Proceedings of the 5th International Conference on Biological Engineering and Medical Science
DOI: 10.54254/2753-8818/99/2025.23012

A Novel Deep Learning Approach for Breast Cancer Ultrasound Image Segmentation

Yuxuan Zeng

Medical College, TianJin University, Tianjin, China
Cub2023@163.com

Abstract: Ultrasound imaging is a widely accessible, cost-effective, and radiation-free modality for early breast cancer screening, particularly valuable in resource-limited settings. In this study, we propose a modified attention U-Net architecture for the automated segmentation of breast lesions in ultrasound images. Leveraging attention gate mechanisms within the U-Net framework, the model adaptively suppresses background noise and enhances lesion-specific feature representation, yielding improved performance over conventional segmentation methods. Evaluated on a dataset comprising 780 annotated breast ultrasound images, the proposed model achieved a test accuracy of 98.24% and demonstrated superior Intersection over Union (IoU) metrics compared to the baseline U-Net. Interpretability analyses using Grad-CAM confirmed that the network effectively localizes diagnostically relevant regions while avoiding false positives. These findings underscore the clinical potential of attention-enhanced deep learning approaches for precise and interpretable breast cancer segmentation in ultrasound imaging.

Keywords: breast cancer, ultrasound image, image segmentation, deep learning, attention U-Net

1. Introduction

Ultrasound imaging plays a pivotal role in the early screening and diagnosis of cancer due to its non-invasive nature, cost-effectiveness, and portability. Compared to high-cost imaging modalities such as Computed Tomography (CT) and Magnetic Resonance Imaging (MRI), ultrasound examinations typically cost less and do not involve ionizing radiation. These advantages make ultrasound an ideal imaging tool for widespread use, especially in resource-limited or primary care settings.

Breast cancer remains the most prevalent malignancy among women worldwide [1]. Its early detection and precise treatment are critical determinants of patient prognosis and survival rates. Early-stage detection can elevate the five-year survival rate to over 90%, whereas late-stage diagnosis often results in survival rates dropping below 25%. This highlights the tremendous clinical value and research significance of utilizing ultrasound imaging for the early and accurate diagnosis of breast cancer.

Integrating ultrasound imaging with artificial intelligence, particularly deep learning techniques, has emerged as a promising direction to enhance diagnostic accuracy. Automated and precise segmentation of tumor regions in ultrasound images can significantly improve diagnostic efficiency and assist clinicians in treatment planning and surgical guidance.

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/99/2025.23012

However, accurate medical image segmentation in clinical and surgical contexts remains a challenging task. Deep learning-based segmentation methods typically require large volumes of high-quality annotated data [2]. Yet, such data is scarce in real-world medical settings, and the annotation process is both time-consuming and resource-intensive. Furthermore, the inherent complexity and heterogeneity of medical images—such as gradual gray-level transitions at tumor boundaries—pose additional difficulties in differentiating between pathological and healthy tissues. Traditional algorithms often struggle to capture these subtle variations effectively.

In recent years, deep learning models have made significant strides in medical image segmentation, with the U-Net architecture standing out as a milestone. Featuring a symmetric encoder-decoder design, U-Net excels at capturing hierarchical semantic features through progressive down-sampling in the encoder and spatial resolution recovery in the decoder via transposed convolutions. A key innovation of U-Net is the use of skip connections, which concatenate multi-scale features from corresponding encoder layers to decoder layers, thereby preserving fine-grained details crucial for accurate segmentation. Its end-to-end structure, without fully connected layers, allows for processing images of arbitrary sizes and has led to its widespread adoption in various medical segmentation tasks.

Despite its success, U-Net still faces limitations when applied to complex medical segmentation scenarios. The conventional skip connection mechanism relies on static feature fusion, which lacks the ability to discriminate between relevant targets and background noise. This shortcoming becomes especially problematic in anatomically complex or low-contrast regions, such as small or indistinct lesions. Additionally, to handle targets of varying sizes, conventional U-Net models often require cascaded architectures, which increase computational cost and reduce generalizability. The inherent inconsistency between low-level texture features and high-level semantic representations can also lead to suboptimal feature fusion, limiting the model's capacity to adapt to heterogeneous and multi-scale targets.

