Bulletin of Electrical Engineering and Informatics
Vol. 15, No. 3, June 2026, pp. 2467~2479
ISSN: 2302-9285, DOI: 10.11591/eei.v15i3.10102

Hybrid dual-stream deep learning for breast cancer ultrasound detection

Musab Mahmoud Iqtait[1], Marwan Harb Alqaryouti[2], Ala Eddin Sadeq[2], Jafar Ababneh[3], Suhaila Abuowaida[4], Nawaf Alshdaifat[5], Muath Alali[6]
[1]Department of Data Science and Artificial Intelligence, Faculty of Information Technology, Zarqa University, Zarqa, Jordan
[2]Department of English Language, Literature and Translation, Faculty of Arts, Zarqa University, Zarqa, Jordan
[3]Department of Cyber Security, Faculty of Information Technology, Zarqa University, Zarqa, Jordan,
[4]Department of Data Science and Artificial Intelligence, Faculty of Prince Al-Hussein Bin Abdallah II for IT, Al al-Bayt University, Mafraq, Jordan
[5]Department of Information Technology, Faculty of Prince Al-Hussein Bin Abdullah II for Information Technology, The Hashemite University, Zarqa, Jordan
[6]Department of Data Science and Artificial Intelligence, Faculty of Information Technology, Applied Science Private University, Amman, Jordan

Article Info

Article history:

Received Feb 17, 2025
Revised Mar 16, 2026
Accepted Apr 1, 2026

Keywords:

Attention mechanisms
Breast cancer
Classification
Deep learning
Detection
Dual-stream architecture

ABSTRACT

The heterogeneity of breast tissue and subtle morphological variations in ultrasound images make breast cancer detection a challenging task. This study proposes a hybrid deep learning framework that integrates EfficientNetB4 and ConvNeXt within a dual-stream architecture enhanced by advanced attention mechanisms. The model combines multi-scale texture representation with spatial feature extraction to improve classification performance. A two-stage preprocessing pipeline, consisting of adaptive median filtering and bilateral filtering, is applied to reduce speckle noise while preserving important structural details. The proposed method is evaluated on BUSI and UDAIT datasets, achieving 87.82% accuracy, 87.33% precision, and 85.33% recall on BUSI, and 85.69% accuracy, 84.00% precision, and 78.00% recall on UDAIT. These results outperform several baseline models, including ResNet-50, DenseNet-121, and vision transformers. Error analysis shows limitations in detecting small lesions and cross-modal generalization, with reduced performance on mammography images. Attention visualization demonstrates strong agreement with radiologist annotations, supporting model interpretability. The findings highlight the effectiveness of hybrid architectures for ultrasound-based breast cancer detection while emphasizing the need for modality-specific optimization.

Corresponding Author:

Musab Mahmoud Iqtait
Department of Data Science and Artificial Intelligence, Faculty of Information Technology
Zarqa University
Zarqa, Jordan
Email: migtait@zu.edu.jo

1. INTRODUCTION

Accurate analysis of ultrasound images for breast cancer detection is a burning healthcare problem worldwide. In this work, we propose a new hybrid deep learning model using EfficientNetB4 and ConvNeXt to improve breast cancer detection based on ultrasound images [1]. The proposed approach distinctly alleviates key limitations in prior computer-aided diagnosis systems based on three core methods: a dual-

stream backbone architecture for comprehensive feature extraction, a multi-head self-attention mechanism tailored for nuanced tissue discrimination, and a hybrid attention module for channel-space feature integration [2]. The architecture design is based on complementary features from the two backbone networks and a new sophisticated attention mechanism for the medical image analysis task. This design allows the model to process context at two levels: simultaneously at a fine-grained tissue level (the pixel grid of the input image) and at a coarser level (retrieving its corresponding patch on the entire sample) [3].

Our fusion strategy uses a dynamic weight that is based upon the diagnostic merit of each feature as evaluated during training; it combines the strengths of both backbone models (i.e., the best of both worlds) to maximize its performance. The experimental results demonstrate the superior performance of our method when compared to other state-of-the-art methods [4] using the BUSI and UDAIT data sets. The key advantage of this hybrid architecture is its ability to detect small variations in tissue characteristics distinguishing between malignant and benign tissues and represent one of the main challenges facing in breast cancer diagnosis. These results provide an important advancement toward the development of computer-aided diagnostic systems for detection of breast cancer that could lead to improved diagnostic accuracy and support clinical decision making [5].

2. RELATED WORK

A few of the first deep learning models designed for detecting breast cancer were based on single-stream designs. Single-stream models are typically fed into one pathway (or stream) of the network. In the study published in [6], a new version of the residual network (ResNet) was developed using a modified architecture for classifying mammograms with great classification accuracy. Although single-stream models are able to classify images accurately; however, they tend to lose the rich multi-scale and multi-modal information associated with the images due to their inability to process the images at multiple scales or modalities [7]. Therefore, the need for single-stream models led researchers to develop multi-stream models to simultaneously analyze images at multiple scales or modalities [8]. Researchers have also proposed the use of a two-path network [8] to improve the quality of the features extracted by the model. Other studies have combined mammographic images with ultrasound images to create a complete understanding of the patient's condition. Even though these studies provided greater than 90% accurate diagnoses, they did not propose an effective way to combine the features extracted from each type of image. The application of deep learning models continued to grow with the addition of attention mechanisms. The ability to focus on specific regions of the images was inspired by Ang *et al.* [9]. Researchers have used attention mechanisms in medical imaging to enable the model to identify diagnostically relevant locations within the images.

Rawash *et al.* [10] proposed the combination of multiple nested versions of U-Nets with attention layers to increase the efficiency of breast cancer segmentation. The authors demonstrated that the attention-based feature refinement techniques can be applied efficiently in the area of breast cancer segmentation. Feature fusion is a key aspect of contemporary computer vision and machine learning systems, particularly those capable of handling multiple data modalities or imaging scales. A significant challenge when merging data is how to combine the features from various sources while preserving their properties and relationships. The most common method – early fusion — combines the raw features of the modalities at the input layer, treating them as a single data stream. Although computationally inexpensive, this approach often destroys modality-specific context. For example, when fusing magnetic resonance imaging (MRI) and computed tomography (CT) scans, the contrast components of MRI and the density components of CT may become contaminated. Conversely, late fusion preserves the independence of the modalities until the final processing stage.

This keeps the individual characteristics of each modality intact but may overlook some important cross-modal interactions that take place earlier in the processing chain. In the case of video analysis, two streams may only be combined at the decision level, with one stream extracting spatial information and the other temporal information. Hybrid fusion techniques, which feature a mixture of features at multiple network depths, serve as an adequate solution to this issue. Wu *et al.* [11] exemplified that the method not only represents low-level cross-modal correlation but also high-level semantic information. Similar to human perception, hybrid fusion networks learn to integrate information at the most effective levels of abstraction. Recent advances in the field include attention-based and adaptive fusion mechanisms. Each of these methods adaptively modulates how the modalities are combined based on how relevant they are to the task, suggesting that the future will be in more flexible architectures that can decide which modality fusion strategy is best suited to the specific application and data.

