

Deep Learning-Based Breast Ultrasound Lesion Segmentation Using Cross-Scale Attention and Edge-Aware Supervision

Sathish Parvatham

sathishphd23@gmail.com

SR University

K. Raj Kumar

SR University

V. Malathy

SR University

Article

Keywords: Breast ultrasonography, lesion segmentation, deep learning, attention U-Net, edge-aware learning, cross-scale attention, computer-aided diagnosis

Posted Date: October 26th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11419824/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Deep Learning-Based Breast Ultrasound Lesion Segmentation Using Cross-Scale Attention and Edge-Aware Supervision

Sathish Parvatham^{1*}, Dr. K. Raj Kumar² and Dr. V. Malathy³

¹ Research Scholar, Electronics and Communication Engineering,
SR University, Warangal, Telangana, India.
[e-mail: sathishphd23@gmail.com]

² Associate Professor, Electronics and Communication Engineering,
SR University, Warangal, Telangana, India.
[e-mail: rajkumar.k@sru.edu.in]

³ Associate Professor, Electronics and Communication Engineering,
SR University, Warangal, Telangana, India.
[e-mail: malathy.v@sru.edu.in]

Abstract

Background: Accurate segmentation of breast lesions in ultrasound is complicated by variations in lesion morphology, speckle noise, and indistinct margins.

Objective: To describe and assess a modified Attention U-Net combining spatial token mixing, squeeze-and-excitation attention, Polar/Log-Polar geometric features, large-kernel context aggregation, cross-scale attention, and auxiliary edge prediction.

Materials and Methods: The architecture was evaluated using the publicly available Breast Ultrasound Images Dataset (BUSI). The submitted manuscript does not specify the number of unique patients, patient-level data partitioning, or independent external validation; these details must be confirmed before submission.

Results: Reported test-set pixel accuracy was 0.9808 and specificity 0.9935, but Dice overlap was 0.3673, intersection-over-union (IoU) was 0.2250, precision was 0.4674, and recall was 0.3026. Table 1 reports ROC-AUC 0.9005, whereas the ROC figure reports 0.9479; this discrepancy requires verification against saved test predictions.

Conclusion: Although the architecture combines multiple context- and edge-sensitive mechanisms, the observed tumor overlap and sensitivity are insufficient to demonstrate reliable clinical lesion delineation. Controlled segmentation-baseline experiments, ablations, and independent validation are required before clinical claims can be made.

Keywords: Breast ultrasonography; lesion segmentation; deep learning; attention U-Net; edge-aware learning; cross-scale attention; computer-aided diagnosis.

1. Introduction

Breast cancer is considered to be one of the most common malignancies among females, and its early diagnosis and characterization are deemed to be critically important for further management. Modern medical imaging, such as breast ultrasound, along with other available radiological methods, has been widely applied for screening and surveillance of suspicious neoplasms [1]. In most cases, such procedures as delineation of the tumor region are essential for identifying the lesion characteristics, estimating its size, and providing the following diagnosis and treatment. However, as breast images usually contain large variety of different tumor appearances, shapes, and borders, the process of automated tumor segmentation may be considered as a challenging one.

Due to the rapid development of deep learning, convolutional neural network (CNN)-based segmentation methods have been widely used in bioinformatics and medical image analysis to extract clinically relevant tumor regions. The methods based on U-Net architecture are one of the most widely used due to their encoder and decoder, which allow for combining context information and spatial details through skip connections [2]. Several attention-based variants of the U-Net have been proposed in the literature to address this issue by assigning higher weights to informative image regions. However, accurate tumor segmentation in breast cancer imaging requires capturing both global context and local details with a particular emphasis on the delineation of irregular tumor boundaries.

Although the existing segmentation networks have made great progress, there are still some challenges in accurate tumor detection [3]. The tumors often demonstrate significant variability in size, intensity, shape, and internal structure; have indistinct boundaries; and may appear similar to normal tissue. At the same time, small tumors can be challenging to differentiate from surrounding tissues, while large tumors may lack specific features, such as distinct tumor boundaries at different scales [4]. Consequently, spatial information can be lost in the process of iterative down-sampling, and artifacts can be introduced during the reconstruction process from features, leading to incomplete tumor regions, false positives and poor localization accuracy.

The necessity for a better segmentation strategy is driven by the need to retain semantic information as well as fine boundary details during the entire segmentation pipeline. While attention mechanisms can improve the representation of salient areas, spatial attention alone might not be enough to recover lost boundary information at varying levels of feature resolution [5]. Furthermore, conventional supervision strategies generally focus on the final segmentation output, giving little supervision to intermediate feature representations about tumor edges. Incorporating multi-scale information and explicit edge-oriented supervision is necessary to learn discriminative contextual features and accurate boundary representations together [6].

In this case, the Modified Attention U-Net with Cross-Scale Edge-Aware Supervision can be considered a beneficial choice for breast cancer tumor segmentation since it allows for incorporating attention-based guidance into the feature refinement process and providing supervision of the boundary information at multiple feature scales. In such a way, with the modified attention mechanism, it becomes possible to pay specific attention to informative tumor features and ignore irrelevant background responses [7]. On the other hand, cross-scale edge-aware supervision can also be introduced to encourage the boundary representation at different levels so as to alleviate the loss of fine spatial resolution during feature encoding/decoding and reconstruction, and better delineation of tumor regions can be achieved by incorporating contextual information, attention guided representations and multi-scale edge features [8].

2. Literature

Mehwish Zafar et al. [9] proposed a U-Net-based encoder–decoder network with residual blocks and a Multi-Cascade Convolutional Block Attention Module (MC-CBAM) for breast lesion segmentation. The MC-CBAM integrated channel and multi-scale spatial attention using 1×1 , 3×3 , and 5×5 convolutions. The network consisted of four encoder and four decoder blocks with skip connections. A combined Tversky and Dice loss was used to improve

segmentation, particularly for small tumors. The framework was evaluated using ultrasound and histopathology datasets.

Rashid Nasimov et al. [10] proposed a Dual-Stream HyperFusionNet (DS-HFN) for nuclei segmentation by separately learning global contextual and boundary-sensitive features. Its dual-stream encoder extracts semantic and edge-focused features, which are fused using an attention-based HyperFeature Embedding Module (HFEM). A dual-decoder structure is incorporated to enhance segmentation accuracy. Additionally, a Gradient-Aligned Loss Function is employed to align predicted segmentation gradients with ground-truth contours. The proposed architecture is designed for robust nuclei segmentation across different digital pathology tissues. It was evaluated using benchmark datasets, including TNBC and MoNuSeg.

Tao Liu et al. [11] proposed ECCA-UNet, a Transformer-based model designed to address challenges in medical image segmentation caused by variable shapes, unclear boundaries, and complex backgrounds. CNN-based methods capture local features effectively but have limitations in modeling long-range dependencies, while Vision Transformers require high computational resources and may lose local details. ECCA-UNet incorporates the Cross-Shaped Window (CSWin) mechanism for efficient long-range dependency modeling. Squeeze-and-Excitation (SE) channel attention is used to enhance feature discrimination. An auxiliary edge-aware branch is introduced to preserve boundary information. The model combines spatial and channel attention with explicit edge supervision for accurate anatomical structure delineation.

Shivpratap Singh Kushwah et al. [12] proposed LTM-UNet to improve context-aware segmentation of ultrasound images. The model combines a Transformer-based encoder with state-space-based decoding within a U-Net framework. An efficient Vision Transformer encoder is used to extract multi-scale global representations. An attention-guided skip-fusion mechanism with spatial-channel attention is incorporated to preserve fine spatial details. A direction-aware state-space decoder is employed to capture long-range dependencies and reconstruct relevant features. The architecture is designed to reduce the semantic gap between encoder and decoder representations. Extensive experiments were conducted on benchmark ultrasound medical imaging datasets.

