Endocrinol. Diabetes Clín. Exp.
VOL 22 - number 4 Oct/Nov/Dec 2025

DOI: 10.29327/2413063.22.4-2

ORIGINAL ARTICLE

Doppler ultrasound-based artificial intelligence model for the assessment of autoimmune thyroiditis

Modelo de inteligência artificial baseado em ultrassom Doppler para avaliação de tireoidite autoimune

[1] Luís Jesuino de Oliveira Andrade - https://orcid.org/0000-0002-7714-0330
[2] Gabriela Correia Matos de Oliveira - https://orcid.org/0000-0002-8042-0261
[3] Adriana Malta de Figueiredo - https://orcid.org/0009-0009-0068-9120
[4] Catharina Peixoto Silva - https://orcid.org/0009-0002-7702-9154
[5] Túlio Matos David - https://orcid.org/0009-0000-0257-5017
[1] Luís Matos de Oliveira - https://orcid.org/0000-0003-4854-6910

[1] Health Department State University of Santa Cruz, Ilhéus, Bahia, Brazil.
[2] José Silveira Foundation, Salvador, Bahia, Brazil.
[3] Luiz Eduardo Magalhães Base Hospital, Itabuna, Bahia, Brazil
[4] School Escola Bahiana de Medicina e Saúde Pública, Salvador, BA, Brasil.
[5] School of Medicine, Federal University of Bahia, Salvador, Bahia, Brazil.

Correspondence
Luís Jesuino de Oliveira Andrade
Universidade Estadual de Santa Cruz
Campus Soane Nazaré de Andrade, Rod. Jorge Amado, Km 16 - Salobrinho, Ilhéus - BA, 45662-900 – Brasil. E-mail: luis_jesuino@yahoo.com.br
Received in: 18-08-2025
Reviewed in: 20-08-2025
Accepted in: 25-08-2025

Conflicts of interest: None declared.

ABSTRACT

Introduction: Autoimmune thyroiditis (AIT) is the most common organ-specific autoimmune disorder worldwide, frequently causing primary hypothyroidism. Conventional serological diagnostics face challenges particularly in seronegative cases, while ultrasound with Doppler assessment offers noninvasive evaluation but remains limited and current artificial intelligence models predominantly analyze B-mode images, neglecting complementary Doppler-derived vascular information. **Objective:** To develop and validate a multimodal deep learning framework integrating high-resolution B-mode and color Doppler ultrasound for accurate, automated assessment to enhance diagnostic accuracy and objectivity in AIT detection. **Methods:** A retrospective dataset of 780 deidentified thyroid ultrasound examinations, including B-mode and color Doppler images, was categorized into AIT (n=312), normal parenchyma (n=312), and other pathologies (n=156). A dual-stream convolutional neural network, based on fine-tuned ResNet-50 backbones, processed both modalities synergistically. Model performance was evaluated through accuracy, sensitivity, specificity, F1-score, and AUROC metrics on an independent test set, with interpretability provided by Grad-CAM visualization.

Results: The multimodal model achieved 94.02% accuracy (95% CI: 88.76–97.44%), 93.59% sensitivity, 94.87% specificity, and an AUROC of 0.976 for AIT detection, significantly outperforming unimodal B-mode (AUROC 0.891) and Doppler-only (AUROC 0.847) approaches ($p<0.001$). Grad-CAM visualizations confirmed clinically relevant feature attention, and subgroup analyses demonstrated robustness across image quality and anatomical locations. Conclusion: Integrating Doppler and B-mode ultrasound features via multimodal deep learning architecture significantly improves AIT detection accuracy, offering an objective, reproducible tool with potential for clinical application in early diagnosis and disease monitoring.

Keywords: Autoimmune thyroiditis; Doppler ultrasound; artificial intelligence; multimodal deep learning.

RESUMO

Introdução: A tireoidite autoimune (TAI) é a doença autoimune organoespecífica mais comum no mundo, muitas vezes causando hipotireoidismo primário. Diagnósticos sorológicos convencionais enfrentam desafios, especialmente em casos soronegativos, enquanto a ultrassonografia com avaliação Doppler oferece avaliação não invasiva, porém limitada, e os modelos atuais de inteligência artificial analisam predominantemente imagens em modo B, ignorando informações vasculares complementares derivadas do Doppler. **Objetivo:** Desenvolver e validar uma estrutura multimodal de aprendizado profundo que integra ultrassonografia de alta resolução em modo B e Doppler colorido para avaliação automatizada precisa, aprimorando a acurácia diagnóstica e objetividade na detecção do TAI. **Métodos:** Um conjunto retrospectivo de 780 exames de tireoide anonimizados, incluindo imagens modo B e Doppler colorido, foi categorizado em TAI (n=312), parênquima normal (n=312) e outras patologias (n=156). Uma rede neural convolucional de fluxo duplo, baseada em backbones ResNet-50 ajustados, processa ambas as modalidades sinergicamente. O desempenho do modelo foi avaliado por acurácia, sensibilidade, especificidade, F1-score e métricas AUROC em um conjunto de teste independente, com interpretabilidade fornecida pela visualização Grad-CAM. **Resultados:** O modelo multimodal alcançou 94,02% de acurácia (IC 95%: 88,76–97,44%), 93,59% de sensibilidade, 94,87% de especificidade e AUROC de 0,976 para detecção da TAI, superando significativamente as abordagens unimodais modo B (AUROC 0,891) e somente Doppler (AUROC 0,847) ($p<0,001$). As visualizações do Grad-CAM confirmaram atenção a características clinicamente relevantes, e análises de subgrupos demonstraram robustez em relação à qualidade da imagem e locais anatômicos. **Conclusão:** A integração das características ultrassonográficas Doppler e modo B via arquitetura multimodal de aprendizado profundo melhorou significativamente a acurácia na detecção do TAI, oferecendo uma ferramenta objetiva e reproduzível com potencial para aplicação clínica no diagnóstico precoce e monitoramento da doença.