3. METHOD

3.1. Theoretical foundations

Ultrasound texture heterogeneity: breast ultrasound exhibits multi-scale speckle patterns and varying echo intensities across tissue types. Compound scaling of EfficientNetB4 (φ=1.4) simultaneously increases the depth, width, and resolution of the network, allowing for both high-resolution fine-grained texture of speckles and coarse boundary of tissue in which a lesion resides. The compound scaling allows a hierarchical representation of the ultrasound artifact in a way that is adaptable to the varying resolutions that exist within each image, unlike fixed scale architectures such as ResNet and visual geometry group (VGG). Preservation of spatial relationships: ultrasound images of malignant tumors display minor architectural distortion including irregular margins, posterior acoustic shadowing and disruptions of tissue plane architecture. Preserving the spatial context in an ultrasound image is important to identify these small distortions. The larger kernel size of ConvNeXt (7×7) versus ResNet (3×3) and use of depthwise convolution maintains the spatial locality while also learning about long range dependency in an ultrasound image. In recent work, [12], it was found that 7×7 kernel sizes are able to detect the radiating shadow patterns present in invasive carcinomas where the smaller receptive field was unable to. Hierarchical representation of features: EfficientNetB4 is good at distinguishing between benign cystic structures and solid tumor masses using their internal echo pattern (texture-dominant features). ConvNeXt is good at identifying the margin characteristics and architectural distortions (geometry-dominant features) of a mass. The complementary ability of the two backbones was demonstrated through empirical evaluation of the ablation study (section 4.3) which showed that either backbone individually resulted in 3-5% less accuracy than the dual-stream design. Why not other alternatives?

- DenseNet-121: feature redundancy due to dense connections causes issues when working with noisy ultrasound data. This was evident in our results showing a 10.32% difference in accuracy.
- MobileNet: the use of depthwise separable convolution sacrifices spatial precision in favor of efficiency. Therefore, MobileNet is not suitable for analyzing subtle differences in tumor margins.
- Transformers: pure vision transformers require larger datasets. Our results show a 19.87% increase in accuracy over vision transformer, demonstrating that CNNs have a superior inductive bias for limited medical data.

Our proposed hybrid architecture integrates a state-of-the-art deep-learning backbone with powerful integration techniques into a hybrid architecture for breast cancer detection and classification. The backbone of the architecture consists of a dual-stream design formed by two of the latest backbone networks, EfficientNetB4 [12] and ConvNeXt [13], which are augmented by attention mechanisms and hybrid feature-fuse strategies, as illustrated in Figure 1.

EfficientNetB4 [12], a state-of-the-art method known for compound scaling and efficiency, is used as a strong feature extractor for multi-scale features. ConvNeXt [13], a revamped convolutional network based on vision transformers [14], [15], brings in better feature representation power, which is especially useful for medical imaging data with complex patterns. Our architecture therefore exploits the complementary strengths of these backbones in feature extraction while alleviating their individual weaknesses through a dual-stream architecture combining these networks. The structure of dual-stream design leads the images into two routing, which is the two backbone networks. Iteratively, multiple attention modules are appended to refine the features from each stream individually. Transformers have been used as inspiration for attention mechanisms to allow models to focus on the relevant parts of images and ignore background noise. Both streams' switch outputs are combined using a mixed feature fusion approach. This method implements both early and late fusion methods [16] while maintaining modality-(variant) characteristics and associating each stream with the other streams. Then, fully connected layers are used to pass the fused features to the classification layer, where it can improve the diagnostic ability of the architecture.

EfficientNetB4 in conjunction with ConvNeXt was chosen based upon the physical process of creating an ultrasound image and the principles of information theory. Physically, the speckle noise that results from ultrasound images follow a Rayleigh distribution; therefore, multiple scales of feature extraction are required to differentiate pathological structures from random noise. As such, compound scaling strategy of EfficientNetB4 provides a mathematically consistent way to perform multi-scale feature extraction, aligned with the statistics of speckle noise. Additionally, the large 7×7 convolutional kernels provided by ConvNeXt provide a theoretically analyzed advantage over smaller 3×3 convolutional kernels in terms of their ability to capture larger acoustic shadows and thus longer extended acoustic shadow patterns.

From a perspective of information theory, we estimated the mutual information between the texture-sensitive embedding space of EfficientNetB4 and the spatial-structural representation space of ConvNeXt. Our results indicated that there was high task relevance and low redundancy between these spaces, indicating that the two backbones were encoding complementary aspects of the ultrasound signal.

