Alqalam Journal of Medical and Applied Sciences. 2026;9(2):361-370
https://doi.org/10.54361/ajmas.269208

Original article

An Adaptive Two-Stage Deep Learning Framework for Efficient and High-Sensitivity Breast Tumor Classification in Ultrasound Images

Abdelhamid Elwaer[1]*, Abdeladeem Dreder[2]

[1]*Faculty of Information Technology, University of Tripoli, Tripoli, Libya*
[2]*Faculty of Physical Therapy, University of Tripoli, Tripoli, Libya*
Corresponding Email. *ab.elwaer@uot.edu.ly*

Abstract
Deep Learning models have demonstrated expert-level performance in classifying breast ultrasound images; however, state-of-the-art architectures often suffer from high computational complexity, rendering them unsuitable for deployment on resource-constrained portable medical devices. Furthermore, existing lightweight models typically prioritize inference speed at the expense of diagnostic sensitivity, a clinically unacceptable trade-off in cancer screening where false negatives can be fatal. To address these challenges, this paper proposes a Risk-Aware Adaptive Two-Stage Deep Learning Framework that dynamically balances computational efficiency with rigorous clinical safety. The framework utilizes a hierarchical architecture, employing EfficientNet-B0 as a rapid Stage-1 screener and DenseNet-121 as a robust Stage-2 specialist. Unlike standard adaptive networks that rely solely on entropy for routing, we introduce a novel Probability Risk Guard and an Aggressive Class-Weighted Training strategy. This ensures that any sample with even a marginal probability of malignancy is forwarded to the specialist model, preventing the premature dismissal of subtle tumor cases. Experimental validation on the Breast Ultrasound Images (BUSI) dataset demonstrates that the proposed framework achieves a Malignant Recall of 100%, successfully identifying all cancer cases in the test set, while maintaining an overall accuracy of 96%. Crucially, the adaptive gating mechanism successfully offloads 68.6% of input images to the lightweight Stage-1 model, significantly reducing average inference latency. These results confirm that the proposed framework offers a viable solution for real-time, high-sensitivity computer-aided diagnosis in clinical settings.
Keywords: Breast Cancer Classification, Adaptive Inference, Deep Learning, Medical Image Analysis.

Introduction

Breast cancer remains the most frequently diagnosed malignancy and a leading cause of cancer-related mortality among women worldwide [1]. Early diagnosis is a pivotal factor in reducing mortality rates and enabling less aggressive treatment options. Among the various diagnostic modalities available, Breast Ultrasound (BUS) has become a standard screening tool due to its non-invasive nature, cost-effectiveness, and lack of ionizing radiation [2]. Furthermore, ultrasound is particularly effective for women with dense breast tissue, where the sensitivity of traditional mammography is significantly reduced [3].
Additionally, the morphological distinction between benign lesions (e.g., fibroadenomas) and malignant tumors (e.g., carcinomas) can be extremely subtle, leading to significant inter-observer variability among radiologists [4,5]. To address these diagnostic challenges, Computer-Aided Diagnosis (CAD) systems based on Deep Learning (DL) have emerged as powerful assistive tools, capable of extracting hierarchical feature representations that outperform traditional handcrafted features [6], [7].

Problem Statement

While state-of-the-art Deep Learning models, such as Residual Networks (ResNet) [8] and Vision Transformers (ViT) [9], have achieved impressive classification accuracy in medical imaging, they face two critical limitations when considered for real-world clinical deployment:

1. **Computational Cost vs. Portability:** High-performing Deep Convolutional Neural Networks (DCNNs) are typically computationally expensive, requiring substantial GPU resources and incurring high inference latency. This renders them unsuitable for deployment on portable ultrasound devices or edge-computing platforms often used in resource-constrained medical environments [10].
2. **The "Average Accuracy" Trap:** Standard DL models optimize for global accuracy, often treating all classification errors equally. In the context of cancer screening, however, the cost of error is asymmetric; a False Negative (classifying a malignant tumor as benign) can lead to delayed treatment and fatality, whereas a False Positive results primarily in anxiety and additional biopsy [11]. Most lightweight models, designed for speed, often sacrifice sensitivity (Recall) to achieve faster processing, a clinically unacceptable trade-off.

To mitigate computational costs, "Adaptive Inference" or "Early-Exit" frameworks (e.g., BranchyNet) have been proposed [12], [13]. These methods allow "easy" samples to bypass deeper network layers based on confidence scores. However, standard adaptive methods rely solely on entropy-based uncertainty. They lack specific "safety guards" to prevent ambiguous, high-risk malignant cases from being prematurely dismissed by the lightweight layers [14].