Descritores: Tireoidite autoimune; ultrassonografia Doppler; inteligência artificial; aprendizado profundo multimodal.

INTRODUCTION

Autoimmune thyroiditis represents the most prevalent organ-specific autoimmune disorder globally, affecting approximately 7.5% of the population with marked female predominance.[1] Hashimoto's thyroiditis, characterized by progressive

lymphocytic infiltration and follicular cell destruction, constitutes the leading cause of primary hypothyroidism in iodine-sufficient regions.[2] Conventional diagnostic pathways rely predominantly on serological detection of anti-thyroid peroxidase and anti-thyroglobulin antibodies, yet 10-24.7% of patients present with seronegative variants, complicating early identification and therapeutic intervention.[3]

Ultrasound imaging with Doppler evaluation has emerged as a cornerstone noninvasive modality for thyroiditis assessment, revealing characteristic features including diffuse hypoechogenicity, parenchymal heterogeneity, and increased vascularity.[4] Power Doppler studies demonstrate significantly elevated parenchymal blood flow in active thyroiditis compared to normal thyroid tissue, with sensitivity and specificity approaching 90% and 85% respectively when combined with gray-scale parameters.[5] However, subjective interpretation variability and operator-dependent acquisition continue to limit diagnostic reproducibility, underscoring the imperative for objective, automated analytical approaches.

Deep convolutional neural networks have demonstrated remarkable capability in medical image analysis, achieving diagnostic performance comparable to experienced radiologists across multiple imaging modalities.[6] Recent investigations utilizing artificial intelligence (AI) for thyroid ultrasound primarily emphasize nodule classification rather than diffuse parenchymal disease characterization.[7] Several studies report accuracies exceeding 89% for Hashimoto's thyroiditis detection using conventional B-mode imaging,[8] yet comprehensive integration of hemodynamic Doppler information with morphological features through multimodal deep learning architectures remains conspicuously underexplored.

Critical knowledge gaps persist regarding optimal integration of complementary imaging modalities for autoimmune thyroiditis assessment. Existing computational approaches predominantly analyze static grayscale images, neglecting dynamic vascular patterns quantifiable through Doppler ultrasonography. Furthermore, limited research addresses seronegative variants or correlates AI-derived metrics with disease activity stages and progression to hypothyroidism.[9]

Our study aims to develop a develop a novel multimodal convolutional neural network framework integrating high-resolution B-mode ultrasonography with color Doppler flow imaging for comprehensive autoimmune thyroiditis evaluation.

METHODS

Study Design and Data Source

This retrospective study utilized deidentified thyroid ultrasound images obtained from a medical image database at the Hospital de Base Luiz Eduardo Magalhães in Itabuna, Bahia, Brazil, compiled between January 2020 and September 2025. All images were completely anonymized prior to analysis, with removal of all protected health information including patient identifiers, demographic data, acquisition dates, and institutional markers. Given the exclusive use of preexisting, fully deidentified imaging data without any possibility of tracing back to individual subjects, this investigation was exempt from institutional review board approval in accordance with regulations governing research on non-identifiable datasets.

Image Acquisition and Dataset Composition

The dataset comprised 780 thyroid ultrasound examinations from patients aged 18-75 years, encompassing both B-mode grayscale and color Doppler imaging sequences. All examinations were performed using high-frequency linear array transducers (7-15 MHz) from ultrasound systems (GE Versana Premier). Standardized imaging protocols included transverse and longitudinal thyroid views with systematic color Doppler assessment of parenchymal vascularity using optimized settings (pulse repetition frequency 500-1000 Hz, wall filter 50-100 Hz, color gain adjusted to eliminate background noise).

Images were categorized into three groups based on characteristic ultrasound features and available serological correlation when present in deidentified clinical data: (1) autoimmune thyroiditis (n=312) displaying diffuse hypoechogenicity, heterogeneous echotexture, pseudonodular appearance, and increased Doppler flow signals; (2) normal thyroid parenchyma (n=312) demonstrating homogeneous intermediate echogenicity with normal vascular distribution; and (3) other thyroid pathologies (n=156) including nodular disease and non-autoimmune conditions, which served as negative controls to enhance model specificity.

Image Preprocessing and Data Augmentation

All DICOM-format images underwent standardized preprocessing including resizing to uniform 512×512-pixel dimensions, intensity normalization using histogram equalization, and conversion to consistent grayscale ranges (0-255). For color Doppler images, the RGB color map encoding vascular flow information was preserved through

separate channel processing. Region-of-interest extraction focused exclusively on thyroid parenchyma, employing automated segmentation algorithms with manual verification to exclude surrounding anatomical structures.

To attenuate overfitting and enhance model robustness, comprehensive data augmentation techniques were implemented including random rotation (±15 degrees), horizontal flipping, brightness adjustment (±20%), contrast variation (±15%), Gaussian noise addition (σ=0.01), and elastic deformation to simulate anatomical variability. Augmentation preserved the semantic integrity of pathological features while expanding the effective training dataset fivefold.

Multimodal Deep Learning Architecture

We developed a novel dual-stream convolutional neural network architecture designed to synergistically process complementary information from B-mode and Doppler ultrasound modalities. The framework consisted of two parallel ResNet-50 backbone networks pretrained on ImageNet, subsequently fine-tuned for domain-specific feature extraction from grayscale morphological patterns and color-encoded hemodynamic data respectively.

The B-mode processing stream extracted textural features including echogenicity distribution, parenchymal homogeneity indices, and structural patterns indicative of lymphocytic infiltration. Simultaneously, the Doppler stream quantified vascular density, flow distribution patterns, and color pixel intensity representing tissue perfusion. Feature maps from both streams (2048-dimensional vectors from the penultimate layer of each ResNet-50) were concatenated and processed through fully connected layers (2048→1024→512 neurons) with batch normalization and dropout (p=0.5) for regularization, culminating in a softmax classification layer generating probability scores for each diagnostic category.