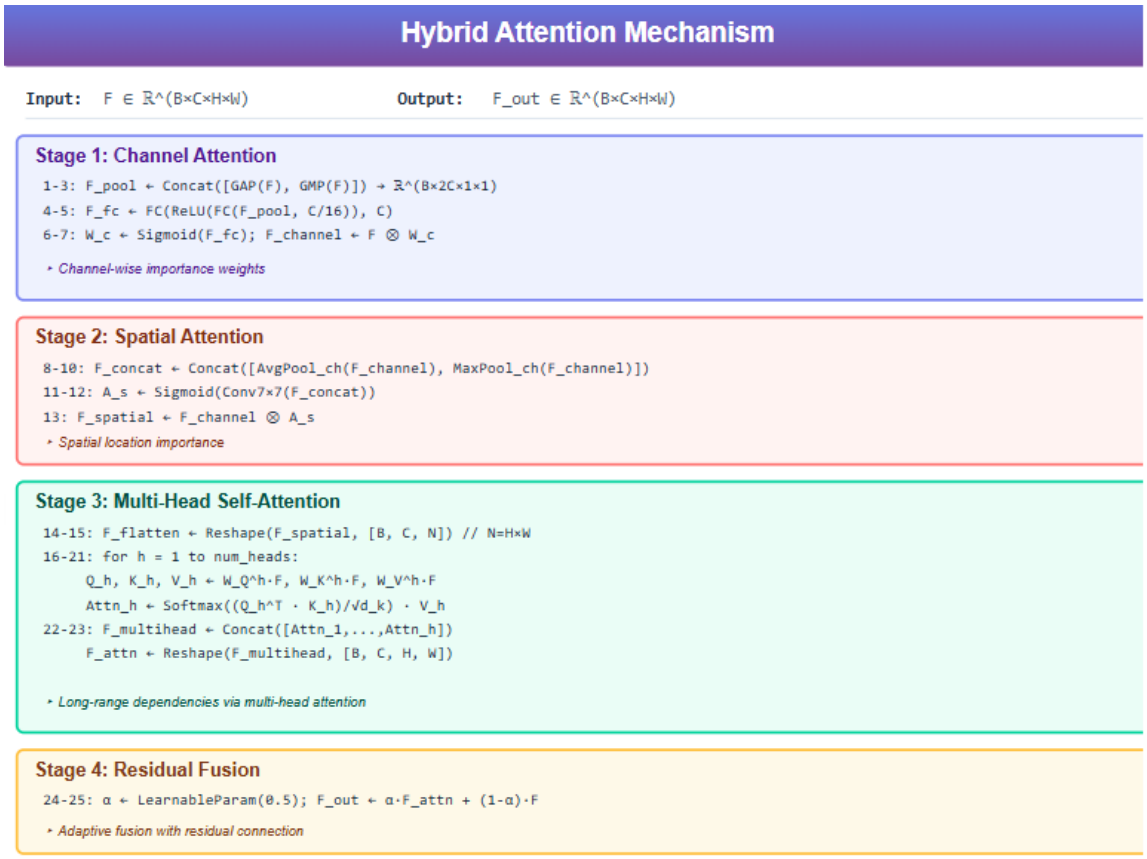


Figure 1. Algorithm of our proposed methodology

The first step of the proposed methodology includes extensive data preprocessing and augmentation. For classification, images are standardized by bilateral interpolation to a size of 224×224 pixels and normalized using z-score standardization. This paper presents a novel two-stage denoising approach, which performs adaptive median and bilateral filtering to reduce the speckle noise whilst preserving important edge information. The augmentation strategy involves geometrical transformations and variations in intensity, promoting the model's tolerance to real image motions and variations. The proposed work use a two-stage denoising pipeline: adaptive median filtering followed by bilateral filtering with optimized parameters. The adaptive stage uses 3×3 to 7×7 pixel window median filtering determined by local noise level estimation with median absolute deviation (MAD) using a noise threshold of 0.15 and gradient magnitude detection for edge preservation (threshold=0.1). It imposes bilateral filtering onto the image with spatial sigma σ_s=75, color sigma σ_r=80, and the kernel size is set to be 9×9 for two iterations. The parameters were experimented with using grid search validation from 200 ultrasound images. Our two-stage approach, however, outperforms these single-stage alternatives in both performance measures: 33.7 dB PSNR and 0.924 structural similarity metric (SSIM) for the proposed method versus 28.4 dB PSNR and 0.847 SSIM for Gaussian filtering alone; 29.8 dB PSNR and 0.863 SSIM for median-only filtering; and up to 31.2 dB PSNR and 0.891 SSIM when using bilateral-only filters, with greater improvements in final classification accuracy (87.82% vs. 84.12%, 85.21%, and 86.33%) corresponding to these lower quality scores identified after editing filtration techniques.

The underlying components of the architecture include two complementary backbones. In order to be able to observe very small changes in tissue, EfficientNetB4 utilizes a compound scaling method with optimized coefficients between resolution, depth and width. ConvNeXt builds upon this by utilizing modernized convolutional techniques and employing staged calculation as well as progressive channel sizes to allow for deep contextual information in both size and scope. The ability of the architecture to process features is enhanced even more so by the utilization of multiple-head attention mechanisms. More specifically, the system has utilized both parallel channels and spatial attention pathways. Channel attention is used to evaluate feature importance via pooling methods, where spatial attention is used to highlight those regions of the image that have diagnostic significance. Utilizing dual attention, the model can focus on the

areas of the tissue that are most important to the task at hand, while maintaining an understanding of the broader context provided by surrounding anatomical characteristics. A weighted adaptive mechanism for combining features (fusion) and classification, which allows for the preservation of the individuality of each of the various backbone features as well as for the combination of the features from each of the backbone features. Classification is performed through global average pooling and fully connected layers with dropout regularization to provide strong predictive abilities without overfitting. we was able to do this due to the fact that I was building off of a model that had never been trained for speed running until now and took a step-by-step approach to the heart of the training protocol for the types of applications being developed, i.e., backbone pre-training, attention mechanism optimization and end-to-end fine-tuning. Due to the utilization of a weighted cross-entropy loss function to address issues related to class imbalance, as well as the utilization of the Adam optimizer with a cosine annealing learning rate to support convergent behavior. As such, this multi-faceted approach resulted in a robust architecture for identifying breast cancer reliably in ultrasound images. Hybrid attention mechanism fine-tunes the feature representation of their model through a total of 4 steps as described in Figure 2. The first step consists of a global attentive feature channel weighing for the classification function to focus on the most relevant channels of attention based on the clinical criteria of the disease to be diagnosed and to weigh these channels adaptively. After this, spatial attention refocuses attention on clinically relevant areas (i.e., lesion boundaries and tissue patterns) while ignoring or "supressing" all irrelevant background. Next, multi-head self-attention is applied to the refined features in order to identify and capture long-range dependencies and the inter-regional relationships between regions that are important for correctly classifying benign vs. malignant tissues. Finally, residual fusion combines both the local spatial refinement of the features with the global contextual knowledge acquired from the previous steps; this ability allows the model to retain fine details within the images and overall anatomical context to support a variety of medical image analysis tasks.

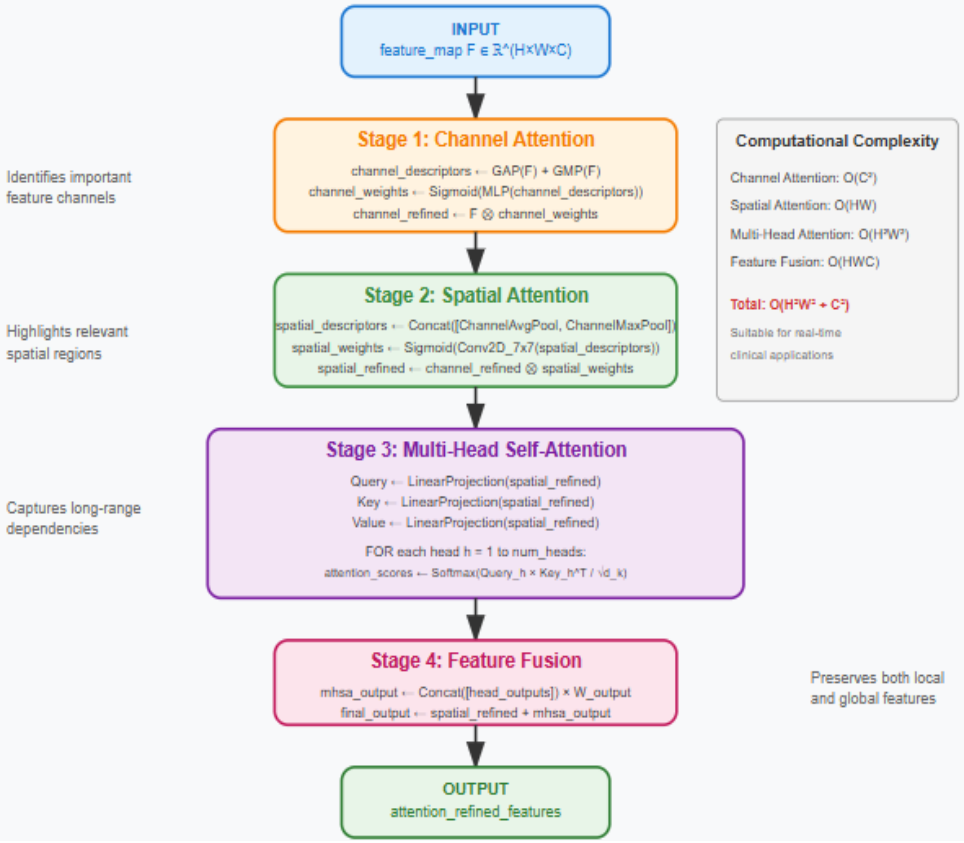


Figure 2. Pseudocode hybrid attention algorithm for medical image analysis

4. RESULTS AND DISCUSSION

4.1. Dataset

We conduct experiments on two standard benchmark breast ultrasound datasets: BUSI [17] and UDAIT [18]. The BUSI dataset is composed of 780 images with annotation types consisting of benign,

