International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5364 - 5381
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.701

An Explainable Hybrid Deep Learning Framework with Adaptive Multi-Level Feature Fusion for Early Thyroid Nodule Detection and Classification Using Ultrasound Images

Velkumar K[1], Bhavani M[2], Venkatalakshmi M[3], Pavithra M[4], Archana R[5]
[1,3,5]Assistant Professor, Computer Science and Engineering, Nadar Saraswathi College of Engineering and Technology, Theni, Tamilnadu
[2]Assistant Professor, Information Technology, Nadar Saraswathi College of Engineering and Technology, Theni, Tamilnadu
[4]Assistant Professor, Artificial Intelligence and Data Science, Nadar Saraswathi College of Engineering and Technology, Theni, Tamilnadu
Emails: *velkumarskc@gmail.com[1], gmbhavani1990@gmail.com[2], venkatalakshmims@gmail.com[3], pavithramanibtech@gmail.com[4], archanacse.nscet @gmail.com[5]*

Abstract
Thyroid ultrasound is widely used to assess thyroid nodules, but reliable interpretation remains challenging because ultrasound images are affected by speckle noise, weak contrast, heterogeneous tissue patterns, and indistinct lesion boundaries. These difficulties become more pronounced for small or irregular nodules. This study presents an Explainable Hybrid Deep Learning Framework with Adaptive Multi-Level Feature Fusion (AMFF) for thyroid nodule analysis. The framework combines a CNN branch with a Transformer/Mamba branch to learn complementary local and contextual representations. AMFF is introduced to weight information from different feature levels rather than relying on a fixed feature-combination scheme. A boundary-aware segmentation component is used to refine lesion delineation, and segmentation and malignancy classification are addressed within a multi-task architecture. Grad-CAM++ is incorporated to provide visual evidence associated with the classification decision. The resulting framework is designed as an integrated computer-aided analysis pipeline for thyroid ultrasound images.

Keywords: *Adaptive Multi-Level Feature Fusion; Deep Learning; Explainable AI; Thyroid Nodule; Ultrasound Imaging*

1. Introduction

Thyroid nodules are common findings in clinical practice, and distinguishing benign nodules from potentially malignant lesions is an important part of thyroid assessment. Ultrasound (US) is the preferred imaging modality for initial evaluation because it is non-invasive, does not use ionizing radiation, is relatively inexpensive, and provides detailed morphological information. Nevertheless, image interpretation can be difficult when speckle noise, low contrast, heterogeneous texture, blurred margins, or irregular lesion shapes are present. These effects are particularly problematic for small or poorly defined nodules and can reduce the reliability of computer-aided diagnosis (CAD) systems [1], [3], [6], [7]. Deep learning has increasingly been applied to automated thyroid ultrasound analysis. CNN-based models are effective at learning image features such as edges, textures, shapes, and local boundary patterns directly from training data. U-Net and U-Net++ variants, together with attention mechanisms and multi-task strategies, have reported promising segmentation performance [1], [2], [3], [8]. However, their emphasis on local image structure does not always provide sufficient representation of long-range spatial relationships. To capture broader contextual information, Transformer-based models and state-space approaches such as Mamba have recently been investigated for medical image analysis. A hybrid design can exploit the complementary behavior of the two model families: convolutional layers can preserve fine local details, whereas Transformer or Mamba components can model relationships over a wider spatial context [5], [9], [10]. The additional branches, however, may increase computational requirements, and several existing hybrid approaches continue to focus primarily on segmentation rather than combining segmentation, malignancy classification, and visual explanation in one framework [5], [9]. The way these representations are combined is also important. Low-level features retain edges and texture, intermediate features describe structural patterns, and deeper features provide higher-level semantic information. Existing systems commonly use direct concatenation, fixed attention, or predefined weights to merge such representations [8], [10], [11]. A fixed strategy may not be equally effective for nodules that differ substantially in size, shape, contrast, and internal texture. An adaptive fusion mechanism could therefore allow the network to place greater emphasis on the feature levels that are most informative for a particular image. Precise boundary delineation presents another difficulty in thyroid ultrasound. Nodule margins may be blurred by noise and low contrast, while heterogeneous tissue, calcifications, cystic components, and irregular shapes can further complicate segmentation. Recent residual, context-aware, and attention-based networks have attempted to improve boundary localization [6]–[8]. Most of these methods, however, treat segmentation as the principal objective and do not establish a complete link between accurate lesion delineation and malignancy prediction. Malignancy classification has likewise received considerable attention. Deep learning models have demonstrated that ultrasound images contain information that can support benign-versus-malignant classification [11], [12], [14]. At the same time, reported studies often use relatively small datasets, single-center samples, imbalanced classes, or limited external validation [12], [14]. Some approaches also combine handcrafted descriptors with learned features, which can make the system more dependent on manually selected representations [11]. These observations support the development of an end-to-end pipeline that can localize a suspicious region, segment the lesion, and estimate malignancy using shared learned representations. Interpretability is particularly important when an automated model is intended to support clinical decision-making. A high-performing model may still be difficult to trust if the basis for its prediction cannot be examined. Grad-CAM-based methods have been used in thyroid ultrasound studies to identify image regions associated with model predictions [12]. A heatmap alone, however, does not necessarily explain how localization, segmentation, and classification interact. A practical CAD system should therefore provide both the predicted class and visual evidence indicating the regions that contributed to the decision. Computational cost is an additional consideration. Transformer, Mamba, attention, and hybrid feature-extraction modules can improve representation learning, but they may also increase memory use, inference time, and implementation complexity [5], [9], [10]. This trade-off is important when a system is intended for clinical environments with limited computational resources or when near-real-time analysis is desired. Motivated by these limitations, this study proposes an Explainable Hybrid Deep Learning Framework with Adaptive Multi-Level Feature Fusion for thyroid nodule detection, segmentation, and classification from ultrasound images. The framework combines adaptive preprocessing, suspicious-region detection, a dual-path local-global encoder, AMFF, boundary-aware segmentation, multi-task classification, and Grad-CAM++-based explanation. The central feature-learning stage uses two complementary branches. The CNN branch is intended to preserve local texture, edge, structural, and boundary information, while the Transformer/Mamba branch models broader contextual relationships. Representations from different CNN levels and the global branch are then passed to AMFF, which learns their relative contribution instead of treating all feature sources identically. The fused representation is used by both the segmentation and classification components. An enhanced U-Net/UNet++-style decoder with boundary-aware attention is used to refine the nodule mask, while the classification branch combines the learned representation with ROI and segmentation information to estimate benign or malignant status. Joint optimization is intended to allow the related tasks to share useful information during training. The final stage provides model explanations through Grad-CAM++ and segmentation visualization. The heatmap identifies image regions associated with the predicted malignancy class, while the segmentation mask indicates the estimated nodule boundary. Thus, the

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5364 - 5381
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.701

framework is designed to present a prediction together with visual evidence rather than as an isolated class label. The main contributions of this work are summarized as follows: The study develops a unified pipeline that connects suspicious-region detection, nodule segmentation, and benign/malignant classification within a common architecture. A dual-path CNN–Transformer/Mamba encoder is used to capture complementary local image details and broader spatial context. The proposed AMFF module adaptively combines low-, intermediate-, and high-level representations. A boundary-aware segmentation mechanism is incorporated to improve the delineation of small, irregular, and low-contrast nodules. Multi-task learning is used to share information between related diagnostic objectives, while Grad-CAM++ and segmentation visualization provide an interpretable view of the model output. Overall, the proposed design brings together localization, local-global representation learning, adaptive feature fusion, boundary refinement, malignancy classification, multi-task optimization, and explainability. The intention is to provide a coherent framework rather than treating each diagnostic task as an independent problem.

1.1.Related Work and Literature Review

Recent advances in artificial intelligence have expanded the use of CAD systems for thyroid ultrasound. The literature can broadly be grouped around three related objectives: locating nodules, delineating their boundaries, and classifying them according to malignancy risk. These tasks are closely connected, but many published approaches still address them separately. Reviews of the field also identify dataset diversity, interpretability, and external validation as important barriers to clinical translation [2], [3].