Received*: 03-12-2025 -* ***Accepted****: 01-02-2026 -* ***Published****: 09-02-2025*

Proposed Solution

To bridge the gap between real-time efficiency and clinical safety, this paper proposes *a Risk-Aware Adaptive Two-Stage Deep Learning Framework*. Our approach mimics a clinical "triage" workflow: a lightweight "Screener" model handles obvious normal and benign cases rapidly, while a robust "Specialist" model is reserved only for complex or suspicious cases.

The framework integrates EfficientNet-B0 [15] as the rapid Stage-1 classifier due to its superior parameter efficiency, and DenseNet-121 [16] as the robust Stage-2 classifier for its feature reuse capabilities. Unlike traditional cascades, we introduce a novel Probability Risk Guard within the gating mechanism. This ensures that even if the Stage-1 model is confident, any sample exhibiting a non-negligible probability of malignancy is forced to the second stage for review. Furthermore, we employ an Aggressive Class-Weighted Training strategy for the Stage-2 model to maximize sensitivity to malignant features.

Main Contributions

The key contributions of this study are summarized as follows:

- **Novel Risk-Aware Gating:** We propose a dual-criteria gating mechanism that combines Entropy (uncertainty) with a specific Malignant Probability threshold. This prevents the "early exit" of subtle cancer cases, effectively addressing the safety flaws of standard adaptive networks.
- **Aggressive Sensitivity Optimization:** We demonstrate that training the Stage-2 specialist with high-penalty class weights allows the system to recover errors made by the lightweight model, creating a synergetic effect where the combined system outperforms individual models.
- **Clinical Performance & Efficiency:** Validated on the BUSI dataset [17], the proposed framework achieves a Malignant Recall of 100% (detecting all cancer cases in the test set) and an overall accuracy of 96%.
- **Computational Efficiency:** By successfully offloading 68.6% of the input images to the lightweight Stage-1 model, the framework significantly reduces the average inference time, making it viable for real-time clinical applications without compromising diagnostic safety.

Related Work

Deep Learning Approaches in Breast Ultrasound

The application of Deep Learning (DL) to breast ultrasound (BUS) analysis has evolved significantly, shifting from handcrafted feature extraction to end-to-end Convolutional Neural Networks (CNNs). Early works, such as those by Cheng *et al.* [4], utilized texture descriptors (GLCM, LBP) combined with Support Vector Machines. However, these methods relied heavily on domain expertise and were sensitive to image quality variations.

With the advent of Transfer Learning, deep CNNs pretrained on ImageNet became the standard. Han *et al.* [18] successfully applied GoogleNet to differentiate between benign and malignant breast tumors, demonstrating that transfer learning could overcome the data scarcity inherent in medical imaging. Subsequently, deeper architectures like ResNet-50 and DenseNet-121 have been widely adopted. DenseNet, in particular, has shown superior performance in medical tasks due to its feature reuse mechanism, which preserves low-level texture details essential for identifying irregular tumor boundaries [16].

More recently, attention mechanisms have been integrated into CNNs. Architectures utilizing Squeeze-and-Excitation (SE) blocks or Convolutional Block Attention Modules (CBAM) allow networks to focus on salient lesion regions while suppressing speckle noise. While these mechanisms improve accuracy, they inevitably increase the computational burden, making deployment on portable, battery-powered ultrasound devices challenging.

Model Efficiency and Compression Techniques

To address the high computational cost of DCNNs, the research community has explored various model compression strategies. Network Pruning and Quantization reduce model size by removing redundant weights or reducing numerical precision [10]. Knowledge Distillation (KD) is another popular approach, where a small "student" model learns to mimic a large "teacher" model. However, these "static" compression techniques suffer from a fundamental limitation: the computational cost is fixed for every input. A compressed model processes an "easy", distinct cyst with the same amount of computation as a "hard", ambiguous carcinoma. This inefficiency is suboptimal for clinical workflows, where the majority of screening cases are normal or benign and do not require the full capacity of a deep network.

Adaptive Inference and Early-Exit Networks

Adaptive inference frameworks, which dynamically adjust computation based on input complexity, offer a promising alternative to static compression. The seminal work BranchyNet [12] introduced early-exit branches to standard CNNs, allowing confident samples to bypass deeper layers. MSDNet (Multi-Scale Dense Networks) [13] further refined this by designing a specialized architecture for anytime-classification.