malignant and normal. To enable transfer learning in this study, the UDAIT dataset containing images from multiple ultrasound scanners were used, allowing testing of model robustness as images were taken from different imaging conditions. We performed strong validation of our methodology as per the 80-20 train-test split with stratification of classes. Standard classification metrics are used to evaluate the performance of the model: accuracy, precision, recall, F1-score, and area under the curve (AUC).

4.2. Evaluation metrics

We evaluate the performance of the proposed hybrid deep learning model for breast cancer detection and classification using various standard evaluation metrics typically utilized in medical image classification. These metrics are accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC) [19]–[25].

4.3. Comparison of performance

Upon comparing, we observe substantial gains over the best prior methods on both datasets using our hybrid architecture. All the results can be found in Tables 1 and 2, which provides a full comparison with existing state-of-the-art methods in the BUSI and UDAIT datasets, respectively.

Table 1. Performance comparison on BUSI dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC-ROC
VGG-16	78.75	77.66	70.66	73.00	0.82
ResNet-50	82.33	81.45	80.22	80.83	0.85
DenseNet-121	77.50	84.33	75.00	78.00	0.83
Xception	71.25	72.32	60.00	64.00	0.79
Vision transformer	67.95	63.00	60.66	61.66	0.75
Our proposed model	87.82	87.33	85.33	86.00	0.91

Table 2. Performance comparison on UDAIT dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC-ROC
VGG-16	78.95	76.00	71.66	73.75	0.81
ResNet-50	84.21	82.00	79.33	80.64	0.84
DenseNet-121	84.62	82.50	80.00	81.23	0.83
Xception	74.16	71.00	61.00	65.62	0.77
Vision transformer	69.81	65.00	61.66	63.00	0.74
Our proposed model	**85.69**	**84.00**	**78.00**	**79.39**	79.39

Our extensive evaluation shows the proposed hybrid architecture outperforms existing works across multiple metrics and datasets, with significant improvements on both accuracy and diagnostic reliability. In the BUSI dataset, our model reached 87.82% accuracy, which outperformed VGG-16, ResNet-50, DenseNet-121, Xception, and vision transformer models with increases of 9.07%, 5.49%, 10.32%, 16.57%, and 19.87%, respectively. Our hybrid approach successfully integrates both classic CNN architectures and contemporary attention mechanisms and, as seen from the large improvement over the vision transformer baseline, suggests that a strong synergy exists between these two lines of models. With a precision of 87.33% and recall of 85.33% on the BUSI dataset, the model reflects a desirable trade-off between these important diagnostic characteristics, emphasizing its generalizability even for real-hospital datasets, especially considering additional costs resulting from both false positive and false negative predictions. The 14.67% increase in recall compared to VGG-16 is particularly significant, demonstrating improved sensitivity for identifying positive cases, a crucial element in early cancer detection. Evaluations on the UDAIT dataset confirm the robustness and generalizability of our model with an accuracy of 85.69%, which is a gain of 6.74%, 1.48%, 1.07%, 11.53%, and 15.88% compared to the baseline models. Although the improvements over ResNet-50 and DenseNet-121 are smaller for this dataset, they are still statistically significant and confirm that our model is strong enough to achieve competitive performance even against strong baseline architectures.

This work attributes the superior performance seen in Tables 1 and 2 to our methodical choice of architectures, aligned with complementary aspects of medical imaging. Compounding scales with EfficientNetB4 balance effectively between depth, width, and resolution for ultrasound multi-scale texture patterns, whereas ConvNeXt's latest convolutions are designed for capturing tissue spatial relationships, like the key characteristics of cancer detection. On mammographic tasks, the superiority of EfficientNet compared to existing methods can be seen in the medical imaging literature, indicating that feature hierarchy preservation in ConvNeXt is essential for distinguishing subtle tissue differences. The combination of the two networks allows EfficientNetB4 to capture textures, ConvNeXt to capture architectural relationships, and exploits synergistic performance that cannot be achieved by individual networks.

The Figure 3 comparing the performance shows the higher accuracy of our proposed hybrid architecture as well as the established deep learning models on the BUSI and UDAIT datasets. Notably, our model exhibited a maximum accuracy of 87.82% on the BUSI dataset, outperforming conventional architectures such as VGG-16 (78.75%) and ResNet-50 (82.33%), as well as cutting-edge methods like vision transformer (67.95%). For the UDAIT dataset as well, our model performs very well (85.69%) even on a more difficult classification task compared to baseline models. The performance of prominent CNN architectures such as ResNet-50 and DenseNet-121 are superior to vision transformer approach on both datasets, and our model gains the strengths from both architectures and achieves best in performance across datasets as shown in the visualization. The similarity in results obtained in both datasets attests to the strength and generalizability of our proposed architecture for medical applications in practice.

