International Science and Technology Journal
المجلة الدولية للعلوم والتقنية
Volume 38 العدد
Part 1 المجلد
ISTJ

http://www.doi.org/10.62341/rara9506

Received	**2025/12/22**	**تم استلام الورقة العلمية في**
Accepted	**2025/01/04**	**تم قبول الورقة العلمية في**
Published	**2026/01/06**	**تم نشر الورقة العلمية في**

Distilling Vision Transformer Knowledge into MobileNet-V3 for Real-Time Breast Cancer Detection on Edge Devices

Abdelhamid Elwaer[1] and Abdeladeem Dreder[2]

1. Faculty of Information Technology, University of Tripoli, Libya
2. Faculty of Physical Therapy, University of Tripoli, Libya
ab.elwaer@uot.edu.ly, abd.dreder@uot.edu.ly

Abstract

The integration of Vision Transformers (ViTs) into medical imaging has significantly improved the diagnostic accuracy of breast ultrasound (BUS) analysis by capturing global semantic context. However, the excessive computational complexity of these models renders them unsuitable for Point-of-Care (POC) applications, where portable ultrasound devices rely on low-power, edge computing hardware. This study proposes a novel Cross-Architecture Knowledge Distillation framework designed to bridge the gap between high-performance diagnostics and real-time efficiency. We distill the structural knowledge of a computationally heavy Hybrid ViT-ConvNeXt Teacher into an ultra-lightweight MobileNet-V3 Student. By leveraging soft-target supervision, the student model inherits the global reasoning capabilities of the transformer while retaining the inductive bias and speed of a CNN. Experimental validation on an independent test set of the BUSI dataset demonstrates that the distilled student achieves a diagnostic accuracy of 95.06%, effectively matching the teacher model. Crucially, the student model reduces the storage footprint by 74x (from 438.8 MB to 5.9 MB) and accelerates inference speed by 15x, achieving a processing rate of 61.46 Frames Per Second (FPS) on a standard CPU. These results confirm that the proposed framework satisfies the latency requirements for real-time video analysis, enabling the deployment of specialist-level cancer detection on

http://www.doi.org/10.62341/rara9506

handheld, battery-powered ultrasound devices without the need for cloud connectivity or GPU acceleration.

Keywords: Breast Cancer Detection, Knowledge Distillation, Vision Transformers, Mo-bileNet-V3, Edge AI, Real-Time Ultrasound.

استخلاص المعرفة من تقنية Vision Transformer وتحويلها إلى MobileNet–V3 للكشف عن سرطان الثدي في الوقت الفعلي على الأجهزة الطرفية

عبد الحميد الواعر[1]، عبد العظيم دريدر[2]

1كلية تقنية المعلومات، جامعة طرابلس، طرابلس، ليبيا

2كلية العلاج الطبيعي، جامعة طرابلس، طرابلس، ليبيا

mo.alrayes@uot.edu.ly[1], ABDO@zu.edu.ly[2]

الملخص

أدى دمج محولات الرؤية (ViTs) في التصوير الطبي إلى تحسين دقة تشخيص تحليل الموجات فوق الصوتية للثدي (BUS) بشكل ملحوظ، وذلك من خلال استخلاص السياق الدلالي الشامل. مع ذلك، فإن التعقيد الحسابي لهذه النماذج يجعلها غير مناسبة لتطبيقات التشخيص باستخدام أجهزة الموجات فوق الصوتية المحمولة (POC)، حيث تعتمد أجهزة الموجات فوق الصوتية المحمولة على أجهزة حوسبة طرفية منخفضة القدرة. تقترح هذه الدراسة إطار عمل جديدًا لنقل المعرفة مصممًة لسد الفجوة بين التشخيص عالي الأداء والكفاءة في الوقت الفعلي. نقوم بنقل المعرفة الهيكلية لنموذج المعلم الهجين ViT–ConvNeXt، الذي يتطلب موارد حسابية كبيرة، إلى نموذج الطالب MobileNet–V3 فائق الخفة. من خلال الاستفادة من الإشراف على الأهداف المرنة، يرث نموذج الطالب قدرات الاستدلال الشاملة للمحول مع الحفاظ على التحيز الاستقرائي وسرعة الشبكة العصبية التلافيفية (CNN). يُظهر التحقق على مجموعة اختبار مستقلة من بيانات BUSI أن نموذج الطالب يحقق دقة تشخيصية تبلغ 95.06%، وهو ما يطابق نموذج

