Enhancing Breast Tumor Segmentation with TL-ESMNet: A Transfer Learning and Ensemble-Based Approach for Mammograms

Satyanarayana Reddy Beram [1,*], R Lalchhanhima [1], and Ksh. Robert Singh [2]

[1] Department of Information Technology, Mizoram University, Mizoram, India
[2] Department of Electrical Engineering, Mizoram University, Mizoram, India
E-mail: mzu22007898@mzu.edu.in (S.R.B.); chhana.mizo@gmail.com (R.L.); robert_kits@yahoo.co.in (K.R.S.)
*Corresponding author

Abstract—**Breast tumor segmentation in mammographic images is challenging because of variations in the tumor size, contrast, and shape. This study introduces a Transfer Learning based Ensemble Net (TL-ESMNet), a novel framework that incorporates adaptive segmentation with attention mechanisms within the Ensemble Net (ESM-Net) stage to address these complexities. Utilizing Transfer Learning (PreTL) with pre-trained models, such as ResNet and VGG16, TL-ESMNet extracts robust feature representations that are critical for precise segmentation. The attention mechanism of the framework, which combines spatial and channel attention, dynamically focuses on relevant regions and channels, enhancing its adaptability to small, low-contrast, or irregularly shaped tumors. Segmentation was performed using U-Net or DeepLabV3, leveraging multi-scale feature extraction to ensure accurate tumor boundary delineation. A dynamic fusion mechanism within ESM-Net merges outputs from specialized models based on image characteristics, providing robustness across diverse tumor presentations. TL-ESMNet is rigorously evaluated on the Digital Database for Screening Mammography (DDSM) and INbreast datasets, achieving Dice Coefficients (DCE) of 0.91 and 0.88, respectively, and demonstrating strong performance across additional metrics, including Precision of 0.92 and 0.89, and Recall of 0.89 and 0.86 on DDSM and INbreast datasets, respectively. Attention mechanisms significantly enhance the segmentation accuracy, particularly in complex cases involving small- or low-contrast tumors. This scalable and efficient framework provides a reliable solution for automated breast tumor segmentation, supporting improved clinical decision making in breast cancer diagnosis.**

Keywords—**breast tumor segmentation, attention mechanism, Ensemble Net (ESM-Net), transfer learning, U-Net, deeplabv3, mammographic images**

I. INTRODUCTION

Breast cancer remains one of the most common cancers globally and a primary contributor to cancer-related deaths among women. As per global health statistics [1], millions of women receive a breast cancer diagnosis each year, highlighting the importance of early detection to improve survival rates. Early-stage diagnosis significantly boosts the five-year survival rate for breast cancer, emphasizing the necessity for reliable diagnostic methods and prompt intervention [2]. Nonetheless, despite advancements in detection and treatment, a significant portion of cases are still identified at later stages, especially in areas with limited access to medical imaging and screening facilities.

Mammography is the widely used modality for breast screening and is effective in identifying the tumors at early stage [3]. However, it faces challenges like the differences in tumor size, shape, and contrast, particularly in dense breast tissue, where the tumors detection become complex. These factors make accurate diagnosis is more challenging and underscores the limitations of the mammography. Overlapping patterns, dense tissue, and low contrast issues also make the process difficult, with a high chance of false positives or false negatives. As the volume of the mammographic data is increasing daily, there is an urgent need for mammogram analysis automation, which help the radiologists to deliver accurate and consistent diagnoses [4]. An adaptable and advanced segmentation framework like our Transfer Learning based Ensemble Net (TL-ESMNet) can assist the clinicians by delivering the consistent, high-quality segmentation results, minimizing the diagnostic inconsistencies, and improving the early detection rates.

One of the most significant technical limitations in breast tumor segmentation is the heterogeneity among the tumor attributes, such as size, shape, and contrast levels. Current segmentation models do not efficiently recognize and segment such complex tumors, especially on low-contrast areas or denser breast tissue. In reality, the tumours can vary significantly, from small, subtle lesions to large, irregularly shaped masses, making it more challenging to create a consistent segmentation approach. The low contrast between tumors and the surrounding tissue in many mammograms, particularly in dense breasts, complicates accurate detection even further [3]. Breast cancer imaging can highly differ based on the type of the modality (i.e., Computed Tomography (CT),

Manuscript received January 2, 2025; revised February 21, 2025; accepted March 5, 2025; published June 12, 2025.

doi: 10.12720/jait.16.6.838-853

Magnetic Resonance Imaging (MRI) etc.) they employed. Traditional film-based mammograms, like those in the DDSM dataset, typically exhibit the lower resolution and image quality when compared to the contemporary digital mammograms, such as those found in the INbreast dataset. These variations add the extra challenges [5], as segmentation algorithms must be able to tackle with the variations in image quality, resolution, and noise, which can impact the accuracy of tumor detection and segmentation.

Due to the complexity of the tumor characteristics and the variations in imaging modalities, there is an increasing need for designing the automated and adoptable segmentation tools [4] to support the radiologists. Manual interpretation process takes more time, which is prone to errors and the results can vary depending on the radiologist's experience. In contrary, the automated segmentation models can deliver the precise and accurate results, which can improve the accuracy of diagnoses, and helps in detecting the tumors from the challenging cases with low contrast or dense tissues. While the traditional segmentation models [4–6] are limited in their ability to handle this diversity in tumor images segmentation, our TL-ESMNet comprised of advanced attention and fusion mechanisms to tackle this diversity efficiently.

The medical image analysis field has been greatly enhanced by the advanced deep learning models, especially in the domain of medical image segmentation. These deep models learn the hidden complex patterns autonomously from the large datasets, often outperforming the traditional models in terms of both accuracy and efficiency. Models like U-Net [7] and DeepLabV3+ [8] have been extensively used in medical image segmentation, ensuring the consistent and reliable outcomes in across various medical applications. Although these models [4, 7, 8] are effective in many cases, they are still facing significant challenges in addressing the complex tumor morphologies—especially irregularly shaped tumors—or when data is scarce, which is often a limitation in medical imaging datasets. To overcome these limitations, attention mechanisms [9] and ensemble models [10] emerged as the popular and reliable solutions. Attention mechanisms allows the models to focus on the most important regions of an image, which enhances the accuracy, especially in dealing with the small, irregular, or low tumors. On other hand, the ensemble models [10] can combine the strengths of various specialized models to improve the robustness and overall performance [11]. Together, these advancements are proven increasing the capabilities of automated breast tumor segmentation, which make it possible to achieve the better diagnostic accuracy, even in challenging scenarios. By utilizing both of these innovations, our TL-ESMNet is offering a systematic model that ensures the reliable tumor segmentation across varying conditions, which helps the clinicians to improve the diagnostic workflows.

Current breast tumor segmentation models [7, 8] are facing the considerable difficulties in identifying and accurately segmenting the small, irregularly shaped tumors, which are challenging to detect. These models are also having the difficulties with low-contrast images and variations in imaging data (i.e., differences between digital and film-based mammograms). A major challenge in current solutions [11] is the ability to generalize the model across different datasets [12] while maintaining the high accuracy in handling the complex cases. Our TL-ESMNet addresses these gaps by incorporating the transfer learning technique, which allows it to adaptively learn from the limited data, and by using the ensemble modelling for flexibility across the diverse imaging conditions. Consequently, there is a pressing need for a more robust and adaptive segmentation framework that can overcome these limitations efficiently. Using these innovative techniques such as transfer learning and ensemble modelling, a new framework can improve the feature extraction and model performance, especially in challenging cases with limited data or complex tumor morphologies.

The prominent contributions of this paper are:

- TL-ESMNet Framework with Attention-Driven Segmentation: We propose TL-ESMNet, an advanced breast tumor segmentation framework that integrates the Transfer Learning, Attention Mechanisms, and Dynamic Fusion within ESM-Net. The inclusion of spatial and channel attention enables the model to focus on critical tumor regions, particularly in low-contrast and irregularly shaped cases.
- Adaptive Feature Extraction and Multi-Scale Learning: By employing pre-trained models like ResNet and VGG16, our TL-ESMNet successfully extracts the prominent features, which reduces the need for extensive annotated datasets. The use of U-Net and DeepLabV3 enables the multi-scale feature learning, which enhances the segmentation for a various tumor types.
- Dynamic Fusion for Robust Segmentation: A novel and dynamic feature fusion mechanism is introduced in ESM-Net. It adaptively combines the outputs from specialized segmentation models, based on tumor characteristics to ensure the accurate segmentation across different tumor sizes and imaging conditions.
- Comprehensive Benchmarking and Validation: Our framework has been thoroughly evaluated using the DDSM and INbreast datasets, showing the better segmentation performance compared to the baseline models based on Dice Coefficient, IoU, and precision-recall metrics.

These contributions offer a promising foundation for the future developments in clinical breast tumor segmentation. It presents a scalable and reliable solution that can assist the radiologists by minimizing the diagnostic inconsistencies and improving the accuracy of early breast cancer detection.

The structure of the paper is as follows: Section II provides a review of related work on breast tumor segmentation, highlighting current deep learning methods and their limitations. Section III details the TL-ESMNet framework, including its architecture and the incorporation of Transfer Learning, Attention Mechanisms, and Dynamic Fusion within ESM-Net. Section IV describes the experimental setup, which includes the datasets (DDSM and INbreast), the steps for data pre-processing, and the evaluation metrics used. Section V showcases the experimental results, compares the performance of TL-ESMNet with baseline models, and discusses the main findings. Finally, Section VI wraps up with a summary of contributions, limitations, and possible future research directions.

II. LITERATURE REVIEW

This section reviews the recent progress in breast tumor segmentation, highlighting the main areas like deep learning approaches, attention mechanisms, and transfer learning. These approaches are used widely to solve the problems like tumor size variation, shape irregularity, and low contrast of mammographic images.