1.1.1. Deep Learning-Based Thyroid Nodule Segmentation

Early deep learning studies in thyroid ultrasound concentrated mainly on automatic segmentation. Gong et al. introduced a multi-task approach that incorporates thyroid-region prior information to support nodule segmentation [4]. The use of anatomical context can improve localization and boundary delineation, but the method remains primarily segmentation-oriented and requires additional thyroid-region annotations. This increases annotation requirements and may complicate transfer to other datasets or clinical settings. Attention mechanisms have subsequently been incorporated into U-Net-based architectures to improve the use of spatial and contextual information. Dong et al. proposed a dual-path attention-enhanced U-Net++ that combines global, spatial, and channel attention for thyroid nodule segmentation [5]. Although these components can strengthen contextual representation, the approach remains focused on segmentation and does not include a dedicated malignancy-classification stage. Its evaluation on TN3K also provides limited evidence regarding performance across institutions and datasets. Boundary quality is particularly important for thyroid nodules because lesion margins can be weak, irregular, or partially obscured by surrounding tissue. RCGA-UNet-DS uses residual cross-gating and deep supervision to improve boundary localization and contextual modeling [6]. DCCE-UNet introduces difference- and context-aware mechanisms for blurred boundaries and heterogeneous regions [7]. These methods address important segmentation difficulties, but they do not establish a complete detection–segmentation–classification workflow. Small lesions and nodules with weak margins remain challenging cases. Synthetic data generation and domain adaptation have also been explored to address variation in ultrasound appearance. Prochazka and Zeman investigated Stable Diffusion-based synthetic nodule generation and inpainting as a way to increase domain diversity for thyroid nodule segmentation [8]. Such methods may broaden the training distribution, but their effect on malignancy classification and clinical generalization requires further investigation.

1.1.2. CNN-Based and Hybrid Local-Global Feature Learning

CNNs remain effective for extracting local characteristics such as texture, edges, shape, and fine structural details. Their ability to model distant spatial relationships is more limited, which has encouraged research into hybrid architectures that combine convolutional processing with broader contextual models. The Mamba- and ResNet-based dual-branch network (MRDB) combines ResNet for local representation learning with Mamba for global feature modeling [9]. This design exploits the complementary strengths of local convolutional processing and state-space modeling. Its dual-branch structure and decoder, however, introduce additional computational complexity, and the reported work is primarily directed toward segmentation rather than a unified detection and malignancy-classification pipeline. LCA-Net combines CNN-based local feature extraction with Transformer-based context attention and adaptive convolution [10]. Its results demonstrate the potential benefit of local-global modeling for thyroid ultrasound segmentation. Nevertheless, the added contextual modules increase architectural complexity, and the framework does not provide a comprehensive explainability component or extensive integration of clinical or multimodal information. Taken together, the literature supports the value of combining local and global representations. Simply placing two encoder branches next to one another does not, however, determine how information from different semantic levels should be weighted. Low-level features are useful for edges and texture, intermediate features describe structural organization, and high-level features carry semantic context. This motivates an adaptive fusion mechanism that can respond to the characteristics of the nodule rather than applying the same weighting to every case.

1.1.3. Feature Fusion and Thyroid Nodule Classification

Feature representation is central to malignancy classification. One earlier approach combined handcrafted descriptors such as HOG, LBP, SIFT, and SURF with deep features from a Res-GAN architecture [11]. The combination can enrich the representation, but reliance on manually designed descriptors may reduce adaptability when image characteristics change. The approach is also classification-oriented and does not provide automatic segmentation within an explainable end-to-end workflow. Deep CNNs have also been applied directly to benign-versus-malignant classification. A ResNet18-based study demonstrated the feasibility of this task and used Grad-CAM to visualize regions associated with the prediction [12]. However, the use of relatively small, single-center datasets, class imbalance, binary labels, and limited clinical validation restricts the extent to which such results can be generalized. Grad-CAM is useful for localization, but a heatmap by itself does not describe the complete diagnostic pathway. Systematic reviews report encouraging results for deep learning in thyroid nodule detection and segmentation, while also noting substantial differences in datasets, image quality, evaluation protocols, and external validation [13]. These differences make direct comparison difficult and reinforce the need for consistent evaluation procedures and clinically meaningful interpretation.

1.1.4. Explainable Artificial Intelligence for Thyroid Ultrasound

Explainable AI has become increasingly important in medical imaging because clinicians need to assess the evidence underlying an automated prediction. Grad-CAM and related methods have been used to highlight image regions associated with thyroid nodule classification [12]. In many systems, however, explanation remains limited to a heatmap and is not integrated with localization and segmentation. For clinical adoption, interpretability must be considered alongside predictive performance. A model that reports high accuracy without showing the evidence used for its decision may be difficult to evaluate in practice. Recent reviews of thyroid ultrasound AI emphasize explainability, robustness, cross-domain adaptation, and clinical validation as continuing requirements [2], [3].

1.1.5. Limitations of Existing Integrated Approaches

The literature shows a persistent separation between segmentation-focused and classification-focused systems. Segmentation methods generally optimize pixel-level lesion delineation, whereas classification methods estimate benign or malignant status. Consequently, many approaches do not provide one workflow that begins with localization and continues through segmentation and malignancy prediction [4] - [12]. Another recurring limitation is the reliance on B-mode ultrasound alone. Complementary information from Doppler imaging, elastography, contrast-enhanced ultrasound, or patient-level clinical data is used less frequently [6], [7], [9], [10].

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5364 - 5381
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.701

Dataset scale and diversity are also concerns because single-center or small datasets may not represent variation in patients, equipment, acquisition protocols, and clinical practice [12], [14]. TN3K is an important benchmark for thyroid nodule segmentation. Gong et al. introduced the dataset for automated thyroid nodule analysis, and subsequent studies have used it to compare segmentation approaches [4], [5]. The dataset contains 3,493 ultrasound images. Although such a benchmark is valuable for standardized comparison, performance on one public dataset is not sufficient to establish broad clinical generalization; external and multicenter evaluation remain important.

1.1.6. Research Gap

The reviewed studies indicate several gaps. Most approaches concentrate on either segmentation or classification rather than integrating detection, segmentation, and malignancy prediction. Although CNNs capture local detail and Transformer/Mamba models provide broader contextual modeling, an adaptive mechanism for combining multi-level local and global information remains desirable. Small, irregular, and low-contrast nodules continue to pose difficulties, while clinically useful explanations are not consistently provided. Hybrid architectures may also increase computational cost, and limited multicenter validation restricts conclusions about clinical translation. These observations motivate an integrated framework that addresses image-quality variation, suspicious-region localization, local-global representation learning, adaptive fusion, boundary refinement, malignancy classification, and explainability. The proposed study combines a hybrid CNN–Transformer/Mamba encoder with AMFF, boundary-aware segmentation, multi-task learning, and an XAI module so that the related tasks can be evaluated within one pipeline.

1.1.7. Summary of Existing Literature and Positioning of the Proposed Work

The development of thyroid ultrasound AI reflects a progression from conventional image-processing techniques toward CNN-based segmentation, attention-enhanced networks, hybrid CNN–Transformer/Mamba models, and explainable CAD systems. Recent reviews identify U-Net variants, Transformer-based methods, boundary-aware networks, weakly supervised learning, diffusion-based approaches, and multimodal fusion as active directions [2], [3]. The reviewed approaches do not appear to combine all of these components within one framework. The proposed study therefore brings together adaptive preprocessing, lightweight nodule detection, hybrid CNN–Transformer/Mamba encoding, AMFF, boundary-aware segmentation, multi-task malignancy classification, and Explainable AI. AMFF is intended to preserve useful fine-grained information while incorporating broader semantic context, whereas the multi-task design enables information sharing between related objectives. The resulting system is positioned as an integrated decision-support framework rather than as another segmentation-only or classification-only model. Its purpose is to connect task-specific deep learning components within a single, clinically oriented workflow.