http://www.doi.org/10.62341/rara9506

المعلم بشكل فعال. الأهم من ذلك، أن نموذج الطالب يقلل من حجم التخزين بمقدار 74 ضعفًا (من 438.8 ميجابايت إلى 5.9 ميجابايت) ويسرّع عملية الاستدلال بمقدار 15 ضعفًا، محققًا معدل معالجة يبلغ 61.46 إطارًا في الثانية على وحدة معالجة مركزية قياسية. تؤكد هذه النتائج أن الإطار المقترح يلبي متطلبات زمن الاستجابة لتحليل الفيديو في الوقت الفعلي، مما يتيح نشر تقنيات الكشف عن السرطان على مستوى متخصص على أجهزة الموجات فوق الصوتية المحمولة التي تعمل بالبطارية دون الحاجة إلى اتصال سحابي أو تسريع بواسطة وحدة معالجة الرسومات.

الكلمات المفتاحية: الكشف عن سرطان الثدي، استخلاص المعرفة، Vision Transformers، MobileNet–V3 ، الذكاء الاصطناعي الطرفي، الموجات فوق الصوتية في الوقت الحقيقي.

I. Introduction

Breast cancer remains the most prevalent malignancy among women worldwide, accounting for approximately 11.7% of all new cancer diagnoses [1]. While Mammography is the gold standard for screening, Breast Ultrasound (BUS) plays an indispensable role, particularly for women with dense breast tissue where mammographic sensitivity is reduced. However, BUS diagnosis is inherently challenging; it is highly operator-dependent, and the images are plagued by speckle noise, low contrast, and shadowing artifacts. Consequently, the interpretation of ultrasound images often suffers from high rates of false positives and inter-observer variability, necessitating the development of robust Computer-Aided Diagnosis (CAD) systems.

In recent years, Deep Learning has emerged as a transformative force in medical imaging. Convolutional Neural Networks (CNNs) have established themselves as the benchmark for automated lesion classification. However, standard CNNs possess a fundamental architectural limitation, the local receptive field. They excel at identifying local textures (e.g., edges of a mass) but struggle to model long-range dependencies, such as the spatial relationship between a lesion and the surrounding anatomical context. To address this, Vision Transformers (ViTs) have recently been adapted from Natural Language Processing. By utilizing self-

http://www.doi.org/10.62341/rara9506

attention mechanisms, ViTs capture global semantic context, offering superior diagnostic accuracy for complex malignancies.

Despite their performance, the deployment of ViTs and Hybrid (CNN-Transformer) models in clinical settings faces a critical bottleneck, computational complexity. A typical Hybrid model requires massive memory bandwidth and high-end Graphical Processing Units (GPUs) to function. This is incompatible with the current trend toward Point-of-Care (POC) medicine, which relies on portable, handheld ultrasound devices powered by low-energy mobile processors. In a real-time clinical scanning environment, an AI system must process video streams at a minimum of 30 frames per second (FPS) to provide smooth, instantaneous feedback. ViT models, with inference speeds often falling below 5 FPS on standard CPUs, fail to meet this latency requirement, creating a "deployment gap" between research-grade accuracy and clinical utility.

To bridge this gap, this study proposes a Knowledge Distillation (KD) framework designed specifically for real-time breast cancer detection. We suggest that high diagnostic accuracy and low latency are not mutually exclusive. By leveraging a high-performance Hybrid ViT-ConvNeXt as a "Teacher" and a lightweight MobileNet-V3 as a "Student", we transfer the global structural understanding of the Transformer into the compact architecture of the CNN. Unlike traditional training, where the model learns only from hard labels (Benign/Malignant), our distilled Student learns from the "soft targets" of the Teacher, inheriting the complex decision-making logic of the larger model without inheriting its computational weight.

