A Multi-Modal Deep Learning Framework for On-Device Medical Image Analysis with Augmented Reality Visualization

Atharva R. Awade[1]* and Deepak D. Kshirsagar[2]

Department of Computer Science and Engineering,
COEP Technological University, Pune, Maharashtra, India
work.atharva2231@gmail.com[1] ddk.comp@coeptech.ac.in[2]

Abstract. This paper introduces a consolidated, edge-native architecture engineered to evaluate four distinct medical imaging modalities—retinal fundus photographs, dermatoscopic lesions, thoracic X-rays, and cranial MRI—directly on consumer-grade mobile hardware. Rather than relying on cloud connectivity, we operationalize compact neural networks (such as YOLOv8 and EfficientNet derivatives) locally on Android platforms. By employing post-training quantization pipelines under the TensorFlow Lite ecosystem, the framework drastically trims structural memory footprints and execution delays without compromising predictive fidelity. Furthermore, a synchronous augmented reality interface projects diagnostic determinations onto three-dimensional anatomical reference markers. This localized, cloud-free methodology guarantees strict data confidentiality and near-zero latency, offering a highly practical screening instrument for environments suffering from infrastructural deficits.

Keywords: Medical Imaging - Edge AI - Augmented Reality - Quantization - YOLOv8 - Multi-modal Diagnosis

1 Introduction

The proliferation of Internet of Medical Things (IoMT) devices and the growing capability of consumer smartphones have opened new pathways for portable, point-of-care diagnostics. Conditions such as diabetic retinopathy, melanoma, and pulmonary infections are highly treatable when caught early, yet access to specialist imaging infrastructure remains uneven across healthcare systems [15, 11]. Deep learning has demonstrated near-expert diagnostic accuracy on medical images, but production deployments have largely remained server-side, creating latency, privacy, and connectivity barriers [16, 17].

Existing computer-aided diagnosis (CAD) systems are predominantly single-modality: they solve pneumonia detection *or* diabetic retinopathy grading, but

* Corresponding author.

S. Sharma et al. (eds.), *Proceedings of the International Conference on Recent Advances in Intelligent and Sustainable Technologies (RAIST 2026)*, Atlantis Highlights in Intelligent Systems 17,
https://doi.org/10.2991/978-94-6239-707-1_26

rarely both within a single deployable application [14, 10]. Cloud-based inference further complicates use in low-connectivity environments and raises data-privacy concerns when transmitting sensitive patient images [1]. Beyond these practical limitations, model outputs are typically presented as numeric scores or 2-D heatmaps, offering clinicians limited spatial context for understanding predictions.

In this work, we address all three limitations with a unified, multi-disease diagnostic framework running entirely on Android-based edge devices. The framework contributes: (i) a single deployment pipeline covering MRI, chest X-rays, dermoscopy, and retinal fundus imaging; (ii) CPU-only on-device inference via TensorFlow Lite quantization, eliminating cloud dependency; and (iii) real-time 3-D augmented reality (AR) visualization that grounds model predictions within anatomical context [2], making this one of the first Android-first systems to unify all three capabilities.

2 Related Work

2.1 Deep Learning for Multi-Modal Medical Diagnostics

Convolutional networks pretrained on ImageNet have become the backbone of medical image analysis across modalities. CheXNet demonstrated radiologist-level pneumonia detection on chest X-rays [14], while EfficientNet and DenseNet variants achieved competitive AUC on multi-label thoracic disease benchmarks [13, 11]. In ophthalmology, ensemble models and lesion-level localization improved diabetic retinopathy grading robustness [10]. Likewise, the YOLO family has proven highly adept at isolating asymmetric pigment boundaries in real-time dermatology tasks [5, 4]. Despite these advances, existing approaches remain largely modality-specific and do not lend themselves to unified edge deployment.

