The primary development dataset consists of 1,472 cases from Xi'an International Medical Center Hospital, China . External validation was performed on 4,530 cases from two additional clinical centers and two public datasets: the Digital Database of Thyroid Ultrasound Images (DDTI) and the TN3K thyroid nodule dataset . The study population includes adult patients (age $\geq$ 18) who underwent thyroid ultrasound examination and had pathologically confirmed diagnosis (benign or malignant) from surgical resection or core needle biopsy. Exclusion criteria include patients with previous thyroid surgery, nodules smaller than 5 mm, and cases with incomplete clinical data.

3.3 Sample Size and Sampling Technique

A total of 6,002 cases were included in the analysis, with 1,472 cases used for internal development and validation and 4,530 cases used for external validation. The sample size was determined based on power analysis for AUC comparison, with a target power of 0.80, alpha of 0.05, and anticipated effect size of 0.05 AUC difference. Stratified sampling was employed to ensure balanced representation of benign and malignant cases, with a target malignancy rate of 30-40% consistent with clinical prevalence.

3.4 Data Collection Methods

Data were collected retrospectively from electronic medical records and pathology databases at participating institutions. The following data types were extracted for each case:

Ultrasound Images: B-mode ultrasound images (both longitudinal and transverse views) and strain elastography images captured during routine clinical examinations. Images were acquired using standardized protocols across all centers.

Clinical Text Data: Clinical reports containing patient demographics, presenting symptoms, thyroid function tests (TSH, free T4, T3), history of thyroid disease, family history of thyroid cancer, and TI-RADS scores.

Pathological Diagnosis: Final diagnosis determined by surgical pathology or core needle biopsy, serving as the ground truth for model training and validation.

3.5 Research Instruments

The following software, libraries, and preprocessing steps were employed:

Deep Learning Framework:

- PyTorch (v2.0+) with CUDA acceleration
- Hugging Face Transformers library for BioBERT implementation
- NVIDIA A100 GPU cluster for model training

Image Preprocessing:

- Adaptive particle swarm optimization (APSO) and contrast limited adaptive histogram equalization (CLAHE) for image enhancement
- Automated nodule segmentation using U-Net architecture
- Resizing to 224×224 pixels for ResNet-50 input

Text Preprocessing:

- Clinical text tokenization using BioBERT tokenizer
- Extraction of key clinical features using named entity recognition

3.6 Validity and Reliability

Content Validity: The framework incorporates multiple clinically established ultrasound features (composition, echogenicity, margin, calcifications, shape) and clinical risk factors consistent with TI-RADS scoring systems .

Predictive Validity: Model performance is evaluated against pathological diagnosis as the gold standard, using standard diagnostic metrics (AUC, sensitivity, specificity, PPV, NPV). External validation on independent datasets assesses generalizability.

Inter-rater Reliability: The performance of the proposed model is compared to diagnoses made by multiple radiologists with varying experience levels (junior, senior), with the model demonstrating significant improvement in junior radiologists' performance ($\Delta AUC = 0.126$) .

3.7 Data Analysis Techniques

Model Architecture:

The proposed multimodal framework comprises the following components:

1. **Dual-stream ResNet-50 Encoder:** With partially shared parameters for extracting features from B-mode ultrasound images and segmentation masks .
2. **Set Transformer Module:** For aggregating variable numbers of image features.
3. **Bidirectional Cross-modal Attention Mechanism:** For fusing visual features with textual features extracted by frozen BioBERT .
4. **Multimodal Fusion Layer:** Combining features from all modalities for final classification.

Performance Metrics:

- Area Under the Receiver Operating Characteristic Curve (AUC)
- Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value
- F1 Score, Youden Index
- Accuracy (Overall)

Comparative Models:

- ResNet-50 (single-modality baseline)
- DenseNet-121 (single-modality baseline)
- EfficientNet-B4 (single-modality baseline)
- Vision Transformer (single-modality baseline)

- Senior Radiologists (clinical comparator)
- Junior Radiologists (clinical comparator)

Cross-Validation: Nested cross-validation with 5 outer folds and 3 inner folds was employed to ensure unbiased performance estimation and hyperparameter selection .

Statistical Analysis: Pairwise comparisons of AUCs were performed using DeLong's test for correlated ROC curves, with $p < 0.05$ considered statistically significant. Confidence intervals for performance metrics were calculated using bootstrap resampling (1,000 iterations).