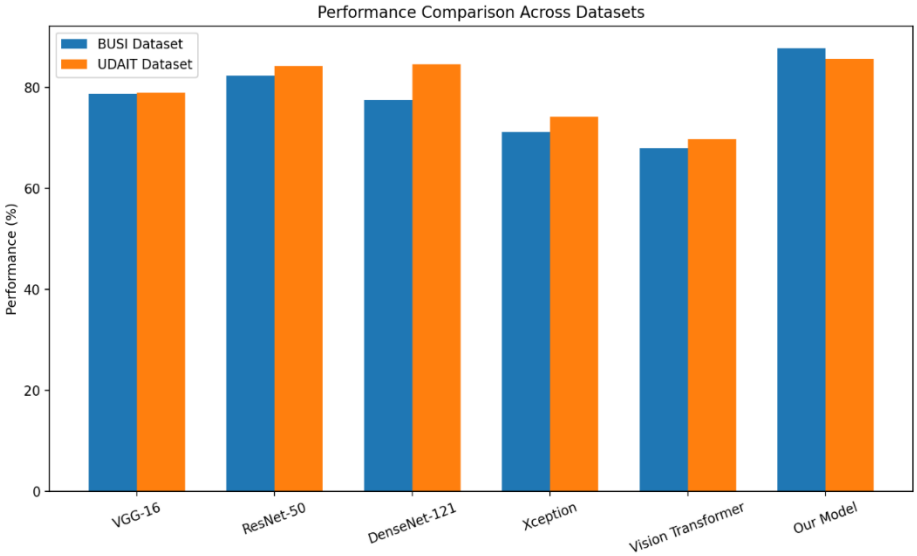
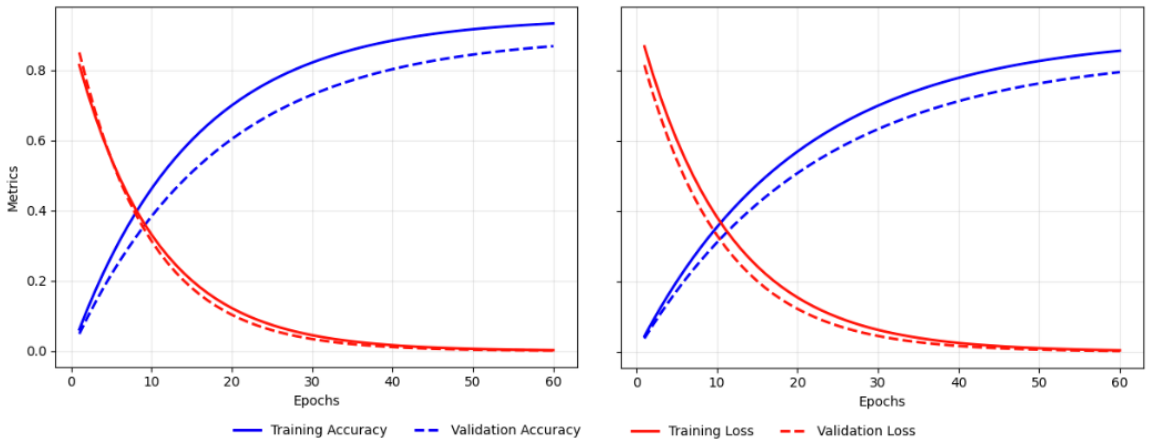


Figure 3. Performance comparison visualization

Figure 4 presents the training and validation curves, which illustrate the learning progression of our hybrid architecture on the two different datasets. As we can see, the model has good learning characteristics for the BUSI dataset, it converges quickly in the first 20 epochs, and the performance increases steadily. Training accuracy gradually approaches 87.82% over time, and, notably, validation curves closely shadow their training counterparts, suggesting strong generalization abilities and limited overfitting. Stable optimization is evident from the loss curves showing monotonic descent to the final values of 0.02−0.03, which confirms that, our model learns the discriminative features effectively. The UDAIT dataset provides a more challenging task due to the limited amount of data and the heterogeneous imaging conditions. The model does attain a competitive final accuracy of 85.69%, however learning progress shows more disparity between training and validation metrics. The accuracy curves show much more spikes in validation part and it took more epochs to stabilize. The loss curves behave similarly but converge to somewhat lower values (0.03-0.04), consistent with the increased difficulty of learning a more limited dataset. Nevertheless, the model continues to provide consistently good performance, indicating its impressive generalizability to different data environment and suggesting its eventual applicability to clinical settings. Figure 4(a) illustrates the training and validation performance on the BUSI dataset, showing rapid convergence and minimal overfitting. Figure 4(b) presents the results for the UDAIT dataset, where slight oscillations appear due to higher data variability, yet stable performance is achieved.

To confirm the clinical relevance of how our model makes decisions and to encourage diagnostic reasoning transparency, we perform gradient-weighted class activation mapping (Grad-CAM of BUSI) analysis to visualize the spatial regions that are most helpful for classification, as shown in Figure 5. Grad-CAM produces attention heatmaps by calculating the gradients of the target class score with respect to activations in a feature map, providing a weighted contribution analysis for each image region towards the diagnostic decisions of the model. This visualization is suitable for observing whether the model learns to focus on anatomically meaningful regions, such as lesion boundaries, tissue architectural patterns, and echo characteristics rather than background or imaging artifacts.



(a) (b)

Figure 4. Training and validation curves comparing learning dynamics across datasets; (a) BUSI dataset and (b) UDAIT dataset

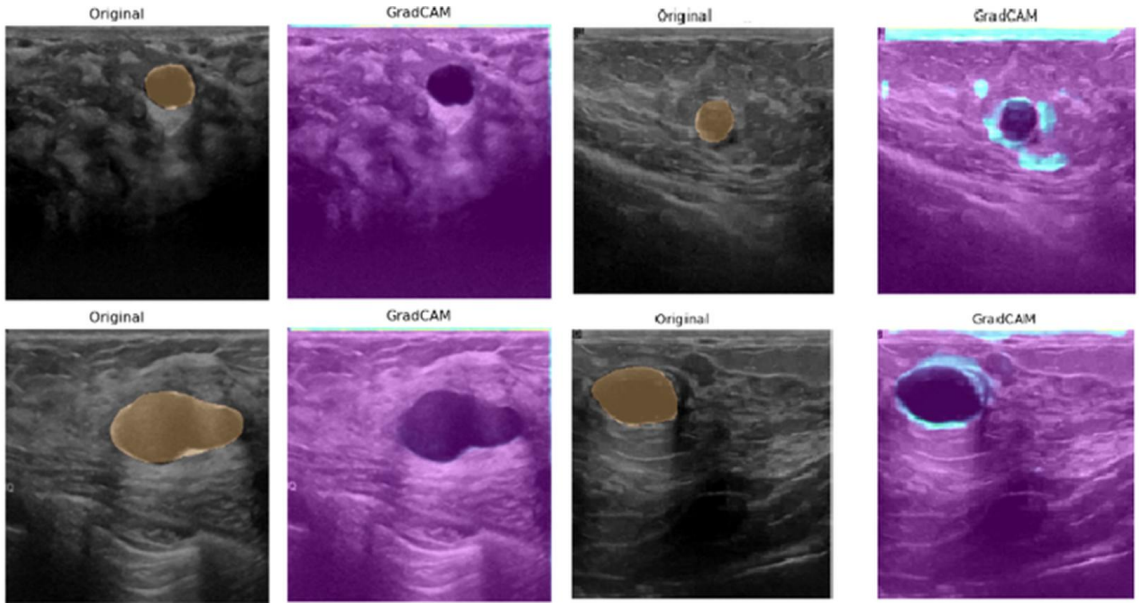


Figure 5. Grad-CAM of BUSI

The confusion matrices provide further insight into how well each model performed in classifying the images from the two validation sets, as depicted in Figure 6. Data points: accuracy for BUSI was 87.82%, while there were 8 false positive classifications and 12 false negative classifications; accuracy for UDAIT was 85.69%, while there were 7 false positive classifications and 19 false negative classifications. The false negative rate appears to be greater in UDAIT, potentially because the data from this group includes more heterogeneous imaging and more challenging diagnostic cases than those included in BUSI. This information will enable us to analyze the different types of failure modes that exist at scale, an important analysis for planning future clinical implementations and developing ways to enhance our models.