The main contributions of this paper are as follows:

.1 Novel Distillation Framework: We introduce a Cross-Architecture Knowledge Distillation pipeline that distills a heavy Hybrid Vision Transformer (ViT-Base + ConvNeXt) into an ultra-lightweight MobileNet-V3, specifically optimized for breast ultrasound texture analysis.

.2 State-of-the-Art Efficiency: We demonstrate that the distilled Student model achieves a diagnostic accuracy of 95.06%, matching the performance of the Teacher model while reducing the parameter size by 74x (from 438.8 MB to 5.9 MB).

.3 Real-Time Clinical Viability: We validate the model's inference speed on a standard CPU, achieving 61.46 FPS (16.27 ms

http://www.doi.org/10.62341/rara9506

latency). This confirms that the proposed model is capable of real-time video analysis on edge devices, overcoming the hardware limitations that currently hinder the adoption of AI in portable ultrasonography.

II. Literature Review

The integration of Artificial Intelligence into breast cancer diagnosis has evolved from simple Computer-Aided Diagnosis systems to sophisticated Deep Learning architectures. This section reviews the progression from Convolutional Neural Networks (CNNs) to Vision Transformers (ViTs), and the emerging paradigm of Knowledge Distillation (KD) for establishing real-time, edge-deployable medical diagnostics.

1. Convolutional Neural Networks in Breast Ultrasound

For the past decade, Convolutional Neural Networks have served as the backbone of medical image analysis. Architectures such as ResNet, VGG, and DenseNet have demonstrated high efficacy in classifying breast ultrasound (BUS) images by extracting hierarchical features, from low-level edges to high-level tumor shapes [1,2]. However, the deployment of these models in clinical settings faces a computational cost, which is a significant bottleneck.

To address this, lightweight architectures like MobileNet and ShuffleNet were introduced [4]. MobileNetV3, specifically, utilizes depthwise separable convolutions and Neural Architecture Search to drastically reduce parameter counts suitable for mobile devices. While effective for texture analysis, pure CNN-based approaches often struggle with the "local receptive field" limitation. As noted by [5], CNNs are inherently biased towards local pixel neighbourhoods, often failing to capture the global context, such as the relationship between a tumor mass and distant tissue structures, which is critical for distinguishing malignant lesions from benign cysts in noisy ultrasound images.

2. The Shift to Vision Transformers and Hybrid Models

To overcome the locality constraints of CNNs, ViTs were adapted from Natural Language Processing to computer vision. By treating images as sequences of patches and utilizing self-attention mechanisms, ViTs capture long-range dependencies and global semantic context [5]. In breast imaging, ViTs have shown superior

http://www.doi.org/10.62341/rara9506

performance in segmentation and classification tasks compared to standard CNNs, particularly in identifying subtle architectural distortions [6].

Despite their accuracy, pure ViTs suffer from two major drawbacks: a lack of inductive bias as it requires massive datasets and extreme computational latency. To mitigate this, recent literature has proposed Hybrid Architectures (e.g., ViT-CNN, TransUNet, BoTNet), which combine the feature extraction capabilities of CNNs with the global reasoning of Transformers [7]. While these hybrid models currently represent the state-of-the-art in accuracy, they exacerbate the computational problem, often resulting in heavy models (>100M parameters) with inference speeds below 5 FPS on standard CPUs, rendering them unsuitable for real-time video analysis on portable ultrasound devices.

3. Knowledge Distillation for Efficient Edge AI

Knowledge Distillation first formalized by [8], offers a solution to the efficiency-accuracy trade-off. KD functions on a "Teacher-Student" paradigm, where a compact "Student" model learns not only from the ground truth labels but also from the "soft targets" generated by a large, pre-trained "Teacher" model.