To address these challenges, this study explores the use of an attention U-Net architecture for breast cancer ultrasound image segmentation. By incorporating attention gating mechanisms into the skip connections, the model selectively fuses encoder features guided by decoder-generated attention maps. This dynamic attention mechanism enhances the model’s ability to suppress background noise, focus on small lesion features, and eliminate the need for complex cascaded structures. As a result, the proposed method achieves efficient and precise end-to-end segmentation tailored for breast ultrasound imaging.

2. Previous works

Early image segmentation techniques primarily relied on manual feature extraction and mathematical modeling. Representative methods include threshold-based segmentation and histogram analysis [3], fuzzy C-means (FCM) clustering based on regional features [4], and the watershed algorithm [5]. These approaches introduced spatial continuity constraints through models such as Markov Random Fields (MRF), which still hold value in specific applications such as multi-organ segmentation in abdominal CT scans. However, their performance is often limited by sensitivity to noise, dependency on manually tuned parameters, and inadequate representation of complex anatomical structures.

Traditional medical image segmentation methods are fundamentally based on hand-crafted features and mathematical models. They typically analyze image intensity, gradients, or regional similarity to separate target structures, but are constrained by noise susceptibility, parameter dependence, and limited ability to handle complex anatomical variability. For example, thresholding methods, such as Otsu’s method, segment images by optimizing inter-class variance on grayscale histograms, showing good generality in bi-modal or multi-modal images, but are easily affected by noise. Watershed algorithms simulate a “flooding” process on the gradient image to produce segmentation boundaries. The marker-controlled watershed variant reduces over-segmentation by

DOI: 10.54254/2753-8818/99/2025.23012

using pre-defined markers and is effective in separating connected tissues or organs. However, it is sensitive to gradient noise and often requires manual intervention. Edge detection techniques like the Canny operator identify anatomical boundaries using gradient magnitude and dual-threshold linking. When combined with sub-pixel interpolation and curve fitting, these methods can yield high-precision boundaries, but they tend to produce fragmented edges and struggle with complex topological structures.

Despite their usefulness in certain controlled scenarios, traditional segmentation methods are severely limited by their shallow feature representation and poor generalization. These approaches typically rely on manually designed grayscale, gradient, or texture features, which are insufficient for modeling low-contrast and multi-scale anatomical structures commonly found in medical images. Moreover, they are highly sensitive to noise, artifacts, and variations in imaging parameters. Traditional algorithms also require manual parameter tuning for different datasets, leading to poor robustness and weak cross-institutional generalization, making them difficult to apply in diverse clinical environments with varying imaging devices and pathological characteristics.

To overcome these limitations, deep learning techniques have been introduced to perform end-to-end feature learning from large-scale data, enabling automatic extraction of hierarchical semantic features. By integrating attention mechanisms and cross-modal fusion strategies, deep learning models have significantly improved segmentation accuracy for complex anatomical structures while greatly reducing the need for manual intervention. Through self-supervised pretraining and lightweight model design, these methods have also enhanced adaptability to noise and cross-center variability, offering new pathways for the clinical application and translational development of medical image segmentation.

3. Dataset and preprocessing

The breast ultrasound dataset used in this study was collected from Baheya Hospital for Early Detection & Treatment of Women's Cancer, located in Cairo, Egypt [6]. The study protocol was approved by the Cairo University Research Ethics Committee under approval number NCI-2018-015. Data acquisition was carried out over a one-year period from January to December 2018, encompassing 600 female patients aged between 25 and 75 years. All ultrasound examinations were performed using a LOGIQ E9 system equipped with an ML6-15-D linear array probe operating at a frequency range of 1–5 MHz. The original images, captured in DICOM format, were subsequently converted to PNG for downstream processing.

The final dataset comprises 780 anonymized ultrasound images, which are categorized into three classes based on pathological results: normal (n = 133), benign (n = 487), and malignant (n = 210). The average image resolution is approximately 500 × 500 pixels, although the original image sizes vary. To ensure uniform input size across all models, all images were resized to a consistent resolution during the preprocessing stage.