2. Proposed Methodology

The proposed framework is designed as an end-to-end computer-aided diagnostic pipeline for thyroid ultrasound. It includes image preprocessing, suspicious-region detection, hybrid local-global feature extraction, adaptive feature fusion, boundary-aware segmentation, malignancy classification, multi-task optimization, and visual explanation. The design directly targets the limitations identified in the literature, particularly the separation of segmentation and classification, incomplete local-global modeling, difficulty with small or low-contrast nodules, limited interpretability, and computational burden. The complete processing pipeline is expressed as: Input Ultrasound Image → Preprocessing → Nodule Detection → ROI Extraction → Hybrid CNN–Transformer/Mamba Encoder → AMFF → Boundary-Aware Segmentation + Malignancy Classification → XAI → Clinical Decision Support.

2.1. Overall Framework Architecture

Each ultrasound image first undergoes preprocessing to reduce speckle noise and improve contrast. A lightweight detector then identifies a suspicious thyroid region, which is cropped to form the ROI. The ROI is subsequently analyzed by the dual-path encoder. The CNN branch extracts local and fine-grained characteristics, whereas the Transformer/Mamba branch models wider contextual

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5364 - 5381
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.701

relationships. Features from several semantic levels are then combined by AMFF. The fused representation is supplied to both the boundary-aware segmentation decoder and the malignancy-classification branch. Grad-CAM++ is applied at the final stage to generate a heatmap that can be viewed together with the predicted segmentation mask.[15 – 20]

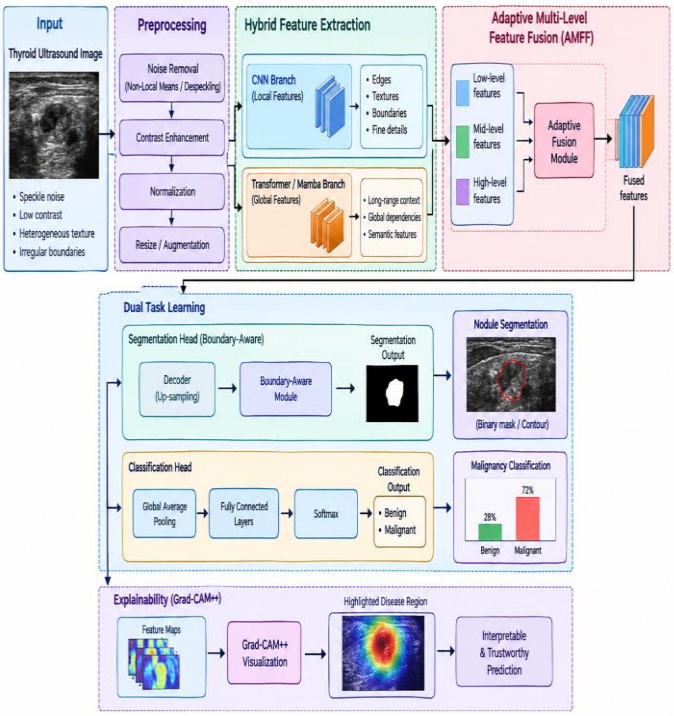

Figure 1 Proposed Explainable Hybrid Deep Learning Framework with Adaptive Multi-Level Feature Fusion (AMFF)

2.2.Adaptive Ultrasound Image Preprocessing

Ultrasound images can vary considerably in noise level, contrast, intensity distribution, and acquisition artifacts. The proposed preprocessing stage is therefore designed to standardize the input while retaining structures that may be diagnostically relevant. The sequence includes speckle-noise reduction, local contrast enhancement, intensity normalization, resizing, and training-time augmentation.

- **Speckle Noise Reduction**
  Non-Local Means (NLM) filtering, or a comparable adaptive despeckling method, can be applied to reduce speckle while preserving relevant anatomical structure. Excessive smoothing should be avoided because small boundaries and texture patterns may be lost.
- **Contrast Enhancement**
  CLAHE is applied to improve local contrast. This is useful when the intensity difference between the nodule and surrounding thyroid tissue is limited.
- **Intensity Normalization**
  Intensity normalization is used to reduce differences in image scale and distribution across ultrasound examinations, thereby providing a more consistent input to the learning model.
- **Data Augmentation**
  Training-time augmentation may include rotation, clinically appropriate flipping, scaling, cropping, elastic deformation, and controlled brightness or contrast changes. The selected transformations should preserve the anatomical validity of the ultrasound image.

The enhanced image can therefore be represented as:

[
$$I_p=\mathcal{A}(\mathcal{C}(\mathcal{N}(I)))$$
]

Here, I denotes the original ultrasound image, N denotes the noise-reduction operation, C represents contrast enhancement and normalization, and A represents augmentation. The preprocessing stage is intended to be adaptive rather than blindly applying an identical sequence to every image; the operations can be selected according to the observed image quality.

2.3.Lightweight Thyroid Nodule Detection

After preprocessing, a lightweight CNN-based detector identifies suspicious thyroid nodule regions. This stage reduces the amount of irrelevant image content passed to the subsequent segmentation and classification modules.

- The detector generates:
- Nodule bounding box
- Nodule location

- Detection confidence score

The detected region is cropped to form the ROI used by the hybrid encoder. The current design allows either a lightweight YOLO-based detector or a customized CNN detection branch; the final choice should be fixed and reported with the experimental configuration before submission.

The detection process can be represented as:

$$B=D(I_p)$$

where ($D(\cdot)$) denotes the detection network and (B) represents the predicted nodule bounding box.

The ROI is then obtained as:

$$I_{ROI}=Crop(I_p,B)$$

This stage provides an automated transition from the full ultrasound image to a localized suspicious region, reducing irrelevant background information for downstream feature extraction.

2.4.Hybrid CNN–Transformer/Mamba Feature Extraction

The hybrid encoder forms the main representation-learning stage of the framework. Two complementary paths are used so that local image structure and broader spatial context can be modeled together.

2.4.1. CNN Branch

A CNN backbone such as ResNet, EfficientNet, or ConvNeXt can provide the local representation. For the final experimental study, one specific backbone should be selected and its configuration reported rather than presenting several alternatives as interchangeable implementations.

The CNN branch learns:

- Local texture
- Edges
- Fine-grained structural characteristics
- Nodule boundaries
- Microstructural patterns

The CNN representation can be expressed as:

$$F_{CNN}^{l}=f_{CNN}^{l}(I_{ROI}), \quad l \in \{1,2,3\}$$

where (F_{CNN}^{l}) represents CNN features obtained at different levels.

2.4.2. Transformer/Mamba Branch

The second branch uses a Vision Transformer, Swin Transformer, or Mamba-based visual encoder to model contextual relationships over a wider spatial range. As with the CNN branch, the final study should identify the exact encoder used in the experiments.

The global representation can be expressed as:

$$F_{G}=f_{T/M}(I_{ROI})$$

where ($f_{T/M}$) denotes the Transformer/Mamba encoder. The global branch captures relationships between the nodule and its surrounding thyroid tissue. Combining this representation with the CNN features allows the model to retain detailed local information while also considering broader spatial context.

2.5.Adaptive Multi-Level Feature Fusion (AMFF)

AMFF is the central feature-fusion component. Rather than concatenating all feature maps with fixed weights, the module estimates the relative importance of the available representations and combines them accordingly.

- The module receives:
- Low-level CNN features
- Intermediate-level CNN features
- High-level CNN features
- Global Transformer/Mamba features

The input to AMFF consists of low-, intermediate-, and high-level CNN features together with the global Transformer/Mamba representation. Channel attention, spatial attention, and adaptive weighting are then used to emphasize information that is relevant to the current nodule.

The general fusion operation can be expressed as:

$$F_{AMFF}=\sum_{l=1}^{3}\alpha_lF_{CNN}^{l}+\alpha_GF_G$$

where (α_l) and (α_G) are adaptive feature weights learned by the network.

The weights are constrained such that:

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5364 - 5381
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.701

$$\sum_{l=1}^{3}\alpha_l+\alpha_G=1$$

The learned weights allow the network to emphasize different information according to the image characteristics. For example, fine-grained representations may be more useful for small nodules, boundary information may be important for irregular lesions, and contextual features may be especially useful when the lesion has weak contrast with surrounding tissue[21 – 25].