Tan Guo et al. [13] proposed SGLF-Net for polyp segmentation, addressing scale variation and ambiguous boundaries in colonoscopy images. The network uses a staged global-to-local strategy to refine segmentation from coarse semantics to fine details. A Swin Transformer Encoder and Dynamic Attentive Decoder (DAD) capture multi-scale contextual features. An Edge-aware Dynamic Attentive Decoder (E-DAD) improves boundary and structural details through edge supervision. The CSSAR module with hybrid attention enables effective cross-scale feature interaction between stages. Experiments were conducted on five public benchmark datasets to evaluate the proposed method.

Zhongzhou Zhang et al. [14] proposed GFNet, an edge-aware network for precise glioma segmentation. The method addresses inaccurate boundaries caused by insufficient edge and global information. It introduces a Dual-path Gradient-guided Training (DGT) strategy to focus on difficult boundary pixels during network optimization. A Fourier Edge-enhancement

Module (FEM) is also developed to enhance edge features using high-frequency spectral information. These components help capture both global context and detailed boundary information for improved segmentation. GFNet was evaluated on the BraTS2020 and Medical Segmentation Decathlon (MSD) glioma and prostate datasets.

Wenbin Yang et al. [15] proposed DGT-MSB-Net for liver tumor segmentation in CT images, addressing irregular tumor shapes, unclear boundaries, and background noise. The network incorporates a Dynamically Guided Transformer (DGT) to model complex tumor structures using deformable sampling and efficient channel modeling. A Multiscale Spatial-Channel Attention (MSCA) mechanism combines shallow boundary information with deep semantic features. Additionally, a Boundary-Aware (BA) module provides explicit supervision to refine tumor boundaries. The proposed framework was evaluated on the LiTS2017 and 3DIRCADb datasets.

Xiaoxuan Ma et al. [16] proposed MAMF-Net for medical image segmentation, addressing limitations in multi-scale feature extraction caused by fixed down-sampling and conventional convolutional encoders. The framework combines CNN and Transformer architectures to capture local features and global contextual information. An Adaptive Atrous Dilated Convolution (AADC) module aggregates features across multiple receptive fields. A Low-level Dual-Channel Attention Gating (LDCAG) mechanism enhances spatial and edge information while reducing background noise. A Multi-Scale Global Channel Attention (MSGCA) module fuses multi-level features with decoder semantics for progressive segmentation refinement. The proposed network was evaluated on the Synapse multi-organ and ACDC cardiac datasets.

3. Proposed Model

The proposed model is an improved Attention U-Net for breast lesion segmentation, which introduces the Token Pool Mixer and SE attention module to refine spatial interactions and boost feature selection in a channel-wise pattern. A Polar/Log-Polar feature branch is introduced to capture geometric information and improve lesion structure representation. The bottleneck is strengthened using a RepLK-SE block with large kernel convolution for extracting wider contextual features. Cross-Scale Attention and CBAM-based attention gates improve skip feature fusion and focus on important lesion regions. Additionally, an edge-aware auxiliary branch is incorporated to learn boundary information along with the final segmentation mask.

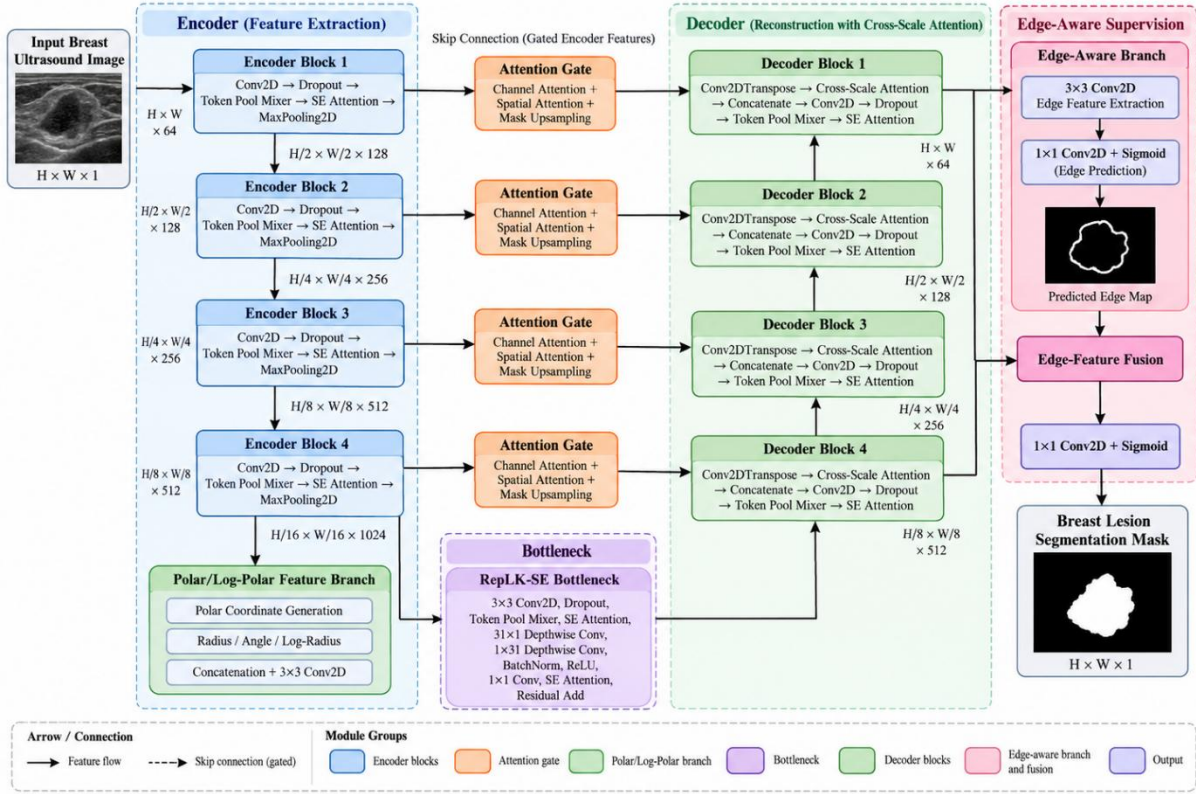


Figure 1: Architecture of Proposed Attention U-Net with Cross-Scale Edge Aware Supervision