The data preprocessing pipeline consisted of the following steps. First, image normalization was performed by resizing all ultrasound images and corresponding mask images to a fixed resolution of 256 × 256 pixels using bicubic interpolation, which preserves geometric structure consistency. The pixel intensity values of the original images were normalized to a range of [0, 1] to mitigate variability introduced by acquisition devices and enhance training stability. This step ensured that all input images shared the same dimensions and standardized value ranges for consistent model training. Second, mask alignment was carried out by strictly matching filenames between original images and their corresponding masks, ensuring spatial consistency of annotated regions. The segmentation masks were saved as single-channel binary images, with tumor regions labeled in pure white (intensity value = 1) and background in pure black (intensity value = 0). For cases with multiple annotation masks, pixel-wise overlays were performed to generate composite annotations,

DOI: 10.54254/2753-8818/99/2025.23012

incorporating consensus from multiple clinical experts. Third, batch loading and dataset construction were implemented by organizing image-mask pairs into class-specific directories. An automated script traversed the subdirectories and extracted matched image-mask pairs. To prevent data leakage, all images from the same patient—regardless of imaging angle or view—were assigned exclusively to either the training or the testing set. This process produced a structured tensor dataset in the format of (number of samples, 256, 256, number of channels), where original ultrasound images retained their three RGB channels and the segmentation masks were represented as single-channel images. Lastly, although not a part of preprocessing in the strict sense, visual verification was conducted as a quality control measure. Random samples were selected for visual inspection by overlaying masks on their corresponding images using color mapping techniques. This step verified the anatomical alignment between annotated regions and actual lesions, ensuring that the data quality met the requirements for supervised learning.

4. Model

The segmentation model used in this study is based on the U-Net architecture, originally proposed in 2015. U-Net is a symmetric encoder–decoder network specifically designed for biomedical image segmentation. The encoder (contracting path) progressively extracts high-level features through repeated modules of 3×3 convolutions (with ReLU activation) followed by 2×2 max pooling, doubling the number of feature channels at each down-sampling step. The decoder (expanding path) restores spatial resolution using 2×2 transposed convolutions, while skip connections concatenate corresponding feature maps from the encoder, preserving fine-grained localization details. Each decoder block uses two 3×3 convolutions to refine the merged features, and the final output layer applies a 1×1 convolution to generate pixel-level class probabilities. The network discards fully connected layers and adopts valid padding, allowing seamless segmentation of large-scale images via an overlapping-tile strategy. Coupled with data augmentation techniques such as elastic deformation, the model can be effectively trained even with limited annotated data.

Although U-Net has been widely adopted in medical image segmentation tasks, its static skip connections—directly concatenating encoder and decoder features—present inherent limitations. Low-level features often contain irrelevant background noise, such as textures of unrelated tissues, which may degrade segmentation accuracy. Moreover, due to the local receptive fields of convolutional layers, the encoder struggles to capture long-range contextual dependencies, limiting the model's sensitivity to small or low-contrast lesions.

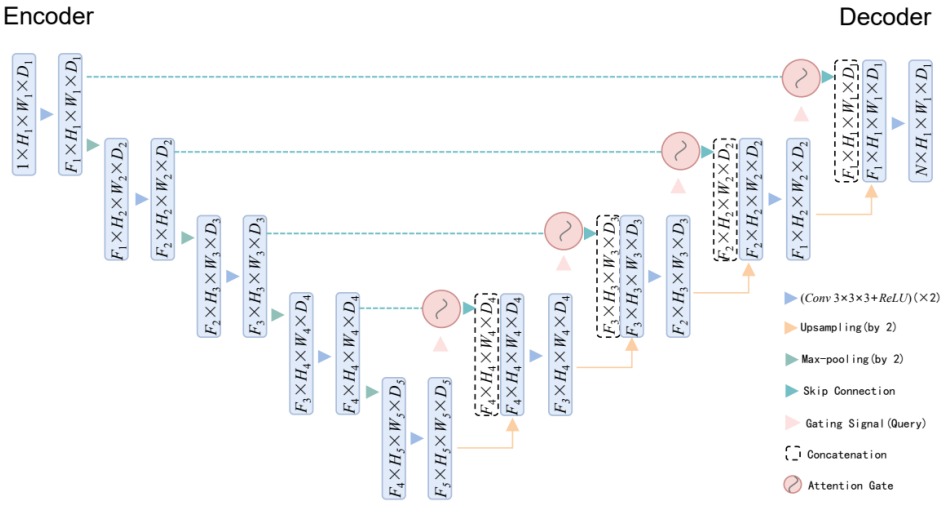

Figure 1: Model architecture

DOI: 10.54254/2753-8818/99/2025.23012

To address these issues, we proposed a modified attention U-Net [7]. Attention U-Net introduces Attention Gates (AGs) into the skip connections. These gates dynamically assign spatial weights to feature maps, enhancing responses in relevant regions while suppressing background noise. The attention mechanism operates in three stages: first, encoder and decoder features are concatenated along the channel dimension and compressed via a 1×1 convolution; second, a Sigmoid function normalizes the attention map to the range [0, 1]; third, the low-level features are point-wise multiplied with the attention map to obtain the weighted output. This mechanism can be embedded at multiple levels of the U-Net architecture, allowing the model to adaptively adjust the contribution of multi-scale features using global context. The attention weights are computed as:

$$A = \sigma\left(W_g^T F_{high} + W_x^T F_{low} + b\right)$$

where Wg and Wx are learnable weight matrices, b is the bias term, and σ denotes the Sigmoid function. The modified model is illustrated in Figure 1.