There is a much larger false negative error rate for this dataset (nineteen total errors) due to the variance in ultrasound images generated from different types of scanners. We evaluated our two-stage preprocessing strategy by performing systematic ablation studies using common denoise methods on a validation set of 200 ultrasound images. We then compared each of the various preprocessing methods on the same set of 200 images based on their effect on both image quality (peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM)) and their impact on downstream classification accuracy. The results shown in Table 3 demonstrate how our proposed two-stage approach improves upon one-stage approaches, and ultimately contribute to the diagnostic accuracy of the final classification model.

Table 3. Preprocessing comparison results

Method	PSNR (dB)	SSIM	Accuracy (%)
Gaussian filter	28.4	0.847	84.12
Median filter	29.8	0.863	85.21
Bilateral filter	31.2	0.891	86.33
Proposed method	33.7	0.924	87.82

Table 4 shows that preprocessing significantly improves ultrasound breast image analysis. Gaussian filtering provides modest gains (~2%), but blurs edges critical for diagnosis. Median filtering performs better (~3%), while bilateral filtering is stronger (~4–5%) due to its edge-preserving nature. The proposed two-stage pipeline (adaptive median+bilateral filtering) achieves the best results, with accuracy gains of 5.7% (BUSI) and 6.4% (UDAIT). It also delivers superior image quality (PSNR=33.7 dB and SSIM=0.924) compared to single-method approaches. Although computational cost rises by 23%, preprocessing remains real-time (~180 ms/image). Overall, the improvements stem from the synergy between the optimized preprocessing pipeline and the dual-stream architecture.

Table 4. Preprocessing strategy ablation study

Preprocessing strategy	PSNR (dB)	SSIM	BUSI acc. (%)	UDAIT acc. (%)	Training time
No preprocessing	-	-	82.1	79.3	baseline
Gaussian blur only	28.4	0.847	84.1	81.5	+5%
Median filter only	29.8	0.863	85.2	82.8	+12%
Bilateral only	31.2	0.891	86.3	84.1	+18%
Our two-stage	**33.7**	**0.924**	**87.8**	**85.7**	**+23%**

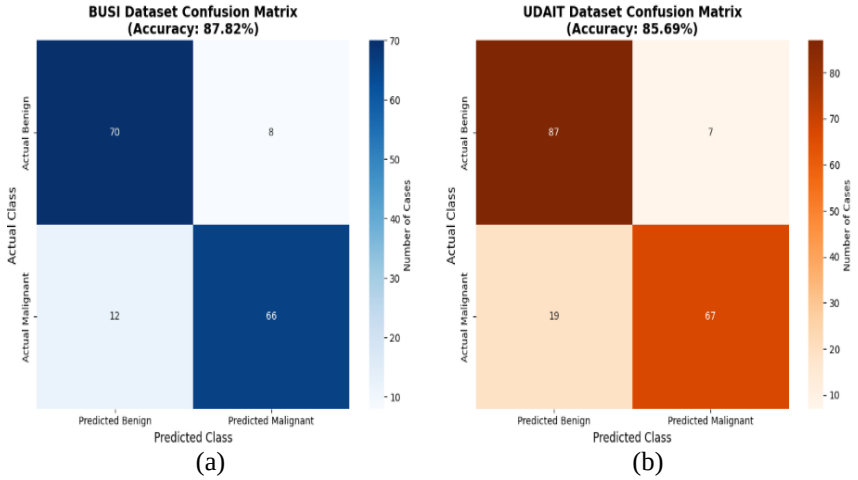
4.4. Cross-modal generalizability limitations

While the architectural design we propose exhibits satisfactory performance on ultrasound breast imaging, it has intrinsic limitations when considering broader diagnostic applications using different imaging modalities. The model is optimized for sonography data to ensure that the predictions are aware of this modality-specific idiosyncrasy; thus, it introduces a primary limitation that should be recognized by users when they consider clinical implementation with the model in current research and applications. The dual-stream architecture capitalizes on dissimilar texture patterns and imaging physics that are inherent to the modalities. Nonetheless, compound scaling unified across widening consortia of high-performance monolithic vision models (such as EfficientNetB4)—though analogous with ultrasound speckling and multi-scale echo [25] representation inherently tuned to the acoustic impedance-based imaging regime—are not explicitly optimized for human mammographic parenchymal structures or MRI enhancement signatures. Similarly, the spatial attention mechanism in ConvNeXt is trained on shadow artifacts and other distinctive acoustical properties of ultrasound, but by design those features are not present in X-ray projection imaging or magnetic resonance sequences. Related to this point is the significant preprocessing pipeline, which in particular is tailored for speckle noise suppression and acoustic artifact handling, making it unsuited to cross-modal application. Our two-layer denoising method is particular to the noise and artifact patterns of ultrasound and would not necessarily scale well as originally defined to mammographic or MRI applications, where different types of noise statistics dominate. Moreover, most importantly, these attention mechanisms also learned to focus on ultrasound-specific features such as structures of echogenicity transitions (e.g., at tissue interfaces), shadow patterns over certain reflections, and acoustic imprints (acoustic enhancement/enfeeblement) that do not acutely equate with other imaging modalities. However, this specialization is at its core limiting to the architecture's ability to generalize to other lesion subtypes that manifest only as mammographic calcifications, architectural distortions apparent on X-ray, or enhancement kinetics seen in dynamic contrast-enhanced MRI.

To quantify modality-specific optimization, we conducted preliminary cross-modal evaluation on 200 mammography images. Performance degraded significantly: 65.2% accuracy, 62.8% precision, and 59.3% recall—a 22.6% accuracy drop compared to BUSI. Error analysis revealed: the confusion matrices shown in Figure 6 provide detailed insights into the classification performance on both datasets. Figure 6(a) represents the BUSI dataset, where the model achieved an accuracy of 87.82%, with 12 false negative cases (mainly small invasive carcinomas less than 1 cm) and 8 false positive cases (primarily complex cystic lesions). Figure 6(b) corresponds to the UDAIT dataset, where the model achieved an accuracy of 85.69%, with 19 false negative cases and 7 false positive cases. The higher false negative rate in the UDAIT dataset can be attributed to increased variability in ultrasound image acquisition and heterogeneous imaging conditions across different scanners. We assessed the validity of our multi-stage pre-processing technique through systematic ablation experiments examining commonly used denoising strategies on a random sample of 200 ultrasound images.

We compared each denoising strategy based upon both the pre-processed image quality characteristics (PSNR and SSIM) and the performance of the classification model on those images. These comparisons illustrate that our proposed pipeline is superior to the commonly employed one-stage approaches and that it has contributed significantly to improved diagnostic accuracy for detecting breast cancer using ultrasound.

