On-Device Deployment of Cerviray AI: Optimization via Knowledge Distillation and Quantization for Mobile Clinical Environments

Byeongin Moon[*1] BLUEIN@AIDOT.AI
Jaeyun Song[*2] SJYUNI105@GMAIL.COM
Dongha Lee[*1] DHLEE@AIDOT.AI
Seongmin Kim[2] NAIAD515@GMAIL.COM
Donghoon Suh[3] SDHWCJ@NAVER.COM
Hansol Choi[†1] SOLCHOI@AIDOT.AI
Jaehoon Jeong[†1] JMAN@AIDOT.AI
[1] *AIDOT Inc., Seoul, South Korea*
[2] *Department of Obstetrics and Gynecology, Korea University Anam Hospital, Seoul, South Korea*
[3] *Department of Obstetrics and Gynecology, Seoul National University Bundang Hospital, Seongnam, South Korea*

Editors: Under Review for MIDL 2026

Abstract

Artificial intelligence has significantly advanced the diagnostic accuracy of Visual Inspection with Acetic acid (VIA) for cervical cancer screening. However, to overcome the GPU dependency of deep learning models and ensure their applicability in point-of-care settings, we present an on-device version of Cerviray AI. By employing Knowledge Distillation (ViT-Base to ViT-Tiny) and INT8 Post-Training Quantization, we successfully migrated the system from an RTX 4060 GPU to a Samsung Galaxy Tab S7 CPU. The optimized model achieves a clinical-grade accuracy of 97.61% (only a 0.24% drop) and an inference speed of 3.4s per image. This work demonstrates the potential of edge-AI in democratizing high-fidelity cancer screening for resource-limited, decentralized settings.

Keywords: Cervical Cancer, On-device AI, Knowledge Distillation, Model Quantization, Vision Transformer.

1. Introduction

Cervical Intraepithelial Neoplasia (CIN) grading is vital for determining treatment strategies. Cerviray AI automates this process, with efficacy validated through clinical trials (Kim et al., 2022, 2023) and its real-world effectiveness demonstrated in a recent prospective field study (Harsono et al., 2025). However, reliance on server-grade GPUs limits its clinical utility in point-of-care (POC) settings, particularly in low- and middle-income countries (LMICs) with unstable connectivity. Additionally, cloud-based processing raises significant data privacy concerns regarding sensitive patient images. To ensure robustness, we optimize the model on the Vision Transformer (ViT) (Dosovitskiy et al., 2021)—via Knowledge Distillation (KD) (Hinton et al., 2015) and Post-Training Quantization (PTQ) (Jacob et al., 2017) to run natively on mobile CPUs, enabling high-fidelity screening without cloud infrastructure.

[*] Contributed equally
[†] Corresponding author

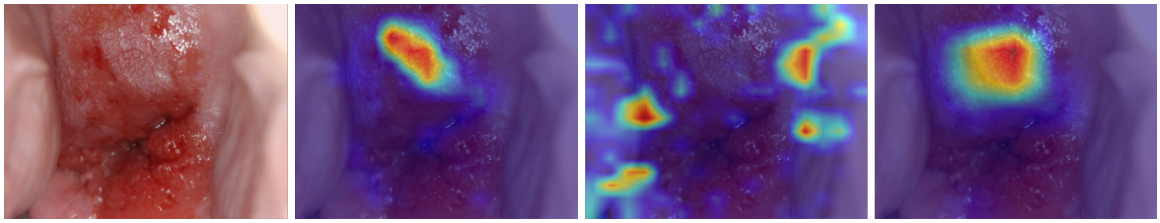

Figure 1: Comparison of attention heatmaps: The proposed optimized model accurately replicates the teacher model's (ViT-Base) diagnostic focus compared to the full fine-tuned(FFT) ViT-Tiny.

2. Dataset and Methodology

We utilized a multicenter dataset of 31,529 cervicography images (IRB approved). The data was split into Training (68%), Validation (16%), and Testing (16%) sets. Each image was labeled using the histopathological gold standard, verified by expert colposcopists.

2.1. Optimization Pipeline

The optimization focused on reducing architectural complexity and bit-precision while preserving spatial features necessary for medical imaging. The overall workflow of our two-stage optimization is illustrated in Figure 2.

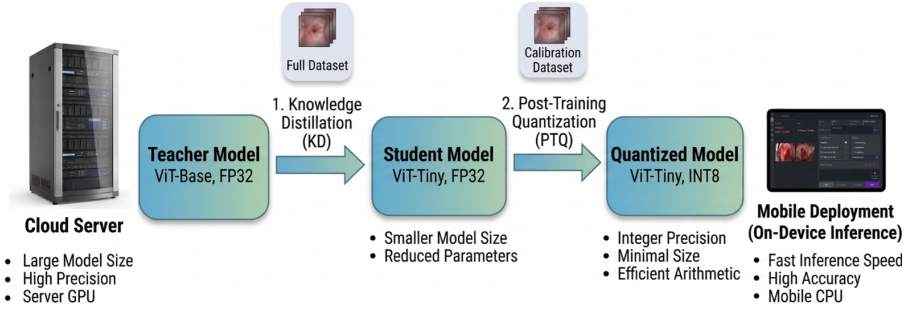

Figure 2: Optimization pipeline

Knowledge Distillation (KD) We distilled the diagnostic expertise of a ViT-Base (Teacher, $\mathcal{T}$) into a ViT-Tiny (Student, $\mathcal{S}$) following the DeiT approach (Touvron et al., 2021). The student was trained using a multi-objective loss function:

$$L_{total} = \alpha L_{CE}(y, \sigma(z_{\mathcal{S}})) + (1-\alpha)\tau^2 KL(\sigma(z_{\mathcal{T}}/\tau), \sigma(z_{\mathcal{S}}/\tau))$$

where L_{CE} is Cross-Entropy loss and KL is Kullback-Leibler divergence. This allows the Student to capture the complex feature representations necessary for subtle CIN grading.