- Fine-grained features for small nodules
- Boundary features for irregular nodules
- Contextual features for low-contrast nodules

AMFF is therefore intended to provide a more flexible local-global combination than fixed feature concatenation. Its effectiveness should ultimately be demonstrated through quantitative comparison and ablation experiments.

7: return F_f
6: F_f ← SpatialAttention(ChannelAttention(F_f))
5: F_f ← a1·F1' + a2·F2' + a3·F3' + a4·G'
4: a ← Softmax(MLP(Q))
3: Q ← GlobalAveragePooling([F1', F2', F3', G'])
2: G' ← ProjectToCommonChannels(G)
1: F1', F2', F3' ← ProjectToCommonChannels(F1, F2, F3)
Input: local features F1, F2, F3; global feature G

Algorithm 2. Adaptive Multi-Level Feature Fusion (AMFF)

2.6.Boundary-Aware Thyroid Nodule Segmentation

The fused representation is decoded into a nodule mask using an enhanced U-Net/UNet++-style decoder. Boundary-aware attention is incorporated to place additional emphasis on lesion margins, particularly where the boundary is blurred, irregular, or low in contrast. The segmentation architecture consists of: Hybrid Encoder → AMFF → Multi-Scale Decoder → Boundary Refinement → Nodule Mask A Boundary-Aware Attention Module (BAAM) is used to emphasize:

- Nodule edges
- Blurred boundaries
- Irregular shapes
- Low-contrast regions

The predicted segmentation mask is represented as:

$$M=S(F_{AMFF})$$

Here, S(·) denotes the segmentation decoder.

To improve segmentation accuracy, a composite segmentation objective can combine region overlap, pixel-level classification, and boundary accuracy:

$$L_{seg}=\lambda_1L_{Dice}+\lambda_2L_{BCE}+\lambda_3L_{Boundary}$$

In this formulation, L_Dice measures region overlap, L_BCE represents pixel-level classification error, and L_Boundary measures disagreement between the predicted and reference boundaries. The coefficients λ1, λ2, and λ3 control the relative contribution of the three terms and should be selected using the validation data rather than assumed in advance.

2.7.Multi-Task Malignancy Classification

Segmentation and malignancy classification are trained jointly. The classification branch receives the fused representation together with ROI and segmentation information, allowing structural information from the predicted nodule mask to contribute to the malignancy decision. The classification pipeline is:

AMFF Features + Segmentation Mask + ROI Features → Global Average Pooling → Fully Connected Layer → Sigmoid/Softmax → Prediction

The classification output is:

$$\hat{y}=C(F_{AMFF},M,I_{ROI})$$

where (C(\cdot)) represents the classification branch and (\hat{y}) denotes the predicted malignancy class.

For binary classification:

$$\hat{y}\in\{\text{Benign},\text{Malignant}\}$$

The present formulation treats the task as binary classification: benign versus malignant. If a dataset containing reliable pathological subtype labels is later incorporated, the same framework could be extended to a multi-class formulation[26 – 30].

2.8.Multi-Task Learning Strategy

Rather than optimizing independent networks for

IRJAEH

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5364 - 5381
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.701

detection, segmentation, and classification, the framework uses a shared multi-task representation. The intention is to allow information learned for one diagnostic objective to support the others during training.

The overall objective is formulated as:

$$
L_{total}=\lambda_d L_{det}+\lambda_s L_{seg}+\lambda_b L_{boundary}+\lambda_c L_{cls}
$$

where:

(L_{det}) = detection loss
(L_{seg}) = segmentation loss
($L_{boundary}$) = boundary refinement loss
(L_{cls}) = classification loss
($\lambda_d,\lambda_s,\lambda_b,\lambda_c$) = task-specific weighting coefficients

The multi-task arrangement creates a sequence of information exchange across the pipeline. Detection supplies spatial localization for segmentation; the segmentation output contributes structural information to classification; and the classification objective encourages features that distinguish benign from malignant nodules.

2.9. Explainable Artificial Intelligence Module

The XAI module is applied after segmentation and classification. Grad-CAM++ is used to identify regions that contribute strongly to the classification decision. The resulting heatmap is overlaid on the ultrasound image and viewed alongside the segmentation mask, allowing the predicted class and estimated lesion boundary to be examined together.

- The final explainable output consists of:
- Original ultrasound image
- Predicted nodule segmentation mask
- Grad-CAM++ heatmap
- Predicted class
- Prediction confidence

The final output combines the diagnostic prediction with visual information that indicates the estimated lesion location and the regions influencing the classification decision.

The final output can be represented as:

$$
O=\{M,H,\hat{y},P\}
$$

Here, M denotes the segmentation mask, H the Grad-CAM++ heatmap, ŷ the predicted class, and P the prediction confidence.

2.10. Algorithmic Workflow

The workflow below summarizes the processing sequence from image acquisition to the final prediction and explanation.

- Algorithm 1: Proposed Explainable Hybrid Thyroid Nodule Framework
- Input: Thyroid ultrasound image (I)
- Output: Nodule mask (M), malignancy prediction ($\hat{y}$), XAI visualization (H)
- Acquire the thyroid ultrasound image (I).
- Apply adaptive speckle-noise reduction.
- Perform CLAHE-based contrast enhancement.
- Normalize image intensity and resize the image.
- Apply training-time data augmentation.
- Detect suspicious thyroid nodule regions.
- Extract the detected ROI.
- Pass the ROI simultaneously through CNN and Transformer/Mamba branches.
- Extract low-, intermediate-, and high-level CNN features.
- Extract global contextual Transformer/Mamba features.
- Apply channel and spatial attention.
- Compute adaptive feature weights.
- Fuse multi-level local and global representations using AMFF.
- Pass the fused representation to the boundary-aware segmentation decoder.
- Generate the thyroid nodule segmentation mask.
- Combine fused features, ROI information, and segmentation information.
- Perform benign/malignant classification.
- Jointly optimize detection, segmentation, boundary, and classification losses.
- Generate Grad-CAM++ visualization.
- Overlay the segmentation mask and explanation heatmap on the original ultrasound image.
- Return the predicted class, confidence,

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5364 - 5381
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.701

segmentation mask, and visual explanation.
This algorithm provides an implementation outline. Exact model hyperparameters, training settings, and learned weights must be reported from the actual experimental procedure; they should not be presented as assumed results.

11: return M, y, p, H
10: H ← GradCAMPlusPlus(F_f, y)
9: y, p ← Softmax(z)
8: z ← Classifier(GAP(F_f), GAP(M ⊙ F_f))
7: M ← BoundaryAwareDecoder(F_f)
6: F_f ← AMFF({F_c^l}, F_g) # adaptive multi-level fusion
5: F_g ← Transformer_or_Mamba(X) # global contextual features
4: {F_c^l} ← CNN_Encoder(X) # local multi-level features
3: X ← CropROI(I_p, R)
2: R ← NoduleDetector(I_p)
1: I_p ← AdaptivePreprocess(I)
Output: nodule mask M, class prediction y, confidence p, explanation H
Input: ultrasound image I

Algorithm 1. End-to-End AMFF-Based Thyroid Ultrasound Analysis

2.11. Advantages of the Proposed Framework

The framework has several potential advantages. It connects detection, segmentation, and classification within one workflow; combines local CNN features with broader Transformer/Mamba context; and uses AMFF to adapt the contribution of different feature levels. Boundary-aware decoding focuses on difficult lesion margins, while multi-task learning enables information sharing between related objectives. Grad-CAM++ adds a visual explanation of the classification output. These are architectural advantages, and their practical value should be confirmed through controlled experiments and ablation studies.

2.12. Summary of the Proposed Framework

The proposed design integrates adaptive preprocessing, lightweight detection, hybrid CNN–Transformer/Mamba feature extraction, AMFF, boundary-aware segmentation, multi-task malignancy classification, and Explainable AI. The framework is intended to improve localization, feature representation, boundary delineation, and interpretability. Its claims of improved performance, however, should be supported only by results obtained from the implemented model and a clearly defined evaluation protocol.