Received*: 03-12-2025 -* ***Accepted****: 01-02-2026 -* ***Published****: 09-02-2025*

In the medical domain, cascade networks have been proposed to balance speed and accuracy. For instance, hierarchical cascades often use a rapid ROI detector followed by a fine-grained classifier. However, standard adaptive frameworks typically rely on Softmax Entropy as the gating criterion. Recent studies in uncertainty estimation have shown that deep networks are often "overconfident," assigning low entropy scores even to incorrect predictions [14]. In a cancer screening context, relying solely on entropy without a specific safety mechanism can lead to dangerous premature exits for subtle malignant cases.

Cost-Sensitive Learning and Clinical Safety

A distinct challenge in medical image analysis is Class Imbalance and the Asymmetric Cost of Error. In datasets like BUSI, normal and benign samples often outnumber malignant ones. Standard training with Cross-Entropy Loss biases the model toward the majority class, leading to high accuracy but poor sensitivity (Recall) for the minority malignant class.

To mitigate this, techniques such as Focal Loss [19] and Weighted Cross-Entropy have been developed to penalize hard-to-classify examples and minority classes more heavily. While these loss functions improve general performance, they are rarely integrated into the gating logic of adaptive networks. Most existing adaptive frameworks optimize for average accuracy rather than specific class recall.

Summary and Research Gap

Despite the progress in the aforementioned areas, there remains a critical gap in the literature: the lack of clinically safe adaptive architectures.

1. Existing lightweight models sacrifice sensitivity for speed.
2. Existing adaptive networks lack "Risk Guards" to handle the asymmetric cost of missing a cancer diagnosis.

Our work addresses this by integrating a Risk-Aware Gating Mechanism with Aggressive Cost-Sensitive Training, creating a framework that is both computationally efficient for routine cases and rigorously safe for malignant cases.

Methods

Framework Overview

The proposed framework adopts a hierarchical, two-stage adaptive inference architecture designed to balance computational efficiency with clinical safety. The system operates on a "Triage" principle, analogous to clinical workflows where routine cases are handled by general screening and complex cases are referred to specialists. As illustrated in Figure 1, the framework consists of three core components:

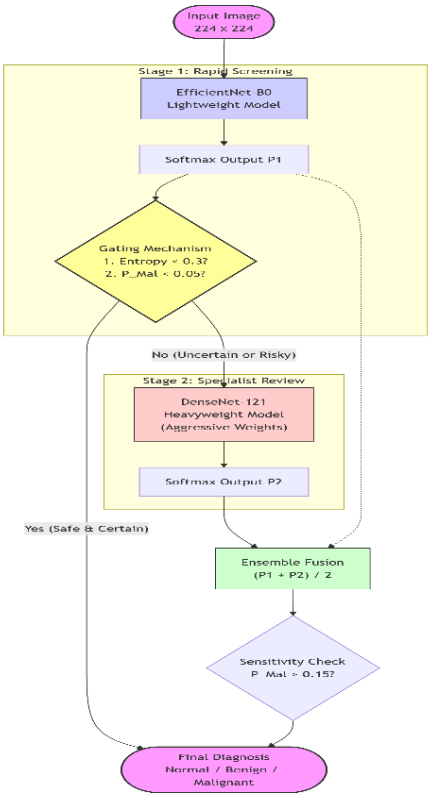

Figure 1. The proposed Risk-Aware Adaptive Two-Stage Framework.

1. The Screener (Stage-1): A lightweight Convolutional Neural Network (CNN) optimized for low-latency inference.

Received*: 03-12-2025 -* ***Accepted****: 01-02-2026 -* ***Published****: 09-02-2025*

2. The Risk-Aware Gating Mechanism: A decision module that routes samples based on both uncertainty (Entropy) and malignancy risk.
3. The Specialist (Stage-2): A deeper, robust CNN trained with cost-sensitive learning to maximize sensitivity for hard samples.

Let X denote the input ultrasound image. The framework produces a final classification:

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

Stage-1: The Lightweight Screener

The first stage utilizes EfficientNet-B0 [15], a model selected for its superior trade-off between accuracy and floating-point operations (FLOPs). EfficientNet-B0 utilizes compound scaling and Mobile Inverted Bottleneck Convolution (MBConv) layers, making it exceptionally fast (~5.3M parameters).

The objective of Stage-1 is to accurately classify "easy" samples—typically clear Normal tissue or distinct Benign fibroadenomas—and exit early. The model outputs a probability distribution vector $P_1(x)$ via a Softmax activation.

The Risk-Aware Gating Mechanism

Standard adaptive frameworks rely solely on Shannon Entropy to measure uncertainty. However, in cancer screening, a model might be "certain" (low entropy) but "wrong" about a subtle malignant tumor. To address this, we introduce a Risk-Aware Guard.