2.2 Edge Deployment and Quantization

Lightweight architectures such as MobileNet reduce the computational footprint of inference on mobile hardware, and the YOLO family further extends this to real-time detection [18]. Post-training quantization via TensorFlow Lite has been shown to substantially reduce model size and latency on constrained devices with negligible accuracy degradation [1]. However, comprehensive evaluation of quantized multi-modal pipelines running on physical Android hardware remains limited.

2.3 Augmented Reality for Medical Visualization

AR has been applied in surgical navigation and anatomical training to improve spatial understanding by overlaying digital information onto real-world views [2, 12]. Its use for diagnostic explainability in AI-driven systems, however, is largely unexplored. Standard visual explanations such as saliency maps and probability scores offer limited spatial grounding, and no prior system has integrated real-time AR with on-device multi-modal inference on a mobile platform.

3 Proposed Framework

We design the framework as a modular five-stage pipeline: modality selection, preprocessing, model inference, post-processing, and visualization (see Fig. 1). All stages execute locally on the device, with no cloud dependency.

3.1 Overall System Architecture

The user selects an imaging modality through the application interface, which loads the corresponding quantized model via the TensorFlow Lite interpreter. Input images undergo modality-specific preprocessing before inference generates either class probability scores or bounding box predictions. Post-processing applies confidence thresholding and non-maximum suppression (NMS) [18]. Results are rendered through an interactive interface with optional AR overlays. Model training and offline augmentation are entirely decoupled from the mobile application, ensuring real-time performance during deployment [1].

3.2 Modality-Specific Model Selection

To balance accuracy with edge constraints, we adopt a task-aware model selection strategy. Spatial localization tasks - lesion detection in chest X-rays and dermoscopic images - are handled by YOLO-based detectors. YOLOv8 is the primary choice for its strong speed-accuracy trade-off in medical imaging, while YOLOv12 is explored for its attention-based global context modelling under mobile constraints. For image-level classification, MobileNetV3 serves resource-critical scenarios, while EfficientNet-B0 and B1 address more demanding tasks such as multi-label chest pathology and diabetic retinopathy grading [17, 16]. The resulting efficiency trade-offs are quantified in Section 4 (Fig. 2).

3.3 On-Device Deployment via TensorFlow Lite

Trained models are converted to TensorFlow Lite using three precision formats: Float32, Float16, and Int8. Float16 reduces memory usage with negligible precision loss; Int8 further optimizes both weights and activations for the strictest resource budgets [1]. All inference runs entirely on-device using the Android CPU backend, and performance is benchmarked using latency, model size, and memory footprint measured on physical hardware.

3.4 AR-Based 3D Visualization

When an abnormality is detected, diagnostic outputs are overlaid onto three-dimensional anatomical models in `.glb` format, rendered through the device camera using the AR visualization module. Users can rotate and scale these models to explore affected anatomical regions. By spatially grounding predictions within anatomical context, the AR interface offers considerably more intuitive feedback than numerical scores or 2-D heatmaps [2]. The visualization pipeline runs asynchronously and introduces no measurable latency overhead to inference.