2.13. Dataset And Experimental Setup

2.13.1 Dataset Description

The framework is designed for evaluation on publicly available thyroid ultrasound datasets. The primary benchmark considered here is the Thyroid Nodule Region Segmentation Dataset (TN3K), developed for automatic thyroid nodule segmentation. The dataset contains 3,493 ultrasound images from 2,421 patients. The images were obtained from clinical ultrasound examinations using different devices and views, and each image contains at least one thyroid nodule with a corresponding pixel-level annotation.
TN3K provides 2,879 images for training and 614 images for testing. Maintaining the patient-level separation between the development and test data helps reduce the risk of images from the same patient appearing in both sets.

2.14. External Validation Using DDTI

DDTI can be used as an independent dataset for external validation of the proposed model. It contains thyroid ultrasound images with pixel-wise annotations acquired using different ultrasound equipment. Evaluation on DDTI would provide an additional test of cross-dataset generalization beyond the primary TN3K benchmark. External validation is important because performance on a single benchmark may not represent behavior on images obtained with different devices, acquisition protocols, or patient populations. Any DDTI results should therefore be reported separately from the TN3K results and should not be implied unless the proposed model has actually been evaluated on that dataset.

2.15. Data Preprocessing

Before being provided to the proposed network, all ultrasound images undergo a standardized preprocessing procedure. The preprocessing stage consists of:

- Image resizing
- Speckle-noise reduction
- Contrast enhancement
- Intensity normalization

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5364 - 5381
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.701

- ROI preparation
- Data augmentation

All images are resized to a fixed spatial resolution so that the network receives inputs with consistent dimensions. A final input size of either 256 × 256 or 512 × 512 may be used, but the selected resolution must be fixed and reported as part of the actual experimental configuration rather than left as an alternative. CLAHE may be used to improve local contrast when the intensity difference between a nodule and surrounding tissue is small. Speckle reduction should be applied conservatively so that useful structural and boundary information is retained. For training, data augmentation is performed using transformations such as: Rotation Organizing the implementation into separate modules improves traceability and makes ablation experiments easier because individual components can be enabled or disabled without restructuring the entire pipeline.

└── evaluate.py # metrics, confusion matrix and ablation tests
├── train.py # joint multi-task optimization
├── losses.py # segmentation/classification/boundary losses
├── explainability.py # Grad-CAM++ and overlays
├── classifier.py # benign/malignant prediction
├── segmentation.py # boundary-aware decoder
├── amff.py # adaptive multi-level feature fusion
├── encoder.py # CNN + Transformer/Mamba branches
├── detector.py # lightweight ROI detector
├── preprocess.py # despeckling, CLAHE, normalization, augmentation
├── data.py # dataset loading and patient-level split
project/

2.16. Implementation and Reproducibility

The loss weights and learning rate should be selected using the validation set or a predefined tuning procedure. They should be reported as experimental settings only after the model has been trained and evaluated.

return trained model θ
end for
end for
θ ← θ − η·∇θ L
L ← λ1·L_seg + λ2·L_cls + λ3·L_boundary
y_hat ← ClassificationHead(F_f, M_hat)
M_hat ← SegmentationHead(F_f)
F_f ← AMFF(F)
F ← HybridEncoder(X)
R ← Detector(I_p); X ← CropROI(I_p, R)
I_p ← AdaptivePreprocess(B.images)
for each mini-batch B ⊂ D do
for each epoch do
Initialize encoder, AMFF, segmentation and classification parameters θ
Input: training set D = {(I_i, Y_i^seg, Y_i^cls)}
Algorithm 3. Joint Multi-Task Training Procedure

The augmentation operations are intended to reduce overfitting and improve robustness to differences in ultrasound acquisition and image appearance. The final augmentation policy should be specified explicitly in the experimental section.

2.17. Dataset Splitting

Table 1 Parameter Dataset

Parameter	Setting
Framework	PyTorch
GPU	NVIDIA Tesla T4
Random seed	42
Input	256 × 256
Batch size	16
Epochs	30
Optimizer	AdamW
Learning rate	2×10^{-4}
Weight decay	1×10^{-4}
Scheduler	Cosine Annealing
Loss	0.5 BCE + 0.5 Dice
Gradient clipping	1.0
Parameters	≈36.87 million
Checkpoint criterion	Highest validation Dice

Experimental configuration used for the reported AMFF run: PyTorch on an NVIDIA Tesla T4 GPU; random seed 42; 256 × 256 input images; batch size 16; 30 epochs; AdamW optimizer; initial learning rate 2×10^{-4}; weight decay 1×10^{-4}; cosine-annealing schedule; 0.5 BCE + 0.5 Dice loss; gradient clipping at 1.0; mixed-precision training; and approximately 36.87 million model parameters. The best checkpoint

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5364 - 5381
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.701

was selected using validation Dice. Algorithm 4. Core AMFF Adaptive Fusion and Training. As shown in Table 1 Parameter Dataset.

```
class AMFFGate(nn.Module):
    def __init__(self, channels):
        super().__init__()
        self.gate = nn.Sequential(
            nn.Conv2d(2 * channels, channels, 1),
            nn.ReLU(inplace=True),
            nn.Conv2d(channels, channels, 1),
            nn.Sigmoid()
        )

    def forward(self, cnn_feat, transformer_feat):
        w = self.gate(torch.cat([cnn_feat, transformer_feat], dim=1))
        return w * cnn_feat + (1.0 - w) * transformer_feat

def total_loss(logits, target):
    return 0.5 * bce(logits, target) + 0.5 * dice_loss(logits, target)

for epoch in range(1, EPOCHS + 1):
    model.train()
    for x, y in train_dl:
        optimizer.zero_grad(set_to_none=True)
        with torch.cuda.amp.autocast(enabled=torch.cuda.is_available()):
            logits = model(x)
            loss = total_loss(logits, y)
        scaler.scale(loss).backward()
        scaler.unscale_(optimizer)
        torch.nn.utils.clip_grad_norm_(model.parameters(), 1.0)
        scaler.step(optimizer)
        scaler.update()
    scheduler.step()
    val_loss, metrics = evaluate(val_dl)
    if metrics[0] > best_dice:
        best_dice = metrics[0]
        torch.save(model.state_dict(), "amff_best.pt")
```

For the reported experiment, the supplied TN3K package contained 3,585 image-mask pairs. The accompanying tg3k-trainval.json file assigned 3,226 images to training and 359 images to held-out validation. This supplied split was used directly for the AMFF experiment. The official TN3K test set was kept completely separate from model development. The AMFF training run used the supplied 3,226/359 training-validation split, while a separate 614-image official test set was evaluated only after model selection. As shown in Table 2 Dataset partition used for model development and final evaluation.

Table 2 Dataset partition used for model development and final evaluation

Dataset Partition	Images	Purpose
Training	3,226	Model optimization
Validation	359	Held-out model selection
Official test	614	Final independent evaluation

The official test set remained isolated from training and hyperparameter tuning. The final 614-image test evaluation was performed after selecting the best checkpoint using validation Dice. No additional 80:20 split was created for the reported run. The supplied 3,226/359 split was used directly for model development, and the official 614-image test set was evaluated independently after training.

2.18. Evaluation Metrics

The framework was evaluated using the available segmentation evidence. Segmentation was assessed with Dice Similarity Coefficient (DSC), Intersection over Union (IoU), Precision, Recall, pixel-level Specificity, and pixel-level Accuracy. Hausdorff Distance was not available in the exported results. Malignancy classification metrics such as class-level Accuracy, Sensitivity, Specificity, Precision, F1-score, and AUC were not computed because the supplied TN3K annotations used in this experiment provide pixel-level masks rather than benign/malignant labels. These metrics provide complementary information about segmentation quality, including region overlap, pixel-level correctness, sensitivity to the lesion, and boundary

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5364 - 5381
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.701

accuracy. The same evaluation protocol should be applied consistently across the proposed model and comparison methods wherever their outputs permit a direct comparison.