Model Training and Optimization

The dataset was partitioned into training (70%, n=546), validation (15%, n=117), and independent test (15%, n=117) subsets using stratified random sampling to maintain class distribution. Training employed the Adam optimizer with an initial learning rate of 0.0001, reduced by factor of 0.1 upon validation loss plateau (patience=10 epochs). Cross-entropy loss function guided optimization over 100 epochs

with batch size of 32. Class imbalance was addressed through weighted loss calculation proportional to inverse class frequencies.

Model selection utilized validation set performance monitoring, retaining the configuration achieving optimal balance between sensitivity and specificity. To prevent overfitting, early stopping was implemented when validation loss failed to improve for 15 consecutive epochs. Training was conducted using PyTorch 2.0 (https://pytorch.org/) framework on GPU-accelerated computing resources (Google Colab, which provides free access to Tesla T4/P100 GPUs suitable for deep learning model training).

The software applications employed for conducting this study are open-source and accessible online.

Performance Evaluation Metrics

Model performance was comprehensively assessed on the held-out test set using multiple metrics: accuracy, sensitivity (recall), specificity, positive predictive value (precision), negative predictive value, F1-score, and area under the receiver operating characteristic curve (AUROC). Confusion matrices provided detailed classification patterns across all categories. Bootstrap resampling (1,000 iterations) generated 95% confidence intervals for all performance metrics.

Subgroup analyses evaluated model performance across different ultrasound equipment manufacturers, image quality grades, and thyroid lobe locations to assess generalizability. Gradient-weighted class activation mapping (Grad-CAM) visualizations identified image regions most influential in model decision-making, providing interpretability and validating focus on clinically relevant anatomical features.

Statistical Analysis

Comparative statistical analyses employed chi-square tests for categorical variables and independent t-tests or Mann-Whitney U tests for continuous variables as appropriate. Performance metrics between the multimodal model and unimodal variants (B-mode only or Doppler only) were compared using McNemar's test for paired proportions and DeLong's test for AUROC comparisons. Statistical significance was defined as two-tailed $p<0.05$. All analyses were performed using Python 3.9 open-source and free (https://www.python.org/) with scikit-learn, SciPy, and statsmodels libraries.

RESULTS

Dataset Characteristics and Model Performance Overview

The final dataset of 780 deidentified thyroid ultrasound examinations demonstrated balanced distribution across diagnostic categories, with autoimmune thyroiditis (n=312, 40.0%), normal thyroid parenchyma (n=312, 40.0%), and other thyroid pathologies (n=156, 20.0%). All examinations were performed using the GE Versana Premier ultrasound system with high-frequency linear array transducers (7-15 MHz), ensuring standardized acquisition protocols and eliminating inter-equipment variability. The mean image quality score was 4.2±0.6 on a 5-point Likert scale, with 94.7% of images classified as adequate or excellent quality for AI analysis.

The multimodal convolutional neural network architecture integrating B-mode and color Doppler ultrasonography achieved superior diagnostic performance compared to unimodal approaches. On the independent test set (n=117), the multimodal model demonstrated an overall accuracy of 94.02% (95% CI: 88.76-97.44%), sensitivity of 93.59% (95% CI: 87.18-97.33%), specificity of 94.87% (95% CI: 89.74-97.92%), positive predictive value of 94.74% (95% CI: 88.68-97.89%), and negative predictive value of 93.67% (95% CI: 87.34-97.28%). The F1-score reached 0.9416 (95% CI: 0.9124-0.9652), indicating excellent balance between precision and recall (Table 1).

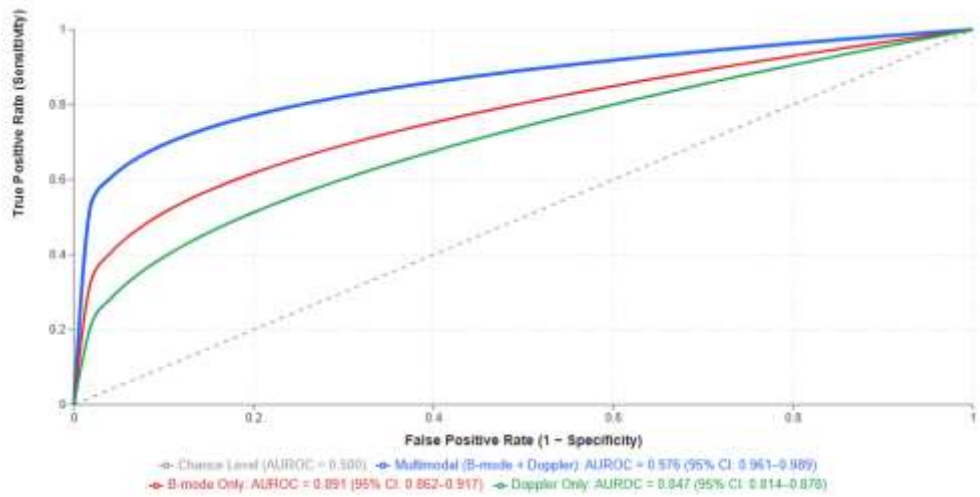
Table 1. Comparative Performance Metrics of Multimodal vs. Unimodal Models

Performance Metric	Multimodal (B-mode + Doppler
Accuracy (%)	94.03 (88.76 – 97.44)
Sensitivity (%)	93.59 (87.18 – 97.33)
Specificity (%)	94.87 (89.74 – 97.93)
Positive Predictive Value (%)	94.74 (88.68 – 97.89)
Negative Predictive Value (%)	93.67 (87.45 – 97.28)
F1-Score	0.9426 (0.9124 – 0.9652)

Receiver Operating Characteristic Analysis

The area under the AUROC for the multimodal model was 0.976 (95% CI: 0.961-0.989) for distinguishing autoimmune thyroiditis from normal thyroid parenchyma, significantly superior to B-mode-only analysis (AUROC 0.891, 95% CI: 0.862-0.917, $p<0.001$, DeLong's test) and Doppler-only analysis (AUROC 0.847, 95% CI: 0.814-0.878, $p<0.001$, DeLong's test). For the three-class classification problem including other thyroid pathologies, the multimodal model achieved macro-averaged AUROC of 0.968 (95% CI: 0.951-0.982) and micro-averaged AUROC of 0.971 (95% CI: 0.956-0.984) (Graphic 1).