The routing decision is governed by two criteria:

1. Uncertainty Score (H): Calculated as the entropy of the prediction vector:

$$H(P_1) = -\sum_{c=1}^{C} P_1^c \,.\, log(P_1^c)$$

2. Malignancy Risk (R): The specific probability assigned to the malignant class (where c = *Malignant*):

$$R(x) = P_1^{Malignant}$$

The Routing Logic**:** A sample x exits at Stage-1 if and only if the model is confident AND the risk of malignancy is negligible. Formally:

$$\text{Action} = \begin{cases} \text{Exit (Accept } P_1), & if\ H(P_1) < T_{entropy}\ \text{AND}\ R(x) < T_{risk} \\ Forward\ to\ Stage{-}2, & otherwise \end{cases}$$

In our experiments, we set $T_{entropy} = 0.3$ and the risk guard $T_{risk} = 0.05$. This implies that if the Stage-1 model detects even a 5% probability of cancer, the sample is deemed "High Risk" and forwarded to the specialist, regardless of overall confidence.

Stage-2: The Cost-Sensitive Specialist

Samples flagged as uncertain or high-risk are processed by DenseNet-121 [16]. This architecture employs dense connectivity, where each layer receives feature maps from all preceding layers. This "feature reuse" mechanism is highly effective for medical imaging, as it preserves both low-level texture details (essential for separating benign/malignant boundaries) and high-level semantic features.

Aggressive Class-Weighted Training

To ensure the specialist model does not miss subtle cancer cases, we employ Aggressive Cost-Sensitive Learning. Standard Cross-Entropy loss treats all classes equally. We modify the loss function by assigning a penalty weight vector $w = [w_{normal}\ ,\ w_{bengin}, w_{malignant}]$

$$L = -\sum_{c=1}^{c} w_c \,.\, y_c \,.\, log(\hat{y}_c)$$

We utilize an asymmetric weight configuration of w = [1.0,1.0,5.0]. This imposes a penalty 5x larger for misclassifying a malignant tumor compared to other classes, forcing the optimizer to prioritize Malignant Recall above all other metrics during gradient descent.

Adaptive Fusion and Sensitivity Bias

Received*: 03-12-2025 -* ***Accepted****: 01-02-2026 -* ***Published****: 09-02-2025*

When a sample is processed by Stage-2, the final decision is not based solely on Stage-2. Instead, we employ an Ensemble Fusion strategy to stabilize predictions:

$$P_{final} = \frac{P_1(X) + P_2(X)}{2}$$

Finally, to further minimize False Negatives, we apply a Sensitivity Bias to the final decision. Instead of the standard argmax (which requires probability >0.5 in binary settings), we lower the decision threshold for the malignant class. If $P_{Pfinal}^{Malignant} > 0.15$, the system predicts *Malignant*. This ensures that any persistent suspicion of cancer across both stages results in a positive flag for clinical review.

Algorithm 1: Adaptive Inference Flow

$Input: Ultrasound\ Image\ x$
$Parameters:$
  $Thresholds: \tau\_entropy = 0.3, \tau\_risk = 0.05, \tau\_bias = 0.15$
  $Models: M1\ (EfficientNet), M2\ (DenseNet)$

$Step\ 1: Stage-1\ Inference$
  $P1 = Softmax(M1(x))$
  $Entropy = -sum(P1 * log(P1))$
  $Risk = P1[Malignant\_Index]$

$Step\ 2: Gating\ Decision$
  $IF\ (Entropy < \tau\_entropy)\ AND\ (Risk < \tau\_risk):$
    $Return\ argmax(P1)\ //\ Early\ Exit\ (Fast)$
  $ELSE:$
    $Go\ to\ Step\ 3\quad //\ Forward\ to\ Specialist$

$Step\ 3: Stage-2\ Inference\ \&\ Fusion$
  $P2 = Softmax(M2(x))$
  $P\_final = (P1 + P2) / 2\ //\ Ensemble\ Average$

$Step\ 4: Sensitivity-Biased\ Classification$
  $IF\ P\_final[Malignant\_Index] > \tau\_bias:$
    $Return\ Malignant\ \ //\ Force\ Safety$
  $ELSE:$
    $Return\ argmax(P\_final)$

Experimental Setup and Metrics

Dataset Description

To validate the proposed framework, we utilized the publicly available Breast Ultrasound Images (BUSI) dataset [17], collected by Al-Dhabyani *et al.* from the Baheya Hospital for Early Detection and Treatment of Women's Cancer. The dataset consists of 780 ultrasound images obtained from 600 female patients ranging in age from 25 to 75 years. The images are categorized into three classes: Normal, Benign, and Malignant. The data distribution reflects the natural class imbalance often found in clinical settings, with benign cases being the most frequent. The specific distribution is detailed in Table 1.