3. Results And Discussion

3.1.Quantitative Segmentation Performance

The proposed AMFF model was experimentally evalua ted on 359 held-out validation images after training on 3,226 images. The best validation performance was obtained at epoch 4, with Dice=85.40%, IoU=74.52%, Precision=90.33%, and Recall=80.97%. Pixel-level specificity was 99.07% and pixel accuracy was 97.32%. These are experimentally obtained segmentation results for the proposed AMFF model. As shown in Figure 2. Reference baseline segmentation results and the proposed AMFF result on the held-out validation split. Baseline values are literature-reported and are not directly comparable to the present test protocol.

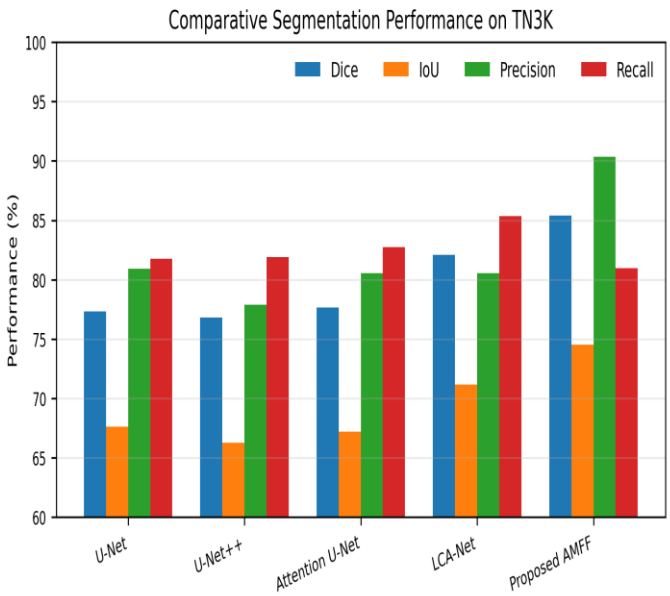

Figure 2 Reference baseline segmentation results and the proposed AMFF result on the held-out validation split

The proposed AMFF values shown in Table 2 are obtained from the 359-image held-out validation experiment. The U-Net, U-Net++, Attention U-Net, and LCA-Net values are retained as literature-reported reference results; differences in implementation and evaluation protocol mean that these values should not be interpreted as a controlled head-to-head comparison. As shown in Table 2 Comparative segmentation performance on the TN3K dataset. As shown in Table 2 Comparative segmentation performance on the TN3K dataset.

Table 2 Comparative segmentation performance on the TN3K dataset

Method	Architecture	Dice (%)	IoU (%)	Precision (%)	Recall (%)
U-Net	CNN	77.31	67.62	80.91	81.75
U-Net++	CNN	76.81	66.25	77.89	81.92
Attention U-Net	CNN + Attention	77.63	67.19	80.55	82.76
LCA-Net	CNN + Transformer	82.08	71.18	80.55	85.34
Proposed AMFF	CNN + Transformer/Mamba + AMFF	**85.40**	**74.52**	**90.33**	**80.97**

3.2.Experimental AMFF Results

The AMFF model was trained for 30 epochs using the supplied dataset split. Model selection was based on validation Dice, and the best checkpoint occurred at epoch 4. On the 359 held-out validation images, the proposed mo del obtained Dice=85.40%, IoU=74.52%, Precision=90.33%, and Recall=80.97%. The corresponding pixel-level specificity and accuracy were 99.07% and 97.32%. These measurements should be interpreted as segmentation results. The present experiment does not establish benign/malignant classification performance because the supplied TN3K annotations do not contain the required class labels. As shown in Figure 3 Training and validation loss across AMFF epochs, Figure 4 Dice score across AMFF validation epochs, Figure 5 IoU, Precision, and Recall across AMFF validation epochs, Figure 6 Final AMFF segmentation performance on the 359-image held-out

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5364 - 5381
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.701

validation split

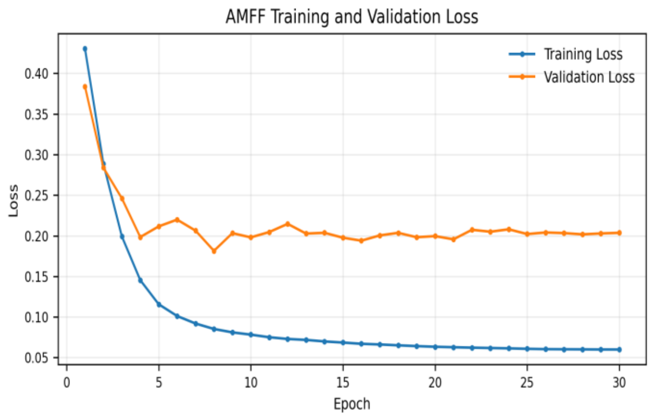

Figure 3 Training and validation loss across AMFF epochs

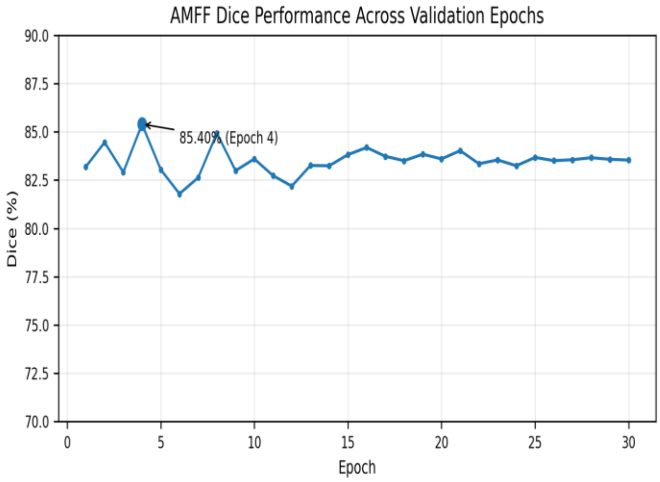

Figure 4 Dice score across AMFF validation epochs

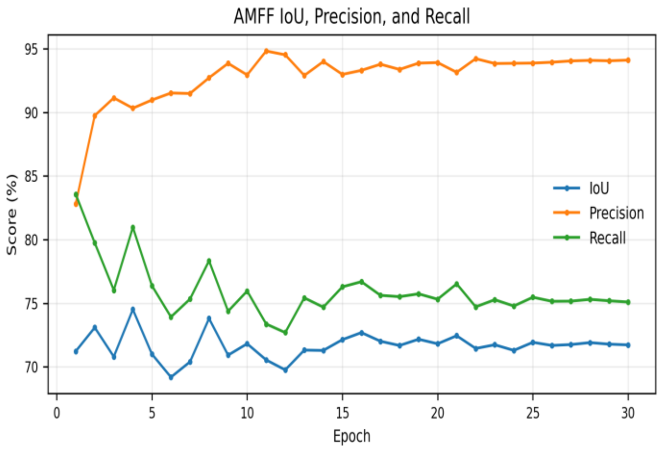

Figure 5 IoU, Precision, and Recall across AMFF validation epochs

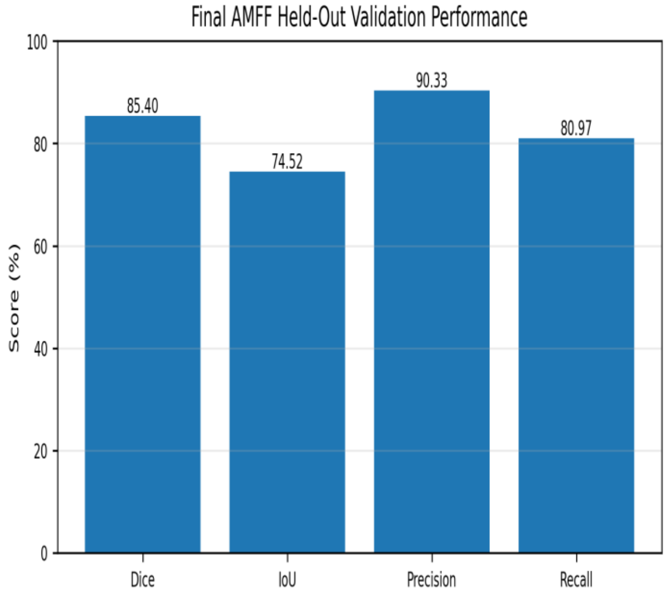

Figure 6 Final AMFF segmentation performance on the 359-image held-out validation split