A. Deep Learning-Based Breast Tumor Segmentation

Former breast tumor segmentation models were depended on traditional techniques like thresholding and region growing. The main limitation of these approaches is, they score less performance while dealing with complex tumor boundaries, low contrast, and irregular shapes. Shrivastava and Bharti [13] discussed about these challenges of region-growing methods and pointing out their reduced accuracy when segmenting the intricate tumor features in mammograms. Park *et al.* [14] developed a deep learning model for 3D breast cancer segmentation in DCE-MRI using the weak annotations. This approach improves the segmentation accuracy and efficiency, by enhancing the tumor detection and segmentation with minimal usage of annotations.

The introduction of deep learning has greatly enhanced the accuracy of breast tumor segmentation. Ronneberger *et al.* [7] has demonstrated the U-Net's encoder-decoder framework has particularly in capturing the fine details, which makes a prominent model in the field of medical image analysis. Subsequent research further improved the U-Net's segmentation performance by introducing the spatial and channel wise attention mechanisms to augment the model's attention towards the low-contrast and small tumors [15]. However, these models still encounter issues with the extreme diversity in tumor shapes which are constrained when there is a lack of training data, which is a common scenario in medical datasets.

B. Enhancing Segmentation with Transfer Learning, Attention Mechanisms, and Multi-Scale Features

Transfer learning was emerged as an influential tool in medical imaging, particularly for tumor segmentation. Bi *et al.* [16] demonstrated the value of pre-trained methodologies, such as ResNet and VGG16, initially trained on large-scale datasets like ImageNet, in enhancing performance for specialized tasks.

Studies have applied transfer learning successfully to breast tumor segmentation. Gomez-Flores and Pereira [17] evaluated Convolutional Neural Network (CNN) architectures, including DeepLabV3+ with ResNet as a backbone, for segmenting breast ultrasound images, finding that fine-tuning improved feature extraction in low-contrast tumor regions. Similarly, Leong *et al.* [18] proposed a transfer learning framework using ResNet50 for segmenting microcalcifications in mammography, achieving notable improvements in accuracy and robustness in early cancer detection. By applying transfer learning and fine-tuning on VGG16, Falconí *et al.* [19] were able to classify mammograms into Breast Imaging Reporting and Data System (BI-RADS) categories, showing the improvement in classification accuracy with the Curated Breast Imaging Subset of the Digital Database for Screening Mammography (CBIS-DDSM) dataset.

While transfer learning has played a vital role in segmentation, current models [17–19] fall short of addressing various imaging conditions. To address these gaps, TL-ESMNet integrates the transfer learning with multi-scale and attention-based mechanisms, by improving its adaptability across different the imaging contexts.

Zhang *et al.* [20] has employed a spatial and channel attention mechanisms to focus on COVID-19 impacted regions and accentuate relevant feature channels, leading to improve the segmentation accuracy to capture the subtle characteristics. Zhao *et al.* [21] created an adaptive channel and multi-scale spatial context network that dynamically alters attention on both channel and spatial fronts, facilitating better tumor segmentation even when faced with low signal-to-noise ratios. A multi-view stereoscopic attention network for 3D tumor classification in automated breast ultrasound was proposed by Ding *et al.* [22], achieving better diagnostic accuracy. This model effectively improves feature extraction and tumor differentiation for improved breast cancer detection. Zhang *et al.* [23] presented DenseATT-Net, model which integrates various attention modules, such as Convolutional Block Attention Module (CBAM) and Squeeze-and-Excitation (SE) blocks, which improve the segmentation of breast masses in automated breast ultrasound.

Multi-scale feature extraction plays a key role in medical image segmentation by allowing the models to capture intricate and large-scale details within the complex images. Gao and Almekkawy [24] showed that multi-level feature extraction methods, like nested U-Net models, which can enhance the segmentation performance of tumors with different sizes and intricate edges, especially with heterogeneous datasets. By introducing the multi-scale attention features, Poudel and Lee [25] provent that the dynamic capture of various levels of semantic information can improve the segmentation accuracy in noisy and challenging medical imaging contexts.

Wang *et al.* [26] employed a multi-scale U-Net-based architecture for brain tumor segmentation with greater accuracy by considering global context and fine details, which play a vital role in identifying the unusual shapes of tumors. Zhang *et al.* [27] proposed a U-shaped network with optimized multi-scale feature extraction for lesion segmentation, highlighting the strategies which improves the boundary preservation in complex datasets like gastric cancer.

To improve the segmentation accuracy, TL-ESMNet combines the multi-scale feature extraction process with attention mechanisms model. This allows the model to dynamically concentrate on important areas based on the tumor characteristics, which includes shape, size, and contrast.

C. Need for Dynamic Fusion and Research Gaps

Although deep learning, transfer learning, and attention-based segmentation have advanced, existing models still encounter challenges:

- Inconsistent segmentation performance for small, low-contrast, and irregularly shaped tumors.
- Limited generalizability across datasets, making models dataset-dependent and reducing their clinical applicability.
- Static fusion methods that lack adaptability, leading to inefficient segmentation in varied imaging conditions.

Fusion methods have been suggested to improve segmentation through the integration of outputs from multiple models. Zou *et al.* [28] introduced to multi-level semantic fusion method for interactive image segmentation has shown significant improvements in both edge refinement and segmentation accuracy. Huang *et al.* [29], in contrast, proposed a dynamic adaptive network which tailors the model by assigning the weights according to the specific characteristics of the input image, which is enhancing segmentation accuracy. Qi *et al.* [30] presented Multi-scale Dynamic Fusion Network (MDF-Net), a multi-scale dynamic fusion network that designed specifically for segmenting the breast tumors in ultrasound images. Additionally, Sultan and Zubair [31] employed a combination of bioinformatics and machine learning methods to discover the common breast cancer biomarkers among various populations. Their study identifies the crucial genetic markers that could improve the early detection and personalized treatment strategies. Furthermore, these studies show that adapting fusion mechanisms to different image modalities and views can significantly boost the segmentation accuracy in challenging medical imaging scenarios.

1) Summary of advancements

The field of breast tumor segmentation has seen considerable advancements, with models like U-Net [7], DeepLabV3 [8], and transformer-based architectures enhancing segmentation accuracy. Transfer learning [17] has played a crucial role in improving feature extraction, while attention mechanisms enable models to concentrate on areas relevant to tumors. Furthermore, multi-scale feature extraction has enhanced segmentation for tumors of varying sizes and complexities.

2) Identified gaps and the need for TL-ESMNet

Despite the progress in the field, there are still some notable gaps:

- Existing models fail to adapt to small, low-contrast, and irregularly shaped tumors.
- Static fusion methods limit segmentation accuracy under varying tumor imaging conditions.
- Generalization across datasets is a challenge, requiring models to be fine-tuned for specific datasets.

TL-ESMNet addresses these limitations by integrating transfer learning, attention-driven segmentation, and dynamic fusion, ensuring:

- Accurate segmentation across varied tumor morphologies.
- Adaptive fusion of multi-scale features, leading to better generalization.
- Improved segmentation consistency in both high-resolution digital mammograms (INbreast) and lower-quality film-based mammograms (DDSM).

III. TL-ESMNET FRAMEWORK

A. Overview of the Architecture

TL-ESMNet framework improves the breast tumor segmentation by incorporating the adaptive attention mechanisms into its architecture. Using the multi-scale feature extraction and attention mechanisms, our framework effectively handle the common challenges in mammographic segmentation, such as variations in tumor size, low contrast, and irregular shapes. As shown in Fig. 1, the TL-ESM-Net framework features three primary stages that enhance the segmentation accuracy are: Transfer Learning (PreTL), Attention Mechanisms, and Dynamic Fusion within TL-ESM-Net.

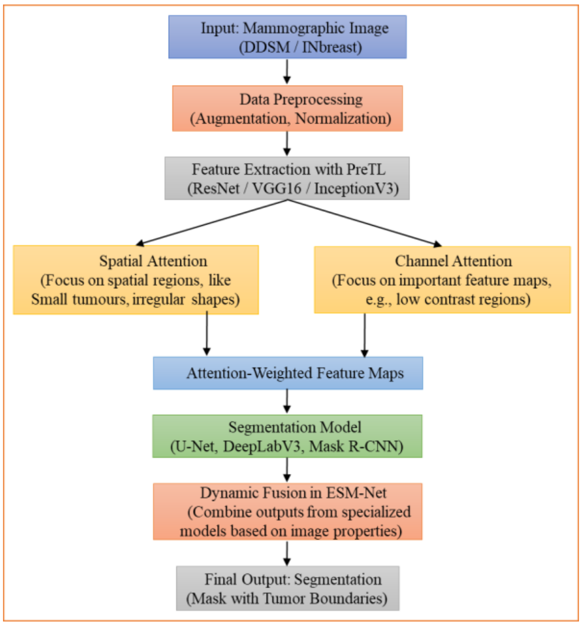

Fig. 1. TL-ESMNet framework for breast tumor segmentation.

Our framework process begins with PreTL, where pre-trained models, such as ResNet [16, 19], are used to extract high-level features like edges and textures from mammographic images. These extracted features are used as basis for the segmentation, minimizing the need for large training datasets while preserving critical characteristics of the image.

After feature extraction, the Spatial and Channel Attention Mechanisms function together. They concentrate on the most critical regions and features, to enhance the model's ability to capture the key elements. Spatial Attention [15] focuses on identifying the likely tumor areas, making it particularly effective for segmenting small or irregularly shaped tumors. On the other hand, the Channel Attention prioritizes the important feature channels in low-contrast regions, to accurately capture the subtle tumor boundaries.