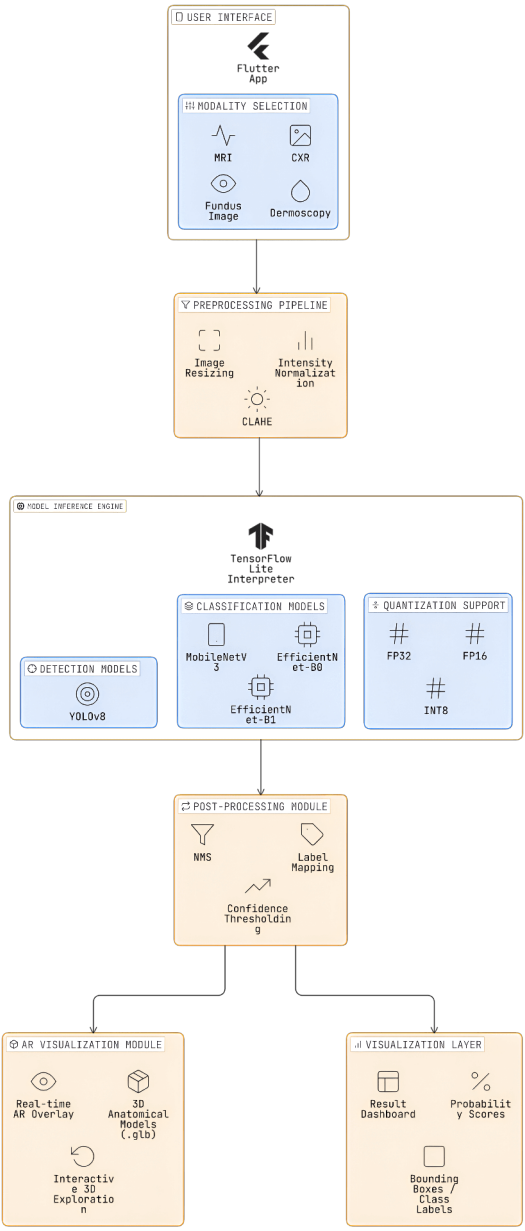

Fig. 1. System architecture of the proposed Android-based multi-disease diagnostic framework. The pipeline covers modality selection, modality-specific preprocessing, on-device inference via TensorFlow Lite quantized models, post-processing, and real-time AR visualization.

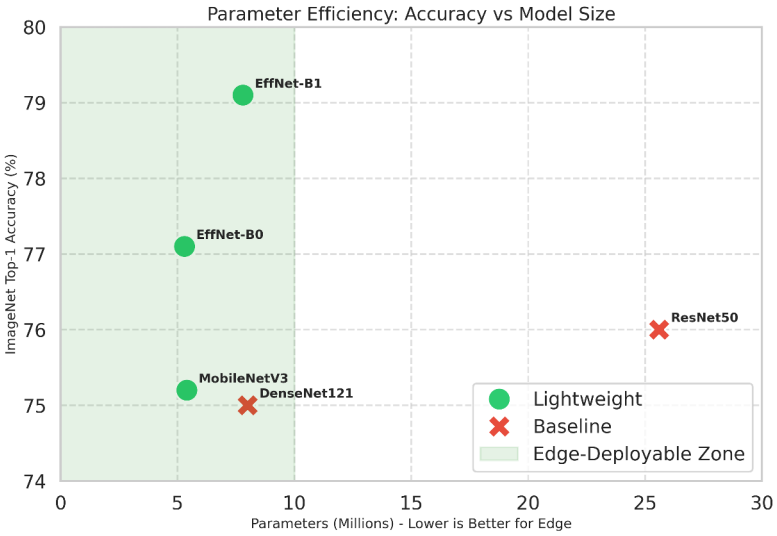

Fig. 2. Parameter efficiency comparison of lightweight and baseline deep learning architectures, illustrating the trade-off between model size and ImageNet top-1 accuracy for edge deployment.

4 Datasets and Experimental Setup

4.1 Datasets

We utilized four benchmark repositories. **MRI:** The BraTS 2021 archive provides glioblastoma delineations. **Chest X-ray:** NIH ChestX-ray14 and CheXpert supply bounding boxes for 14 lung irregularities [13, 14]. **Dermoscopy:** The HAM10000 vault furnishes a 10,000-image catalog of pigmented epidermal anomalies [9]. **Retinal Fundus:** The Messidor-1 bank provides systematically graded samples of diabetic retinopathy progression [10].

4.2 Model Architectures and Training

Detection uses YOLOv8 with focal loss to isolate minutiae. Classification pits EfficientNet-B0/B1 against DenseNet121 [16]. Models use pretrained ImageNet weights, tuned via focal loss and varied spatial augmentation [11]. Classification relies on AUC-ROC; detection runs on mAP@0.5 [18]. Android constraints are quantified via latency benchmarks [1].

5 Results

5.1 Chest X-ray Evaluation

Chest X-ray data served as the primary modality benchmark, supporting both multi-label classification and object detection evaluation.