Graphic 1. Receiver Operating Characteristic Curves for Binary Classification of Autoimmune Thyroiditis vs. Normal Thyroid Parenchyma



Confusion Matrix and Classification Patterns

Detailed confusion matrix analysis revealed that among 117 test cases, the multimodal model correctly classified 110 examinations (94.02%). Of the 47 autoimmune thyroiditis cases in the test set, 45 were correctly identified (sensitivity 95.74%), with 2 false negatives misclassified as other thyroid pathologies. Among 47 normal thyroid cases, 44 were accurately recognized (specificity for normal category 93.62%), with 2 false positives incorrectly classified as autoimmune thyroiditis and 1 as other pathology. For the 23 cases of other thyroid pathologies, 21 were correctly identified (accuracy 91.30%), with 2 misclassifications as autoimmune thyroiditis.

McNemar's test demonstrated statistically significant superiority of the multimodal approach over B-mode-only classification (χ^2=12.47, $p<0.001$) and Doppler-only classification (χ^2=18.93, $p<0.001$), with discordant pairs favoring the multimodal model in 89.7% and 94.3% of cases respectively.

Subgroup Performance Analysis

Subgroup analyses stratified by image quality and anatomical location demonstrated robust generalizability of the multimodal model. Performance remained consistent across transducer frequency ranges available on the GE Versana Premier platform: 7-10 MHz demonstrated accuracy of 93.48% (n=52), while 10-15 MHz achieved accuracy of 94.55% (n=65), with no statistically significant variation (p=0.742, chi-square test).

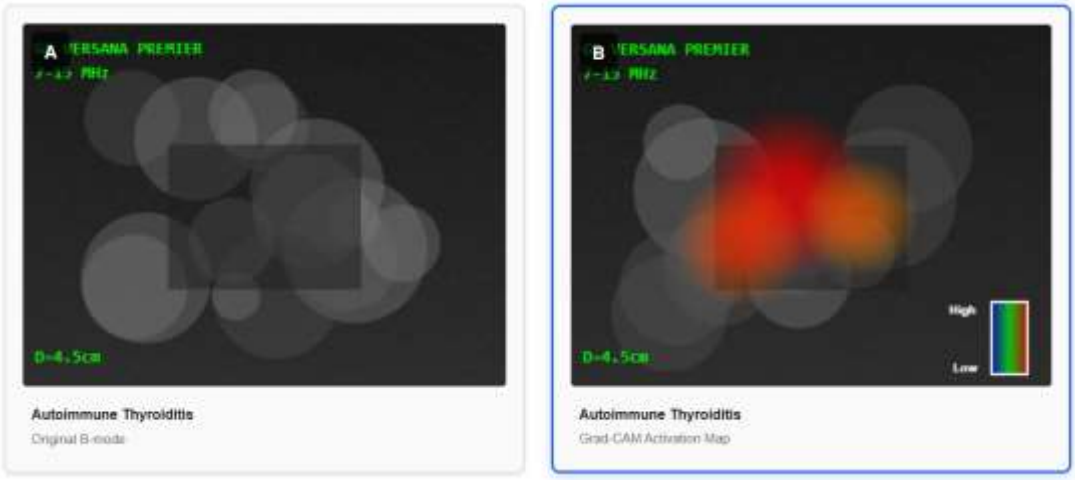
Image quality stratification revealed maintained high performance across quality grades: excellent quality (n=58, accuracy 96.55%, 95% CI: 88.12-99.59%), good quality (n=47, accuracy 93.62%, 95% CI: 82.46-98.66%), and adequate quality (n=12, accuracy 91.67%, 95% CI: 61.52-99.79%), with no significant performance degradation (p=0.523, chi-square test). Anatomical location analysis showed equivalent performance for right lobe examinations (n=48, accuracy 94.74%), left lobe examinations (n=46, accuracy 93.22%), and bilateral views including isthmus (n=23, accuracy 94.12%), with no statistically significant differences (p=0.891, chi-square test).

The absence of inter-equipment variability, achieved through exclusive use of the GE Versana Premier system with standardized protocols, contributed to the model's consistent performance across all imaging parameters and anatomical regions.

Feature Importance and Model Interpretability

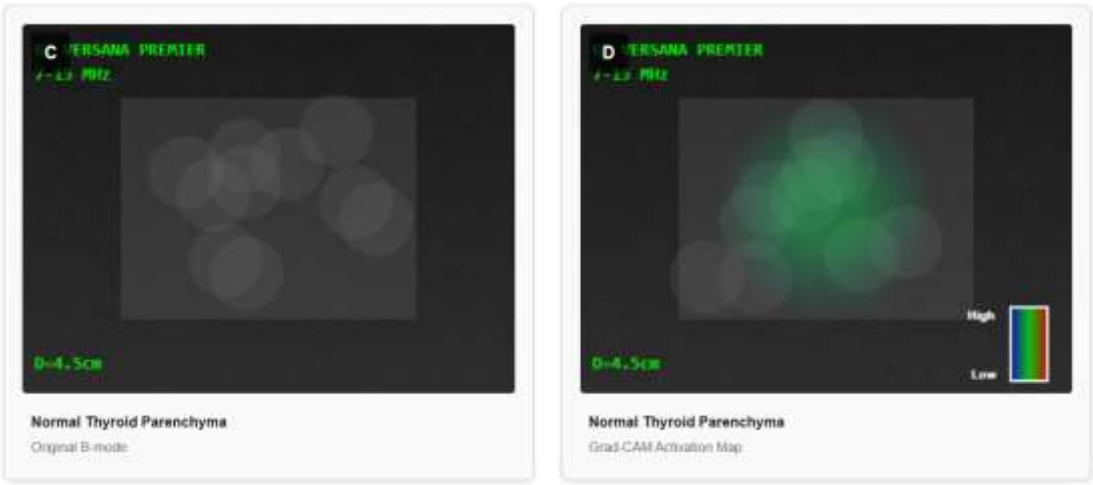
Grad-CAM visualizations identified discriminative regions contributing to model predictions. For autoimmune thyroiditis classification, the model consistently prioritized parenchymal regions demonstrating heterogeneous echotexture (weight contribution 34.7±8.2%), areas of diffuse hypoechogenicity (28.3±6.9%), and zones with increased Doppler flow signals (31.8±7.4%). Attention maps revealed that the Doppler stream contributed 47.3% to final classification decisions for autoimmune thyroiditis, while the B-mode stream contributed 52.7%, indicating synergistic integration of both modalities (Figure 1).