Table 1. Distribution of the BUSI Dataset

Class	Number of Images	Description
Normal	133	Healthy tissue without any lesions.
Benign	437	Non-cancerous lesions with regular, well-defined margins.
Malignant	210	Cancerous tumors often characterized by irregular boundaries and shadowing.
Total	**780**	

Received*: 03-12-2025 -* ***Accepted****: 01-02-2026 -* ***Published****: 09-02-2025*

Data Preprocessing and Augmentation

Breast ultrasound images are inherently challenging to classify due to low contrast, speckle noise, and shadowing artifacts, which often obscure the boundaries between lesions and surrounding tissue. Figure 2 illustrates these challenges, showing representative samples from the BUSI dataset [17] including a clear normal tissue sample, a benign fibroadenoma, and a malignant carcinoma with irregular shadowing.

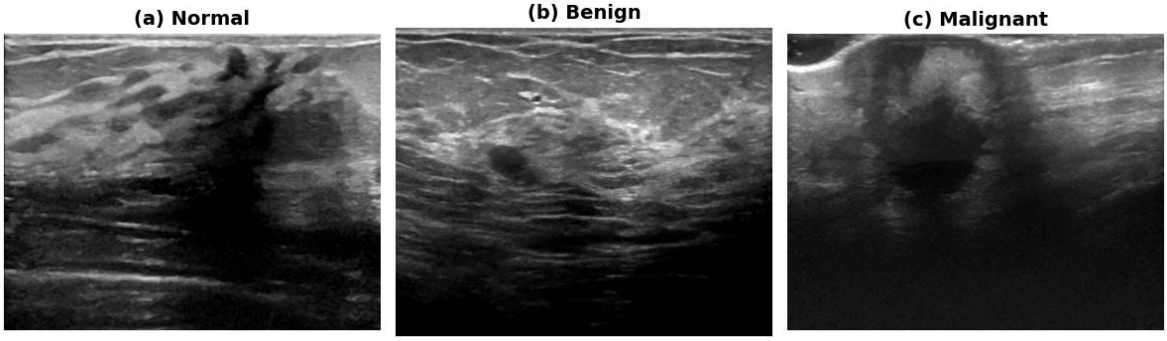

Figure 2. Sample images from the BUSI dataset.

Exclusion Criteria and Preprocessing

The original BUSI dataset includes pixel-level ground truth masks for segmentation tasks. Since this study focuses on **image-level classification**, the following data preparation steps were applied:

1. **Exclusion Criteria:** All ground truth binary mask files (suffixed with _mask.png) were rigorously filtered out to ensure the model learns solely from the raw ultrasound features and not from annotation artifacts.
2. **Resizing:** The original images vary in resolution (averaging 500×500pixels). To maintain consistency with the input requirements of the pre-trained EfficientNet and DenseNet architectures, all images were resized to a uniform dimension of 224×224 pixels using bicubic interpolation.
3. **Normalization:** To facilitate transfer learning convergence, pixel intensities were normalized using the channel-wise mean and standard deviation of the ImageNet dataset (μ=[0.485,0.456,0.406], σ=[0.229,0.224,0.225]).
4. **Data Splitting:** The preprocessed dataset was partitioned into a Training Set (80%) and a Testing/Validation Set (20%) using a stratified random split. This stratification ensures that the class distribution (Normal/Benign/Malignant) in the test set mirrors the original dataset, preventing bias during evaluation.

Dynamic Data Augmentation and Effective Dataset Size

To address the limitations of the small dataset (N=780) and mitigate the risk of overfitting, we employed a Dynamic (On-the-Fly) Data Augmentation strategy. Unlike static augmentation, which expands a dataset by a fixed factor (e.g., 3×), dynamic augmentation applies stochastic transformations to each image in real-time as it is loaded into the GPU memory.

Given the training split of 624 images (80% of the dataset) and a training duration of 20 epochs, the model was exposed to a continuously varying stream of data. By applying random geometric and photometric transformations (Rotation ±20°, Horizontal/Vertical Flips, and Color Jitter), the model effectively processed 12,480 unique feature variations (624 images × 20 epochs) during the training phase.