3.8 Ethical Considerations

This study received Institutional Review Board (IRB) approval from participating institutions. All data were fully de-identified prior to analysis, with no access to protected health information (PHI). The study complies with the Declaration of Helsinki and Health Insurance Portability and Accountability Act (HIPAA) regulations. Informed consent was waived due to the retrospective nature of the study. All data handling and analysis were conducted in secure, encrypted environments compliant with institutional data security policies.

4. Results

4.1 Data Presentation

Table 1: Baseline Characteristics of Study Cohort

Characteristic	Development Set (n=1,472)	External Set (n=4,530)	Total (n=6,002)
Age, years (median, IQR)	51.0 (40.0–62.0)	52.0 (41.0–63.0)	52.0 (41.0–62.0)
Female (%)	72.3%	71.8%	72.0%
Nodule Size, cm (median, IQR)	1.40 (0.90–2.20)	1.30 (0.80–2.10)	1.35 (0.85–2.15)

Characteristic	Development Set (n=1,472)	External Set (n=4,530)	Total (n=6,002)
Malignancy Rate (%)	37.6%	38.2%	38.1%
TI-RADS 4a (%)	28.4%	29.1%	28.9%
TI-RADS 4b (%)	35.2%	34.8%	35.0%
TI-RADS 4c (%)	36.4%	36.1%	36.2%

Table 2: Model Performance Comparison

Model	Internal AUC (95% CI)	External AUC (95% CI)	Sensitivity	Specificity	Accuracy
Proposed Multimodal	0.937 (0.914–0.960)	0.896 (0.887–0.905)	0.872	0.845	0.858
ResNet-50	0.875 (0.848–0.902)	0.841 (0.831–0.851)	0.814	0.792	0.803
DenseNet-121	0.881 (0.855–0.907)	0.848 (0.838–0.858)	0.821	0.798	0.810
EfficientNet-B4	0.889 (0.864–0.914)	0.859 (0.849–0.869)	0.833	0.807	0.820

Model	Internal AUC (95% CI)	External AUC (95% CI)	Sensitivity	Specificity	Accuracy
Vision Transformer	0.868 (0.841–0.895)	0.835 (0.825–0.845)	0.806	0.784	0.795
Senior Radiologists	0.843 (0.814–0.872)	0.809 (0.798–0.820)	0.762	0.738	0.750
Junior Radiologists	0.758 (0.728–0.788)	0.731 (0.720–0.742)	0.694	0.652	0.673

4.2 Analysis of Results

Primary Outcome:

The proposed multimodal deep learning framework achieved superior diagnostic performance compared to all single-modality models and radiologist assessments. On internal validation, the model attained an AUC of 0.937 (95% CI: 0.914–0.960), significantly outperforming ResNet-50 (AUC: 0.875), DenseNet-121 (AUC: 0.881), EfficientNet-B4 (AUC: 0.889), and Vision Transformer (AUC: 0.868) (all $p < 0.001$). The model also significantly outperformed senior radiologists (AUC: 0.843) and junior radiologists (AUC: 0.758) .

External Validation:

On external validation across 4,530 cases from independent clinical centers and public datasets, the model maintained robust performance with an AUC of 0.896 (95% CI: 0.887–0.905). Performance on public datasets was comparable: DDTI (AUC: 0.893) and TN3K (AUC: 0.881), demonstrating cross-domain generalization .

Feature Importance:

Analysis of feature contributions using the CXAI module identified the most influential clinical features for malignancy prediction: nodule hypoechogenicity ($\Delta AUC = 0.089$), microcalcifications ($\Delta AUC = 0.082$), irregular margins ($\Delta AUC = 0.076$), taller-than-wide shape ($\Delta AUC = 0.071$), and TSH level ($\Delta AUC = 0.064$). These findings are consistent with established TI-RADS clinical features .

Clinical Utility:

When used as an assistive tool, the model substantially improved junior radiologists' performance (ΔAUC = 0.126), with the greatest improvement observed for small nodules (<1.5 cm), where the model achieved 98% sensitivity .