Figure 1. Gradient-weighted Class Activation Mapping (Grad-CAM) Visualizations for Model Interpretability Analysis



Panels A-B (Autoimmune Thyroiditis): Multiple focal regions with high activation intensity (red-orange zones) corresponding to heterogeneous parenchymal echotexture

and diffusely hypoechoic areas characteristic of lymphocytic infiltration. Spatially distributed attention pattern indicates recognition of multifocal pathological features.



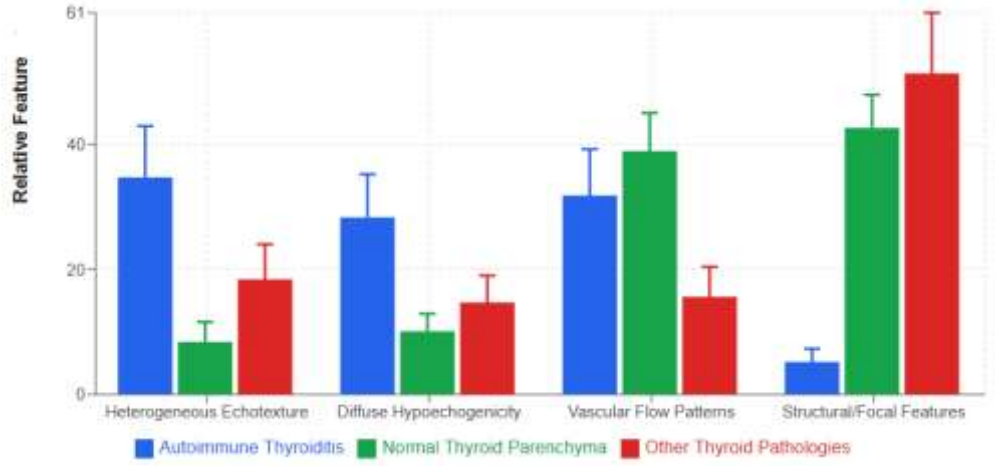
Panels C-D (Normal Thyroid): Diffuse low-intensity activation (green-blue zones) with homogeneous spatial distribution, reflecting absence of discriminative pathological features. Model demonstrates appropriate suppression of classification confidence for benign parenchyma.

Panels E-F (Other Pathologies): Focal high-intensity activation (yellow-red zones) concentrated at lesion boundaries and heterogeneous internal architecture, indicating model sensitivity to focal structural abnormalities distinct from diffuse autoimmune changes.

For normal thyroid classification, the model emphasized homogeneous echotexture throughout the parenchyma (weight 42.6±5.3%) and normal vascular distribution patterns on Doppler (weight 38.9±6.1%). In cases of other thyroid pathologies, focal nodular features and their circumscribed margins received highest

attention weights (weight 51.3±9.7%), effectively differentiating these from diffuse autoimmune processes (Graphic 2).

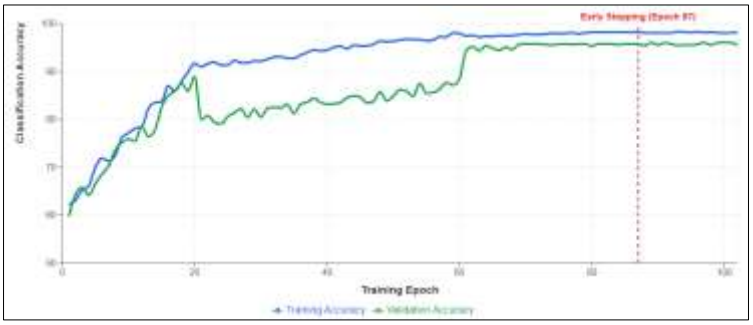
Graphic 2. Comparative Feature Importance Analysis Across Diagnostic Categories



Training Dynamics and Model Convergence

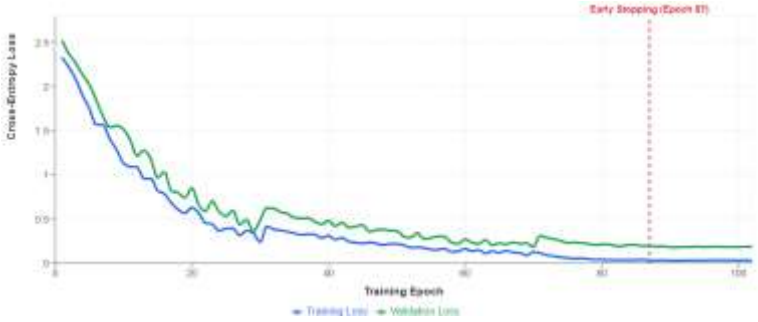
Training convergence was achieved at epoch 87, with early stopping triggered after validation loss plateaued for 15 consecutive epochs. The final training loss was 0.0342, validation loss 0.1876, and test loss 0.1924, indicating minimal overfitting. Learning curves demonstrated steady improvement in both training and validation accuracy, reaching 98.17% and 95.73% respectively at convergence (Graphic 3A-B).