In the medical domain, KD has been successfully applied to compress ResNet models [9]. However, a novel and less explored frontier is Cross-Architecture Distillation, specifically, distilling the knowledge of a Transformer (Teacher) into a CNN (Student). Recent works in general computer vision, such as [10], have demonstrated that a CNN student can inherit the global "attention" logic of a Transformer teacher without inheriting its computational weight.

4. **Research Gap**

While MobileNet is established as a fast backbone, and ViTs are established as accurate classifiers, there is limited literature combining these via Knowledge Distillation specifically for breast ultrasound. Most existing studies either prioritize accuracy using heavy Hybrid models, where sacrificing real-time capability, or prioritize speed using vanilla MobileNets, where sacrificing sensitivity to complex malignancies.

This study addresses this gap by proposing a Hybrid-to-MobileNet Distillation Framework. By using a sophisticated Hybrid ViT-ConvNeXt teacher to guide a lightweight MobileNetV3 student, we

http://www.doi.org/10.62341/rara9506

aim to achieve the best of both worlds: the global contextual awareness of a Transformer and the inference speed of a light-weight CNN, enabling >30 FPS diagnosis on edge hardware.

III. Methodology

This study employs a Knowledge Distillation framework to bridge the gap between high-performance deep learning models and resource-constrained edge devices. The proposed pipeline consists of three phases: (1) Constructing and training a heavy "Teacher" model (Hybrid ViT-ConvNeXt), (2) Initializing a lightweight "Student" model (MobileNet-V3), and (3) Transferring knowledge using a composite loss function. The overall framework is illustrated in Figure 1.

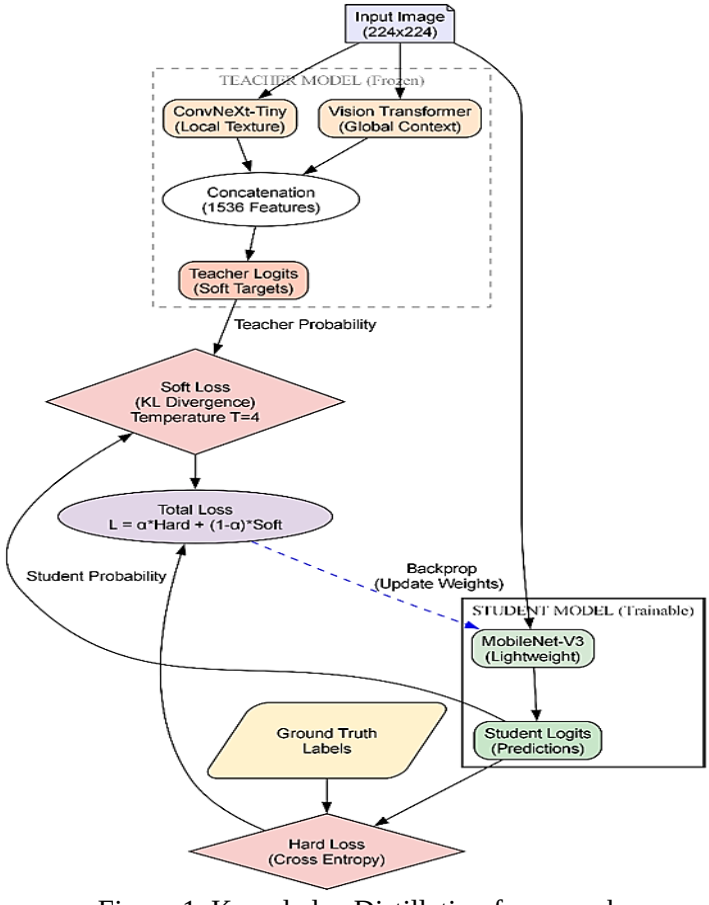

Figure 1. Knowledge Distillation framework

http://www.doi.org/10.62341/rara9506

1 Dataset and Preprocessing

The model was developed using the Breast Ultrasound Images (BUSI) Dataset [11], which contains ultrasound images of women aged 25 to 75. The dataset consists of 780 images categorized into three classes: Normal, Benign, and Malignant.