The parameters of the attention gates are optimized in an end-to-end fashion using backpropagation. Gradients are first computed with respect to the attention map, and then propagated to the corresponding encoder and decoder features. The use of the Sigmoid function naturally constrains gradient magnitude, preventing issues such as exploding or vanishing gradients during training. Parameter updates follow the standard stochastic gradient descent (SGD) algorithm or its variants, such as Adam, and are iteratively refined as:

$$W^{(t+1)} = W^{(t)} - \eta \cdot \frac{\partial \mathcal{L}}{\partial W^{(t)}},$$

where η denotes the learning rate.

In this study, the Adam optimizer was used with a default learning rate of 0.001 to ensure smooth convergence. Binary cross-entropy was selected as the loss function to quantify the difference between predicted and ground truth masks, as appropriate for binary classification. The number of filters in the convolutional layers was progressively increased from 32 to 1024, improving the network's ability to learn complex features. Dropout was applied to both encoder and decoder modules to mitigate overfitting, with dropout rates set to 0.1, 0.2, and 0.3 respectively. Batch normalization was applied after each attention gate to further stabilize training. The batch size was set to 8, and the model was trained for 20 epochs, with the number of steps per epoch determined by the training data volume. Model performance was evaluated using accuracy and Intersection over Union (IoU). Accuracy measures overall classification performance, while IoU quantifies segmentation precision by evaluating the overlap between predicted and true labels.

5. Results



Figure 2: The good segmentation results

DOI: 10.54254/2753-8818/99/2025.23012

To comprehensively evaluate the performance of the proposed attention U-Net model on the task of breast cancer ultrasound image segmentation, we conducted a series of quantitative assessments on the test dataset. Evaluation metrics included Accuracy and IoU, which are widely adopted in biomedical image segmentation to measure both classification effectiveness and the spatial overlap between predicted and ground truth masks. The experimental results show that the proposed attention U-Net achieved an accuracy of 98.24% on the test set, significantly outperforming the conventional U-Net model, which yielded approximately 92% accuracy. This demonstrates that incorporating the attention mechanism substantially enhances segmentation performance, especially in challenging scenarios involving complex backgrounds or blurred boundaries. The predicted masks exhibited high consistency with the ground truth, confirming the effectiveness of attention gates in improving the model's focus on critical lesion regions. Figure 2 illustrates the model's impressive segmentation capability, where the predicted mask almost perfectly aligns with the actual lesion area, demonstrating precise boundary localization. Furthermore, in Figure 1, a dark region adjacent to the tumor was not mistakenly classified as a lesion, showcasing the model's robustness and ability to avoid false positives.

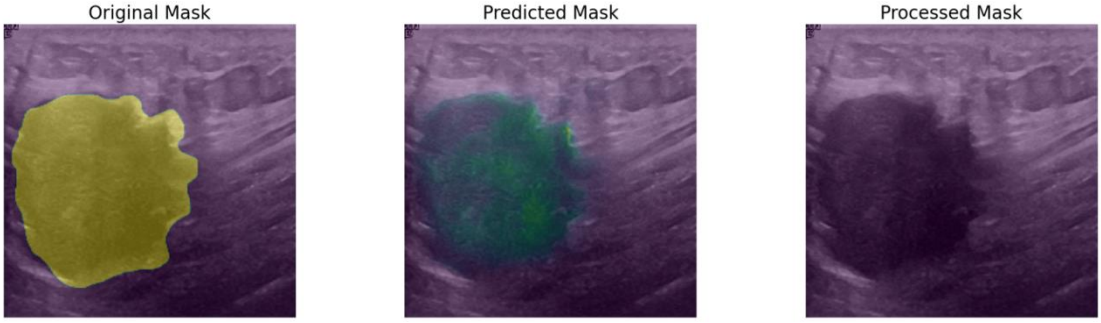

Figure 3: The bad segmentation results

Despite its strong overall performance, the model still exhibits certain limitations. In the case shown in Figure 3, although the predicted region nearly coincides with the true lesion, the model's output was affected by a threshold that was set too high. As a result, the model assigned a low probability to this region, leading to it being labeled only in green on the prediction map, rather than as a definitive lesion. This highlights the impact of threshold selection on segmentation outcomes, particularly for borderline or low-probability regions. Optimizing the threshold—potentially through adaptive or dynamic thresholding strategies—could further improve the model's sensitivity to subtle or low-contrast lesions.