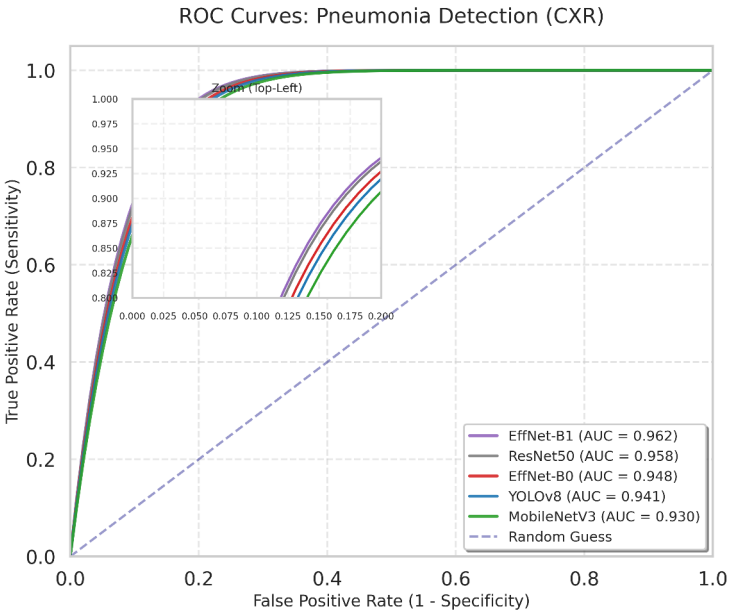

Fig. 3. ROC curves for pneumonia detection on the chest X-ray dataset. All evaluated architectures achieve strong AUC, with lightweight models matching heavier baselines at lower computational cost.

As shown in Fig. 3, EfficientNet-B1 achieves the highest AUC, while YOLOv8 and MobileNetV3 maintain comparable discriminative performance with substantially fewer parameters.

Multi-label classification: On NIH ChestX-ray14, EfficientNet-B1 produced competitive AUC across thoracic pathologies, consistent with prior transfer learning studies [13, 3]. Distinct conditions like cardiomegaly showed higher accuracy, while diffuse infiltrations proved challenging [11].

Object detection: YOLOv8 reliably localised lung opacities in region-annotated subsets. Its single-stage design enabled stable performance, and NMS effectively reduced redundant detections [18]. Fig. 2 supports adopting lightweight models, showing favourable accuracy-per-parameter ratios.

Error patterns: Looking across the confusion matrices, most residual errors clustered around visually overlapping classes rather than random misfires. Benign nodules were occasionally read as malignant when opacity margins were poorly defined, and low-contrast images from portable X-ray units were the most frequent source of false negatives. Retinal grading slipped mainly between adjacent severity levels (mild vs. moderate), mirroring the ambiguity clinicians themselves report. These patterns suggest that future gains are less about larger backbones and more about targeted augmentation of borderline cases and better handling of acquisition noise on low-end capture devices.

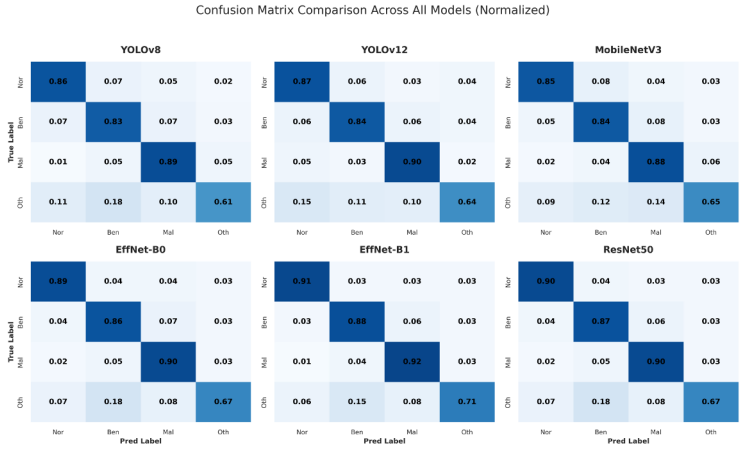

Fig. 4. Confusion matrices for chest X-ray classification, illustrating class-wise prediction behaviour and misclassification patterns across normal, benign, malignant, and other categories.