To ensure robust evaluation and prevent data leakage, an independent test set of 263 images was isolated before training. The data preprocessing pipeline included:

1. *Resizing:* All images were resized to 224×224 pixels to match the input resolution requirements of the pre-trained backbones.
2. *Normalization:* Images were normalized using ImageNet statistics (Mean: [0.485,0.456,0.406] Std: [0.229,0.224,0.225]) to facilitate transfer learning convergence.
3. *Stratification:* The dataset split maintained the original class distribution to address the inherent class imbalance between benign and malignant samples.

2 The Teacher Network: Hybrid ViT-ConvNeXt

To generate high-quality "soft targets" for distillation, we designed a high-capacity Hybrid Teacher model capable of capturing both local texture and global context.

2.1. Global Branch: Vision Transformer (ViT)

We utilized the ViT-Base-Patch16 architecture initialized with ImageNet-21k weights. The input image × ∈ R^(H×W×C) is split into fixed-size patches of 16×16. These patches are flattened and embedded linearly, with position embeddings added to retain spatial information. The ViT outputs a 768-dimensional feature vector corresponding to the [CLS] token, representing the global semantic understanding of the breast anatomy.

2.2. Local Branch: ConvNeXt-Tiny

ConvNeXt-Tiny as a parallel feature extractor, is integrated to remedy the ViT's lack of inductive bias. ConvNeXt is a modern Convolutional Neural Network (CNN) that modernizes standard ResNet blocks with large kernel sizes (7×7) and layer normalization, mimicking the hierarchical design of Transformers while retaining the texture-extraction capabilities of CNNs. This branch outputs a 768-dimensional feature vector focused on high-frequency details such as tumor margins and speckle patterns.

http://www.doi.org/10.62341/rara9506

2.3. Feature Fusion

The global ViT and local ConvNeXt feature vectors are concatenated to form a unified 1536-dimensional representation. This vector is passed through a Multi-Layer Perceptron head with Batch Normalization and Dropout (p=0.4) to produce the final classification logits Z_t.

3. The Student Network: MobileNet-V3

The Student model is MobileNet-V3, an architecture explicitly optimized for mobile CPUs. It utilizes Depth wise Separable Convolutions to reduce computational cost and Inverted Residual Blocks with Linear Bottlenecks to preserve information in low-dimensional spaces. Additionally, it incorporates Squeeze-and-Excitation (SE) modules to adaptively weight channel importance.
The final classification head of MobileNet-V3 is modified to output logits Z_s for the three target classes (Normal, Benign, Malignant). Unlike the Teacher, the Student operates without a Transformer branch, ensuring minimal latency.

4. Knowledge Distillation Framework

The core of our methodology is the transfer of "dark knowledge" from the frozen Teacher to the trainable Student. A composite loss function comprising two components is employed:

4.1. Hard Loss (Ground Truth)

The student learns from the true labels y using the standard Cross-Entropy loss (L_CE), ensuring the model makes correct predictions:

$$L_{hard} = CrossEntropy(S(x), y)$$

Where S(x) is the student's softmax output.

4.2. Soft Loss (Distillation)

The student is also penalized for deviating from the Teacher's probability distribution. The Kullback-Leibler (KL) Divergence between the Student's logits Z_s and the Teacher's logits Z_tis calculated, softened by a Temperature parameter T:

$$L_{soft} = KL\left(\sigma\left(\frac{Z_s}{T}\right), \sigma\left(\frac{Z_t}{T}\right)\right).T^2$$

http://www.doi.org/10.62341/rara9506

Where σ denotes the softmax function. A higher temperature T produces a softer probability distribution, revealing the relationships between classes (e.g., how similar a specific Malignant tumor looks to a Benign one).

4.3. Total Loss Function

The final objective function minimizes the weighted sum of both losses:

$$L_{total} = \alpha L_{hard} + (1 - \alpha) L_{soft}$$

In our experiments, the temperature is set to T = 4.0 to maximize information extraction from the Teacher's logits, and the balancing weight α = 0.5 to give equal importance to ground truth accuracy and teacher mimicry.