Figure 4: Grad-CAM for good results

DOI: 10.54254/2753-8818/99/2025.23012

To enhance interpretability and validate the model's decision-making process, we employed Gradient-weighted Class Activation Mapping (Grad-CAM) as an additional evaluation method [8]. Grad-CAM is a gradient-based visualization technique that generates class-discriminative heatmaps by computing the gradient of the target class score with respect to the final convolutional feature maps. Its core mechanism involves globally averaging the gradients to obtain channel-wise weights, which are then combined with the activation maps and filtered by a ReLU function to retain positively contributing regions. This method can be seamlessly applied to any CNN-based model without architectural modifications. As shown in Figures 4, Grad-CAM successfully highlights the lesion areas with high-intensity regions on the heatmaps, clearly demonstrating that the model has learned to focus on diagnostically relevant features. In Figure 4, even though attention is observed above the lesion region, the model does not misclassify it as a tumor, further reflecting its robustness and selective attention capacity.

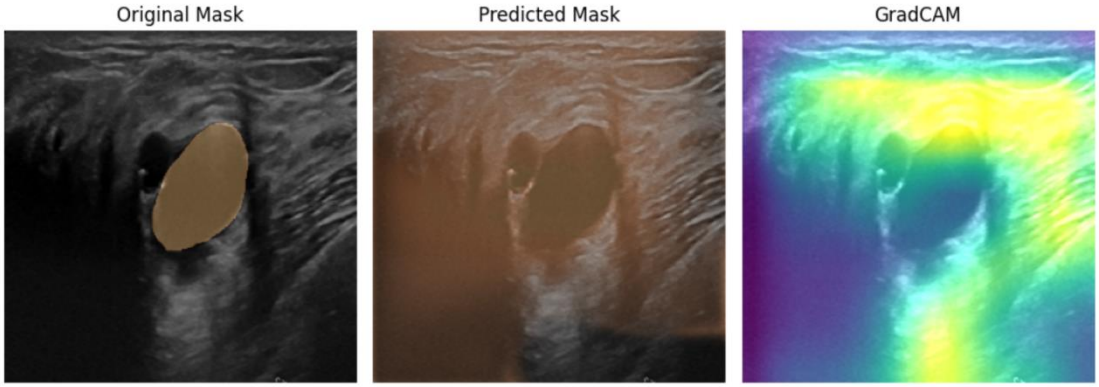

Figure 5: Grad-CAM for bad results

However, in certain cases such as that shown in Figure 5, the attention distribution revealed by Grad-CAM does not fully align with the ground truth. The model assigns higher attention weights to surrounding background tissue rather than the lesion itself. This misalignment could impair segmentation accuracy in certain challenging scenarios. Such cases suggest future directions, including the integration of multi-scale attention mechanisms or anatomical priors, to improve the model's localization precision and overall segmentation reliability.

6. Discussion and conclusion

In this study, we proposed a modified attention U-Net for the task of automated breast cancer segmentation in ultrasound images. By integrating attention gate mechanisms into the U-Net architecture, the model adaptively focuses on salient lesion areas while suppressing irrelevant background noise. This enhancement significantly improves the model's ability to detect small or low-contrast lesions while maintaining high segmentation accuracy. Experimental results demonstrate that our model outperforms the traditional U-Net in terms of accuracy and IoU, showing superior robustness and generalization under complex imaging conditions. Furthermore, we employed Grad-CAM to evaluate model interpretability and visual attention. The resulting heatmaps confirm that the model effectively attends to pathological regions, accurately distinguishing them from non-lesion areas. This visual validation reinforces the reliability and clinical usability of the proposed method, especially in high-stakes medical settings.

Furthermore, we employed Grad-CAM to evaluate model interpretability and visual attention. The resulting heatmaps confirm that the model effectively attends to pathological regions, accurately

distinguishing them from non-lesion areas. This visual validation reinforces the reliability and clinical usability of the proposed method, especially in high-stakes medical settings.

