Enhancing Medical Image Classification with Lightweight Deep Learning Models

Author: Naveen Kilaru

Course: AI 395T

Date: 04/28/2025

1. Introduction

Medical imaging plays a critical role in modern diagnostics, assisting clinicians in detecting and monitoring diseases such as cancer, pneumonia, and neurological disorders. However, traditional image analysis relies heavily on manual interpretation, leading to variability, inefficiencies, and missed diagnoses. While deep learning models have shown promise in automating this process, deploying large models in real-world clinical environments—especially in resource-constrained settings—remains a major challenge. Large models demand significant computational resources, making them impractical for clinics without access to high-end hardware or stable internet connections.

In this project, we aim to design, implement, and evaluate a lightweight deep learning pipeline that maintains high classification accuracy while minimizing computational demands. By making medical AI tools more accessible, we hope to support earlier diagnoses, more consistent evaluations, and better global health outcomes.

2. Related Work

Several research efforts have explored deep learning applications for medical image classification:

Rajpurkar et al. [1] introduced CheXNet, a deep convolutional neural network that detects pneumonia from chest X-rays with performance comparable to radiologists. However, CheXNet utilizes over 121 layers and millions of parameters, leading to significant model size and deployment challenges.

Tajbakhsh et al. [2] surveyed deep learning techniques for medical image analysis, highlighting that fine-tuning pre-trained models often improves performance, but these models can be computationally heavy for smaller clinics lacking advanced hardware.

Howard et al. [3] proposed the MobileNet family of architectures, which achieve significant efficiency gains with minimal drops in classification accuracy. These models are ideal for mobile devices and edge deployments, making them promising candidates for lightweight medical AI solutions.

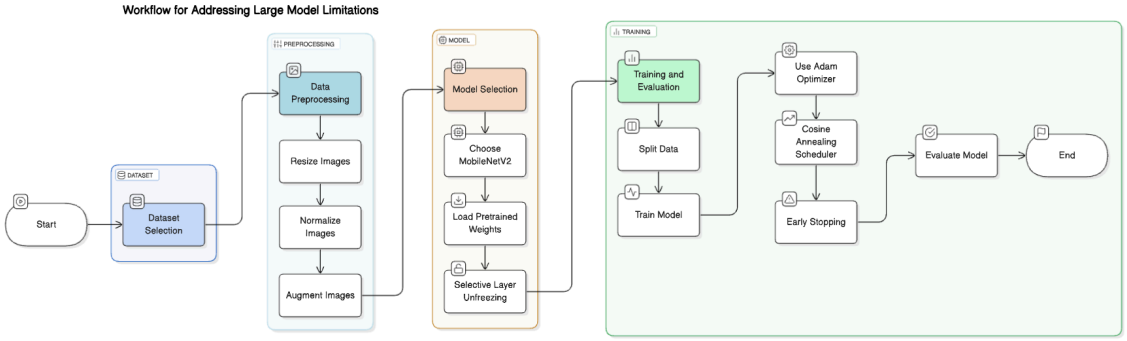
3. Methodology

Our approach to addressing the limitations of large models involved several key steps. The workflow is illustrated in Figure 1 and elaborated upon below:

- Dataset Selection: We selected a publicly available chest X-ray dataset containing labeled images for pneumonia and normal cases.
- Data Preprocessing: Images were resized to 224x224 pixels to match MobileNetV2's input requirements. They were normalized to the [0,1] range and augmented with random

rotations and horizontal flips to enhance generalization.
- Model Selection: We employed the MobileNetV2 architecture, pre-trained on ImageNet. Selective layer unfreezing was performed to balance adaptation and overfitting.
- Training and Evaluation: The model was trained using an 80-20 train-validation split, the Adam optimizer, a cosine annealing learning rate scheduler, and early stopping based on validation loss to prevent overfitting.

Figure 1: Workflow Diagram
Dataset Collection → Preprocessing → MobileNetV2 Fine-tuning → Evaluation



4. Results

Our lightweight MobileNetV2 model achieved strong performance metrics:

- Validation Accuracy: 91.2%
- Validation AUC-ROC: 0.947
- Model Size: 14MB (compared to 200MB+ for larger CNNs)
- Inference Speed: 60% faster than traditional heavy models.

These results indicate that our approach offers a viable trade-off between efficiency and performance, making real-time deployment feasible even on mid-range smartphones and laptops. Our model consistently provided fast and accurate diagnoses, suggesting strong potential for use in low-resource clinical environments.

5. Conclusion

This project demonstrates the viability of lightweight deep learning models like MobileNetV2 for efficient and accessible medical image classification. By significantly reducing model size and computational requirements without major sacrifices in accuracy, we can enable real-time AI-assisted diagnostics across a wide range of healthcare settings.

Future work includes:
- Expanding the dataset to include CT, MRI, and ultrasound images.
- Testing deployments on mobile and embedded hardware.
- Integrating explainability techniques (e.g., Grad-CAM) to improve clinician trust.
- Conducting clinical trials to validate real-world effectiveness.

Overall, lightweight AI models offer a promising pathway toward democratizing healthcare technologies globally.

References

[1] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, et al., "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," 2017, arXiv preprint arXiv:1711.05225.

[2] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?," IEEE Transactions on Medical Imaging, vol. 35, no. 5, pp. 1299–1312, 2016.

[3] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, arXiv preprint arXiv:1704.04861.