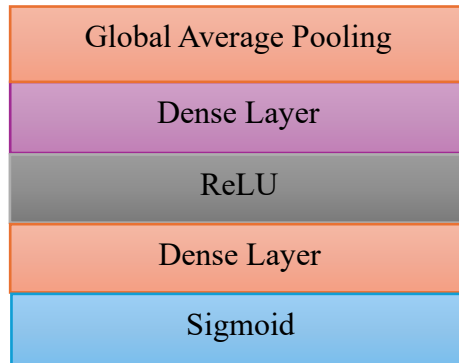


Figure 2(a): SE Attention Block

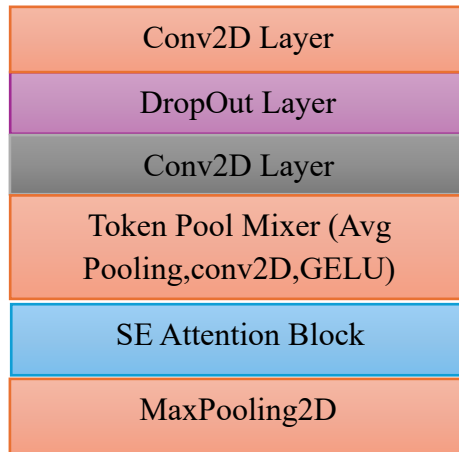


Figure 2 (b): Encoder Block

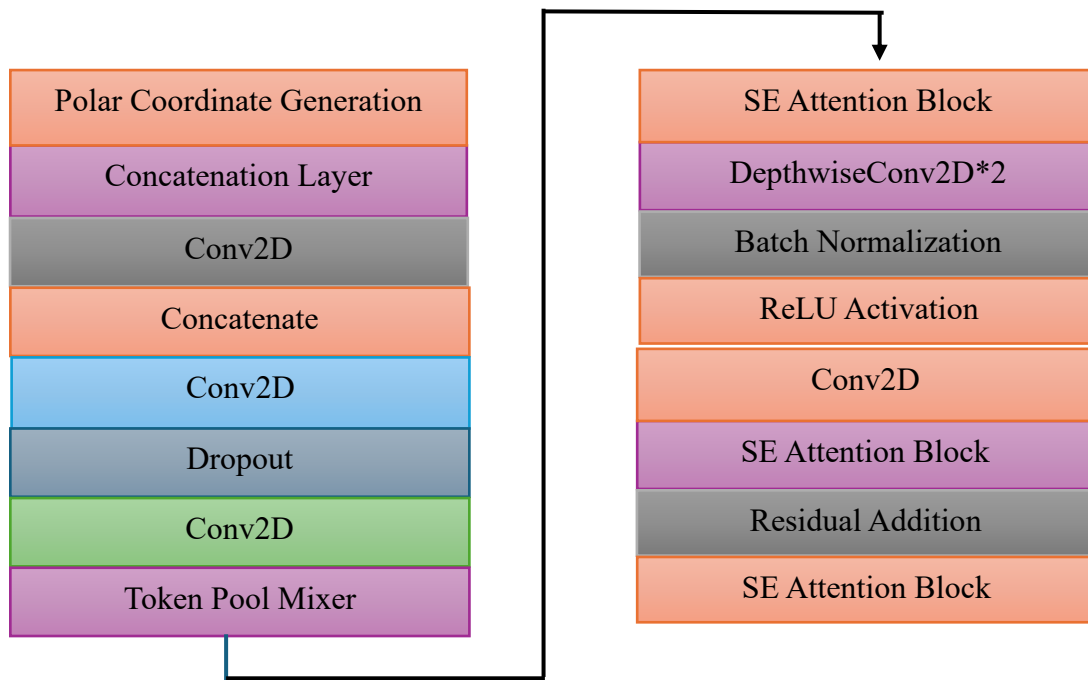


Figure 2(c): Attention Gate for Channel-Spatial Feature Refinement

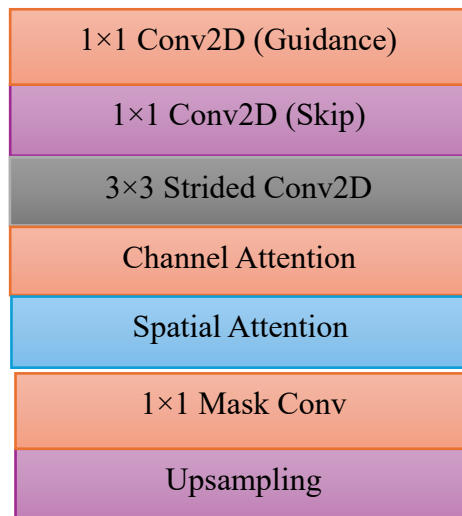


Figure 2 (d): Attention Gate Block

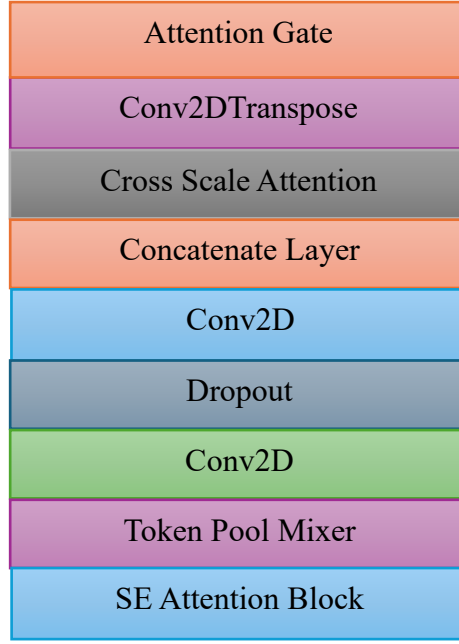


Figure 2(e): Decoder Block

Figure 2: Internal Layers of Proposed model Architecture

- **Input Layer**

The Input Layer receives the breast image and defines its spatial dimensions and number of channels. The image is directly provided to the first encoder block without an additional transformation inside the model. The initial feature representation is therefore the input image itself.

$$F_0 = X, X \in \mathbb{R}^{H \times W \times C} \quad (1)$$

where F_0 denotes the initial feature representation, X denotes the input breast image, H represents image height, W represents image width, and C represents the number of input channels.

- **Convolutional Feature Extraction Layer**

The Convolutional Feature Extraction Layer uses 3×3 Conv2D operations with ReLU activation to extract local spatial, texture, intensity, and structural information. The convolution applies learnable filters to the input feature representation and generates transformed feature maps. These features form the basic representation used by subsequent feature-enhancement mechanisms.

$$F_c = \text{ReLU}(W_c * F + b_c) \quad (2)$$

where F is the input feature map, W_c represents the learnable convolution kernel, $*$ denotes the convolution operation, b_c represents the convolution bias, and $\text{ReLU}(\cdot)$ is the Rectified Linear Unit activation function.

- **Dropout Layer**

The Dropout Layer regularizes the model by zeroing out some feature activations during the training stage. The binary random mask is applied to the features representation, and the non-zero ones are scaled up to have the same expected magnitude. This trick allows reducing the dependence of the subsequent model layers on the individual neurons during the training stage and hence reduce overfitting.

$$F_d = \frac{M \odot F}{1 - p}, M_i \sim \text{Bernoulli}(1 - p) \quad (3)$$

where F is the input feature map, F_d is the dropout output, M is the binary dropout mask, denotes $\odot$ element-wise multiplication, p is the dropout rate, and $\text{Bernoulli}(1 - p)$ determines whether an activation is retained.

- **Token Pool Mixer**

The Token Pool Mixer conducts the feature mixing operation using the 3×3 average pooling followed by subtracting the original feature representation. The resultant difference is transformed using the 1×1 convolution with GELU activation function and added back through the residual connection resulting in the spatial interaction.

$$\begin{aligned} P &= \text{AvgPool}_{3 \times 3}(F) \quad (4) \\ F_{\text{TPM}} &= F + \text{GELU}(W_{1 \times 1} * (P - F) + b) \quad (5) \end{aligned}$$

where P is the average-pooled feature representation, F is the original feature map, $P - F$ represents the pooled feature difference, $W_{1 \times 1}$ is the learnable 1×1 convolution kernel, b is the convolution bias, $\text{GELU}(\cdot)$ is the Gaussian Error Linear Unit activation, and F_{TPM} is the Token Pool Mixer output.

- **SE Channel Attention Layer**

The SE Channel Attention Layer can be seen as performing channel-wise recalibration by applying Global Average Pooling followed by two Dense operations. The first Dense layer reduces the channel dimensionality, while the second Dense layer recovers it and applies a Sigmoid activation to produce channel importance weights. These weights are multiplied with the input feature map to highlight informative channels.

$$z_c = \frac{1}{HW} \sum_{i=1}^H \sum_{j=1}^W F_{i,j,c} \quad (6)$$

$$\begin{aligned} a &= \sigma(W_2 \text{ReLU}(W_1 z + b_1) + b_2) \quad (7) \\ F_{SE} &= F \odot a \quad (8) \end{aligned}$$

where z_c is the global descriptor of channel c , W_1 and W_2 are the weights of the two Dense layers, b_1 and b_2 are their biases, a is the channel attention vector, $\sigma(\cdot)$ denotes the Sigmoid activation, and F_{SE} is the recalibrated feature representation.

- **Max Pooling Layer**

The Max Pooling Layer is responsible for downsampling the input images spatially by a factor of 2 in both dimensions using the 2×2 pooling window. The layer only lets the strongest activation from each pool through to the next layer, thus allowing the encoder part of the model to learn increasingly broad contextual representations at larger scales.

$$F_p(i, j, c) = \max_{(m,n) \in \mathcal{N}_{2 \times 2}} F(2i + m, 2j + n, c) \quad (9)$$

where F is the input feature map, F_p is the pooled feature map, i, j represent the output spatial coordinates, c denotes the channel index, and $\mathcal{N}_{2 \times 2}$ represents the 2×2 pooling neighborhood.