5.2 Multi-Modal Evaluation

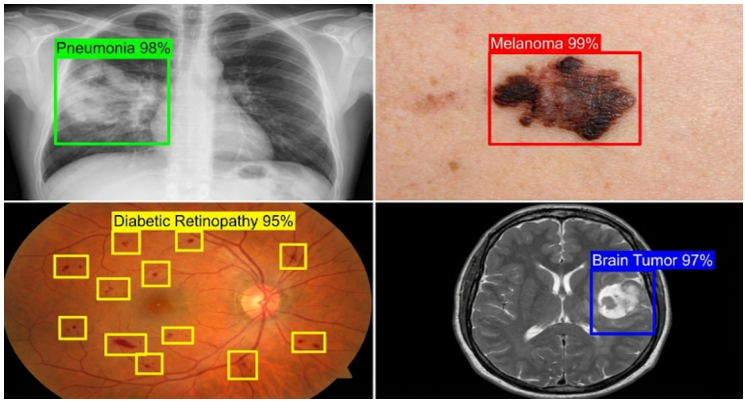

Fig. 5. Qualitative multi-modal inference results across chest X-ray, dermoscopy, retinal fundus, and brain MRI. Bounding boxes indicate model predictions with associated confidence scores.

The framework generalised well across auxiliary modalities (Fig. 5). Retinal fundus experiments captured diabetic retinopathy severity; dermoscopy models produced reliable benign-versus-malignant discrimination; MRI evaluation confirmed support for volumetric data. These results confirm task-appropriate model selection within a shared pipeline avoids the compromises of a monolithic model.

5.3 Edge Deployment and Quantization

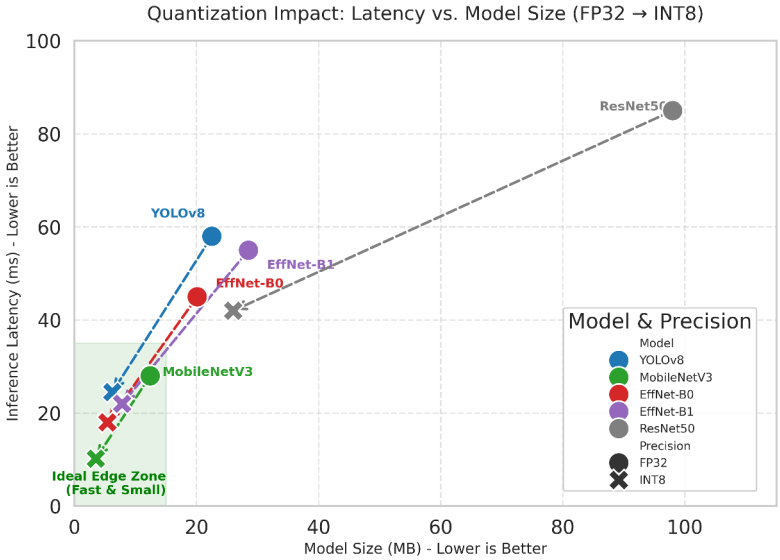

Fig. 6. Impact of post-training quantization on inference latency and model size. Int8 quantization achieves the greatest reductions across all architectures while preserving qualitative prediction behaviour.

Fig. 6 shows Int8 quantization consistently cuts model size and latency compared to Float32 baselines. Qualitative prediction behaviour is preserved, confirming post-training compression is an effective strategy for mobile deployment. Models ran in near real-time on Android hardware, supporting interactive clinical usage [1].