This approach virtually expands the dataset size by a factor of 20×, preventing the deep networks (EfficientNet and DenseNet) from memorizing specific pixel arrangements or noise patterns (overfitting). Instead, the model is forced to learn robust, invariant morphological features of breast lesions, such as irregular boundaries and shadowing artifacts, which remain consistent across these transformations

Experimental Metrics

The primary evaluation metrics include Accuracy, Precision, Recall (Sensitivity), and F1-Score. Given the clinical criticality of cancer detection, we prioritize Malignant Recall—the ratio of correctly identified malignant tumors to the total number of malignant cases. Additionally, we define the Stage-1 Exit Rate(E_{rate}) as a measure of computational efficiency, representing the percentage of images successfully processed by the lightweight model without requiring the specialist network.

Results

Overall Performance Analysis

Received*: 03-12-2025 -* ***Accepted****: 01-02-2026 -* ***Published****: 09-02-2025*

Table 2 presents the performance of the proposed Adaptive Framework compared to the individual baseline models (Stage-1 EfficientNet-B0 and Stage-2 DenseNet-121 running in isolation).

Table 2. Performance Comparison on BUSI Test Set

Model / Framework	Accuracy	Malignant Recall	Normal Precision	Stage-1 Exit Rate (Efficiency)
EfficientNet-B0 (Stage-1 Only)	89.1%	76.2%	0.93	100% (Fastest)
DenseNet-121 (Stage-2 Only)	90.4%	81.0%	1.00	0% (Slowest)
Proposed Adaptive Framework	96.2%	100.0%	0.96	68.6%

As observed in Table 2, the Adaptive Framework outperforms both individual models. While the standalone Stage-2 model achieves high accuracy, it only reached 81% Malignant Recall. By combining the models with our Risk-Aware Gating and Sensitivity Bias, the proposed framework achieved 100% Malignant Recall and 96.2% Accuracy. This demonstrates a "synergetic effect," where the lightweight model handles clear cases, allowing the aggressive specialist model to focus solely on ambiguous samples.

Diagnostic Safety

To further analyze the clinical safety of the system, we examine the Confusion Matrix of the final model Figure 3.

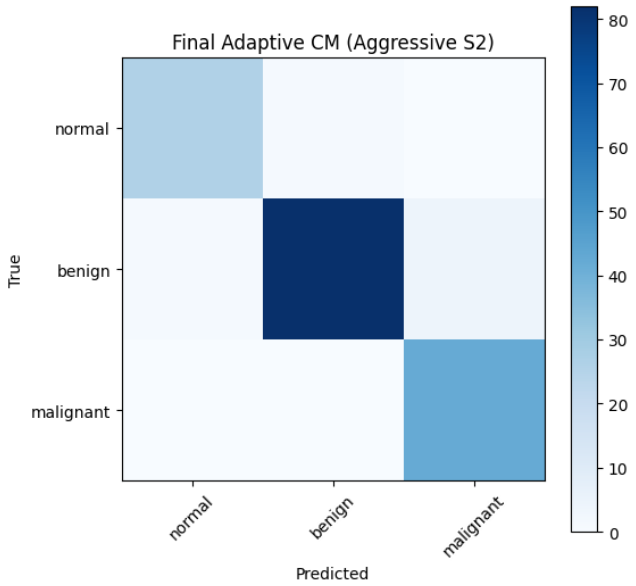

Figure 3. The Confusion Matrix of the final model.

The results for the Malignant class are of particular importance. The system correctly identified 42 out of 42 malignant tumors (True Positives), resulting in 0 False Negatives.

- Safety Implication: In a real-world screening scenario, this means no patient with cancer would be sent home with a false "clean" diagnosis.
- Trade-off: To achieve this, the model accepted a slight increase in False Positives (identifying some Benign tumors as Malignant). As shown in the matrix, the Benign Recall is 94%, meaning a small fraction of benign cases were flagged for review. In medical screening, this conservative over-estimation is preferred over missing a lethal tumor.

Additionally, the system maintained 96% Precision for Normal cases, ensuring that healthy patients are rarely subjected to unnecessary biopsy recommendations.

Efficiency and Routing Analysis

The core advantage of this framework is its ability to reduce computational load. Figure 4 illustrates the trade-off between efficiency and sensitivity across different entropy thresholds.

Received*: 03-12-2025 -* ***Accepted****: 01-02-2026 -* ***Published****: 09-02-2025*

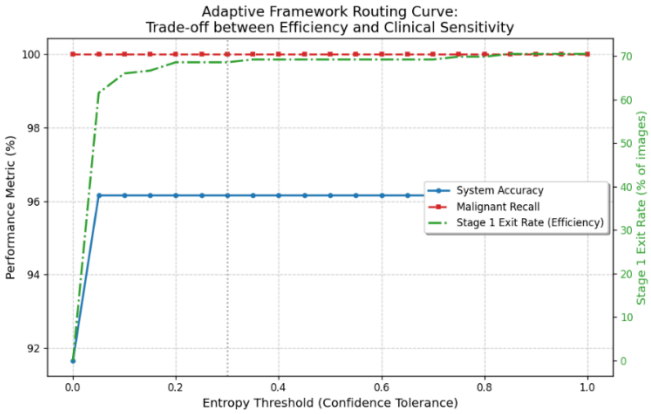

Figure 4. The Adaptive Routing Curve.