Following the attention feature adaption, the enhanced feature maps were passed through segmentation model (i.e. U-Net [7] or DeepLabV3 [8]) to produce the pixel-wise segmentation masks. With the application of multi-scale feature extraction, these models can efficiently detect different tumor sizes and shapes, which is leading to the precise and complete segmentation results. Finally the Dynamic Fusion [28] in TL-ESM-Net combines the segmentation outputs based on the key image features. This technique assures that the final segmentation is flexible and adaptable to various tumor characteristics and imaging conditions. By integrating the PreTL, attention mechanisms, and dynamic fusion, this framework offers a scalable and efficient way to enhance the segmentation precision, especially in complex and challenging mammographic cases.

B. Transfer Learning (PreTL) for Feature Extraction

PreTL within the TL-ESMNet framework functions as a prominent module for feature extraction. It leverages the pre-trained models like ResNet and VGG16, which were trained on extensive datasets like ImageNet [16]. These pre-trained models perfectly captures the low-level (e.g., texture and edge) and high-level (e.g., objects) features from images, which are crucial for the segmentation process. For the breast tumor segmentation, these pre-trained models are fine-tuned on mammography images, to adapt to the specific domain of the medical imaging.

1) Model selection

We selected the models like ResNet, VGG16, and InceptionV3 because they prove their efficiency in generalizing and capturing the important features needed to accurately segment the tumors. Each of these models uses multiple convolutional layers, where each layer progressively captures the more complex features, which are key to differentiate the tumors from their background tissue.

Let $x \in \Re^{H \times W \times C}$ represent the input mammogram image, with dimensions $H \times W \times C$ where H, W , C denote the height, width, and number of channels, respectively. The feature maps generated by the convolutional layers capture key tumor characteristics [17], serving as the foundation for segmentation and are represented as:

$$F_l(x) = \sigma(W_l \times x + b_l)$$

where F_l is the feature map in the l, W_l and b_l are the learned weights and biases of the l-th layer, and σ is the activation function. These layers capture features like edges, shapes, and textures that are pertinent to identifying tumor regions.

2) Fine-tuning strategy

To adapt these pre-trained models for mammographic images, a fine-tuning process was implemented [19]. The initial (lower) layers, which capture general features, are kept frozen to maintain their learned weights, while the deeper layers, which learn more specialized and task-specific features, are retrained using mammographic data. Let θ_{frozen} represent the parameters of the frozen layers and θ_{tuned} represent the parameters of the layers being fine-tuned. The final loss function for model training can be represented as

$$L = \frac{1}{N}\sum_{i=1}^{N} L_{CE}(y_i, \hat{y}_i)$$

where L_{CE} is the cross-entropy loss between the true label y_i and the predicted label $\hat{y}_i$ for each pixel i in the mammogram and N is the pixels count in the image. The model parameters θ_{tuned} were updated using back propagation with gradient descent.

After fine-tuning, the pre-trained model was generated feature maps that captured the key visual information about the tumor-like structures. These maps were helped the model to focus on potential tumor regions and reduced the interference from background noise [15]. Each feature map F_k from the k-th layer encapsulates key information related to the presence of tumor-like structures, such as

$$F_k = [F_{k,1}, F_{k,2}, \ldots, F_{k,c}]$$

where $F_{k,c}$ is the c-th feature channel in the k-th layer and C is the channels count in that layer. These feature maps serve as inputs to the attention mechanism, guiding the subsequent segmentation process by highlighting tumor regions and reducing irrelevant background information.

C. Attention Mechanism

Following the feature extraction phase, the Attention Mechanism in TL-ESMNet guides the models focus on the most important regions and feature channels. As a result, segmentation is improved, especially for small, low-contrast, or irregular tumors [11]. By applying both Spatial and Channel Attention, the framework ensures that the critical areas and features receive high priority, thereby improving the tumor detection and segmentation.

1) Spatial attention

Spatial Attention highlights important spatial areas in the image, directing the focus of the model to potential tumor regions, especially in low-contrast or dense tissues [9].

From the input feature map $F \in \Re^{H \times W \times C}$, Spatial Attention $S \in \Re^{H \times W}$ uses both average and max pooling to capture global and local spatial dependencies as follows:

$$S(i,j) = \sigma(Conv_{2D}([AvgPool_c(F) + MaxPool_c(F)]))$$

where $\sigma(\cdot)$ is the sigmoid activation function, $+$ denotes concatenation, $AvgPool_c$ and $MaxPool_c$ is the global average and max pooling across the channel dimension, emphasizing both globally and locally important spatial regions. The spatially modulated feature map F' is then computed by element-wise multiplication of the original feature map F and spatial attention map S:

$$F'(i,j,c) = F(i,j,c) \cdot S(i,j)$$

This mechanism ensures that the relevant spatial regions, such as small or low-contrast tumor areas, are amplified for the model's attention.

2) Channel attention

Channel Attention [32] adjusts the importance of each feature channel, which is crucial for distinguishing tumors with subtle characteristics. For the input feature map $F \in \Re^{H \times W \times C}$, channel attention computes a channel-wise attention map, $C \in \Re^{C}$, through global pooling across spatial dimensions as follows:

$$C(c) = \sigma\Big(MLP\big(AvgPool_{spatial}(F)_c + MaxPool_{spatial}(F)_c\big)\Big)$$

Here, $AvgPool_{spatial}(F)_c$ and $MaxPool_{spatial}(F)_c$ represent the global average pooling and max pooling operations across the spatial dimensions for channel c, respectively, reducing the feature map to a single value per channel. Multi-Layer Perceptron (MLP) [21] processes the averaged and maximized pooled values per channel, assigning weights that highlight the channels essential for tumor segmentation. The final feature map after channel attention, F'' is computed by applying the channel attention map to the feature map:

$$F''(i,j,c) = F'(i,j,c) \cdot C(c)$$

Channel Attention ensures that subtle tumor boundaries are emphasized by prioritizing channels linked to critical features, such as density variations.

3) Final attention-weighted feature map

The final attention-modulated feature map F'' combines both spatial and channel attention, emphasizing tumor-specific regions to improve segmentation precision. This weighted feature map is represented as

$$F''(i,j,c) = F(i,j,c) \cdot S(i,j) \cdot C(c)$$

This combined map is passed to the segmentation model, ensuring that areas crucial for accurate tumor segmentation are dynamically highlighted, thus enhancing the performance, especially in complex cases with low contrast or irregular shapes.

D. Segmentation Model

After enhancing the feature maps with the Attention Mechanism, TL-ESMNet employs U-Net and DeepLabV3 architectures for segmentation, allowing it to generate accurate tumor boundary masks. These models were chosen for their proven effectiveness in medical imaging, particularly in dealing with small and irregularly shaped tumors.

1) U-Net architecture

U-Net [7] employs encoder-decoder network with skip connections to facilitate multi-scale combination of high-level semantic data and precise spatial information. The encoder reduces features step by step using convolution layers, and the decoder increases for spatial resolution reconstruction to preserve key details for precise segmentation of tumor boundaries. In TL-ESMNet, attention-weighted feature maps $F'' \in \Re^{H \times W \times C}$ serve as the U-Net's input, which improves the network's ability to focus on subtle tumor boundaries while maintaining spatial accuracy. The encoder applies several convolution operations.

$$F_l = \sigma(Conv_l(F'')) \quad for\ layer\ 'l'$$

where $Conv_l$ represents the l-th convolution layer in the encoder and σ is the activation function (typically ReLU). These features are progressively down sampled using max pooling, whereas the decoder performs up sampling to reconstruct the spatial resolution of the image:

$$U_l = UpSample(F_l) \text{ for up-sampling in layer } l$$

The U-Net structure enhances the network's ability to focus on subtle tumor boundaries while maintaining spatial accuracy to handle shape irregularities.

2) DeepLabV3 architecture

DeepLabV3 [8] has Atrous Spatial Pyramid Pooling (ASPP) for detecting multi-scale contexts, required to segment the tumors of varied sizes and shapes. The ASPP uses various dilation rates in the convolution operation to address the feature maps of multiple scales so that DeepLabV3 can detect localized features and extensive context within a single framework. This model leverages the ASPP module that processes the attention-weighted feature map F'' at multiple scales as follows:

$$F_{asap} = \sum_{r \in \{r_1, r_2, r_3\}} Conv_{atrous}^{T}(F'')$$

where r_1, r_2, r_3 are different dilation rates applied to the atrous convolution filters, and F_{aspp} is the multiscale output feature map. The attention-weighted feature map input to DeepLabV3 enables the model to focus on important areas, making it particularly effective for segmenting various tumor shapes such as small tumor regions and larger tumor structures in mammographic images.

3) Integration of U-Net and DeepLabV3

In TL-ESMNet, attention-enhanced feature maps $F'' \in \Re^{H \times W \times C}$ are processed by U-Net and DeepLabV3 simultaneously. Each model independently generates pixel-wise segmentation masks [18], denoted as M_{U-Net} and $M_{DeepLab}$, respectively. The final segmentation output was a weighted fusion of the masks.

$$M_{combined}(i,j) = \alpha \cdot M_{U-Net}(i,j) + (1-\alpha) \cdot M_{DeepLab}(i,j)$$

where α is a weighting factor between 0 and 1 that controls the contribution of each model and $M_{combined}(i,j)$ is the final segmentation mask at pixel (i,j). This integration leveraged U-Net's fine detail capture and DeepLabV3's contextual awareness, producing a more accurate segmentation mask adapted to diverse tumor characteristics.

4) Multi-scale feature extraction for irregular shapes

The integrated outputs from U-Net and DeepLabV3 ensure that features at multiple scales are fused, allowing the final segmentation mask to capture both small irregularly shaped tumors and more expansive regions [21]. This fused multiscale feature map F_{multi} combines the complementary strengths of both architectures, resulting in enhanced segmentation performance for complex tumor structures, as follows:

$$F_{multi} = \sum_{s \in \{s_1, s_2, s_3\}} \mathrm{Re}\,sample(F_s'')$$

By effectively combining the features at different scales, this approach can enhances the segmentation accuracy, particularly for the intricate tumor shapes.