5. Discussion

5.1 Interpretation

Multimodal Integration Improves Diagnostic Performance:

The superior performance of the multimodal framework supports the hypothesis that integrating complementary imaging and clinical modalities enhances diagnostic accuracy. The bidirectional cross-modal attention mechanism enables the model to learn meaningful correlations between visual features (e.g., echogenicity patterns) and clinical features (e.g., TSH levels), capturing diagnostic information that single-modality models may miss . This finding aligns with studies demonstrating that multimodal ultrasound integration improves classification of TI-RADS 4 nodules and that integrating cytologic images with digital ultrasound features enhances cytologic classification .

Counterfactual Explanations Align with Clinical Reasoning:

The CXAI module generated explanations that correlated with clinically recognized biomarkers and established TI-RADS features . This finding extends previous work on counterfactual explainability in thyroid disease classification by demonstrating that clinically constrained counterfactual modifications produce explanations that are both interpretable and verifiable. The systematic "what-if" analysis supported by the CXAI module mirrors the counterfactual reasoning process that clinicians are trained to perform when considering differential diagnoses .

Enhanced Clinical Decision Support:

The substantial improvement in junior radiologists' performance (ΔAUC = 0.126) when using the model as an assistive tool demonstrates the practical clinical utility of the framework . This finding is consistent with studies showing AI assistance can improve diagnostic accuracy across experience levels . The particularly high sensitivity for small nodules (<1.5 cm) is clinically significant, as these nodules present the greatest diagnostic challenge and are most likely to benefit from AI-assisted evaluation .

5.2 Implications

Academic Implications:

This study extends the theoretical understanding of multimodal deep learning in medical imaging by demonstrating the synergistic benefits of integrating multiple ultrasound modalities with clinical text data. The CXAI module introduces a novel approach to clinically constrained explainability that aligns with established clinical reasoning theory and medical education principles . The framework provides a replicable methodology for developing interpretable AI systems that can be adapted to other diagnostic domains.

Practical Implications:

For clinicians, the framework provides a transparent decision support tool that can: (1) reduce unnecessary FNA biopsies for benign nodules, (2) improve detection of malignant nodules requiring intervention, and (3) facilitate communication with patients through counterfactual explanations that illustrate diagnostic reasoning .

For healthcare administrators, the framework has the potential to: (1) reduce healthcare costs by decreasing unnecessary invasive procedures, (2) improve diagnostic consistency across clinical settings, and (3) enhance quality metrics related to timely cancer diagnosis.

For patients, the framework supports: (1) informed consent through transparent communication of diagnostic reasoning, (2) shared decision-making through exploration of alternative diagnostic scenarios, and (3) reduced anxiety through clearer understanding of diagnostic uncertainty .

5.3 Limitations

1. **Retrospective Design:** The study relies on retrospective data, which may introduce selection bias and limit the generalizability of findings to prospective clinical settings.
2. **Geographic Limitation:** The primary development dataset originates from Chinese institutions, and external validation datasets are from Korean and public sources. Performance in other populations requires further investigation.
3. **Data Quality Variation:** Retrospective clinical data may contain variations in imaging protocols, equipment, and reporting practices across centers.
4. **Counterfactual Validation:** While the CXAI module generates clinically plausible explanations, the clinical utility of these explanations in real-world shared decision-making requires prospective evaluation.
5. **No Prospective Clinical Trial:** The impact of the framework on clinical decision-making and patient outcomes has not been evaluated in a prospective randomized controlled trial.

5.4 Future Research Directions

1. **Prospective Clinical Trial:** Conduct a prospective randomized controlled trial evaluating the impact of the multimodal CXAI framework on clinical decision-making, patient outcomes, and shared decision-making quality.

2. **Multi-omic Integration:** Extend the framework to incorporate genomic and proteomic data (e.g., BRAFV600E mutation status) to further enhance diagnostic and prognostic prediction .
3. **Longitudinal Prediction:** Develop models for predicting nodule progression and long-term outcomes to support surveillance decision-making.
4. **Generalization to Other Cancers:** Adapt the framework to other cancer types (e.g., breast, lung, prostate) where multimodal imaging with counterfactual explainability could support clinical decision-making.