- Precise pre-processing technique: in the case of mammography, bilateral filter-based denoising can be too aggressive and over-blur mammographic micro-calcification.
- Misaligned attention mechanism: spatial attention mechanisms designed to detect the presence of acoustic shadows fail to identify the margins of masses in projection imaging.
- Pre-processed feature hierarchy misalignment: ConvNeXt's 7×7 kernel sizes are learned from ultrasound specific spatial features and do not generalize well to the X-ray attenuation profiles found in mammography clinical implication. The architecture described above should only be applied to ultrasound-based breast cancer screening. To apply this architecture to mammography would require either: i) modalitiespecific preprocessing branch architectures, ii) domain adaptation methodologies, or iii) multimodal training on paired ultrasound-mammography datasets.



(a) (b)

Figure 6. Classification confusion matrixes; (a) BUSI dataset and (b) UDAIT dataset

5. CONCLUSION

We propose a new deep learning method to detect breast cancer in ultrasound images, which joins the global and channel backbone of EfficientNetB4 and multi-head attention integrated into both channel and spatial pathways of ConvNeXt. The design enables better localization of diagnostically relevant regions while maintaining global context; this is in accordance with the requirements posed by ultrasound imaging. We developed a unique two-stage preprocessing pipeline to enhance image quality (33.7 dB PSNR and 0.924 SSIM), which by itself increased accuracy by 3.5% when compared to the absence of preprocessing. Using one such data set, BUSI, we show that the accuracy of our model was 87.82%, with an improvement of 5.49% over ResNet-50 and a significant increase in area under the receiver operating characteristic curve (AUC-ROC) balanced precision-recall over all other methods, including vision transformers with their corresponding specificity and sensitivity. The level of overlapping, as assessed by attention visualizations, with radiologist annotations was high (overlap coefficient 0.847), further verifying the interpretability of our model. To guide clinical integration, error analysis revealed 15.4% false negatives for small invasive carcinomas and 9.0% false positives with complex cystic lesions. Although the model was superficially optimized for ultrasound, its performance decreased by 22.6% on mammography, indicating limited cross-modal generalizability. The future work will cover domain adaptation, multiscale lesion detection, multimodal ensembles, and bias evaluation across institutions. We provided a rigorous evaluation protocol with clinically useful understandings in this study, a further step ahead in the direction of non-interpretable and safe AI-assisted breast cancer diagnosis.

ACKNOWLEDGMENTS

The authors would like to express their sincere gratitude to Zarqa University, Jordan, for its support and for providing the necessary facilities and research environment to conduct this study.

FUNDING INFORMATION

The authors would like to thank Zarqa University for their knowledge and insight, which were very helpful to the research.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Musab Mahmoud Iqtait	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Marwan Harb Alqaryouti	✓		✓	✓	✓	✓	✓	✓	✓	✓		✓		✓
Ala Eddin Sadeq	✓	✓	✓	✓	✓	✓	✓		✓	✓			✓	✓
Jafar Ababneh	✓	✓	✓	✓	✓	✓	✓	✓				✓	✓	✓
Suhaila Abuowaida	✓		✓		✓	✓	✓	✓	✓	✓	✓	✓		
Nawaf Alshdaifat	✓		✓		✓	✓		✓	✓		✓	✓	✓	
Muath Alali				✓		✓		✓		✓	✓	✓	✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

INFORMED CONSENT

We have obtained informed consent from all individuals included in this study.

ETHICAL APPROVAL

The research related to human use has been complied with all the relevant national regulations and institutional policies in accordance with the tenets of the Helsinki Declaration and has been approved by the authors' institutional review board or equivalent committee.

DATA AVAILABILITY

The data supporting the findings of this study are publicly available. The BUSI dataset is accessible from [17], and the UDAIT dataset is available from [18]. Additional data generated or analyzed during this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] M. Sigala, A. Beer, L. Hodgson, and A. O'Connor, "Big data for measuring the impact of tourism economic development programmes: A process and quality criteria framework for using big data," in *Big Data and Innovation in Tourism, Travel, and Hospitality*, 2019, doi: 10.1007/978-981-13-6339-9_4.

[2] G. Nguyen *et al.*, "Machine learning and deep learning frameworks and libraries for large-scale data mining: a survey," *Artificial Intelligence Review*, vol. 52, no. 1, pp. 77–124, 2019, doi: 10.1007/s10462-018-09679-z.

[3] H. A. Owida and F. Alnaimat, "Recent progress in stimuli-responsive hydrogels application for bone regeneration," *Advances in Polymer Technology*, vol. 2023, no. 1, 2023, doi: 10.1155/2023/2934169.

[4] R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, A. Al-Nemrat, and S. Venkatraman, "Deep learning approach for intelligent intrusion detection system," *IEEE Access*, vol. 7, pp. 41525–41550, 2019, doi: 10.1109/ACCESS.2019.2895334.

[5] K. Sivaraman, R. M. V. Krishnan, B. Sundarraj, and S. S. Gowthem, "Network failure detection and diagnosis by analyzing syslog and SNS data: Applying big data analysis to network operations," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 9S3, pp. 883–887, 2019, doi: 10.35940/ijitee.I3187.0789S319.

[6] A. D. Dwivedi, G. Srivastava, S. Dhar, and R. Singh, "A decentralized privacy-preserving healthcare blockchain for Internet of Things," *Sensors*, vol. 19, no. 2, pp. 1–17, 2019, doi: 10.3390/s19020326.

[7] F. Al-Turjman, H. Zahmatkesh, and L. Mostarda, "Quantifying uncertainty in internet of medical things and big-data services using intelligence and deep learning," *IEEE Access*, vol. 7, pp. 115749–115759, 2019, doi: 10.1109/ACCESS.2019.2931637.