- **Polar Coordinate Generation Layer**

The Polar Coordinate Generation Layer creates normalized spatial coordinate grids corresponding to the dimensions of the deep encoder feature map. The implementation uses normalized coordinate values ranging from -1 to 1 , providing a spatial reference for geometric feature generation. These coordinates are subsequently used to derive radial, angular, and logarithmic radial information.

$$x, y \in [-1, 1] \quad (10)$$

where x represents the normalized horizontal coordinate and y represents the normalized vertical coordinate. The interval $[-1, 1]$ represents the normalized spatial coordinate range used in the implementation.

- **Radius, Angle and Log-Radius Generation Layer**

The Radius, Angle and Log-Radius Generation Layer derives three geometric descriptors from the normalized coordinate grids. Radial distance represents the distance from the coordinate origin, angular position represents orientation, and logarithmic radius provides an additional scale-sensitive geometric representation. These descriptors provide explicit geometric information to the feature-learning process.

$$r = \sqrt{x^2 + y^2} \quad (11)$$

$$\theta = \text{atan2}(y, x) \quad (12)$$

$$r_{\log} = \log(\max(r, \epsilon)) \quad (13)$$

where r is the radial distance, θ is the angular coordinate, $r_{\log}$ is the logarithmic radial coordinate, x and y are normalized spatial coordinates, atan2 calculates the angular position while preserving the correct quadrant, and ϵ is a small positive value used to avoid numerical instability.

- **Polar/Log-Polar Feature Fusion Layer**

The Polar/Log-Polar Feature Fusion Layer concatenates the original encoder feature with the generated radius, angle, and log-radius representations. The combined representation is processed using a 3×3 Conv2D with 64 filters and ReLU activation. The resulting geometry-enhanced feature is subsequently concatenated with the original deep encoder feature before entering the bottleneck.

$$F_g = \text{Concat}[F, r, \theta, r_{\log}] \quad (14)$$

$$F_{\text{polar}} = \text{ReLU}(W_{\text{polar}} * F_g + b_{\text{polar}}) \quad (15)$$

$$F_{\text{fused}} = \text{Concat}[p, 4F_{\text{polar}}] \quad (16)$$

where F_g is the geometry-augmented feature, F is the original encoder feature, $r, \theta, r_{\log}$ are the geometric descriptors, W_{polar} is the convolution kernel, b_{polar} is its bias, F_{polar} is the geometry-enhanced feature, and p represents the fourth encoder output supplied to the Polar branch.

- **Bottleneck Feature Encoding Layer**

The Bottleneck Feature Encoding Layer processes the fused representation using 3×3 Conv2D operations with 512 filters, dropout, Token Pool Mixer, and SE attention. The convolutional transformations progressively produce a high-level semantic representation at the deepest level of the network. This representation is then provided to the RepLK-SE refinement stage.

$$F_{b1} = \text{ReLU}(W_{b1} * F_{\text{fused}} + b_{b1}) \quad (17)$$

$$F_{b2} = \text{ReLU}(W_{b2} * \text{Dropout}(F_{b1}) + b_{b2}) \quad (18)$$

where F_{b1} and F_{b2} are intermediate bottleneck features, W_{b1} and W_{b2} are the corresponding convolution kernels, and b_{b1} and b_{b2} are their biases.

- **RepLK Large-Kernel Feature Extraction Layer**

The RepLK Large-Kernel Feature Extraction Layer uses sequential 31×1 and 1×31 DepthwiseConv2D operations to capture broad spatial dependencies. The factorized large-

kernel structure provides an expanded receptive field without directly applying a full 31×31 depthwise convolution. This allows the bottleneck to capture broader contextual and structural information.

$$F_{31 \times 1} = DWConv_{31 \times 1}(F) \quad (19)$$

$$F_{LK} = DWConv_{1 \times 31}(F_{31 \times 1}) \quad (20)$$

where $DWConv_{31 \times 1}$ represents the 31×1 depthwise convolution, $DWConv_{1 \times 31}$ represents the 1×31 depthwise convolution, $F_{31 \times 1}$ is the intermediate directional feature, and F_{LK} is the resulting large-kernel feature representation.

- **RepLK-SE Residual Refinement Layer**

The RepLK-SE Residual Refinement Layer applies Batch Normalization and ReLU activation to the large-kernel representation, followed by a 1×1 pointwise convolution and SE channel attention. The refined representation is then combined with the original feature through residual addition. This integrates broad spatial context with channel-wise feature recalibration.

$$F_r = \text{ReLU}(BN(F_{LK})) \quad (21)$$

$$F_p = W_{1 \times 1} * F_r + b_p \quad (22)$$

$$F_a = SE(F_p) \quad (23)$$

$$F_{\text{RepLK-SE}} = F + F_a \quad (24)$$

where F_r is the normalized and activated large-kernel feature, $BN(\cdot)$ denotes Batch Normalization, F_p is the pointwise-projected feature, F_a is the SE-refined representation, and $F_{\text{RepLK-SE}}$ is the residual output

- **Attention Gate Layer**

The Attention Gate uses the decoder feature as guidance to refine the corresponding encoder skip feature. The decoder feature and skip feature are projected into compatible feature spaces, while the skip representation is downsampled before being combined with the decoder guidance. The combined representation is subsequently processed by channel and spatial attention.

$$G = W_g * X + b_g \quad (25)$$

$$S_p = \text{ReLU}(W_s * S + b_s) \quad (26)$$

$$S_d = \text{ReLU}(W_d * S_p + b_d) \quad (27)$$

$$Z = G + S_d \quad (28)$$

where X is the decoder feature, S is the encoder skip feature, G is the projected decoder guidance, S_p is the projected skip feature, S_d is the downsampled skip representation, W_g, W_s, W_d are learnable convolution kernels, b_g, b_s, b_d are their biases, and Z is the combined Attention Gate representation.

- **Channel Attention Layer**

The Channel Attention Layer within the Attention Gate uses Global Average Pooling and Global Max Pooling to obtain complementary channel descriptors. These descriptors pass through shared Dense transformations and are combined before Sigmoid activation. The resulting channel attention weights are applied to the feature representation to emphasize informative channels.

$$z_{\text{avg}} = \text{GAP}(Z), z_{\text{max}} = \text{GMP}(Z) \quad (29)$$

$$A_c = \sigma \left(\text{MLP}(z_{\text{avg}}) + \text{MLP}(z_{\text{max}}) \right) \quad (30)$$

$$Z_c = Z \odot A_c \quad (31)$$

where z_{avg} and z_{max} are the global average and maximum channel descriptors, GAP and GMP denote Global Average Pooling and Global Max Pooling, $\text{MLP}(\cdot)$ represents the two Dense transformations, A_c is the channel attention vector, and Z_c is the channel-refined representation.

- **Spatial Attention Layer**

The Spatial Attention Layer determines the spatial importance of different feature locations after channel refinement. Channel-wise average and maximum activation maps are calculated and concatenated before being processed by a 7×7 convolution with Sigmoid activation. The resulting spatial attention map is applied to the feature representation.

$$M_{\text{avg}}(i, j) = \frac{1}{C} \sum_{c=1}^C Z_c(i, j, c) \quad (32)$$

$$M_{\text{max}}(i, j) = \max_c Z_c(i, j, c) \quad (33)$$

$$A_s = \sigma(W_{7 \times 7} * \text{Concat}[M_{\text{avg}}, M_{\text{max}}] + b_s) \quad (34)$$

$$Z_s = Z_c \odot A_s \quad (35)$$

where M_{avg} and M_{max} are the spatial average and maximum activation maps, C is the number of channels, $W_{7 \times 7}$ is the spatial convolution kernel, A_s is the spatial attention map, and Z_s is the spatially refined feature representation.

- **Attention Mask Generation Layer**

The Attention Mask Generation Layer converts the spatially refined feature representation into a single-channel spatial mask using a 1×1 Conv2D with Sigmoid activation. The resulting mask contains spatial importance values between zero and one. It is subsequently upsampled and applied to the original encoder skip feature.

$$M = \sigma(W_m * Z_s + b_m) \quad (36)$$

where Z_s is the spatially refined feature, W_m is the 1×1 mask-generation kernel, b_m is its bias, $\sigma(\cdot)$ is the Sigmoid activation, and M is the generated single-channel attention mask.