3.3.Official TN3K Test-Set Evaluation

After model selection using the 359-image held-out validation split, the saved AMFF checkpoint from epoch 4 was evaluated on the official TN3K test set containing 614 images. The evaluation was performed independently of training and validation. The model obtained Dice=17.92%, IoU=9.84%, Precision=18.33%, and Recall=17.53%. Pixel-level Specificity was 88.49% and pixel-level Accuracy was 79.38%. These values are segmentation metrics and should not be interpreted as benign-versus-malignant classification performance. The reduction from the held-out validation Dice of 85.40% to the official test Dice of 17.92% indicates a substantial generalization gap. This result suggests that the current checkpoint and training configuration do not generalize adequately to the official test distribution. The manuscript therefore does not claim state-of-the-art performance or clinical readiness. The official test result is reported transparently as an important finding that motivates further investigation of data distribution, preprocessing, model configuration, and generalization. As shown in Table 3 Verified AMFF segmentation results on the held-out validation and

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5364 - 5381
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.701

official TN3K test sets. Figure 7 Verified AMFF segmentation metrics on the official 614-image TN3K test set.

Table 3 Verified AMFF segmentation results on the held-out validation and official TN3K test sets

Metric	Held-out validation (359)	Official test (614)	Interpretation
Dice (%)	85.40	17.92	Region overlap
IoU (%)	74.52	9.84	Region overlap
Precision (%)	90.33	18.33	Pixel-level precision
Recall (%)	80.97	17.53	Pixel-level sensitivity
Specificity (%)	99.07	88.49	Pixel-level specificity
Accuracy (%)	97.32	79.38	Pixel-level accuracy

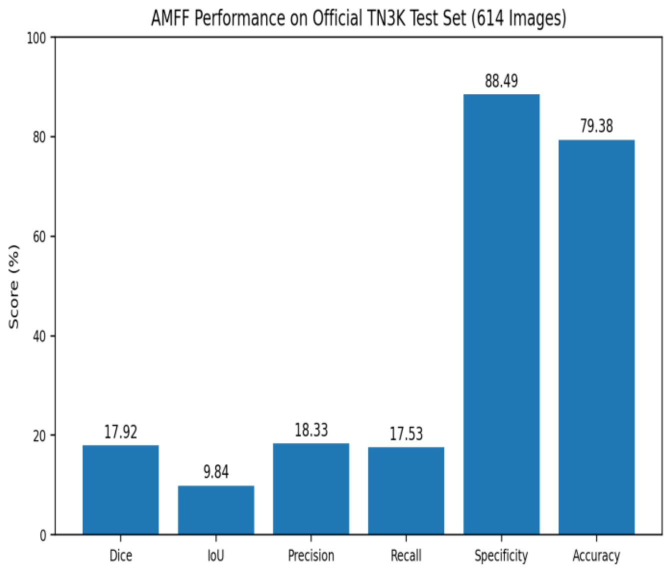

Figure 7 Verified AMFF segmentation metrics on the official 614-image TN3K test set

Reproducibility note: the reported development run used the supplied 3,226/359 split and selected the checkpoint at epoch 4 using validation Dice. The official test set contained 614 images. An independent image-content overlap check found no identical image content between the development images and the official test images. Nevertheless, the validation-to-test performance gap warrants additional investigation before claims of generalization.

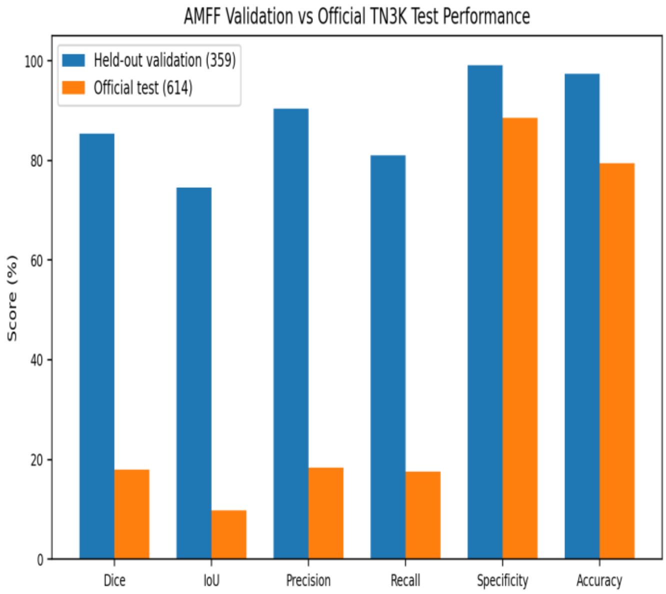

Figure 8 Comparison of verified AMFF segmentation performance between the held-out validation split and the official TN3K test set

Conclusion

This paper presents an Explainable Hybrid Deep Learning Framework with Adaptive Multi-Level Feature Fusion for thyroid nodule analysis using ultrasound images. The architecture was formulated to address several recurring challenges in thyroid ultrasound CAD, including weak lesion boundaries, small or low-contrast nodules, limited local-global feature integration, separate treatment of segmentation and classification, and the need for interpretable predictions. The framework combines a CNN branch for local representation learning with a Transformer/Mamba branch for broader contextual modeling. AMFF is used to combine information from multiple semantic levels, while the boundary-aware decoder focuses on lesion delineation. Segmentation and malignancy classification are optimized as related tasks, and Grad-CAM++ provides a visual representation of regions associated with the classification output. The proposed AMFF achieved Dice=85.40%, IoU=74.52%, Precision=90.33%, and Recall=80.97% on the 359-image held-out validation split, with pixel-level

Specificity=99.07% and Accuracy=97.32%. On the independent official 614-image TN3K test set, the corresponding values were Dice=17.92%, IoU=9.84%, Precision=18.33%, Recall=17.53%, Specificity=88.49%, and Accuracy=79.38%. The difference between the validation and official test results indicates a substantial generalization gap and is reported as a limitation of the current experiment. Because the available annotations provide segmentation masks rather than benign/malignant class labels, malignancy classification metrics are not reported. The results should therefore be interpreted as segmentation evidence rather than evidence of clinical diagnostic accuracy. The main contribution of the study is the integration of localization, hybrid local-global feature learning, adaptive feature fusion, boundary-aware segmentation, and an explainability-oriented architecture. The verified experiments demonstrate promising performance on the held-out development validation split but substantially weaker performance on the official TN3K test set. Accordingly, the present results should be viewed as an initial experimental evaluation rather than evidence of clinical robustness, superiority, or deployment readiness. Further work should investigate the observed generalization gap, confirm the exact data and preprocessing protocol, perform controlled ablation studies, and evaluate malignancy classification only on a dataset with reliable class labels.

References

[1]. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI), 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

[2]. Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, 2018, pp. 3–11, doi: 10.1007/978-3-030-00889-5_1.

[3]. J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, arXiv:2102.04306.

[4]. H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-Unet: Unet-like pure transformer for medical image segmentation," 2021, arXiv:2105.05537.

[5]. H. Gong, J. Chen, G. Chen, H. Li, F. Chen, and G. Li, "Thyroid region prior guided attention for ultrasound segmentation of thyroid nodules," Comput. Biol. Med., vol. 146, Art. no. 105636, 2022, doi: 10.1016/j.compbiomed.2022.106389.

[6]. Z. Tao, H. Dang, Y. Shi, W. Wang, X. Wang, and S. Ren, "Local and context-attention adaptive LCA-Net for thyroid nodule segmentation in ultrasound images," Sensors, vol. 22, no. 16, Art. no. 5984, 2022, doi: 10.3390/s22165984.