Each pixel in the final mask $M_{combined}$ is assigned a probability of being part of the tumor, with additional post-processing steps that refine the mask to reduce the noise and boost the accuracy. This comprehensive approach makes the TL-ESMNet highly effective in detecting the small and irregular tumors, while also preserving the necessary context for the consistent and reliable segmentation across varying imaging conditions. The integration of multi-scale features from U-Net and DeepLabV3 guarantees that, our TL-ESMNet excels in both detail and adaptability, ultimately producing the robust and reliable segmentation results.

E. Dynamic Fusion in ESM-Net

Based on the segmentation outputs from U-Net and DeepLabV3, the Dynamic Fusion Mechanism [33] in TL-ESMNet intelligently merges the advantages of both models. In contrast to static fusion techniques, this dynamic method modifies the influence of each model according to specific image features like tumor size, shape, and contrast, thereby providing customized segmentation results.

1) Dynamic fusion mechanism overview

The dynamic fusion mechanism adaptively weighs the outputs from U-Net and DeepLabV3 according to their strengths [30]. U-Net excels in segmenting smaller, well-defined tumors, whereas DeepLabV3 captures broader, irregularly shaped tumors more effectively. Let the segmentation masks from U-Net and DeepLabV3 be denoted as $M_{U-Net}(i,j)$ and $M_{DeepLab}(i,j)$, respectively, for a pixel at position (i,j), respectively. The Dynamic Fusion Mechanism computes the final segmentation mask, $M_{fused}(i,j)$ based on the following weighted sum:

$$\begin{aligned} M_{fused}(i,j) = {} & W_{U-Net}(i,j) \cdot M_{U-Net}(i,j) \\ & + W_{DeepLab}(i,j) \cdot W_{DeepLab}(i,j) \end{aligned}$$

where $w_{U-Net}(i,j)$ and $w_{DeepLab}(i,j)$ are the dynamically computed weights assigned to U-Net and DeepLabV3 outputs, respectively.

2) Weight assignment based on image characteristics

The dynamic weights of $w_{U-Net}(i,j)$ and $w_{DeepLab}(i,j)$ are not fixed and are determined by factors such as tumor size and contrast [29]. For small, well-defined tumors, the dynamic weight favors M_{U-Net}, whereas for larger, irregular tumors, it shifts to emphasize $M_{DeepLab}$. This adaptive weighting was computed using a logistic function, as follows:

$$w_{U-Net}(i,j) = \frac{1}{1 + \exp(-\gamma(size_{tumor}(i,j) - T))}$$

where γ is a scaling parameter and T is the threshold for tumor size. A similar equation can be applied for DeepLabV3, with the weight favoring larger tumors:

$$w_{DeepLab}(i,j) = 1 - w_{U-Net}(i,j)$$

These adaptive weighting functions facilitate the smooth adjustments in the weight distribution, which enables the framework to customize the fusion process according to the specific requirements of each image.

3) Improvement in segmentation accuracy and robustness

By leveraging the unique strengths of U-Net and DeepLabV3, a dynamic fusion mechanism offers several benefits.

- Adaptability to Tumor Size: U-Net is prioritized for small tumors and DeepLabV3 is prioritized for large, irregular shapes.
- Handling Contrast Variability: Adjusts weights based on image contrast and optimizes segmentation accuracy in low-contrast regions.

- Balancing Detail with Context: Merges U-Net's fine detail with DeepLabV3's broader context, ensuring that both small and large tumors are accurately segmented.

This adaptive fusion in TL-ESMNet enhances the segmentation performance, providing robust and accurate tumor delineation across diverse imaging conditions.

IV. Experimental Setup

A. Datasets

The effectiveness of the TL-ESMNet framework was evaluated using two well-known mammogram datasets, DDSM [34] and Inbreast [35], chosen for their ability to provide diverse imaging conditions, breast densities, and tumor characteristics that challenge and validate the segmentation models.

The performance of the TL-ESMNet framework was assessed using two widely recognized mammogram datasets are DDSM [34] and Inbreast [35]. These datasets were selected because they offer a variety of imaging conditions in breast densities, and tumor characteristics that help challenge and validate the models.

DDSM: This database [34] contains around 2620 cases with more than 10,000 images for both malignant and benign lesions. Each image is annotated with respective pixel-level ground truth (mask), which provides the detailed information for segmentation purpose. The DDSM film-based images were digitized at different resolutions, with metadata detailing about the breast density and abnormality descriptions. This makes the DDSM as an ideal choice for evaluating the proposed model's ability to generalize across different image quality variations and lesion types.

INbreast Dataset: Consisting of 115 cases with 410 high-resolution digital mammography images (up to 3328×4084 pixels), Inbreast [35] contains detailed annotations of various lesion types, including masses, calcifications, and architectural distortions. The high resolution supports the precise segmentation of small structures, such as calcifications, providing a challenging benchmark for detecting the fine tumor details.

These datasets are comprehensively tested the adaptability and accuracy of TL-ESMNet in diverse imaging scenarios, offering robust validation across the lesion types and imaging resolutions.

B. Data Preprocessing

Several data augmentation techniques were used in preprocessing to enhance the adaptability of TL-ESMNet across datasets and simulating different imaging conditions. These included random rotations between –15° and 15°, as well as horizontal flips with a 50% probability, to account for potential variations in mammogram orientation. Contrast and brightness adjustments were applied to account for the intensity variations between the images. Gaussian noise was introduced to assist the model in handling the noise that often appears in mammograms due to various imaging protocols.

To ensure consistency, all images were normalized to a standard intensity range and resized to uniform dimensions. This approach allowed for seamless processing of both DDSM and INbreast images, even though they had different resolutions. This preprocessing technique boosts the model's strength and broad applicability under various imaging modalities.

C. Training Setup

The training process for TL-ESMNet utilized a combination of binary cross-entropy and Dice loss to tackle with class imbalance and enhance segmentation accuracy. Binary cross-entropy focused on the disparity between the tumor and background classes, whereas the Dice loss improved accuracy in small tumor areas by increasing the overlap between the predicted and actual segmentation masks [19].

For evaluating model stability, 3-fold cross-validation was employed, splitting the data into 70% for training, 15% for validation, and 15% for testing with uniform class distribution. In every fold, the model was retrained with distinct subsets for validation and testing, and the average performance across the folds provided a comprehensive evaluation. For optimization, the Adam optimizer was used, starting with an initial learning rate of 0.001. The learning rate decayed by a factor of 0.1 every 30 epochs to maintains stability. A batch size of 16 was used to strike a balance between the memory usage and efficiency, particularly for high resolution images of the INBreast dataset [35]. TL-ESMNet was trained for a maximum of 100 epochs, and early stopping was used to avoid overfitting.

All experiments were executed on an NVIDIA Tesla V100 GPU with 32 GB of memory, providing the necessary resources for efficient training and processing of high-resolution mammographic data. This framework was aimed to achieve the precise and dependable segmentation under a range of imaging conditions.

D. Evaluation Metrics

TL-ESMNet segmentation performance was evaluated according to a set of significant metrics [36] in an attempt to provide the accuracy and reliability of a comprehensive assessment under the various medical imaging conditions:

- DCE and IoU: These metrics evaluate the overlap between the predicted and actual tumor regions. DCE is particularly sensitive to small regions, while IoU provides an efficient assessment by penalizing both false positives and false negatives.
- Precision and Recall: Precision evaluates how accurately the tumor pixels are predicted, aiming to reduce false positives. Recall, on the other hand, prioritizes detecting the true positives to minimize the missed tumor regions.
- F1-Score: It is the harmonic mean of Precision and Recall, offers a balanced assessment of detection accuracy and the rate of false detections.
- Pixel Accuracy (PA): This metric calculates the proportion of correctly classified pixels, offering an overall indication of the segmentation quality.

All these metrics collectively allow for overall assessment of TL-ESMNet segmentation quality with both global accuracy and local information over an extensive range of tumors and imaging settings. Such an extensive group of robust measures provides the model with confidence while operating in a clinical setting.

E. Evaluation Procedure

A rigorous testing procedure was employed in order to validate and prove the stability of TL-ESMNet segmentation.

- 3-Fold Cross-Validation: To minimize overfitting and guarantee consistent performance, we performed 3-fold cross-validation. Each fold was assessed separately, and the average performance across all folds was reported.
- Multiple Runs: Each experiment was conducted five times, using different random initializations and data shuffles. The results were averaged across these runs to account for variations and enhance the stability of the reported metrics.
- Baseline Comparison: The performance of TL-ESMNet was compared against the baseline models such as U-Net [7], DeepLabV3 [8], and Attention U-Net [22], utilizing the same metrics to ensure a consistent and fair evaluation.
- Quantitative and Qualitative Analysis: Besides quantitative measures, qualitative assessment was also conducted by visually inspecting the TL-ESMNet segmentation result against ground truth masks. Performance on difficult cases, like low-contrast or malformed tumors, was evaluated.

This evaluation strategy offered a rigorous examination of TL-ESMNet's performance on different aspects of tumors and imaging conditions.

F. Baseline Methods

The performance of TL-ESMNet was compared against the benchmark models for medical image segmentation, each chosen based on its novel approach and well-documented performance in the field of segmentation.

- U-Net: it is recognized for its encoder-decoder architecture [7], which excels at capturing intricate details, making it an ideal choice for segmenting small structures in medical images.
- DeepLabV3+: Includes an encoder-decoder model to the original DeepLabV3 [8], enabling it to learn the multi-scale context and achieves better performance when there are diversely shaped and sized tumors.
- Swin-Unet: Utilizes Swin Transformers [37] instead of traditional convolutions, enhancing global context capture through self-attention mechanisms.
- TransUNet [38]: It integrates the transformer-based encoders with U-Net decoders, utilizing both global context and precise spatial segmentation.
- nnU-Net [39]: Automatically adjusts itself according to the dataset, fine-tuning U-Net architectures to enhance performance across various segmentation tasks.
- Attention R2U-Net [40]: Incorporates recurrent residual units and attention mechanisms into U-Net, enhancing the focus on important areas for better segmentation accuracy.