- **Attention-Gated Skip Connection Layer**

The generated attention mask is upsampled by a factor of two and applied element-wise to the original encoder skip feature. This selectively retains relevant spatial information and suppresses less useful responses before decoder fusion. When batch normalization is enabled, the gated feature is subsequently normalized.

$$M_u = \text{UpSample}_2(M) \quad (37)$$

$$S_g = S \odot M_u \quad (38)$$

$$S_{\text{gate}} = \text{BN}(S_g) \quad (39)$$

where M_u is the upsampled attention mask, UpSample_2 denotes two-times spatial upsampling, S is the original encoder skip feature, S_g is the gated skip feature, and S_{gate} is the normalized gated representation.

- **Transposed Convolution Upsampling Layer**

The Transposed Convolution Upsampling Layer uses a 3×3 Conv2DTranspose with stride two to increase the spatial resolution of the decoder representation. The operation performs learnable upsampling while simultaneously transforming the feature representation. It is used progressively throughout the decoder to restore the spatial resolution lost during encoder downsampling.

$$F_u = \text{ReLU}(W_T \star F + b_T) \quad (40)$$

where F is the decoder input feature, W_T is the transposed-convolution kernel, b_T is its bias, $\star$ denotes the transposed convolution operation, and F_u is the resulting upsampled decoder feature.

- **Query Projection Layer**

The Query Projection Layer applies a 1×1 Conv2D to the decoder feature to generate the query representation for Cross-Scale Attention. This projection transforms the decoder feature into

the attention representation space. The resulting query provides the information used to retrieve relevant contextual information from the encoder representation.

$$Q = W_q * F_u + b_q \quad (41)$$

where F_u is the upsampled decoder feature, W_q is the 1×1 query-projection kernel, b_q is its bias, and Q is the resulting query representation.

- **Key/Value Projection Layer**

The Key/Value Projection Layer applies a 1×1 Conv2D to the attention-gated skip feature to generate the representation used as both key and value in the Multi-Head Attention operation. Thus, the implementation uses the same projected representation for K and V . This provides encoder-derived contextual information for the decoder attention mechanism.

$$K = V = W_k * S_{\text{gate}} + b_k \quad (42)$$

where S_{gate} is the attention-gated encoder feature, W_k is the key/value projection kernel, b_k is its bias, K is the key representation, and V is the value representation.

- **Spatial Resolution Control Layer**

The Spatial Resolution Control Layer limits the spatial dimensions used by Cross-Scale Attention to a maximum dimension of 32. When the feature map exceeds this limit, the query and key/value representations are resized using bilinear interpolation before attention computation. This reduces the computational and memory requirements of spatial attention while preserving the overall feature structure.

$$(H_s, W_s) = f(H, W, \text{max_spatial} = 32) \quad (43)$$

$$Q_s = \text{Resize}_{\text{bilinear}}(Q, H_s, W_s) \quad (44)$$

$$K_s = \text{Resize}_{\text{bilinear}}(K, H_s, W_s) \quad (45)$$

where H, W are the original spatial dimensions, H_s, W_s are the controlled attention-grid dimensions, Q_s and K_s are the resized query and key representations, and $\text{max_spatial} = 32$ is the spatial limit specified in the implementation.

- **Cross-Scale Multi-Head Attention Layer**

The Cross-Scale Multi-Head Attention Layer establishes interaction between the decoder query and encoder-derived key/value representation. The implementation uses four attention heads, with each head computing scaled dot-product attention before the head outputs are combined. This allows the decoder to selectively retrieve relevant contextual information from the encoder feature.

$$A_h = \text{softmax}\left(\frac{Q_h K_h^T}{\sqrt{d_h}}\right) V_h \quad (46)$$

$$F_{\text{CSA}} = \text{MHA}(Q_s, K_s, V_s) \quad (47)$$

where Q_h, K_h, V_h are the query, key, and value representations for attention head h , d_h is the dimensionality of an attention head, A_h is the attention output of an individual head, and F_{CSA} is the combined Multi-Head Attention output. The implementation uses four attention heads.

- **Cross-Scale Residual Fusion Layer**

The cross-scale attention output is resized to the original decoder spatial dimensions and added to the original decoder feature through a residual connection. This preserves the original decoder representation while incorporating contextual information obtained from the encoder. The resulting feature is then used for multi-scale feature fusion.

$$F_{\text{att}} = \text{Resize}_{\text{bilinear}}(F_{\text{CSA}}, H, W) \quad (48)$$

$$F_{\text{CR}} = F_u + F_{\text{att}} \quad (49)$$

where F_{CSA} is the Cross-Scale Attention output, F_{att} is the resized attention representation, F_u is the original upsampled decoder feature, and F_{CR} is the cross-scale residual feature.

- **Multi-Scale Feature Concatenation Layer**

The Multi-Scale Feature Concatenation Layer combines the cross-scale refined decoder feature with the attention-gated encoder skip feature through channel-wise concatenation. This combines high-level semantic information from the decoder with fine-grained spatial information from the encoder. The resulting feature representation is supplied to the decoder refinement block.

$$F_{\text{MS}} = \text{Concat}[F_{\text{CR}}, S_{\text{gate}}] \quad (50)$$

where F_{CR} is the cross-scale refined decoder feature, S_{gate} is the attention-gated encoder feature, Concat denotes channel-wise concatenation, and F_{MS} is the resulting multi-scale feature representation.

- **Decoder Feature Refinement Layer**

The Decoder Feature Refinement Layer processes the multi-scale fused representation using two 3×3 Conv2D operations, Dropout, Token Pool Mixer, and SE attention. The convolutional operations refine the reconstructed features, while Token Pool Mixer enhances spatial interaction and SE attention performs channel recalibration. This functional structure is repeated across the four decoder stages with different filter dimensions.

$$F_1 = \text{ReLU}(W_1 * F_{\text{MS}} + b_1) \quad (51)$$

$$F_2 = \text{ReLU}(W_2 * \text{Dropout}(F_1) + b_2) \quad (52)$$

$$F_{\text{dec}} = \text{SE}(\text{TPM}(F_2)) \quad (53)$$

where F_{MS} is the multi-scale fused feature, F_1 and F_2 are intermediate convolutional representations, W_1, W_2 and b_1, b_2 are convolution parameters, $\text{TPM}(\cdot)$ denotes Token Pool Mixer processing, $\text{SE}(\cdot)$ denotes SE channel attention, and F_{dec} is the refined decoder representation.

- **Edge Feature Extraction Layer**

The Edge Feature Extraction Layer operates on the final decoder feature $d4$ using a 3×3 Conv2D with 32 filters and ReLU activation. It extracts boundary-sensitive features associated with local structural transitions in the reconstructed representation. These features are then provided to the auxiliary edge prediction layer.

$$F_e = \text{ReLU}(W_e * d4 + b_e) \quad (54)$$

where $d4$ is the final decoder feature, W_e is the 3×3 edge-feature convolution kernel, b_e is its bias, and F_e represents the extracted edge-sensitive feature map.

- **Edge Prediction Conv2D + Sigmoid Layer**

The Edge Prediction Layer converts the extracted boundary-sensitive features into a single-channel edge probability map using a 1×1 Conv2D with one filter and integrated Sigmoid activation. The output represents the predicted boundary probability at each spatial location. This forms the auxiliary edge output of the network.

$$E = \sigma(W_{\text{edge}} * F_e + b_{\text{edge}}) \quad (55)$$

where F_e is the extracted edge feature, W_{edge} is the 1×1 edge-prediction kernel, b_{edge} is its bias, $\sigma(\cdot)$ is the Sigmoid activation, and E is the single-channel predicted edge probability map.

- **Edge-Feature Fusion Layer**

The Edge-Feature Fusion Layer combines the predicted edge map with the final decoder feature through channel-wise concatenation. This introduces explicitly predicted boundary information into the main segmentation pathway. The resulting fused representation is subsequently processed by the final segmentation output layer.

$$F_{\text{EdgeFuse}} = \text{Concat}[d, 4E] \quad (56)$$

where d_4 is the final decoder representation, E is the predicted edge map, Concat denotes channel-wise concatenation, and F_{EdgeFuse} is the resulting edge-enhanced feature representation.