Post-Training Quantization (PTQ) Following KD, the model was converted from FP32 to INT8 using Static PTQ (Gholami et al., 2021). We used a calibration set of 400

images balanced across the four CIN grades. The quantization mapping is defined as:

$$Q(x) = \text{clip}(\lfloor x/S \rceil + Z, q_{min}, q_{max})$$

where S is the scaling factor and Z is the zero-point. This reduces memory bandwidth and computational load on the ARM-based mobile CPU.

3. Experiments and Results

3.1. Implementation and Hardware

Models were trained on an NVIDIA RTX 4060 GPU (8GB VRAM). For clinical validation, we deployed the optimized model on a Samsung Galaxy Tab S7 (Qualcomm Snapdragon 865+, 8GB RAM), reflecting the actual *Cerviray Expert* hardware used in POC environments.

Table 1: Comprehensive performance comparison of model optimization strategies. The proposed ViT-Tiny (KD+Static PTQ) achieves the optimal balance between inference speed and diagnostic accuracy.

Metric	**ViT-Base (Baseline)**	**ViT-Base**	**ViT-Tiny**	**ViT-Tiny**	**ViT-Tiny**	**ViT-Tiny (Proposed)**
Input Size	512	512	224	224	224	**224**
Training	FFT	FFT	FFT	KD	KD	**KD**
Post-Training Quantization	—	INT8 Dynamic	—	—	INT8 Dynamic	**INT8 Static**
Model Size	344 MB	88 MB	23 MB	23 MB	6.5 MB	**6.5 MB**
Latency (Server/GPU)	1.5 sec	—	1.1 sec	1.1 sec	—	—
Latency (Mobile/CPU)	—	8.2 sec	—	—	3.8 sec	**3.4 sec**
Accuracy (%)	97.85%	97.79%	86.94%	97.61%	96.53%	**97.48%**
AUC	0.986	0.985	0.965	0.983	0.976	**0.981**

FFT: Full Fine-Tuning, KD: Knowledge Distillation

3.2. Performance Analysis

As summarized in Table 1, our proposed ViT-Tiny (KD+Static PTQ) model achieves a significant reduction in model size (344 MB to 6.5 MB) compared to the teacher model. While the transition to a mobile CPU environment increases latency to 3.4s, it remains well within the clinical threshold for real-time diagnostics. Crucially, the integration of KD effectively bridges the accuracy gap, recovering the student's baseline performance from 86.94% to 97.4%, with only a negligible 0.24% drop from the teacher model. This quantitative recovery is qualitatively reinforced by the attention heatmaps in Figure 1. This confirms that the combined KD and Static PTQ approach successfully preserves diagnostic integrity while enabling efficient on-device deployment.

4. Conclusion

We successfully engineered an on-device version of Cerviray AI by optimizing its core architecture for mobile CPUs. By leveraging advanced model compression, we bridged the gap between server performance and mobile accessibility, enabling democratized cancer screening in resource-limited settings.

Acknowledgments

This work was conducted by AIDOT Inc. The authors are grateful to Korea University Anam Hospital and Seoul National University Bundang Hospital for their essential support in providing the clinical datasets, which were accessed and used under the approval of their respective Institutional Review Boards (IRBs).

References

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W. Mahoney, and Kurt Keutzer. A survey of quantization methods for efficient neural network inference, 2021. URL https://arxiv.org/abs/2103.13630.

Ali Budi Harsono, Hadi Susiarno, Dodi Suardi, et al. Results comparison of cervical cancer early detection using Cerviray ® with VIA test. *BMC Research Notes*, 18(30), 2025. doi: 10.1186/s13104-025-07086-6. URL https://doi.org/10.1186/s13104-025-07086-6.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference, 2017. URL https://arxiv.org/abs/1712.05877.

Seongmin Kim, Hwajung Lee, Sanghoon Lee, Jae-Yun Song, Jae Lee, and Nak-Woo Lee. Role of artificial intelligence interpretation of colposcopic images in cervical cancer screening. *Healthcare*, 10:468, 03 2022. doi: 10.3390/healthcare10030468.

Seongmin Kim, Hyonggin An, Hyun Cho, Kyung-Jin Min, Jinhwa Hong, Sanghoon Lee, Jae-Yun Song, Jae Lee, and Nak-Woo Lee. Pivotal clinical study to evaluate the efficacy and safety of assistive artificial intelligence-based software for cervical cancer diagnosis. *Journal of Clinical Medicine*, 12:4024, 06 2023. doi: 10.3390/jcm12124024.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention, 2021. URL https://arxiv.org/abs/2012.12877.