These models present a comprehensive benchmark that represents different segmentation strategies, showcasing the strengths of TL-ESMNet in effectively dealing with complex and varied tumor features.

V. Results and Discussion

A. Quantitative Results Comparison

The quantitative results from the 3-fold cross-validation shown in Tables I and II for the Inbreast [35] and DDSM [34] datasets highlight the impressive segmentation abilities of TL-ESMNet under various imaging conditions. TL-ESMNet recorded an average Dice Coefficient (DCE) of 0.88 on the Inbreast dataset and 0.91 on the DDSM, demonstrating its flexibility with both smaller and larger datasets.

1) INbreast dataset cross-validation

On the INbreast dataset [35], TL-ESMNet effectively addressed the variability inherent in limited data (see Table I). DCE scores ranged from the 0.86 to 0.91 across folds, demonstrating its capacity to adapt to varying tumor characteristics. The model performed best with Fold 2, accurately delineating complex tumor boundaries with an Intersection over Union (IoU) of 0.85, attributed to its attention mechanisms and dynamic fusion that prioritize regions essential for segmentation. The Precision and Recall metrics for the INbreast dataset show notable variability, with precision values between 0.87 and 0.92, and Recall values ranging from 0.83 to 0.89. This performance variability across different folds emphasizes how the TL-ESMNet design effectively tackles the challenges posed by small datasets, consistently reaching an average Pixel Accuracy (PA) of 94%.

TABLE I. Detailed 3-Fold Cross-Validatio n Results for TL-ESMNet Segmentation on the INbreast Dataset

Fold	Dice Coefficient (DCE)	Intersection Over Union (IoU)	Precision	Recall	F1-Score	Pixel Accuracy (PA)
Fold 1	0.88	0.82	0.89	0.86	0.87	0.94
Fold 2	0.91	0.85	0.92	0.89	0.90	0.95
Fold 3	0.86	0.80	0.87	0.83	0.85	0.93
Average	**0.88**	**0.82**	**0.89**	**0.86**	**0.87**	**0.94**

TABLE II. DETAILED 3-FOLD CROSS-VALIDATION RESULTS FOR TL-ESMNET SEGMENTATION ON THE DDSM DATASET

Fold	Dice Coefficient (DCE)	Intersection over Union (IoU)	Precision	Recall	F1-Score	Pixel Accuracy (PA)
Fold 1	0.92	0.86	0.93	0.9	0.91	0.96
Fold 2	0.89	0.84	0.9	0.87	0.88	0.95
Fold 3	0.91	0.85	0.92	0.89	0.9	0.96
Average	**0.91**	**0.85**	**0.92**	**0.89**	**0.9**	**0.96**

2) DDSM dataset cross-validation

Conversely, the DDSM dataset [34] yielded more stable results due to its larger and more diverse dataset. As illustrated in Table II, TL-ESMNet consistently achieved high performance across all folds, with DCE values ranging from 0.89 to 0.92 and an average IoU of 0.85, demonstrating its effectiveness in accurately capturing complex tumor boundaries. The model's attention mechanisms improved the focus on the tumor regions, while dynamic fusion enabled it to adapt effectively to various tumor shapes and contrasts, resulting in an average Pixel Accuracy of 96%.

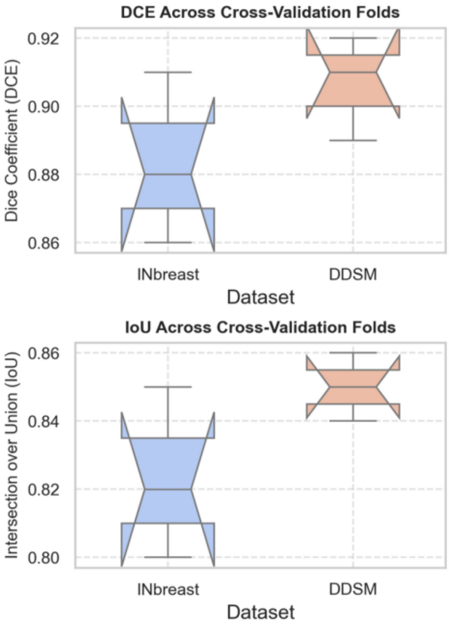

Fig. 2. Segmentation performance comparison of TL-ESMNet across cross-validation folds: DCE and IoU on INbreast and DDSM datasets.

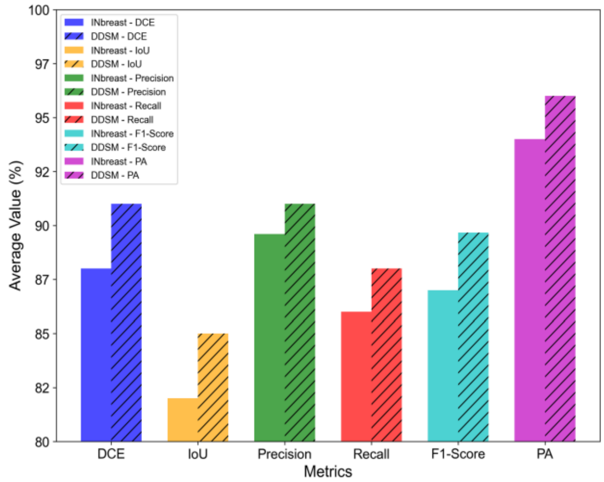

Fig. 3. Comparison of TL-ESMNet's average segmentation metrics between INbreast and DDSM datasets.

Fig. 2 displays the box plots of DCE and IoU across the folds, illustrating the performance variability and stability of TL-ESMNet over multiple cross-validation runs. Each box represents the distribution of DCE or IoU values across the three folds for each dataset. Similarly, the average comparison results presented in Fig. 3 emphasize the robustness of the TL-ESMNet when applied to a larger dataset, underscoring its capacity for generalization and reliability.

3) Comparison with baseline methods

Tables III and IV compare TL-ESMNet with six baseline models (U-Net [7], DeepLabV3+ [8], Swin-Unet [37], TransUNet [38], nnU-Net [39], and Attention R2U-Net [40]). TL-ESMNet segmentation performance consistently surpassed these models, particularly on the INbreast dataset, where the data scarcity become challenge for the other models.

TABLE III. BREAST TUMOR SEGMENTATION PERFORMANCE COMPARISON OF TL-ESMNET WITH BASELINE MODELS ON THE INBREAST DATASET (3-FOLD AVERAGE RESULTS)

Model	Dice Coefficient (DCE)	Intersection over Union (IoU)	Precision	Recall	F1-Score	Pixel Accuracy (PA)
TL-ESMNet	0.88	0.82	0.89	0.86	0.87	0.94
U-Net	0.77	0.71	0.79	0.74	0.76	0.88
DeepLabV3+	0.80	0.74	0.82	0.78	0.79	0.90
Swin-Unet	0.81	0.75	0.83	0.79	0.80	0.91
TransUNet	0.78	0.72	0.80	0.76	0.77	0.89
nnU-Net	0.79	0.73	0.81	0.77	0.78	0.89
AttenR2U-Net	0.80	0.74	0.82	0.78	0.79	0.90

TABLE IV. COMPARISON OF TL-ESMNET WITH BASELINE MODELS ON THE DDSM DATASET (3-FOLD AVERAGE RESULTS)

Model	Dice Coefficient (DCE)	Intersection over Union (IoU)	Precision	Recall	F1-Score	Pixel Accuracy (PA)
TL-ESMNet	0.91	0.85	0.92	0.89	0.90	0.96
U-Net	0.84	0.78	0.85	0.82	0.83	0.92
DeepLabV3+	0.87	0.82	0.88	0.85	0.86	0.93
Swin-Unet	0.89	0.83	0.9	0.87	0.88	0.94
TransUNet	0.88	0.82	0.89	0.85	0.87	0.94
nnU-Net	0.86	0.80	0.87	0.84	0.85	0.93
AttenR2U-Net	0.87	0.81	0.88	0.85	0.86	0.93

a) Baseline models comparison on inbreast dataset

On the INbreast dataset, which features high-resolution images but has a limited number of samples, TL-ESMNet consistently surpassed the baseline models in all metrics (see Table III and Fig. 4). The Dice Coefficient (DCE) for TL-ESMNet was 0.88, significantly higher than the best-performing baseline, Swin-Unet [37], which achieved 0.81. Other models, such as U-Net [7] and TransUNet [38], struggled more with limited data, with DCE values of 0.77 and 0.78, respectively.

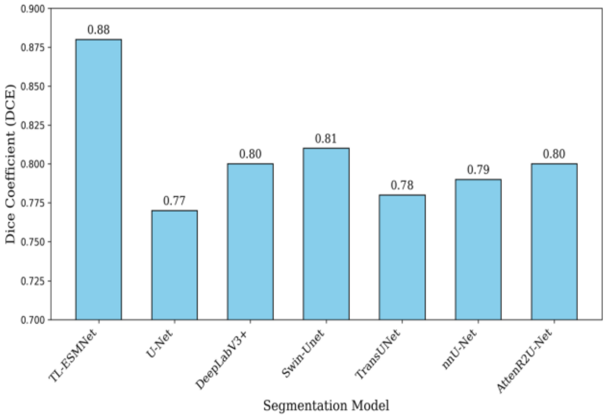

Fig. 4. Comparison of breast tumor segmentation Dice Coefficient (DCE) across models for INbreast dataset.