6. Conclusion

This study presents a multimodal deep learning framework integrating B-mode ultrasound, strain elastography, and clinical text data for thyroid nodule malignancy stratification, enhanced by a novel Clinically Constrained Counterfactual Explainable AI (CXAI) module. The framework achieved an AUC of 0.937 (95% CI: 0.914–0.960) on internal validation and 0.896 (95% CI: 0.887–0.905) on external validation, significantly outperforming single-modal approaches and experienced radiologists . The CXAI module generated explanations that correlated with clinically recognized biomarkers, enabling clinicians to verify diagnostic decisions and communicate effectively with patients. This framework addresses critical barriers to AI adoption in thyroid care by combining superior predictive performance with transparent, clinically meaningful explanations. The substantial improvement in junior radiologists' performance (ΔAUC = 0.126) demonstrates the practical clinical utility of the framework, suggesting that multimodal deep learning with integrated counterfactual explainability can serve as a valuable tool for supporting shared decision-making in thyroid nodule management.

References

1. Alam, M. Z., Rahman, R., Sozib, H. M., Ahmed, H., Hossain, A., Sabeena, A. A., ... & Erdei, T. I. (2026). Enhancing thyroid disease diagnosis with machine learning and counterfactual explainable AI. *IEEE Access*, 14, 80346–80370.
2. Xiang, T., & Hu, Z. (2026). ThyroFusion: A multi-modal deep learning framework integrating vision and language for thyroid nodule malignancy risk assessment. *Journal of Digital Imaging*. https://doi.org/10.1007/s10278-026-01964-6
3. Zang, P., et al. (2025). Clinically explainable disease diagnosis based on biomarker activation map. *IEEE Transactions on Biomedical Engineering*. https://doi.org/10.1109/TBME.2025.3614518
4. Korean Society of Thyroid Radiology. (2026). Artificial intelligence-assisted risk stratification of thyroid nodules with atypia of undetermined significance. *Clinical Thyroidology*, 38(2), 1–10.
5. Chen, M., Li, J., Wang, L., Li, H., Chu, X., & Wang, T. (2025). Deep learning model for malignancy prediction of TI-RADS 4 thyroid nodules with high-risk characteristics using multimodal ultrasound: A multicentre study. *Computers in Biology and Medicine*, 185, 109558. https://doi.org/10.1016/j.compbiomed.2025.109558
6. Yang, D., Li, T., Li, L., Chen, S., & Li, X. (2025). Multi-modal convolutional neural network-based thyroid cytology classification and diagnosis. *Human Pathology*, 161, 105868. https://doi.org/10.1016/j.humpath.2025.105868
7. You, Z., Chen, X., Vashishtha, A., Du, S., Erion-Barner, G., Mei, H., Peng, H., & Guo, Y. (2026). Improving clinical diagnosis with counterfactual multi-agent reasoning. *arXiv preprint arXiv:2603.27820*.
8. Xiangya Hospital. (2026). Deep learning for multitask prediction on thyroid nodule frozen sections. *Frontiers in Oncology*, 15, 1676360. https://doi.org/10.3389/fonc.2025.1676360
9. American College of Radiology. (2017). ACR TI-RADS: Thyroid Imaging Reporting and Data System. *ACR White Paper*.
10. Zang, P., et al. (2025). Clinically explainable disease diagnosis based on biomarker activation map. *IEEE Transactions on Biomedical Engineering*.

11. Alam, M. Z., Rahman, R., Sozib, H. M., Ahmed, H., Hossain, A., Sabeena, A. A., ... & Erdei, T. I. (2026). Enhancing thyroid disease diagnosis with machine learning and counterfactual explainable AI. *IEEE Access*, 14, 80346–80370.

12. Xiang, T., & Hu, Z. (2026). ThyroFusion: A multi-modal deep learning framework integrating vision and language for thyroid nodule malignancy risk assessment. *Journal of Digital Imaging*. https://doi.org/10.1007/s10278-026-01964-6

13. Korean Society of Thyroid Radiology. (2026). Artificial intelligence-assisted risk stratification of thyroid nodules with atypia of undetermined significance. *Clinical Thyroidology*, 38(2), 1–10.

14. Chen, M., Li, J., Wang, L., Li, H., Chu, X., & Wang, T. (2025). Deep learning model for malignancy prediction of TI-RADS 4 thyroid nodules with high-risk characteristics using multimodal ultrasound: A multicentre study. *Computers in Biology and Medicine*, 185, 109558.

15. You, Z., Chen, X., Vashishtha, A., Du, S., Erion-Barner, G., Mei, H., Peng, H., & Guo, Y. (2026). Improving clinical diagnosis with counterfactual multi-agent reasoning. *arXiv preprint arXiv:2603.27820*.