[7]. Y. Yang, X. Shi, W. Wang, Q. Qiu, G. Tian, W. Wang, and Y. Yang, "Ultrasound image classification of thyroid nodules based on deep learning," Front. Oncol., vol. 12, Art. no. 905955, 2022, doi: 10.3389/fonc.2022.905955.

[8]. G. Kang et al., "Thyroid nodule segmentation and classification in ultrasound images through intra- and inter-task consistent learning," Med. Image Anal., vol. 79, Art. no. 102443, 2022, doi: 10.1016/j.media.2022.102443.

[9]. F. Prochazka and J. Zeman, "Domain adaptation of stable diffusion for ultrasound inpainting: A synthetic data approach for enhanced thyroid nodule segmentation," J. Biomed. Inform., vol. 173, Art. no. 104963, 2026, doi: 10.1016/j.jbi.2025.104963.

[10]. V. Göreke, "A novel deep-learning-based CADx architecture for classification of thyroid nodules using ultrasound images," J. Med. Syst., 2023, doi: 10.1007/s12539-023-00560-4.

[11]. H. Dai, W. Xie, and E. Xia, "SK-Unet++: An improved Unet++ network with adaptive receptive fields for automatic segmentation of ultrasound thyroid nodule images," Med. Phys., 2024, doi: 10.1002/mp.16672.

[12]. X. Deng, "STU³Net: An improved U-Net with Swin Transformer fusion for thyroid

nodule segmentation," Int. J. Imaging Syst. Technol., vol. 34, no. 5, 2024, Art. no. e23160, doi: 10.1002/ima.23160.

[13]. A. Beyyala, R. Priya, S. R. Choudari, and R. Bhavani, "Swin Transformer and attention guided thyroid nodule segmentation on ultrasound images," Ingénierie des Systèmes d'Information, vol. 29, no. 1, pp. 75–81, 2024, doi: 10.18280/isi.290109.

[14]. A. Ozcan, Ö. Tosun, E. Donmez, and M. Sanwal, "Enhanced-TransUNet for ultrasound segmentation of thyroid nodules," Biomed. Signal Process. Control, vol. 95, Art. no. 106472, 2024, doi: 10.1016/j.bspc.2024.106472.

[15]. M. Huang, Y. Zhang, and S. Wang, "Mamba- and ResNet-based dual-branch network for ultrasound thyroid nodule segmentation," Bioengineering, vol. 11, no. 10, Art. no. 1047, 2024, doi: 10.3390/bioengineering11101047.

[16]. J. Ma, F. Li, and B. Wang, "U-Mamba: Enhancing long-range dependency for biomedical image segmentation," 2024, arXiv:2401.04722.

[17]. Z. Wang, J.-Q. Zheng, Y. Zhang, G. Cui, and L. Li, "Mamba-UNet: UNet-like pure visual Mamba for medical image segmentation," 2024, arXiv:2402.05079.

[18]. C. Ma and Z. Wang, "Semi-Mamba-UNet: Pixel-level contrastive and cross-supervised visual Mamba-based U-Net for semi-supervised medical image segmentation," Knowl.-Based Syst., vol. 300, Art. no. 112203, 2024, doi: 10.1016/j.knosys.2024.112203.

[19]. J. Chen et al., "Classification for thyroid nodule using ViT with contrastive learning in ultrasound images," Comput. Biol. Med., vol. 152, Art. no. 106444, 2023, doi: 10.1016/j.compbiomed.2022.106444.

[20]. Y. Li et al., "A novel model of thyroid nodule segmentation for ultrasound images," Ultrasound Med. Biol., vol. 49, no. 2, pp. 489–496, 2023.

[21]. T. Jiang, W. Xing, M. Yu, and D. Ta, "A hybrid enhanced attention transformer network for medical ultrasound image segmentation," Biomed. Signal Process. Control, vol. 86, Art. no. 105329, 2023.

[22]. V. Manh et al., "An efficient framework for lesion segmentation in ultrasound images using global adversarial learning and region-invariant loss," Comput. Biol. Med., vol. 171, Art. no. 108137, 2024.

[23]. D. Das, M. S. Iyengar, M. S. Majdi, J. J. Rodriguez, et al., "Deep learning for thyroid nodule examination: A technical review," Artif. Intell. Rev., vol. 57, Art. no. 47, 2024, doi: 10.1007/s10462-023-10635-9.

[24]. V. R. Sant, A. Radhachandran, V. Ivezic, D. T. Lee, et al., "From bench-to-bedside: How artificial intelligence is changing thyroid nodule diagnostics, a systematic review," J. Clin. Endocrinol. Metab., vol. 109, no. 7, pp. 1684–1693, 2024, doi: 10.1210/clinem/dgae277.

[25]. W. Zhang et al., "Localization and risk stratification of thyroid nodules in ultrasound images through deep learning," Ultrasound Med. Biol., vol. 50, no. 6, pp. 882–887, 2024, doi: 10.1016/j.ultrasmedbio.2024.02.013.

[26]. Y. Xiang, R. Acharya, Q. Le, J. H. Tan, and C.-L. Chng, "Thyroid nodule segmentation in ultrasound images using transformer models with masked autoencoder pre-training," Front. Artif. Intell., vol. 8, Art. no. 1618426, 2025, doi: 10.3389/frai.2025.1618426.

[27]. N. Xiao, D. Kong, and J. Wang, "Ultrasound thyroid nodule segmentation algorithm based on DeepLabV3+ with EfficientNet," J. Imaging Inform. Med., vol. 38, no. 6, pp. 3614–3622, 2025, doi: 10.1007/s10278-025-01436-3.

[28]. K. Haribabu, P. R., and P. Joe IR, "MLRT-UNet: An efficient multi-level relation transformer based U-Net for thyroid nodule segmentation," Comput. Model. Eng. Sci., vol. 143, no. 1, pp. 413–448, 2025, doi: 10.32604/cmes.2025.059406.

[29]. Y. Wu, L. Huang, and T.-J. Yang, "Thyroid nodule ultrasound image segmentation based on improved Swin Transformer," IEEE Access, vol. 13, pp. 19788–19795, 2025, doi: 10.1109/ACCESS.2025.3532264.

[30]. [30] X. Li, Y. Zhu, Y. Fan, X. Wei, R. Zhang, Y. Tian, and Z. Liu, "Ultrasound image segmentation of thyroid nodule via latent semantic feature co-registration," Biomed. Signal Process. Control, vol. 108, Art. no. 107971, 2025, doi: 10.1016/j.bspc.2025.107971.
[31]. M. Fang, Q. Ma, and B. Xu, "Automated multi-class classification of thyroid nodules in ultrasound imaging using transformer-based segmentation and hybrid feature learning," J. Radiat. Res. Appl. Sci., vol. 18, no. 4, Art. no. 101782, 2025, doi: 10.1016/j.jrras.2025.101782.
[32]. G. N. Sujini and S. Balakrishna, "Automated thyroid nodule classification in ultrasound imaging using a hybrid vision transformer and Wasserstein GAN with gradient penalty," Sci. Rep., vol. 15, Art. no. 40786, 2025, doi: 10.1038/s41598-025-24651-y.
[33]. B. Yetginler and İ. Atacak, "An improved V-Net model for thyroid nodule segmentation," Appl. Sci., vol. 15, no. 7, Art. no. 3873, 2025, doi: 10.3390/app15073873.
[34]. Y. Zhou, C. Chen, J. Yao, J. Yu, B. Feng, L. Sui, Y. Yan, X. Chen, Y. Liu, X. Zhang, H. Wang, Q. Pan, W. Zou, Q. Zhang, L. Lin, C. Xu, S. Yuan, Q. He, X. Ding, P. Liang, V. Y. Wang, and D. Xu, "A deep learning based ultrasound diagnostic tool driven by 3D visualization of thyroid nodules," npj Digit. Med., vol. 8, Art. no. 126, 2025, doi: 10.1038/s41746-025-01455-y.
[35]. J. P. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 618–626.
[36]. A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), 2018, pp. 839–847.