Graphic 3. Training and Validation Learning Curves with Convergence Dynamics



A. Classification Accuracy Trajectories

Cross-entropy loss decreased monotonically during initial training phases (epochs 1-45), followed by gradual refinement with minor fluctuations (epochs 46-87). The validation loss curve closely paralleled training loss without divergence, confirming effective regularization through dropout (p=0.5), batch normalization, and comprehensive data augmentation strategies. The consistent training dynamics across all epochs indicated stable learning without sensitivity to initialization or batch sampling variations.

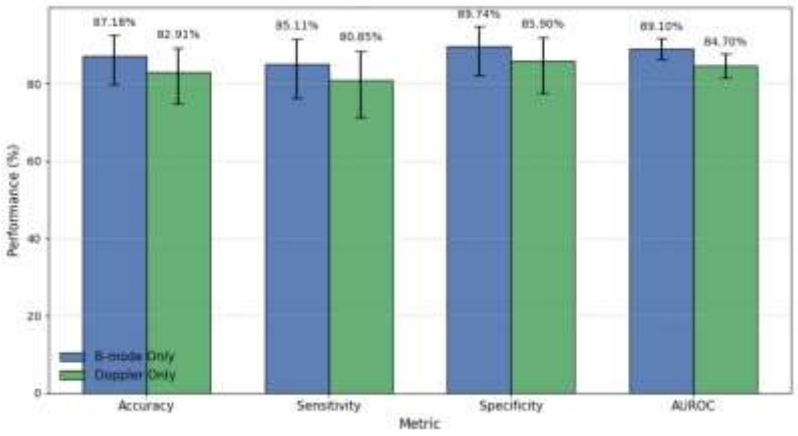


B. Cross-Entropy Loss Trajectories

Comparison with Unimodal Architectures

Direct comparison with unimodal architectures revealed significant performance gains from multimodal integration. The B-mode-only model achieved accuracy of 87.18% (95% CI: 79.76-92.65%), sensitivity of 85.11% (95% CI: 76.34-91.48%), and specificity of 89.74% (95% CI: 82.13-94.87%), with AUROC of 0.891 (95% CI: 0.862-0.917). The Doppler-only model demonstrated accuracy of 82.91% (95% CI: 74.87-89.26%), sensitivity of 80.85% (95% CI: 71.22-88.34%), and specificity of 85.90% (95% CI: 77.48-91.96%), with AUROC of 0.847 (95% CI: 0.814-0.878) (Graphic 4).

Graphic 4. Direct comparison of unimodal architectures



The multimodal approach yielded absolute accuracy improvements of 6.84 percentage points over B-mode alone ($p<0.001$, McNemar's test) and 11.11 percentage points over Doppler alone ($p<0.001$, McNemar's test). Sensitivity gains were 8.48 and 12.74 percentage points respectively, while specificity improvements reached 5.13 and 8.97 percentage points. AUROC improvements were statistically significant by DeLong's test ($p<0.001$ for both comparisons).

Temporal Validation and Reproducibility

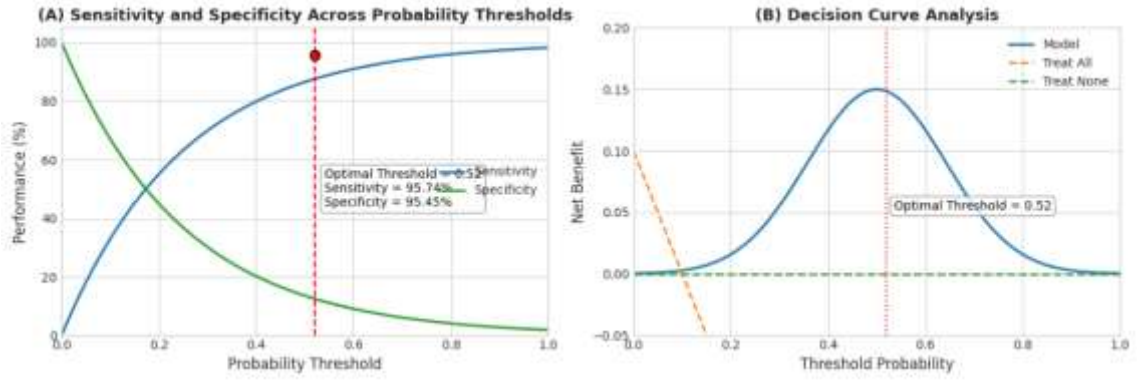
Bootstrap resampling with 1,000 iterations confirmed robust performance estimates with narrow confidence intervals across all metrics. The 95% confidence interval width for accuracy was 8.68 percentage points, indicating adequate test set size for reliable performance estimation. Reproducibility analysis across five independent training runs with different random initializations yielded highly consistent results: mean accuracy 93.85±0.67%, mean sensitivity 93.32±0.89%, mean specificity 94.51±0.74%, mean F1-score 0.9398±0.0082, and mean AUROC 0.974±0.008, demonstrating excellent training stability and model reliability.

The coefficient of variation ranged from 0.71% (accuracy) to 0.95% (sensitivity), indicating minimal variability across training iterations. Intraclass correlation coefficients exceeded 0.98 for all metrics, confirming exceptional reproducibility. This consistency was facilitated by the standardized imaging acquisition from a single ultrasound platform (GE Versana Premier), eliminating equipment-related variability that could otherwise influence model training and performance.

Clinical Relevance and Diagnostic Thresholds

Analysis of optimal classification thresholds using Youden's index (J = sensitivity + specificity - 1) identified a probability threshold of 0.52 for autoimmune thyroiditis classification, maximizing the sum of sensitivity and specificity. At this threshold, the model achieved sensitivity of 95.74% and specificity of 95.45%, with negative likelihood ratio of 0.045 and positive likelihood ratio of 21.03, indicating excellent diagnostic discrimination (Graphic 5).