The IoU metric showed a similar trend, with TL-ESMNet achieving 0.82, while the baselines ranged from 0.71 (U-Net [7]) to 0.75 (Swin-Unet [37]). This indicates that the TL-ESMNet is better than its counterparts at segmenting tumors with precise boundaries, even with limited training samples. The model's superior Precision of 0.89 and Recall of 0.86 also suggest that TL-ESMNet manages to minimize the both false positives and false negatives, which assures more reliable segmentation outputs. Baseline models exhibited greater variations in these metrics. U-Net [7] and TransUNet [38] showed lower recall values (0.74 and 0.76, respectively), underscoring their difficulty in handling the limited data from the INbreast dataset.

b) Baseline models comparison on DDSM dataset

On the larger DDSM dataset [34], while all models performed more consistently, TL-ESMNet maintained a lead with a DCE of 0.91, compared to 0.89 for Swin-Unet [37] and 0.88 for TransUNet [38]. TL-ESMNet's ability to adjust to different imaging conditions using dynamic fusion and attention mechanisms made it especially good at segmenting tumors with complex boundaries. As a result, it achieved a higher IoU and Pixel Accuracy compared to the baselines (see Table IV and Fig. 5). U-Net [7] demonstrated reasonable performance on DDSM, with a DCE of 0.84. Despite this, it was outperformed by the ore advanced architectures, like DeepLabV3+ [8] and Swin-Unet [37], which achieved the DCE values of 0.87 and 0.89, respectively.

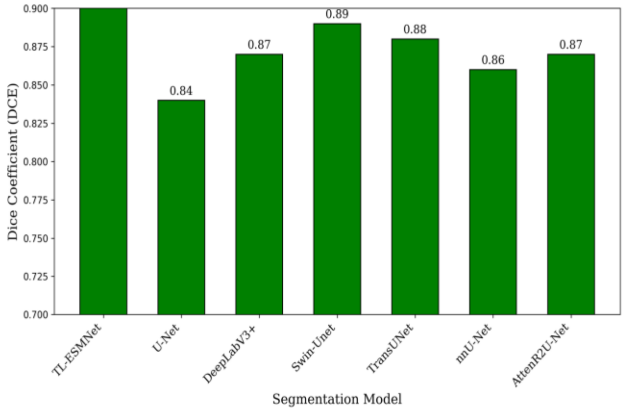

Fig. 5. Comparison of Breast tumor segmentation Dice Coefficient (DCE) across models for DDSM dataset.

The IoU values showed a similar trend, with TL-ESMNet achieving a score of 0.85, while the baseline models ranged from 0.78 (U-Net [7]) to 0.83 (Swin-Unet [37]). Additionally, the Precision and Recall values for all models were higher for the DDSM dataset [34], due to the increased number of training examples. However, TL-ESMNet maintained a small lead, with a precision of 0.92 and recall of 0.89, compared to Swin-Unet's 0.90 and 0.87, and TransUNet's 0.89 and 0.85. All models exhibited high Pixel Accuracy (PA), between 92% and 96%, with TL-ESMNet leading in this comparison as well.

Tables III and IV indicate that TL-ESMNet has a clear edge over the baseline models on both datasets, especially on the Inbreast dataset (see Table III). The baseline models struggle with the limited data, which results in lower performance in terms of segmentation accuracy and overlap metrics. On the DDSM dataset (see Table IV), where more training data is available, TL-ESMNet still maintains its lead, though the gap between models is smaller. The segmentation performance results from the baseline models shown in Fig. 6 highlight the TL-ESMNet's robustness and adaptability. It performs exceptionally well in both limited and extensive data scenarios, demonstrating the reliable segmentation accuracy across a diverse array of tumor characteristics.

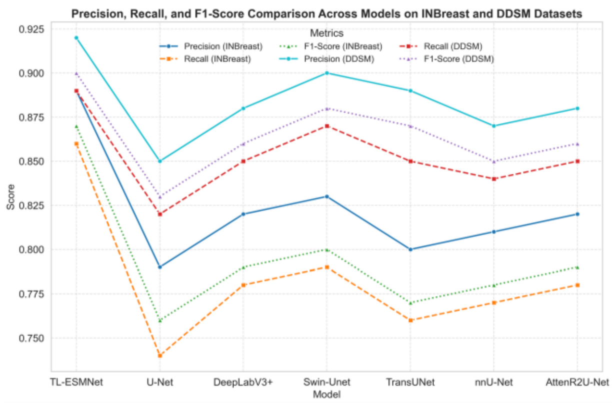

Fig. 6. Breast tumor segmentation performance comparisons of precision, recall, and F1-Score across segmentation models on INBreast and DDSM datasets.

B. Qualitative Results Comparison

This section presents qualitative results comparing TL-ESMNet with same six baseline models on the Inbreast [35] and DDSM [34] datasets, showing its performance across challenging tumor types. The chosen examples included small, irregularly shaped, and low-contrast tumors to illustrate the superior adaptability and segmentation accuracy of TL-ESMNet in diverse cases.

1) Visual comparisons with ground truth

Fig. 7 shows examples of segmentation outputs from each model, along with the ground truth, allowing for a clear comparison of TL-ESMNet's performance against the baseline models. TL-ESMNet consistently produces the segmentation masks that closely align with the ground truth data, especially in scenarios involving low contrast or irregular shapes. With the use of attention mechanisms and dynamic fusion, TL-ESMNet shows the improved boundary definition, effectively in highlighting the tumor areas that might otherwise be overlooked. In low-contrast tumor cases from the INbreast dataset, TL-ESMNet's attention focused on critical regions led to well-defined tumor boundaries segmentation, whereas the baseline models often blurred or partially missed these areas.

2) Performance on small and irregular tumors

In complex cases, TL-ESMNet's capability to deal with small and irregular tumors is remarkable, as shown in Figs. 7 and 8. For small granular tumor structures, models like U-Net [7] and DeepLabV3+ [8] often miss finer details, leading to incomplete segmentations. Swin-Unet [37] and TransUNet [38] offer improvements through transformer-based mechanisms, but still face limitations due to the lack of a specialized focus on subtle regions. In contrast, TL-ESMNet utilizes its spatial and channel attention mechanisms to dynamically prioritize tumor-specific features. This results in more accurate segmentation of small and irregular boundaries, especially in cases from the INbreast dataset, where other models encounter difficulties with low contrast.

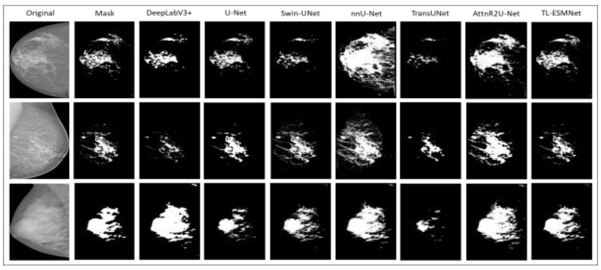

Fig. 7. Qualitative comparison of segmentation accuracy on irregularly shaped tumors in the INbreast dataset.

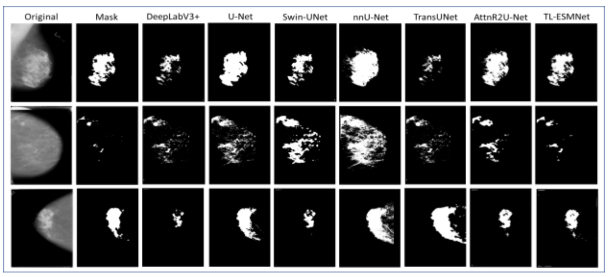

Fig. 8. Qualitative comparison of segmentation accuracy on irregularly shaped tumors in the DDSM dataset.

3) Limitations and areas for improvement

Although TL-ESMNet generally performs well, a few cases reveal areas for further enhancement, particularly in extremely low-contrast images. For instance, TL-ESMNet sometimes slightly over segments the regions with a very low boundary contrast, extending the tumor area beyond its actual borders. Similar difficulties were also observed in the baseline models under these conditions. These limitations suggest that enhancing the attention mechanisms of TL-ESMNet to distinguish finer boundary differences could improve its accuracy. Additionally, using extra preprocessing techniques like contrast normalization might aid in preserving boundary clarity.

From the total qualitative analysis, we find that TL-ESMNet demonstrates robustness in handling diverse segmentation tasks. Improving attention mechanisms and adding preprocessing steps could make it even more reliable for clinical use.

C. Detailed Discussion of Results

This section offers an in-depth discussion of TL-ESMNet's quantitative results, linking them to the objectives of the study, and showcasing its contributions to breast tumor segmentation.

1) Effectiveness of attention mechanism

Spatial and channel attention mechanisms are crucial for the effectiveness of TL-ESMNet, particularly when dealing with low-contrast and complex scenarios. These mechanisms enable the model to dynamically concentrate on important areas, accurately define subtle tumor boundaries, and improve feature selection in densely packed tissue regions. This method enables TL-ESMNet to achieve high segmentation accuracy, especially in situations where regular models like U-Net [7] have difficulty capturing intricate details. By focusing on tumor-relevant areas using spatial and channel attention, as illustrated in Fig. 9, the attention mechanisms facilitate accurate segmentation, giving TL-ESMNet a significant edge in difficult scenarios.

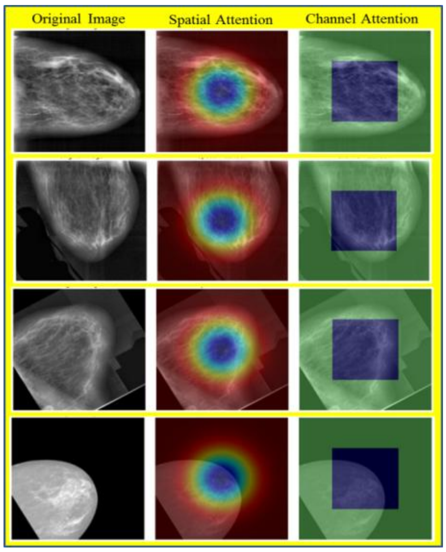

Fig. 9. Visualization of spatial and channel attention mechanisms in TL-ESMNet.