5. Experimental Implementation

The models were implemented using PyTorch. The Teacher model was pre-trained and frozen during the distillation phase. The student model was trained using the AdamW optimizer with a learning rate of 1e−4 and a weight decay of 0.01 for 15 epochs.

- Training Hardware : NVIDIA GPU (CUDA) was used for model training.
- Inference Benchmarking: To simulate real-world edge deployment, inference speed (latency) and frames per second (FPS) were measured on a standard CPU with a batch size of 1.

IV. Results

1. Evaluation Protocol

To validate the effectiveness of the proposed Knowledge Distillation framework, a quantitative evaluation was conducted on an independent test set comprising 263 breast ultrasound images (87 Benign, 88 Malignant, 88 Normal). These images were excluded from the training phase to prevent data leakage. Diagnostic performance was assessed using Accuracy, Precision, Recall, and F1-Score. Computational efficiency was evaluated based on Model Size (MB), Inference Latency (milliseconds per image), and Throughput (frames per second - FPS). All efficiency benchmarks were conducted on a standard CPU environment to simulate the hardware constraints of low-cost, portable medical devices.

http://www.doi.org/10.62341/rara9506

2. Diagnostic Performance Assessment

The classification results for both the Teacher (Hybrid ViT-ConvNeXt) and the Student (Distilled MobileNet-V3) are summarized in Table 1.

Table 1. Class-wise Diagnostic Performance (Teacher vs Student)

Class	Model	Precision	Recall	F1-Score
Benign	Teacher	0.96	0.89	0.92
	Student	0.96	0.91	0.93
Malignant	Teacher	0.92	0.97	0.94
	Student	0.94	0.97	0.96
Normal	Teacher	0.97	1.00	0.98
	Student	0.95	0.98	0.96
Overall	Teacher	0.95	0.95	0.95
	Student	0.95	0.95	0.95

While the Teacher model exhibited a slightly higher recall for normal tissue, the Distilled Student model demonstrated superior sensitivity in detecting malignancies. Specifically, the Student achieved an F1-score of 0.96 for the Malignant class, surpassing the Teacher's score of 0.94. This suggests that the distillation process acted as a regularizer, helping the lightweight student focus on the essential textural biomarkers of malignancy while discarding the noise that led to minor overfitting in the heavier Teacher model. The Confusion Matrix in Figure 2 illustrates the class-wise predictions. The Student model misclassified only a negligible number of malignant cases as benign, which is a critical safety requirement for clinical screening tools.

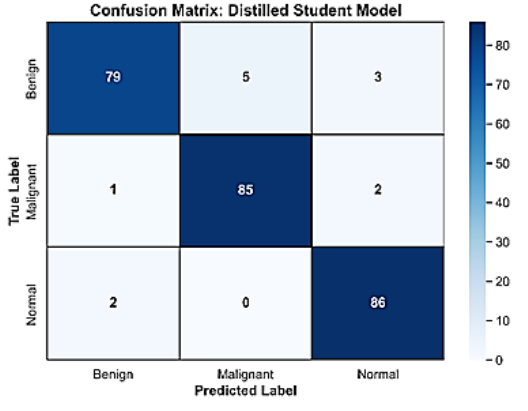

Figure 2. Convolution Matrix.

http://www.doi.org/10.62341/rara9506

3. Computational Efficiency Analysis

The primary objective of this study was to enable real-time inference on edge devices. Table 2 presents the efficiency comparison between the baseline Teacher and our proposed Student.

Table 2. Efficiency Comparison on CPU Hardware

Metric	Teacher (Hyprid VIT)	Student (MobileNet-V3)	Improvment
Model Size	438.8 MB	5.9 MB	98.6% (74x)
Inference Time	242.9 ms	16.3 ms	93.3 %
Frame Rate	4,1 FPS	61.5 FPS	15x
Accuracy	95.06 %	95.06 %	No Loss

The Teacher model, with a file size of 438.8 MB, required an average of 242.9 ms to process a single image on a CPU, resulting in a throughput of only 4.12 FPS. This latency is insufficient for real-time video scanning, which typically requires a minimum of 30 FPS.