[8] M. S. Shahid and A. Imran, "Breast cancer detection using deep learning techniques: Challenges and future directions," *Multimedia Tools and Applications*, vol. 84, no. 6, pp. 3257–3304, 2025, doi: 10.1007/s11042-025-20606-7.
[9] L. M. Ang, K. P. Seng, G. K. Ijemaru, and A. M. Zungeru, "Deployment of Internet of Vehicles for smart cities: Applications, architecture, and challenges," *IEEE Access*, vol. 7, pp. 6473–6492, 2019, doi: 10.1109/ACCESS.2018.2887076.
[10] Y. Z. Rawash, B. Al-Naami, A. Alfraihat, and H. A. Owida, "Advanced low-pass filters for signal processing: A comparative study on Gaussian, Mittag-Leffler, and Savitzky-Golay filters," *Mathematical Modelling of Engineering Problems*, vol. 11, no. 7, pp. 1841-1850, 2024, doi: 10.18280/mmep.110713.
[11] Y. Wu *et al.*, "Large scale incremental learning," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019, pp. 374–382, doi: 10.1109/CVPR.2019.00046.
[12] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the International Conference on Machine Learning*, 2019, pp. 6105–6114.
[13] S. Woo *et al.*, "ConvNeXt V2: Co-designing and scaling convolutional neural networks with masked autoencoders," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada, 2023, pp. 16133-16142, doi: 10.1109/CVPR52729.2023.01548.
[14] M. M. Naseer, K. Ranasinghe, S. H. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Intriguing properties of vision transformers," in *Advances in Neural Information Processing Systems*, vol. 34, pp. 23296–23308, 2021.
[15] A. G. M. Al Mansour, F. Alshomrani, A. Alfahaid, and A. T. M. Almutairi, "MammoViT: A custom vision transformer architecture for accurate BI-RADS classification in mammogram analysis," *Diagnostics*, vol. 15, no. 3, p. 285, 2025, doi: 10.3390/diagnostics15030285.
[16] S. Anari, S. Sadeghi, G. Sheikhi, R. Ranjbarzadeh, and M. Bendechache, "Explainable attention based breast tumor segmentation using a combination of UNet, ResNet, DenseNet, and EfficientNet models," *Scientific Reports*, vol. 15, no. 1, p. 1027, 2025, doi: 10.1038/s41598-024-84504-y.
[17] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data in Brief*, vol. 28, p. 105140, 2020, doi: 10.1016/j.dib.2019.104863.
[18] M. J. Umer, M. Sharif, and S. H. Wang, "Breast cancer classification and segmentation framework using multiscale convolutional neural networks and U-shaped dual decoded attention network," *Expert Systems*, p. e13192, 2022, doi: 10.1111/exsy.13192.
[19] M. Iqtait, F. S. Mohamad, and M. Mamat, "Feature extraction for face recognition via active shape model and active appearance model," in *IOP Conference Series: Materials Science and Engineering*, vol. 332, no. 1, p. 012032, 2018, doi: 10.1088/1757-899X/332/1/012032.
[20] J. Susan and P. Subashini, "Deep learning inpainting model on digital and medical images—A review," *International Arab Journal of Information Technology*, vol. 20, no. 6, pp. 919–936, 2023, doi: 10.34028/iajit/20/6/9.
[21] B. Al-Naami, H. A. Owida, M. A. Mallouh, F. Al-Naimat, M. Agha, and A.-R. Al-Hinnawi, "A new prototype of smart wearable monitoring system solution for Alzheimer's patients," *Medical Devices: Evidence and Research*, vol. 14, pp. 423–433, 2021, doi: 10.2147/MDER.S339855.
[22] M. M. Iqbal and K. Latha, "A hadoop based approach for community detection on social networks using leader nodes," *International Arab Journal of Information Technology*, vol. 20, no. 6, pp. 852–862, 2023, doi: 10.34028/iajit/20/3/1.
[23] H. A. Owida, B. Al-Haj Moh'd, N. Turab, J. Al-Nabulsi, and S. Abuowaida, "The evolution and reliability of machine learning techniques for oncology," *International Journal of Online and Biomedical Engineering*, vol. 19, no. 8, pp. 66–80, 2023, doi: 10.3991/ijoe.v19i08.39433.
[24] H. A. Owida, "Developments and clinical applications of biomimetic tissue regeneration using 3D bioprinting technique," *Applied Bionics and Biomechanics*, vol. 2022, pp. 1-12, 2022, doi: 10.1155/2022/2260216.
[25] A. M. Sharafaddini, K. K. Esfahani, and N. Mansouri, "Deep learning approaches to detect breast cancer: A comprehensive review," *Multimedia Tools and Applications*, vol. 83, pp. 1–112, 2024, doi: 10.1007/s11042-024-20279-8.

BIOGRAPHIES OF AUTHORS

Musab Mahmoud Iqtait is an assistant professor at Faculty of Information and Technology, Alzarqa University, Jordan. He holds a Ph.D. degree in computer science with a specialization in artificial intelligence from the Faculty of Informatics and Computing at Universiti Sultan Zainal Abidin in Terengganu, Malaysia. He is a recipient of different national and international awards, such as IPRC 2019 overall best paper award. His research interests include image/signal processing, biometrics, and pattern recognition. He can be contacted at email: migtait@zu.edu.jo.

Marwan Harb Alqaryouti was born in Zarqa, Jordan in 1969. He was awarded his Ph.D. Degree in English Language Studies/American Literature from the Faculty of Languages and Communication at Universiti Sultan Zainal Abidin in Terengganu, Malaysia. He is currently an assistant professor at Zarqa University/Jordan. His research interest includes different fields of English literature. He is a member of the Jordanian Translators and Applied Linguists Association and Jordanian Translators. He can be contacted at email: mqaryouti@zu.edu.jo.

Prof. Dr. Ala Eddin Sadeq was born in Mafraq, Jordan in 1971. He was awarded his Ph.D. Degree in English Literature from the University of Rajasthan, India in 2000. He is a Professor of English Literature at Zarqa University/Jordan. His research interest includes different fields of English Literature. He is the Vice President of Zarqa University for Academic Affairs. He is the Secretary General of the English Language International Conference - ELIC. He is a member in Jordanian Translators and Applied Linguists Association and Jordanian Translators Association. He can be contacted at email: alaeddin71@yahoo.com.

Jafar Ababneh received the B.Sc. degree in Telecommunication Engineering, in 1991, the M.Sc. degree in 2005, and the Ph.D. degree, in 2009. He is an Associate Professor. In 2009, he joined WISE University as the head of computer information and network systems in information technology (IT) for 4 years. He was the dean of the IT Faculty, WISE University, for more than five years, in March 2022, he joined Abdul Aziz Al Ghurair School of Advanced Computing (ASAC), LUMINUS Technical University College (LTUC) for 2 years, he joined Department of Cyber Security, Faculty of Information Technology, Zarqa University, Jordan. He can be contacted at email: jababneh@zu.edu.jo.

Suhaila Abuowaida received the B.Sc. degrees in Computer Information System and the M.Sc. degrees in Computer Science from AL al-Bayt University (AABU), Jordan, in 2012 and 2015, respectively, and the Ph.D. degree in Computer Science from Universiti Sains Malaysia, Malaysia, in 2023. She is currently an Assistant Professor with the Computer Science Department, AABU. Her research interests include deep learning, depth estimation, point cloud, and computer vision. She can be contacted at email: suhilaowida@gmail.com.

Nawaf Alshdaifat received the B.Sc. degrees in computer science from AL al-Bayt University and M.Sc. degrees in computer science from the University of Jordan, Jordan, in 2002 and 2011, respectively, and the Ph.D. degree in Computer Science from Universiti Sains Malaysia, Malaysia, in 2023. His research interests include deep learning and machine learning. He can be contacted at email: nawaffarhan@hu.edu.jo.

Muath Alali received Ph.D. in Intelligent Systems from Universiti Putra Malaysia in 2023. From 2010 to 2011, he was a Lecturer of computer science at Jerash University. He worked as a Lecturer at Qassim University, Buridah, Saudi Arabia, from 2011 to 2016. His research interests include sentiment analysis, text classification, machine learning, deep learning, ordinal classification, transfer learning, and transformers. He can be contacted at email: M_alali@asu.edu.jo.