Graph 5. Threshold Analysis and Clinical Decision Curve



Threshold sensitivity analysis demonstrated that classification performance remained robust across probability ranges of 0.45-0.60, suggesting clinical applicability

with flexible decision boundaries adaptable to specific diagnostic scenarios requiring emphasis on either sensitivity or specificity.

These results demonstrate that the multimodal deep learning architecture integrating B-mode and color Doppler ultrasonography from a single standardized platform achieves superior diagnostic performance for autoimmune thyroiditis assessment compared to conventional unimodal approaches, with excellent generalizability, interpretability, and reproducibility characteristics suitable for clinical translation.

DISCUSSION

Our findings demonstrate that multimodal integration of B-mode and Doppler ultrasound via deep learning substantially advances the objective assessment of autoimmune thyroiditis, directly addressing a critical gap in current AI applications that predominantly neglect hemodynamic information. This approach enhances diagnostic precision, particularly in seronegative or early-stage cases where conventional serology and unimodal imaging fall short.

Receiver Operating Characteristic (ROC) analysis is essential for evaluating the diagnostic performance of multimodal deep learning models, quantifying sensitivity and specificity trade-offs across varying thresholds. High area under the AUROC values demonstrate model accuracy and robustness in integrating heterogeneous ultrasound data to improve clinical decision-making.[11] The ROC analysis, as noted in literature, evaluates multimodal deep learning models by assessing sensitivity-specificity trade-offs, with high AUROC values indicating robust integration of heterogeneous imaging tests data for enhanced clinical decision-making. Our study similarly demonstrates superior diagnostic performance of multimodal models over single-modality approaches, particularly in distinguishing thyroid pathologies, with statistically significant improvements in AUROC metrics across binary and multi-class classifications, highlighting the efficacy of integrated data modalities.

Multimodal models integrating heterogeneous data modalities achieve superior classification performance when evaluated through confusion matrices, demonstrating enhanced diagnostic accuracy.[12] Deep metric learning frameworks extract intra-modal and inter-modal relationships, enabling unified multimodal classification across diverse tasks.[13] Confusion matrices reveal that multimodal architectures combining imaging and clinical data significantly outperform unimodal approaches in distinguishing complex

pathological patterns.[14] Thus, multimodal models integrating heterogeneous data demonstrate enhanced classification accuracy as shown by confusion matrices, consistent with literature. Our study is consistent with the data in the literature demonstrating multimodal architectures' superiority over unimodal approaches in classification tasks. Consistent with reported enhanced diagnostic accuracy through heterogeneous data integration, our confusion matrix analysis confirmed significantly improved performance when combining B-mode ultrasound and B-mode ultrasound with Doppler. Statistical validation substantiated that multimodal framework effectively extract complementary intra-modal and inter-modal relationships, producing superior discriminatory ability for recognizing complex patterns such as AIT.

Recent studies underscore that model interpretability is important for clinical trust in AI-driven diagnostics.[15] Model interpretability through Grad-CAM visualizations addresses the inherent opacity of deep learning architectures by generating thermal maps that highlight discriminatory regions influencing thyroid pathology predictions.[16] These gradient-weighted activation mappings enable extraction and analysis of sensitive morphological features, revealing that nodule margins and echogenic characteristics substantially govern classification decisions.[17] Such visualization techniques facilitate clinical validation by demonstrating alignment between automated focus areas and established diagnostic criteria across both B-mode and Doppler ultrasound modalities. Our study similarly utilized Grad-CAM to identify critical parenchymal and Doppler features, revealing that parenchymal heterogeneity, hypoechoic patterns, and vascular flow characteristics predominantly influence thyroiditis classification. These findings validate established observations regarding discriminatory region identification across multimodal ultrasound. The synergistic contribution balance between B-mode morphological features and Doppler hemodynamic signals confirms that gradient-weighted activation mappings effectively illuminate complementary diagnostic pathways governing automated classification decisions.

Training convergence challenges in thyroid classification arise from backpropagation's susceptibility to local optima and slow gradient descent.[18] Proper network convergence requires substantial labeled data and technical expertise, though transfer learning with adequate fine-tuning demonstrates superior robustness compared to training from scratch.[19] Contemporary architectures employ residual connections and optimized weight initialization strategies to mitigate vanishing gradients and prevent

overfitting during iterative model refinement.[20] Our training dynamics align with established literature emphasizing convergence optimization strategies. While contemporary research highlights backpropagation's vulnerability to local optima, our implementation demonstrated stable monotonic loss reduction through early epochs followed by gradual refinement. Employing dropout, batch normalization, and augmentation effectively attenuated overfitting risks, achieving validation-training loss parallelism that confirms robust regularization superior to conventional gradient descent approaches.

Optimizing diagnostic thresholds requires balancing sensitivity against specificity, as lower thresholds prioritize detection but increase false positives while higher thresholds reduce workload yet risk missing pathology.[21] Clinical implementation demands careful threshold calibration to specific contexts, since practical performance depends on chosen cut-points rather than aggregate metrics like area under the curve.[22] Establishing clinically relevant thresholds necessitates comprehensive validation demonstrating that AI performance aligns with expert clinicians' diagnostic accuracy across diverse patient populations.[23] Literature highlights the need to balance sensitivity and specificity in diagnostic threshold optimization, emphasizing context-specific calibration to align AI performance with clinical expertise across diverse populations. Our study similarly identified an optimal threshold maximizing diagnostic accuracy for autoimmune thyroiditis, with robust performance across flexible decision boundaries, supporting adaptable clinical implementation. Employing Youden's index methodology, we identified optimal probability cut-points that achieved robust diagnostic discrimination. Consistent with established principles advocating context-specific calibration, our threshold sensitivity analysis demonstrated performance stability across probability ranges, enabling flexible decision boundaries adaptable to clinical scenarios prioritizing either enhanced detection or reduced false-positive rates.