- **Final 1×1 Conv2D + Sigmoid Segmentation Output Layer**

The Final Segmentation Output Layer uses a 1×1 Conv2D with one filter and integrated Sigmoid activation to convert the edge-fused representation into the final pixel-wise tumor probability map. The 1×1 convolution performs the final channel projection, while the Sigmoid activation converts the projected values into probabilities between zero and one. Therefore, the final output is a single Conv2D layer with Sigmoid activation, exactly as implemented in the code.

$$\hat{Y} = \sigma(W_m * F_{\text{EdgeFuse}} + b_m) \quad (57)$$

where F_{EdgeFuse} is the concatenated decoder and edge representation, W_m is the final 1×1 Conv2D kernel, b_m is its bias, $\sigma(\cdot)$ is the integrated Sigmoid activation, and $\hat{Y}$ represents the final single-channel tumor segmentation probability map.

The proposed Model enhances the baseline Attention U-Net by introducing several feature-learning and attention mechanisms aimed at improving breast-lesion segmentation. The encoder is modified with a Token Pool Mixer and Squeeze-and-Excitation (SE) attention, allowing spatial feature interaction and channel-wise feature recalibration while retaining residual information. A Polar/Log-Polar branch is additionally incorporated at the deeper encoder level to provide explicit geometric information through radius, angle, and logarithmic-radius representations, which strengthens the representation of lesion shape and structure. At the bottleneck, a RepLK-SE block using factorized 31×1 and 1×31 depthwise convolutions expands the receptive field to capture broader contextual information without using a computationally heavy full 31×31 convolution. These modifications go beyond traditional convolution encoder and are meant to account for variations in lesion size, structure and spatial properties described in the document.

The proposed Model modifies the decoder and skip-fusion pathway to further enhance the localization of the relevant lesion regions and boundaries. Each stage of the decoder makes use of an attention gate, which makes use of channel attention and spatial attention mechanisms to enable selective retention of informative encoder features, while a Cross-Scale Attention facilitates the interaction of the decoder representation with encoder-derived features in advance of multi-scale fusion. The attention operation is limited in space to a maximum grid size of 32, thereby reducing the computational and memory requirements for the cross-attention operation, and the resulting information is integrated using residual fusion. In addition, an auxiliary edge branch extracts boundary-sensitive features from the final decoder representation and produces an edge probability map, which is concatenated with the decoder features before the final 1×1 convolution generates the tumor segmentation mask. Thus, the main contribution of the proposed modification is the combination of enhanced spatial mixing,

channel recalibration, geometric feature encoding, large-receptive-field contextual extraction, cross-scale attention, attention-gated skip connections, and edge-feature fusion within a single segmentation framework, providing both contextual and boundary-sensitive representations for more precise lesion delineation.

4. Experimental results

We evaluated the proposed network on the Breast Ultrasound Images Dataset (BUSI) cited as reference [17]. The metrics below are reproduced from the original analysis and have not been recalculated. The present description does not establish image and patient counts by partition, whether splitting preceded augmentation, the number of random seeds, or whether subjects overlap between training and test sets. These reporting items are required for reproducibility and should be completed from the original experiment logs. Pixel-level accuracy and specificity were interpreted in the context of lesion-to-background imbalance, with lesion Dice and IoU treated as primary segmentation outcomes.

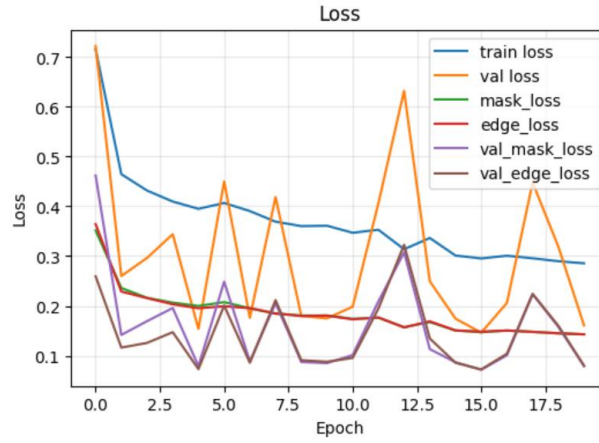


Figure (3): Training and Validation Loss Curves of the Proposed Hybrid Attention U-Net

Figure 3 presents training and validation loss trajectories over 20 epochs. The displayed training loss generally decreases, whereas the validation loss fluctuates substantially. The curves alone do not establish generalization performance; the training/validation split, stopping rule, selected checkpoint, and uncertainty across runs should be reported. The separate mask and edge losses should be defined in the Methods, including their respective weights.

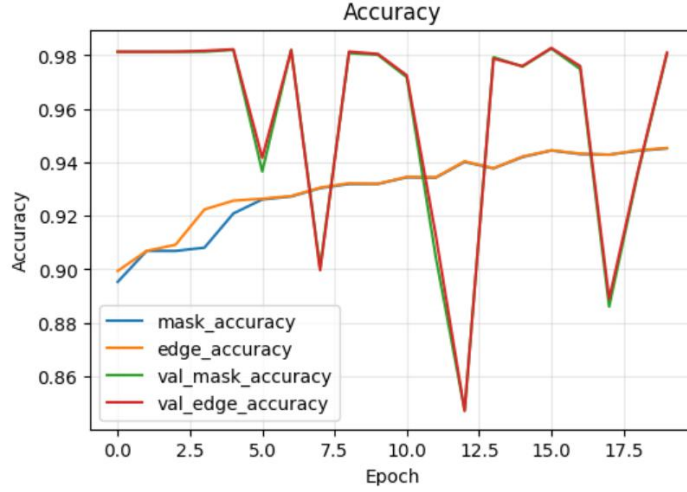


Figure (4): Training and Validation Accuracy Curves of the Proposed Hybrid Attention U-Net

Figure 4 shows mask- and edge-related training and validation accuracy across 20 epochs. The relatively high accuracy values should not be interpreted as evidence of accurate tumor segmentation, because non-lesion pixels may comprise most of the image. Overlap measurements (Dice and IoU), lesion sensitivity, and per-image performance are more informative for this task.

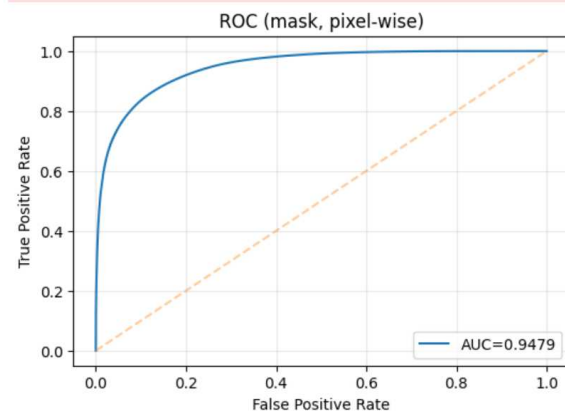


Figure (5): Pixel-Wise ROC Curve for Tumor Segmentation

Figure 5 displays a pixel-wise ROC curve annotated with AUC 0.9479. However, Table 1 and the abstract give ROC-AUC 0.9005. The discrepancy must be resolved using the saved ground-truth masks and unthresholded test-set probability maps, with the pooling/averaging convention and test split explicitly stated. Until verified, neither value should be emphasized as a definitive result.