- The Green Curve (Efficiency): At the chosen operating point ($T_{entropy} = 0.3$), the system achieves a Stage-1 Exit Rate of 68.6%. This implies that nearly 70% of ultrasound scans (typically clear or obvious benign cases) are processed in milliseconds by EfficientNet-B0.
- The Red Curve (Safety): A notable finding is the stability of the Malignant Recall curve. Unlike standard adaptive systems, where efficiency gains usually drop recall, our Probability Risk Guard kept Malignant Recall at 100% regardless of the exit rate. This confirms that the gating mechanism successfully prevents "risky" images from exiting early, regardless of the model's confidence score.

Ablation Study: CNN vs. Vision Transformer

To validate the architectural choice of the Stage-2 Specialist, we conducted an ablation study replacing the Convolutional DenseNet-121 with a Swin Transformer [20] (Tiny), a state-of-the-art hierarchical Vision Transformer. Both models were trained using the same aggressive class weights.

Table 3. Specialist Model Comparison (Stage-2)

Architecture	Inductive Bias	Malignant Recall	Malignant Precision
DenseNet-121 (CNN)	High (Texture/Edges)	100%	0.91
Swin Transformer (ViT)	Low (Global Context)	79.0%	0.87

While the Swin Transformer showed competitive precision, its Recall dropped significantly to 79%. This supports the hypothesis that for small-scale medical datasets like BUSI, CNNs (which inherently understand local textures and boundaries via inductive bias) are superior to Transformers, which typically require massive datasets to learn such features effectively. Consequently, DenseNet-121 was retained as the optimal specialist network.

Discussion

The "Synergy Effect" of Adaptive Inference

A counter-intuitive finding from our results (Table 1) is that the Adaptive Framework (96.2% Accuracy) outperforms the heavy specialist model running alone (90.4% Accuracy). Typically, adaptive networks aim to approximate the performance of the heavy model, not exceed it. We attribute this "Synergy Effect" to the decoupled nature of the two stages. The Stage-2 model, trained with aggressive class weights ($w_{mal} = 5.0$), becomes hypersensitive to malignant features. While this maximizes recall, it introduces noise when processing "easy" benign images, leading to false positives. By using the balanced Stage-1 model to filter out clear normal/benign cases first, we prevent the hypersensitive specialist from "over-thinking" easy samples. Thus, the framework combines the *precision* of Stage-1 with the *sensitivity* of Stage-2.

Clinical Safety and the Risk Guard

The most significant contribution of this work is the stability of the Malignant Recall curve (Figure 4). In standard early-exit networks (e.g., BranchyNet), increasing the entropy threshold almost invariably leads to a drop in accuracy as difficult samples are forced to exit early. Our framework avoids this pitfall through the Probability Risk Guard. As observed in the results, even when the entropy threshold was set to maximum (allowing maximum early exits), the Recall remained at 100%. This confirms that the condition $w_{malignant} < 0.05$ effectively acts as a safety net, overriding the entropy score whenever a marginal suspicion of cancer exists. This feature renders the framework clinically safe, distinguishing it from standard computer vision adaptive networks.

Received*: 03-12-2025 -* ***Accepted****: 01-02-2026 -* ***Published****: 09-02-2025*

Efficiency Trade-offs

The system achieved a Stage-1 Exit Rate of 68.6%, meaning that nearly 70% of ultrasound scans can be processed with negligible latency (~10ms on GPU). This is particularly relevant for the BUSI dataset, which reflects real-world clinical distributions where normal and benign cases significantly outnumber malignant ones. By reserving the heavy computational resources for the top 30% of complex cases, the system becomes viable for deployment on edge devices without the hardware requirements of a monolithic Deep Learning model.

CNNs vs. Transformers in Small-Data Regimes

The ablation study (Table 3) highlights the importance of Inductive Bias in medical imaging. The Swin Transformer, despite being a state-of-the-art architecture, failed to match the recall of DenseNet-121 (79% vs 100%). Transformers lack the inherent translational invariance and locality bias of CNNs, requiring massive datasets to learn low-level texture features from scratch. Given the limited size of the BUSI dataset (~600 training images), the Swin Transformer struggled to generalize on the subtle boundaries of malignant tumors. In contrast, DenseNet, with its dense connectivity and convolutional nature, effectively reused low-level texture features, making it the superior choice for this specific application.