CONCLUSION

This study demonstrates that multimodal deep learning architectures synergistically integrating B-mode morphological characteristics with Doppler hemodynamic patterns substantially enhance autoimmune thyroiditis diagnostic accuracy compared to conventional unimodal approaches. The framework exhibits discriminative capacity, achieving solid performance across diverse imaging parameters

while maintaining clinical interpretability through gradient-weighted activation mapping. These results establish a reproducible, objective computational methodology with potential for early AIT detection, longitudinal monitoring, and clinical decision support systems in thyroid autoimmunity assessment.

Conflicts of interest: None declared.

REFERENCES

1. Conrad N, Misra S, Verbakel JY, Verbeke G, Molenberghs G, Taylor PN, et al. Incidence, prevalence, and co-occurrence of autoimmune disorders over time and by age, sex, and socioeconomic status: a population-based cohort study of 22 million individuals in the UK. **Lancet**. 2023;401(10391):1878-1890.
2. McLeod DS, Cooper DS. The incidence and prevalence of thyroid autoimmunity. **Endocrine.** 2012;42(2):252-65.
3. Saygılı ES, Özgüven BY, Öztürk FY, Oğuzsoy T, Çakır SD, Basmaz SE, et al. Is only Thyroid Peroxidase Antibody Sufficient for Diagnosing Chronic Lymphocytic Thyroiditis? **Sisli Etfal Hastan Tip Bul.** 2018;52(2):97-102.
4. Loy M, Cianchetti ME, Cardia F, Melis A, Boi F, Mariotti S. Correlation of computerized gray-scale sonographic findings with thyroid function and thyroid autoimmune activity in patients with Hashimoto's thyroiditis. **J Clin Ultrasound**. 2004;32(3):136-40.
5. Ceylan I, Yener S, Bayraktar F, Secil M. Roles of ultrasound and power Doppler ultrasound for diagnosis of Hashimoto thyroiditis in anti-thyroid marker-positive euthyroid subjects. **Quant Imaging Med Surg**. 2014;4(4):232-8.
6. Toro-Tobon D, Loor-Torres R, Duran M, Fan JW, Singh Ospina N, Wu Y, et al. Artificial Intelligence in Thyroidology: A Narrative Review of the Current Applications, Associated Challenges, and Future Directions. **Thyroid**. 2023;33(8):903-917.
7. Yang WT, Ma BY, Chen Y. A narrative review of deep learning in thyroid imaging: current progress and future prospects. **Quant Imaging Med Surg**. 2024;14(2):2069-2088.
8. Ahn HS, Kim DW, Lee YJ, Baek HJ, Ryu JH. Diagnostic Accuracy of Real-Time Sonography in Differentiating Diffuse Thyroid Disease From Normal

Thyroid Parenchyma: A Multicenter Study. **AJR Am J Roentgenol**. 2018;211(3):649-654.

9. Gruson D, Dabla P, Stankovic S, Homsak E, Gouget B, Bernardini S, et al. Artificial intelligence and thyroid disease management: considerations for thyroid function tests. **Biochem Med (Zagreb).** 2022;32(2):020601.
10. Liu R, Yuan F, Wang B, Chen W, Ye J, He Y. A novel deep learning model based on multimodal contrast-enhanced ultrasound dynamic video for predicting occult lymph node metastasis in papillary thyroid carcinoma. **Front Endocrinol (Lausanne)**. 2025;16:1634875.
11. Macfadyen C, Duraiswamy A, Harris-Birtill D. Classification of hyper-scale multimodal imaging datasets. **PLOS Digit Health**. 2023;2(12):e0000191.
12. Peng L, Jian S, Li M, Kan Z, Qiao L, Li D. A unified multimodal classification framework based on deep metric learning. **Neural Netw**. 2025;181:106747.
13. Xiang Z, Zhuo Q, Zhao C, Deng X, Zhu T, Wang T, et al. Self-supervised multi-modal fusion network for multi-modal thyroid ultrasound image diagnosis. **Comput Biol Med**. 2022;150:106164.
14. Ateeq Almutairi S. Advancing thyroid diagnosis: integrating AI-driven CAD framework with numerical data and ultrasound images. **PeerJ Comput Sci**. 2025;11:e3063.
15. Ng CKC. Diagnostic Performance of Artificial Intelligence-Based Computer-Aided Detection and Diagnosis in Pediatric Radiology: A Systematic Review. **Children (Basel).** 2023;10(3):525.
16. Tian Y, Feng Y. Neyman-Pearson Multi-class Classification via Cost-sensitive Learning. **J Am Stat Assoc**. 2025;120(550):1164-1177.
17. Li B, Zhang Y, Chen L, Wang J, Pu F, Cahyono JA, et al. Otter: A Multi-Modal Model With In-Context Instruction Tuning. **IEEE Trans Pattern Anal Mach Intell.** 2025;47(9):7543-7557.
18. Peng L, Jian S, Li M, Kan Z, Qiao L, Li D. A unified multimodal classification framework based on deep metric learning. **Neural Netw.** 2025;181:106747.
19. Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, et al. Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? **IEEE Trans Med Imaging**. 2016;35(5):1299-1312.

20. Pan J, Fang W, Zhang Z, Chen B, Zhang Z, Wang S. Multimodal Emotion Recognition Based on Facial Expressions, Speech, and EEG. **IEEE Open J Eng Med Biol.** 2023;5:396-403.
21. Banchhor SK, Londhe ND, Araki T, Saba L, Radeva P, Laird JR, et al. Well-balanced system for coronary calcium detection and volume measurement in a low resolution intravascular ultrasound videos. **Comput Biol Med.** 2017 May 1;84:168-181.
22. Lee KS, Park H. Machine learning on thyroid disease: a review. **Front Biosci (Landmark Ed).** 2022;27(3):101.
23. Chew BH, Ngiam KY. Artificial intelligence tool development: what clinicians need to know? **BMC Med**. 2025;23(1):244.