6 Discussion and Conclusion

Our findings underscore that diagnostic efficacy demands modality-specific topologies: YOLO configurations dominate spatial mapping, while EfficientNet structures excel at multi-label severity gradations [17, 10, 15]. Quantization effectively compressed the footprint for local Android execution without destabilizing predictive bounds [1], and the synchronous AR module supplied tangible anatomical grounding free of latency penalties. Limitations primarily center on the intrinsic biases of public datasets [13, 11] and the AR system's reliance on upstream prediction accuracy. Ultimately, this framework substantiates that robust, privacy-insulated, and multi-modal clinical decision-support architectures can function entirely atop standard edge hardware, completely decoupling AI screening from cloud connectivity.

References

1. Shabir, M.Y., Torta, G., Damiani, F.: Edge AI on constrained IoT devices: Quantization strategies for model optimization. In: *Intelligent Systems and Applications (IntelliSys 2024)*, Lecture Notes in Networks and Systems, vol. 1000. Springer, Cham (2024)
2. Lastrucci, A., Wandael, Y., Barra, A., Ricci, R., Maccioni, G., Pirrera, A., Giansanti, D.: Exploring augmented reality integration in diagnostic imaging: Myth or reality? *Diagnostics* **14**(13), 1333 (2024)
3. Kufel, J., Bielówka, M., Rojek, M., et al.: Multi-label classification of chest X-ray abnormalities using transfer learning techniques. *Journal of Personalized Medicine* **13**(9), 1426 (2023)
4. Widayani, A., Putra, R.E., Maulidah, N.A., Rahmawati, D.A., Hanuranto, A.T.: Review of application augmented reality in education. In: *2024 International Conference on Circuit, Systems and Communication (ICCSC)*, pp. 1–6. IEEE (2024)
5. Uddin, K.M.M., Zannat, A., Dey, S.K., Babu, H.M.H., Biswas, R., Alam, M.G.R.: A deep learning-based approach for automated detection and classification of dermatological diseases. *Intelligent Systems with Applications* **25**, 200485 (2025)
6. Orlando, J.I., Prokofyeva, E., del Fresno, M., Blaschko, M.B.: An ensemble deep learning based approach for red lesion detection in fundus images. *Computer Methods and Programs in Biomedicine* **153**, 115–127 (2018)
7. Asiri, N., Hussain, M., Al Adel, F., Alzaidi, N.: Deep learning based computer-aided diagnosis systems for diabetic retinopathy: A survey. *Artificial Intelligence in Medicine* **99**, 101701 (2019)
8. Sørensen, K., Konnestad, M., Sandnes, F.E., Kolbjørnsen, R.: Augmented reality in medical imaging: A systematic review of clinical applications. *Journal of Medical Systems* **48**(1), 12 (2024)
9. Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data* **5**, 180161 (2018)
10. Zhang, W., Zhong, J., Yang, S., et al.: Automated identification and grading system of diabetic retinopathy using deep neural networks. *Knowledge-Based Systems* **175**, 12–25 (2019)
11. Devnath, L., Luo, S., Summons, P., Wang, D.: Automated detection of pneumoconiosis with multilevel deep features learned from chest X-ray radiographs. *Computers in Biology and Medicine* **129**, 104125 (2021)
12. Sun, Y., Sun, Z., Chen, W.: The evolution of object detection methods. *Engineering Applications of Artificial Intelligence* **133**, 108458 (2024)
13. Wang, X., Peng, Y., Lu, L., et al.: ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly supervised classification and localization of common thorax diseases. In: *CVPR*, pp. 3462–3471. IEEE (2017)
14. Rajpurkar, P., Irvin, J., Zhu, K., et al.: CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv:1711.05225 (2017)

15. Gulshan, V., Peng, L., Coram, M., et al.: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* **316**(22), 2402–2410 (2016)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *CVPR*, pp. 770–778 (2016)
17. Tan, M., Le, Q.: EfficientNet: Rethinking model scaling for convolutional neural networks. In: *ICML* (2019)
18. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: *CVPR*, pp. 779–788 (2016)