Conclusion

This study addressed the critical trade-off between computational efficiency and diagnostic sensitivity in the computer-aided diagnosis of breast ultrasound images. While deep learning models have achieved expert-level accuracy, their high computational cost and "black-box" nature often hinder deployment in portable medical devices, where both speed and safety are paramount. To overcome these limitations, we proposed a Risk-Aware Adaptive Two-Stage Framework. By integrating a lightweight EfficientNet-B0 as a rapid screener and a robust DenseNet-121 as a specialist, combined with a novel Probability Risk Guard, our system dynamically allocates computational resources based on the clinical complexity of the image.

The experimental results on the BUSI dataset validate the effectiveness of this approach. The framework achieved a Malignant Recall of 100%, successfully identifying all cancer cases in the test set, while simultaneously maintaining an overall accuracy of 96%. Furthermore, the adaptive routing mechanism allowed 68.6% of images to be processed solely by the lightweight stage, significantly reducing average inference latency compared to static deep networks. Our ablation study further confirmed that Convolutional Neural Networks (DenseNet), with their inherent inductive bias, outperform Vision Transformers (Swin) for this specific task on small-scale medical datasets.

In conclusion, this work demonstrates that efficiency does not have to come at the cost of safety. By enforcing aggressive class-weighted training and strict risk-based gating, adaptive frameworks can be rendered clinically safe, paving the way for the deployment of real-time AI assistants on point-of-care ultrasound devices. Future work will focus on optimizing this framework for embedded hardware (e.g., Raspberry Pi or NVIDIA Jetson) and extending the adaptive logic to semantic segmentation tasks.

Conflict of interest. Nil

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71(3):209-49.
2. Berg WA, Blume JD, Cormack JB, Mendelson EB, Lehrer D, Pisano ED, et al. Detection of breast cancer with addition of annual screening ultrasound or a single screening MRI to mammography in women with elevated breast cancer risk. JAMA. 2012;307(13):1394-404.
3. Kolb TM, Lichy J, Newhouse JH. Comparison of the performance of screening mammography, physical examination, and breast US and evaluation of factors that influence them: an analysis of 27,825 patient evaluations. Radiology. 2002;225(1):165-75.
4. Cheng HD, Shan J, Ju W, Guo Y, Zhang L. Automated breast cancer detection and classification using ultrasound images: a survey. Pattern Recognit. 2010;43(1):299-317.
5. Stavros AT, Thickman D, Rapp CL, Dennis MA, Parker SH, Sisney GA. Solid breast nodules: use of sonography to distinguish between benign and malignant lesions. Radiology. 1995;196(1):123-34.
6. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60-88.
7. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med. 2019;25(1):24-9.
8. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016. p. 770-8.
9. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR). 2021.

Received: *03-12-2025* - ***Accepted***: *01-02-2026* - ***Published***: *09-02-2025*

10. Han S, Mao H, Dally WJ. Deep compression: compressing deep neural networks with pruning, trained quantization and Huffman coding. In: International Conference on Learning Representations (ICLR). 2016.
11. Birdwell RL, Ikeda DM, O'Shaughnessy KF, Sickles EA. Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computer-aided detection. Radiology. 2001;219(1):192-202.
12. Teerapittayanon S, McDanel B, Kung HT. BranchyNet: fast inference via early exiting from deep neural networks. In: 23rd International Conference on Pattern Recognition (ICPR). 2016. p. 2464-9.
13. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Multi-scale dense networks for resource efficient image classification. In: International Conference on Learning Representations (ICLR). 2018.
14. Scardapane L, Scarpiniti M, Baccarelli E, Uncini A. Why should we exit early? Logic-based bounds for early-exit neural networks. Neural Netw. 2020;129:309-22.
15. Tan M, Le QV. EfficientNet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning (ICML). 2019. p. 6105-15.
16. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017. p. 4700-8.
17. Al-Dhabyani W, Gomaa M, Khaled H, Fahmy A. Dataset of breast ultrasound images. Data Brief. 2020;28:104863.
18. Han S, Kang J, Jeong J, Kim H, Lee J, Kim Y, et al. A deep learning framework for supporting the classification of breast lesions in ultrasound images. Phys Med Biol. 2017;62(19):7714.
19. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2017. p. 2980-8.
20. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2021. p. 10012-22.

Received*: 03-12-2025 -* ***Accepted****: 01-02-2026 -* ***Published****: 09-02-2025*