Despite its promising performance, the proposed model has several limitations. Firstly, fixed thresholding may lead to under-segmentation in low-probability regions, reducing sensitivity to borderline lesions. Secondly, the model relies entirely on supervised learning and requires a large volume of high-quality annotated data, which is expensive and labor-intensive to obtain in medical contexts. This dependence limits its scalability and cross-institutional adaptability. Additionally, while Grad-CAM enhances model interpretability to some extent, its spatial resolution is limited, making it difficult to fully trace the semantic reasoning paths within the network.

To address these challenges, future work may explore several directions. First, transfer learning can alleviate the dependence on large labeled datasets [9]. Pretraining on large-scale medical datasets or related segmentation tasks and fine-tuning on breast ultrasound images can improve feature representation, accelerate convergence, and enhance performance in low-data regimes. Second, multi-modal data fusion could be introduced by combining ultrasound with other modalities such as elastography, histopathology, or clinical textual records. This could enable richer feature representation through image-text joint modeling or cross-modal alignment, improving diagnostic precision. Third, with the rapid advancement of large language models (LLMs) such as GPT [10] and LLaMA [11] in healthcare applications, future systems could integrate LLMs with vision models to build intelligent image–language collaborative frameworks. For instance, prompt-based interaction could guide the segmentation focus toward specific lesion types, or LLMs could generate clinically relevant explanations and decision support based on segmentation results. Lastly, optimization of the model architecture through self-supervised learning, lightweight design, and adaptive thresholding strategies could further enhance deployment efficiency and cross-platform adaptability, accelerating the clinical translation of AI-assisted diagnosis systems.

In summary, the proposed modified attention U-Net demonstrates excellent performance in terms of segmentation accuracy, robustness, and interpretability for breast ultrasound images, underscoring its strong potential for clinical applications. With future integration of heterogeneous data sources, advanced pretraining strategies, and multimodal collaborative mechanisms, this work lays a solid foundation for building more comprehensive and trustworthy intelligent diagnostic systems for breast cancer.

References

[1] *M. Arnold et al., "Current and future burden of breast cancer: Global statistics for 2020 and 2040," The Breast, vol. 66, pp. 15–23, Dec. 2022, doi: 10.1016/j.breast.2022.08.010.*

[2] *D. Gupta, "Image segmentation keras : Implementation of segnet, FCN, UNet, PSPNet and other models in keras," Jul. 25, 2023, arXiv: arXiv:2307.13215. doi: 10.48550/arXiv.2307.13215.*

[3] *S. S. Al-amri, N. V. Kalyankar, and K. S. D, "Image segmentation by using threshold techniques," May 21, 2010, arXiv: arXiv:1005.4020. doi: 10.48550/arXiv.1005.4020.*

[4] *K.-S. Chuang, H.-L. Tzeng, S. Chen, J. Wu, and T.-J. Chen, "Fuzzy c-means clustering with spatial information for image segmentation," Computerized Medical Imaging and Graphics, vol. 30, no. 1, pp. 9–15, Jan. 2006, doi: 10.1016/j.compmedimag.2005.10.001.*

[5] *A. S. Kornilov and I. V. Safonov, "An overview of watershed algorithm implementations in open source libraries," Journal of Imaging, vol. 4, no. 10, Art. no. 10, Oct. 2018, doi: 10.3390/jimaging4100123.*

[6] *W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," Data in Brief, vol. 28, p. 104863, Feb. 2020, doi: 10.1016/j.dib.2019.104863.*

[7] *O. Oktay et al., "Attention U-net: Learning where to look for the pancreas," May 20, 2018, arXiv: arXiv:1804.03999. doi: 10.48550/arXiv.1804.03999.*

[8] *R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," Int J Comput Vis, vol. 128, no. 2, pp. 336–359, Feb. 2020, doi: 10.1007/s11263-019-01228-7.*

[9] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," J Big Data, vol. 3, no. 1, p. 9, May 2016, doi: 10.1186/s40537-016-0043-6.

[10] X. Liu et al., "GPT understands, too," AI Open, Aug. 2023, doi: 10.1016/j.aiopen.2023.08.012.

[11] A. Grattafiori et al., "The Llama 3 Herd of Models," Nov. 23, 2024, arXiv: arXiv:2407.21783. doi: 10.48550/arXiv.2407.21783.