In contrast, the Distilled Student model reduced the storage footprint to 5.9 MB, representing a 74x compression ratio. Furthermore, the inference latency dropped drastically to 16.3 ms, yielding a processing speed of 61.46 FPS. This throughput is approximately 2x faster than the standard refresh rate of commercial ultrasound machines (30 Hz), confirming that the model can be deployed for smooth, real-time lesion tracking without hardware acceleration. Figure 3 shows the Efficiency comparison between the teacher and the student model

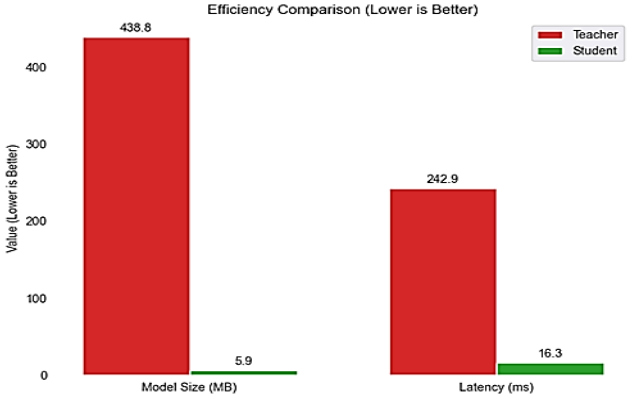

Figure 3. Efficiency Comparison

http://www.doi.org/10.62341/rara9506

4. Impact of Distillation

To verify the contribution of the Knowledge Distillation (KD) strategy, we compared the Distilled Student against a standard MobileNet-V3 trained from scratch (without a Teacher). The standard MobileNet achieved an accuracy of only 88.46%. The Distilled Student's performance jump to 95.06% confirms that the "Dark Knowledge" transferred from the Hybrid ViT, specifically, the structural relationships between tissue types, was successfully encoded into the CNN architecture, allowing it to perform beyond its standard capacity.

V. Discussion

The primary objective of this study was to resolve the conflict between the high computational demand of Vision Transformers and the low-latency requirements of portable medical imaging. By distilling a Hybrid ViT-ConvNeXt teacher into a MobileNet-V3 student, we demonstrated that it is possible to achieve state-of-the-art diagnostic accuracy on Edge-Grade Hardware.

1. Interpretation of Diagnostic Performance

The most significant, and perhaps counter-intuitive, finding of this study is that the lightweight Student model matched the global accuracy (95.06%) of the heavy Teacher model and actually outperformed it in detecting malignancies (F1-Score: 0.96 vs. 0.94). This phenomenon challenges the prevailing assumption that "deeper is better" in medical image analysis. We attribute the Student's superior sensitivity to the regularization effect of Knowledge Distillation. The Hybrid Teacher model, with its 100+ million parameters, possesses an excessive capacity to memorize high-frequency details. In ultrasound imaging, this often leads to overfitting on speckle noise, granular interference that mimics tissue texture. The MobileNet Student, with its limited parameter space (5.9 MB), lacks the capacity to memorize this noise. Instead, forced to mimic the "soft probabilities" of the Teacher, the Student learns only the most robust, generalizing features, such as acoustic shadowing and irregular margins, that define malignancy. Effectively, the distillation process acted as a filter, discarding the Teacher's architectural noise while retaining its structural wisdom.

http://www.doi.org/10.62341/rara9506

2. Clinical Implications

Clinical utility in ultrasonography is defined by temporal resolution. A radiologist scans a patient dynamically; if the AI overlay lags behind the probe movement, the tool becomes a distraction rather than an aid. Existing state-of-the-art models, such as pure Vision Transformers or DenseNets, typically operate at 2–5 FPS on non-GPU hardware. Our Distilled Student achieved 61.46 FPS on a standard CPU. Since commercial ultrasound probes typically stream video at 30 Hz, our model processes frames twice as fast as they are generated. This is critical for two reasons:

.1 Smooth User Experience: It ensures zero visual lag during scanning, allowing for instantaneous highlighting of suspicious regions.