2) Impact of dynamic fusion

Dynamic fusion in TL-ESMNet improves the segmentation reliability by adaptively adjusting the influence of U-Net [7] and DeepLabV3 [8] according to the tumor characteristics, as shown in Fig. 10. The fusion mechanism assigns higher weights to the U-Net for small and irregular tumors, as it can capture fine details. For larger tumors, DeepLabV3 is preferred, as its multi-scale feature extraction performs better. Furthermore, when contrast conditions vary, the model adjusts the fusion weights dynamically to enhance the segmentation accuracy:

- Low Contrast: U-Net receives a higher weight since it preserves tumor boundaries effectively, compensating for weak intensity variations that DeepLabV3 may overlook.
- Medium Contrast: The fusion weights are balanced, leveraging U-Net's fine-grained segmentation and DeepLabV3's contextual awareness for optimal performance.
- High Contrast: DeepLabV3 takes precedence as it efficiently segments well-defined tumor regions, while U-Net complements by refining boundary details.

The adaptive fusion mechanism allows the TL-ESMNet to effectively capture the tumors across various imaging conditions and improving the segmentation accuracy where other models are facing challenges.

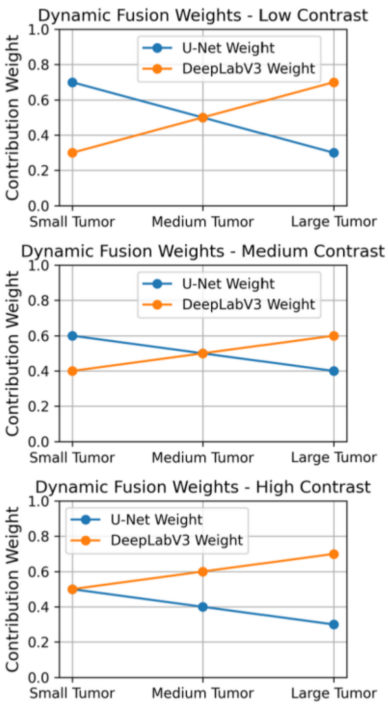

Fig. 10. Dynamic fusion weights of u-net and deeplabv3 based on tumor size and contrast.

3) Generalization across datasets and imaging modalities

The TL-ESMNet design shows impressive generalization on both the INbreast and DDSM datasets, attaining the high accuracy even with the variations in imaging protocols, resolutions, and types of lesions. With consistently high precision and recall across both datasets, the model demonstrates its versatility and robustness. This makes it effective in processing both high-resolution digital images and lower-resolution film-based mammograms. This ability to generalize demonstrates that TL-ESMNet is well-suited for real-world applications, where a variety of imaging data is frequently encountered.

While this study mainly looks at the mammographic breast tumor segmentation, we recognize how important it would be to expand the framework to the other imaging techniques like MRI and ultrasound [41].

Applying TL-ESMNet to different imaging modalities create several challenges. Domain adaptation is a major concern because each imaging modality has distinct characteristics related to contrast, resolution, and noise levels. For instance, mammograms are high-resolution grayscale images that display clear contrast variations, while MRI provides multi-sequence soft-tissue contrast. On the other hand, ultrasound images are often influenced by speckle noise, which makes segmentation more challenging. In addition, the pre-trained models like ResNet and VGG16, which were originally trained on natural images, might need some extra fine-tuning to better capture the unique tissue characteristics of medical imaging.

Despite these challenges, we believe that TL-ESMNet's attention-driven segmentation and dynamic fusion mechanisms will be advantageous across different modalities. In MRI, the spatial and channel attention modules have the potential to improve soft-tissue contrast differentiation, leading to better segmentation accuracy. Similarly, in ultrasound imaging, the attention mechanisms could help to filter out the noise and make thetumor boundaries clearer, is addressing a common challenge in CNN-based methods.

4) Comparison with baseline models

When compared to baseline models, TL-ESMNet clearly outperforms them. This is particularly evident in its ability to handle limited and low-contrast data, such as in the INbreast dataset [35]. With advanced attention and fusion mechanisms, TL-ESMNet focuses on the important regions and adapt to different tumor characteristics. It outperforms models like Swin-Unet [37] and DeepLabV3+ [8], which don't have these adaptive features. With its advanced attention and fusion mechanisms, TL-ESMNet effectively targets relevant areas and adjusts to different tumor characteristics, surpassing the models like Swin-Unet [37] and DeepLabV3+ [8], which do not possess these adaptive features.

Although TL-ESMNet shows impressive segmentation performance, there are still some limitations to consider. TL-ESMNet facing the challenges with extremely low-contrast mammograms, where the tumor boundaries are hard to distinguish. In addition, its dependence on manually labeled datasets leads to the variability in ground truth annotations. Moreover, the relatively small size of the INbreast dataset presents a risk of overfitting. In the end, TL-ESMNet meets the study objectives by delivering accurate and flexible segmentation across different tumor types and imaging conditions. This underscores its ability to improve clinical outcomes in breast tumor diagnosis.

5) Computational cost analysis

To assess the computational efficiency of TL-ESMNet, its training time and GPU memory usage were compared against the baseline models on the DDSM and INbreast datasets. As shown in Fig. 11, TL-ESMNet takes more time to train due to its attention mechanisms and dynamic fusion, which require more computational resources. On the DDSM dataset, TL-ESMNet took approximately 24.7 hours for training, while on the INbreast dataset, it required 18.3 hours, demonstrating a moderate increase compared to U-Net and DeepLabV3+.

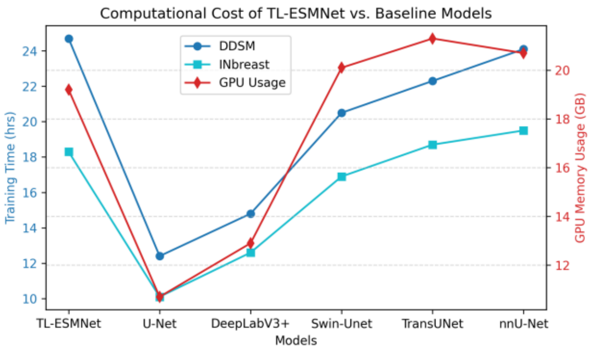

Fig. 11. Computational cost of TL-ESMNet *vs.* baseline models.

GPU memory usage analysis revealed that the TL-ESMNet consumes 19.2 GB, placing it between conventional CNN models (U-Net: 10.7 GB, DeepLabV3+: 12.9 GB) and transformer-based architectures (Swin-Unet: 20.1 GB, TransUNet: 21.3 GB). The moderate computational footprint ensures that the model remains feasible on high-performance GPUs, while providing superior segmentation accuracy. Although TL-ESMNet requires higher computational resources, its dynamic fusion and attention-based segmentation greatly enhance performance, particularly for low-contrast and irregular tumor cases. Future optimizations, such as model compression techniques (e.g., quantization, mixed precision training), could further improve the efficiency while maintaining the accuracy.

Thus, TL-ESMNet balances computational cost and segmentation precision, making it a clinically viable solution for the automated breast tumor segmentation.

6) Ablation studies

The ablation studies were conducted to assess the individual contributions of the Transfer Learning (PreTL), Attention Mechanisms, and Dynamic Fusion within the TL-ESMNet (see Fig. 12) framework. The results demonstrate that each component can significantly enhances segmentation performance, and removing any one of them leads to a notable decline in the accuracy.

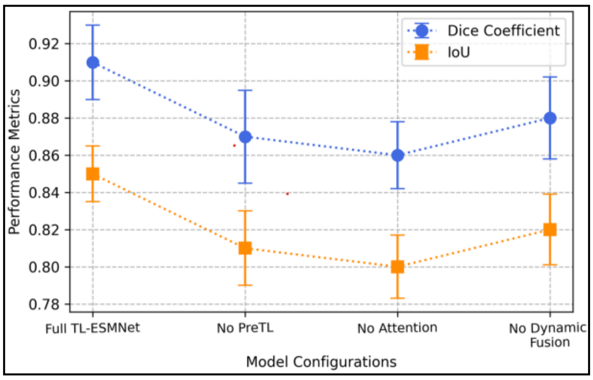

Fig. 12. Ablation studies: Impact of removing components on TL-ESMNet.

Without Transfer Learning (PreTL):

- When PreTL was removed, Dice Coefficient (DCE) dropped by approximately 3–4%, indicating that the pre-trained feature extraction plays a crucial role in identifying the tumor boundaries, particularly in low-data scenarios like INbreast.
- The model struggled to converge efficiently, requiring more epochs to achieve the comparable performance.

Without Attention Mechanisms:

- Removing the spatial and channel attention caused a drop in segmentation accuracy, especially in the case of low-contrast and irregularly shaped tumors, where the attention mechanisms are critical for refining the feature focus.
- DCE decreased by 4–5%, with IoU and Recall metrics also suffering, highlighting that the importance of the dynamic feature enhancement in improving boundary delineation.

Without Dynamic Fusion:

- When dynamic fusion was disabled, and U-Net and DeepLabV3 were used independently, segmentation performance became inconsistent across tumor types.
- In small, granular tumors, U-Net alone performed better, while DeepLabV3 was superior for larger lesions. However, without fusion, neither model adapted effectively across varying tumor sizes and contrast levels.
- The overall Dice Coefficient dropped by 3–4%, confirming that adaptive weighting is essential for generalizing across diverse tumor characteristics.

The complete TL-ESMNet framework achieves optimal segmentation performance by utilizing the PreTL for robust feature extraction. It also leverages the attention mechanisms for improved focus on tumor regions and dynamic fusion for adaptive decision-making. Eliminating any of these components results in reduced accuracy and flexibility. This underscores the value of an integrated approach.