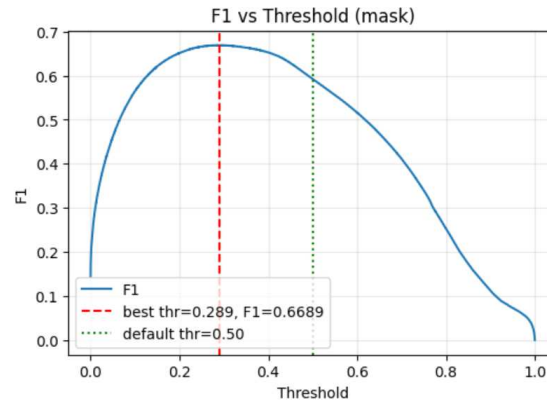


Figure (6): F1-Score Variation with Classification Threshold for Tumor Segmentation

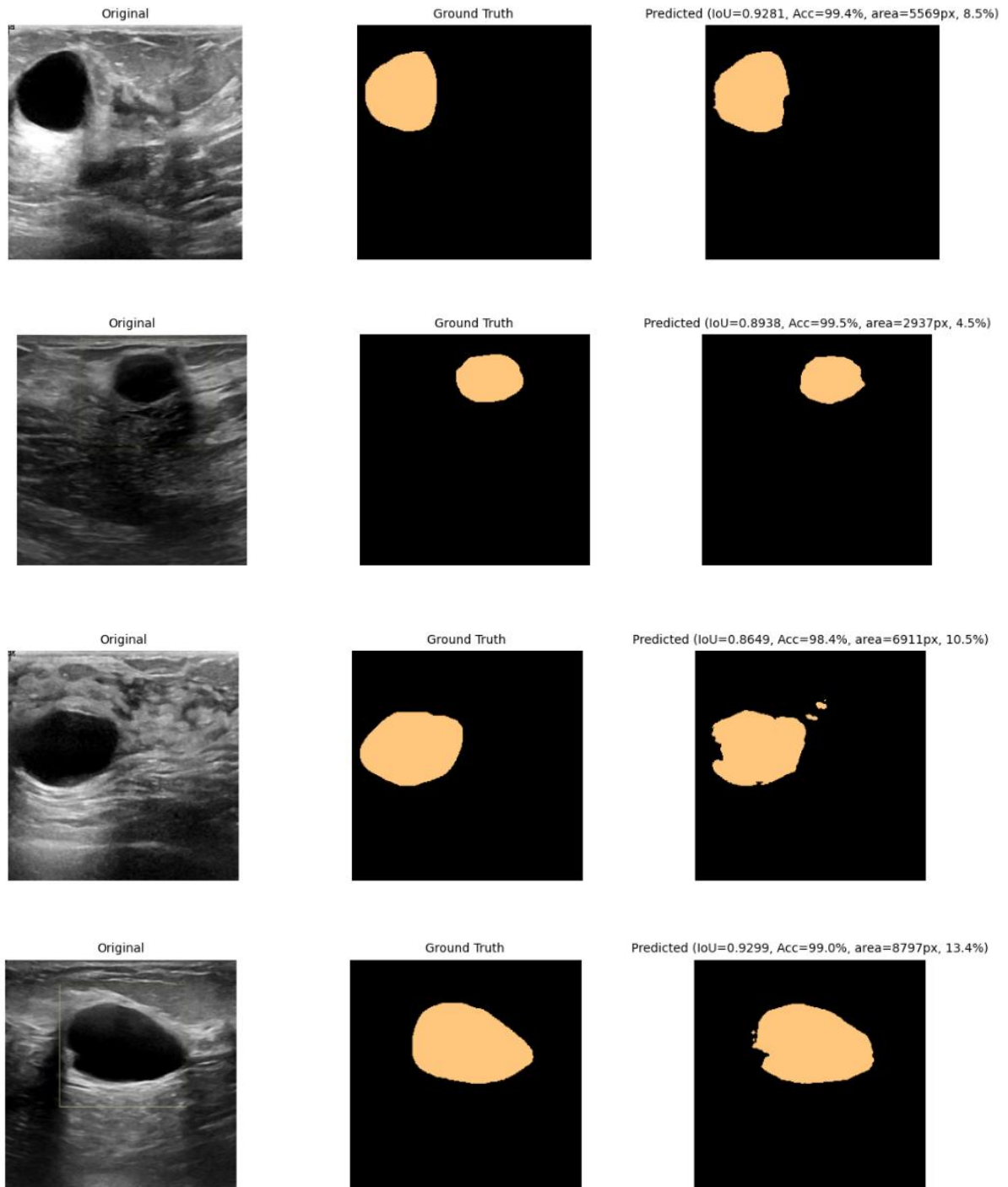
Figure 6 displays F1 as a function of segmentation threshold and identifies an apparent maximum near threshold 0.289 with F1 approximately 0.6689. Table 1 reports F1 0.3673, and the figure indicates a different value at threshold 0.50. These results may reflect different threshold, sample-aggregation, or metric definitions, but the manuscript does not establish the reason. The evaluation protocol must report the threshold chosen on validation data, the fixed threshold applied to the held-out test set, and whether F1 was calculated per image or from aggregated pixel counts. Test-set-based threshold optimization must not be used to report unbiased performance.

Table (1): Segmentation Performance Metrics

Metric	Value
Accuracy	0.9808
IoU	0.2250
Dice Score	0.3673
Precision	0.4674
Recall	0.3026
Specificity	0.9935
F1 Score	0.3673
ROC-AUC	0.9005

Table 1 summarizes the originally reported test metrics: pixel accuracy 0.9808, specificity 0.9935, IoU 0.2250, Dice 0.3673, precision 0.4674, recall 0.3026, and ROC-AUC 0.9005. High accuracy and specificity principally indicate correct classification of background pixels; by contrast, the low Dice, IoU, and recall show inadequate correspondence with annotated tumor

regions. The results therefore do not support a claim of precise breast-lesion delineation. The difference between F1 and Dice values shown in different parts of the manuscript also requires reconciliation of the metric definitions and averaging protocols.



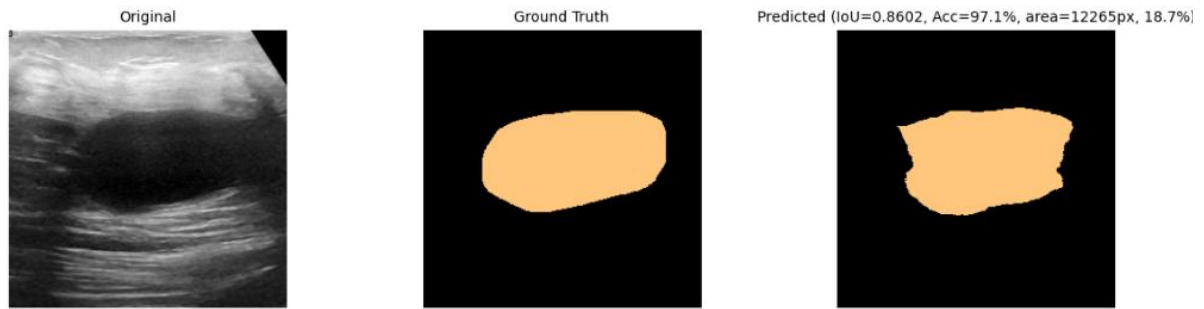


Figure (7): Qualitative Comparison of Ground-Truth and Predicted Breast Lesion Segmentation Results

Figure 7 provides selected examples of original ultrasound images, reference masks, and model predictions. Some displayed examples show visually substantial overlap, but these examples cannot substitute for test-set-level performance. The overall Dice and IoU are low; a representative error-analysis figure should also include typical poor-segmentation cases, missed lesions, false-positive regions, and lesion-size-stratified cases selected by a prespecified rule.

Table (2): Comparative analysis

Methods	Accuracy
ResNet34 [18]	0.6239
CNN [19]	0.6698
Bi-LSTM [20]	0.9316
EfficientNetB3 [21]	0.9367
Proposed Model (Modified Attention U-Net)	0.9807

Table 2 reproduces previously cited accuracy values for ResNet34, CNN, Bi-LSTM, and EfficientNetB3, alongside a proposed-model accuracy of 0.9807 (Table 1 reports 0.9808). Because it is not established that these values arise from identical BUSI partitions, mask definitions, thresholds, or segmentation tasks, this table must not be interpreted as a controlled benchmark or evidence that the proposed segmentation method outperforms those systems. A valid comparison requires implementation or re-evaluation of U-Net, Attention U-Net, and appropriate contemporary BUSI segmentation methods under identical patient-independent data partitions, with Dice, IoU, recall, precision, and uncertainty reported. The original accuracy-only comparison is retained for traceability, not as a demonstrated superiority claim.