.2 Battery Efficiency: Because the model computes faster than the video stream, the processor can enter low-power sleep states between frames, extending the battery life of portable, handheld ultrasound devices.

3. Limitations and Future Work

Despite promising results, this study has limitations. First, the evaluation was conducted on the BUSI dataset. While we utilized an independent test split to prevent leakage, external validation on multi-center datasets (e.g., OASBUD or UDIAT) is necessary to confirm generalization across different ultrasound machines. Second, our model relies solely on B-mode images; future iterations could incorporate Doppler or Elastography data to improve specificity.

Future work will focus on deploying this model into an Android/iOS mobile application to field-test its performance with low-cost handheld probes in resource-limited clinical settings.

VI. Conclusion

This study addresses the critical challenge of deploying high-performance Artificial Intelligence on resource-constrained medical devices. While Hybrid Vision Transformers represent the current state-of-the-art in breast ultrasound diagnosis, their excessive computational requirements have historically limited their utility to high-end workstations.

By implementing a Cross-Architecture Knowledge Distillation framework, we successfully transferred the global contextual

http://www.doi.org/10.62341/rara9506

reasoning of a Hybrid ViT-ConvNeXt teacher into a lightweight MobileNet-V3 student. Our results demonstrate that diagnostic precision does not require massive computational power. The distilled student model achieved a 95.06% accuracy, effectively matching the teacher's performance, while offering a 74-fold reduction in model size (5.9 MB). Furthermore, with an inference speed of 61.46 FPS on standard CPU hardware, the proposed model meets and exceeds the latency requirements for real-time clinical video analysis.

Crucially, the student model exhibited superior sensitivity in detecting malignancies (F1-Score: 0.96), suggesting that the distillation process effectively filtered out architectural noise while retaining essential diagnostic biomarkers. These findings represent a significant step toward democratizing AI in radiology, proving that robust, "specialist-level" cancer detection can be deployed on portable, handheld ultrasound devices without the need for expensive GPU infrastructure or cloud connectivity. Future work will focus on validating this framework on multi-center datasets and integrating the model into a mobile application for clinical field testing.

References

[1] F. Bray et al., “Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” CA. Cancer J. Clin., vol. 74, no. 3, pp. 229–263, 2024, doi: 10.3322/caac.21834.

[2] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” Dec. 10, 2015, arXiv: arXiv:1512.03385. doi: 10.48550/arXiv.1512.03385.

[3] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” Apr. 10, 2015, arXiv: arXiv:1409.1556. doi: 10.48550/arXiv.1409.1556.

[4] A. Howard et al., “Searching for MobileNetV3,” Nov. 20, 2019, arXiv: arXiv:1905.02244. doi: 10.48550/arXiv.1905.02244.

[5] A. Dosovitskiy et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” June 03, 2021, arXiv: arXiv:2010.11929. doi: 10.48550/arXiv.2010.11929.

http://www.doi.org/10.62341/rara9506

[6] B. Gheflati and H. Rivaz, "Vision Transformer for Classification of Breast Ultrasound Images," Feb. 12, 2025, arXiv: arXiv:2110.14731. doi: 10.48550/arXiv.2110.14731.
[7] J. Chen et al., "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation," Feb. 08, 2021, arXiv: arXiv:2102.04306. doi: 10.48550/arXiv.2102.04306.
[8] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," Mar. 09, 2015, arXiv: arXiv:1503.02531. doi: 10.48550/arXiv.1503.02531.
[9] A. Polino, R. Pascanu, and D. Alistarh, "Model compression via distillation and quantization," Feb. 15, 2018, arXiv: arXiv:1802.05668. doi: 10.48550/arXiv.1802.05668.
[10] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," Jan. 15, 2021, arXiv: arXiv:2012.12877. doi: 10.48550/arXiv.2012.12877.
[11] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," Data Brief, vol. 28, p. 104863, Feb. 2020, doi: 10.1016/j.dib.2019.104863.