VI. Conclusion

In this study, we presented TL-ESMNet, a cutting-edge framework for the breast tumor segmentation in mammographic images that combines the adaptive segmentation with attention mechanisms in the ESM-Net stage. By leveraging the Transfer Learning (PreTL) with pre-trained models like ResNet and VGG16, TL-ESMNet is able to extract the high-quality feature representations that are crucial for the precise segmentation. The framework employs spatial and channel attention mechanisms to dynamically emphasize the relevant areas, enhancing its performance in difficult cases, such as small, low-contrast, or irregularly shaped tumors. Our experimental results on the DDSM and INbreast datasets demonstrate the robustness and effectiveness of TL-ESMNet, which achieved impressive segmentation accuracy with a Dice Coefficient (DCE) of 0.91 on the DDSM dataset and 0.88 on the INbreast dataset. Additionally, TL-ESMNet surpassed baseline models like U-Net, DeepLabV3+, Swin-Unet, and TransUNet across various metrics, including Precision and Recall. The integration of adaptive attention mechanisms and dynamic fusion within the ESM-Net played a crucial role in this improved performance, especially in scenarios involving complex tumor shapes or limited data. Future research could investigate the potential of TL-ESMNet in other imaging techniques, like MRI or ultrasound, and enhance attention mechanisms to improve segmentation performance in various medical imaging scenarios. This proposed framework marks a significant step forward in automated breast tumor analysis, which could lead to better diagnostic accuracy and efficiency in clinical practice.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Mr. Satyanarayana Reddy Beram conducted the research and wrote the paper under the guidance of Dr. R Lalchhanhima and Dr. Ksh. Robert Singh. All authors had approved the final version.

References

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

[2] E. *Heer et al.*, "Global burden and trends in premenopausal and postmenopausal breast cancer: A population-based study," *The Lancet Global Health*, vol. 8, no. 8, pp. E1027–E1037, Aug. 2020.

[3] N. Saffari *et al.*, "Fully automated breast density segmentation and classification using deep learning," *Diagnostics*, vol. 10, no. 11, 988, Nov. 2020.

[4] J. Bai *et al.*, "Applying deep learning in digital breast tomosynthesis for automatic breast cancer detection: A review," *Medical Image Analysis*, vol. 71, 102049, July 2021.

[5] V. Romeo *et al.*, "Tumor *segmentation* analysis at different post-contrast time points: A *possible* source of variability of quantitative DCE-MRI parameters in locally advanced breast cancer," *European Journal of Radiology*, vol. 126, 108907, May 2020.

[6] C. Militello *et al.*, "Semi-automated and *interactive* segmentation of contrast-enhancing masses on breast DCE-MRI using spatial fuzzy clustering," *Biomedical Signal Processing and Control*, vol. 71, 103113, Jan. 2022.

[7] O. *Ronneberger*, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*, 2015, pp. 234–241.

[8] H. Liu, Q. *Zhang*, and Y. Liu, "Image segmentation of bladder cancer based on DeepLabv3+," in *Proc. 2021 Chinese Intelligent Systems Conf.*, 2021, pp. 614–621.

[9] N. S. Punn and S. Agarwal, "RCA-IUnet: A residual cross-spatial attention-guided inception U-Net model for tumor segmentation in breast ultrasound imaging," *Machine Vision and Applications*, vol. 33, no. 2, 27, Feb. 2022.

[10] H. Abdel-Nabi, A. Awajan, and M. Ali, "A novel ensemble strategy with *enhanced* cross attention encoder-decoder framework for tumor segmentation in whole slide images," in *Porc. 2022 13th International Conf. on Information and Communication Systems (ICICS)*, 2022, pp. 262–269.

[11] S. L. P. Sitanaboina *et al.*, "Attention 3D-CU-Net: Enhancing kidney tumor *segmentation* accuracy through selective feature emphasis," *IEEE Access*, vol. 11, pp. 139798–139810, Jan. 2023.

[12] Z. Li *et al.*, "Domain generalization for *mammography* detection via multi-style and multi-view contrastive learning," in *Proc. Medical Image Computing and Computer Assisted Intervention–MICCAI 2021*, 2021, pp. 98–108.

[13] N. Shrivastava *and* J. Bharti, "Breast tumor detection in digital mammogram based on efficient seed region growing segmentation," *IETE Journal of Research*, vol. 68, no. 4, pp. 2463–2475, Jan. 2020.

[14] G. E. Park *et al.*, "3D breast cancer segmentation in DCE-MRI using deep learning with weak annotation," *J. Magn. Reson. Imaging*, vol. 59, no. 6, pp. 2252–2262, Aug. 2023.

[15] P. Zarbakhsh, "Spatial attention mechanism and cascade feature extraction in a u-net model for enhancing breast tumor segmentation," *Applied Sciences*, vol. 13, no. 15, 8758, July 2023.

[16] Y. Bi, "The study of transfer learning *effectiveness* on medical brain tumor images," *Applied and Computational Engineering*, vol. 8, pp. 401–410, April 2023.

[17] W. Gomez-Flores and W. C. de Albuquerque Pereira, "A comparative study of pre-trained convolutional neural networks for semantic segmentation of breast tumors in ultrasound," *Computers in Biology and Medicine*, vol. 126, 104036, Nov. 2020.

[18] Y. S. Leong *et al.*, "Microcalcification discrimination in mammography using deep convolutional neural network: Towards rapid and early *breast* cancer diagnosis," *Frontiers in Public Health*, vol. 10, 875305, April 2022.

[19] L. G. Falconi *et al.*, "Transfer *learning* and fine tuning in breast mammogram abnormalities classification on CBIS-DDSM database," *Adv. Sci. Technol. Eng. Syst. J.*, vol. 5, no. 2, pp. 154–165, Mar. 2020.

[20] Y. Zhang, X. Zhang, and W. Zhu, "ANC: Attention network for COVID-19 explainable diagnosis based on convolutional block attention module," *Computer Modeling in Engineering & Sciences*, vol. 127, no. 3, pp. 1037–1058, 2021.

[21] W. Zhao *et al.*, "Adaptive *channel* and multiscale spatial context network for breast mass segmentation in full-field mammograms," *Applied Intelligence*, vol. 51, pp. 8810–8827, April 2021.

[22] W. Ding *et al.*, "*Multi*-view stereoscopic attention network for 3D tumor classification in automated breast ultrasound," *Expert Systems with Applications*, vol. 234, 120969, Dec. 2023.

[23] H. Zhang *et al.*, "DenseATT-net: Densely-connected neural network with intensive attention modules for 3D ABUS mass segmentation," in *Proc. 2022 7th International Conf. on Image, Vision and Computing (ICIVC)*, 2022, pp. 348–353.

[24] Q. Gao and M. Almekkawy, "ASU-net++: A nested u-net with adaptive feature extractions for liver tumor segmentation," *Computers in Biology and Medicine*, vol. 136, 104688, Sep. 2021.

[25] S. Poudel and S.W. Lee, "Deep multi-*scale* attentional features for medical image segmentation," *Applied Soft Computing*, vol. 109, 107445, Sep. 2021.

[26] L. Wang *et al.*, "A multi-scale method based on u-net for brain tumor segmentation," in *Proc. 2022 7th International Conf. on Communication, Image and Signal Processing (CCISP)*, 2022, pp. 271–275.

[27] Y. Zhang, Y. Li, and D. He, "U-shaped network with multi-scale feature extraction for lesion region segmentation in gastric cancer images," in *Proc. 2023 IEEE 3rd International Conf. on Information Technology, Big Data and Artificial Intelligence (ICIBA)*, 2023, pp. 234–238.
[28] R. Zou *et al.*, "An *interactive* image segmentation method based on multi-level semantic fusion," *Sensors*, vol. 23, no. 14, 6394, July 2023.
[29] F. Huang and Z. Li, "Study of two-level P2P model on self-adaptive dynamic network," *Modern Applied Science*, vol. 4, no. 7, pp. 59–64, July 2010.
[30] W. Qi, H. C. Wu, and S. C. Chan, "MDF-net: A multi-scale dynamic fusion network for breast tumor segmentation of ultrasound images," *IEEE Transactions on Image Processing*, vol. 32, pp. 4842–4855, 2023.
[31] Y. Ding *et al.*, "MVFusFra: A multi-view dynamic fusion framework for *multimodal* brain tumor segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 4, pp. 1570–1581, April 2022.
[32] A. G. Balamurugan *et al.*, "Robust brain tumor classification by fusion of deep learning and channel-wise attention mode approach," *BMC Medical Imaging*, vol. 24, no. 1, 147, June 2024.
[33] J. Yang *et al.*, "D-net: Dynamic large kernel with dynamic feature fusion for volumetric medical image segmentation," arXiv preprint, arXiv:2403.10674, 2024.
[34] M. Heath *et al.*, "*Current* Status of the Digital Database for Screening Mammography," in *Computational Imaging and Vision*, Dordrecht: Springer, 1998, ch. 1, pp. 457–460.
[35] I. C. Moreira *et al.*, "Inbreast: Toward a full-field digital mammographic database," *Academic Radiology*, vol. 19, no. 2, pp. 236–248, Feb. 2012.
[36] D. Müller, I. Soto-Rey, and F. Kramer, "Towards a guideline for evaluation metrics in *medical* image segmentation," *BMC Research Notes*, vol. 15, no. 1, 210, June 2022.
[37] H. Cao *et al.*, "Swin-unet: Unet-*like* pure transformer for medical image segmentation," in *Proc. European Conf. on Computer Vision*, 2023, pp. 205–218.
[38] J. Chen *et al.*, "Transunet: *Transformers* make strong encoders for medical image segmentation," arXiv preprint, arXiv:2102.04306, 2024.
[39] F. Isensee *et al.*, "Nnu-net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, Dec. 2020.
[40] S. *Ma*, J. Tang, and F. Guo, "Multi-*task* deep supervision on attention R2U-net for brain tumor segmentation," *Frontiers in Oncology*, vol. 11, 704850, Sep. 2021.
[41] R. A. Dar *et al.*, "Breast cancer detection using deep learning: Datasets, methods, and challenges *ahead*," *Computers in Biology and Medicine*, vol. 149, 106073, Oct. 2022.