5. Conclusion

This work described a modified Attention U-Net for breast ultrasound lesion segmentation integrating token mixing, channel recalibration, Polar/Log-Polar descriptors, large-kernel convolution, cross-scale attention, and an auxiliary edge-learning branch. On the reported BUSI evaluation, pixel accuracy (0.9808) and specificity (0.9935) were high, but tumor-overlap performance was limited (Dice 0.3673; IoU 0.2250), with low recall (0.3026). Consequently, the present findings are preliminary and do not establish clinically dependable tumor boundary localization or superiority over properly matched segmentation baselines. The unresolved ROC-AUC and F1 reporting discrepancies must be checked using original predictions. Future work should focus on improving lesion overlap, conducting component-level ablations and controlled baselines, validating across independent clinical data, and evaluating whether segmentation quality supports meaningful radiological measurements.

Funding Declaration

The authors declare that no financial support or funding was received for the research, authorship, and/or publication of this article

Conflict of interest statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability statement: The dataset analyzed during this study is publicly available as the Breast Ultrasound Images Dataset (BUSI). The dataset was used to develop and evaluate the proposed deep learning-based breast lesion segmentation model. No new clinical imaging dataset was generated during this study.

Dataset link: <https://www.kaggle.com/datasets/aryashah2k/breast-ultrasound-images-dataset>

References

- [1] Iranmakani, Sepideh, Tohid Mortezaazadeh, Fakhrossadat Sajadian, Mona Fazel Ghaziani, Ali Ghafari, Davood Khezerloo, and Ahmed Eleojo Musa. "A review of various modalities in breast imaging: technical aspects and clinical outcomes." *Egyptian Journal of Radiology and Nuclear Medicine* 51, no. 1 (2020): 57.
- [2] Khanh, Trinh Le Ba, Duy-Phuong Dao, Ngoc-Huynh Ho, Hyung-Jeong Yang, Eut-teum Baek, Gueesang Lee, Soo-Hyung Kim, and Seok Bong Yoo. "Enhancing U-Net with spatial-channel attention gate for abnormal tissue segmentation in medical imaging." *Applied Sciences* 10, no. 17 (2020): 5729.
- [3] Gao, Shuo, Jia Liu, Linqian Li, Di Yang, Yafei Miao, Xu Zhang, Qianqian Han, Yasong Shi, Jianguo Wu, and Ke Zhang. "Application of deep learning technology in breast cancer: a systematic review of segmentation, detection, and classification approaches." *BioMedical Engineering OnLine* 25, no. 1 (2026): 19.

- [4] Zhang, Yi, Yiji Mao, Xuanyu Lu, Xingyu Zou, Hao Huang, Xinyang Li, Jiayue Li, and Haixian Zhang. "From single to universal: tiny lesion detection in medical imaging: Y. Zhang et al." *Artificial Intelligence Review* 57, no. 8 (2024): 192.
- [5] Sun, Fuming, Peng Ren, Bowen Yin, Fasheng Wang, and Haojie Li. "CATNet: A cascaded and aggregated transformer network for RGB-D salient object detection." *IEEE Transactions on Multimedia* 26 (2023): 2249-2262.
- [6] Li, Yangyang, Xiang Zheng, Zhenghong Gao, Ronghua Shang, and Licheng Jiao. "Edge-Aware Dual-Stream Dynamic Fusion Network for Building Extraction in Remote Sensing Images." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (2026).
- [7] Zhou, Yuce, Fengqing Jin, Guodong Suo, and Jianlan Yang. "ResViT-GANNet: a deep learning framework for classifying breast cancer histopathology images using multimodal attention and GAN-based augmentation." *BMC Medical Imaging* 25, no. 1 (2025): 401.
- [8] Wang, Dongyue, Weiyu Zhao, Kaixuan Cui, and Yi Zhu. "VMC-UNet: A Vision Mamba-CNN U-Net for Tumor Segmentation in Breast Ultrasound Image." *International Journal of Imaging Systems and Technology* 34, no. 6 (2024): e23222.
- [9] Zafar, Mehwish, Javaria Amin, Muhammad Sharif, and Mrim M. Alnfai. "Hybrid Loss-Driven Attention Residual U-Net with Multiscale Cascaded CBAM for Breast Cancer Segmentation." *Cognitive Computation* 18, no. 1 (2026): 27.
- [10] Nasimov, Rashid, Kudratjon Zohirov, Adilbek Dauletov, Akmalbek Abdusalomov, and Young Im Cho. "From Pixels to Precision—A Dual-Stream Deep Network for Pathological Nuclei Segmentation." *Bioengineering* 12, no. 8 (2025): 868.
- [11] Liu, Tao, JianLong Hu, and MengYu Zhao. "Ecca-unet: edge-aware and channel-enhanced cross-attention network for medical image segmentation." *Multimedia Systems* 31, no. 6 (2025): 450.
- [12] Kushwah, Shivpratap Singh, Santosh Prakash Chouhan, Narinder Singh Pun, and Mahua Bhattacharya. "LTM-UNet: Linear Transformer–Mamba with Attention-Based U-Net for Context-Aware Breast Ultrasound Image Segmentation." *Diagnostics* 16, no. 12 (2026): 1888.
- [13] Guo, Tan, Wenhan Zhang, Fulin Luo, Chuan Fu, Weiwen Wu, and Lei Zhang. "SGLF-Net: Staged Global-to-Local Cross-Scale Fusion Network for Colonoscopic Polyp Segmentation." *IEEE Journal of Biomedical and Health Informatics* (2026).
- [14] Zhang, Zhongzhou, Hui Yu, Zhongxian Wang, Zhiwen Wang, Jingfeng Lu, Yan Liu, and Yi Zhang. "Gradient-guided network with Fourier enhancement for glioma segmentation in multimodal 3D MRI." *IEEE Journal of Biomedical and Health Informatics* 28, no. 11 (2024): 6778-6790.
- [15] Yang, Wenbin, and Xinyue Guo. "DGT-MSB-Net: A Transformer-Based Framework With Boundary Awareness for Accurate Liver Tumor Segmentation in CT Images." *International Journal of Imaging Systems and Technology* 36, no. 4 (2026): e70376.

- [16] Ma, Xiaoxuan, and Sirui Wang. "MAMF-Net: A medical image segmentation network with multiple attention and multi-scale feature fusion." *Signal, Image and Video Processing* 20, no. 9 (2026): 483.
- [17] A. Shah, "Breast Ultrasound Images Dataset: Breast ultrasound images for classification, detection & segmentation," Kaggle, 2021. [Online]. Available: <https://www.kaggle.com/datasets/aryashah2k/breast-ultrasound-images-dataset>
- [18] Wang, Qiucheng, He Chen, Gongning Luo, Bo Li, Haitao Shang, Hua Shao, Shanshan Sun, Zhongshuai Wang, Kuanquan Wang, and Wen Cheng. "Performance of novel deep learning network with the incorporation of the automatic segmentation network for diagnosis of breast cancer in automated breast ultrasound." *European radiology* 32, no. 10 (2022): 7163-7172.
- [19] Tsochatzidis, Lazaros, Panagiota Koutla, Lena Costaridou, and Ioannis Pratikakis. "Integrating segmentation information into CNN for breast cancer diagnosis of mammographic masses." *Computer Methods and Programs in Biomedicine* 200 (2021): 105913.
- [20] Aburass, Sanad, Osama Dorgham, and Jamil Al Shaqsi. "A hybrid machine learning model for classifying gene mutations in cancer using LSTM, BiLSTM, CNN, GRU, and GloVe." *Systems and Soft Computing* 6 (2024): 200110.
- [21] Jasrotia, Heena, Chandan Singh, and Sukhjeet Kaur. "EfficientNet-Based Attention Residual U-Net With Guided Loss for Breast Tumor Segmentation in Ultrasound Images." *Ultrasound in Medicine & Biology* (2025).