

<Conference abbreviation>

<Conference Series name>

<Volume number and Year> <DOI Number>

A Review of Lightweight and Explainable AI for Clinical Edge Deployment

Shital Ajit Dumbre¹, Dr. Arati Kale²

^{1,2}*School of Computing, MIT Art, Design and Technology University, Loni, Maharashtra, India, shital.dumbre@mituniversity.edu.in, arati.kale@mituniversity.edu.in*

Abstract.

Clinical artificial intelligence models are highly accurate in their diagnostic outcomes, but have difficulty in their practical application because of the high computing resources and non-interpretable nature. The paper will solve this issue of empowering effective and reliable AI towards clinical edge conditions using limited resources.

To address this, a systematic comparative study of lightweight AI methods such as pruning, quantization, knowledge distillation and sparse architectures is discussed and explainable AI approaches such as SHAP, LIME and Grad-Cam. Performance, efficiency, and interpretability are evaluated using a single evaluation framework.

It is analyzed that lightweight models can reach a reduction of up to 80-95 percent in model size and enormous latency cuts and still ensure a high level of performance within a narrow performance error. These results justify the implementation of AI models to edge images like microcontrollers, smartphones, and embedded health care frameworks to the application of diagnostics and remote monitoring and clinical decision support.

Keywords. Clinical AI; Model Compression; Knowledge distillation; explainable AI; Edge computing;

1. INTRODUCTION

Deep learning has enhanced medical diagnosis, particularly in imaging and patient monitoring, but application in clinical practice is a challenge. Models with high capacity that are trained to be as accurate as possible may need to compute on the cloud and cannot meet the on-device, real-time, and privacy requirements of regulations like GDPR and HIPAA. This has seen a shift towards deployment-first AI, where efficiency, energy consumption, and compatibility with various edge devices are critical [1]. Meanwhile, the black-box characteristics of deep models prevent clinical trust, as clinicians require clear and comprehensible explanations to make serious judgments. These issues have made lightweight and explainable AI models more popular, which can be implemented on edge devices, including wearables, handheld scanners, and smartphones. This survey evaluates compact model design, compression and sparsity, explainability, real-world healthcare applications and open challenges in 2021-2025.

2. RELATED WORKS

2.1. Compression Methods

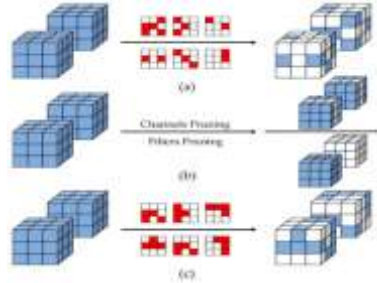


Fig. 2.1.1. CNN pruning techniques: non-structured, structured, and pattern pruning.

Pruning, quantization, and knowledge distillation are the key methods that are required to compress models and deploy AI to edge devices. These techniques can greatly cut down the size of models, but do not impact performance: pruning removes unnecessary weights to build sparse networks, quantization transfers values of high precision to lower-bit representations to save memory, and distillation transfers knowledge in large teacher networks to smaller student networks [4]. Recent healthcare research reports a good performance, with models scaled down to 1518 percent of their initial size and have up to 56 inference times faster with little to no loss of accuracy [5], [6]. Practically, the best performance is achieved by merging the techniques (e.g., pruning with quantization) and allows reducing the number of parameters by up to 90 percent with only a slight decrease in performance, allowing edge deployment into healthcare environments to be a practicable reality [9].

2.2. Sparse Architectures and Small Models with XAI

Lightweight model design provides an alternative to conventional compression by constructing efficient architectures based on fundamentals. Factorized convolutions and scaling in models like MobileNet, EfficientNet, and ShuffleNet can reduce the number of parameters and computation by a large margin compared to heavy networks like VGG-16 [10]. Recent health signal architectures such as MoE-TinyMed show that sparse architectures can be more efficient than larger dense models with fewer active parameters, emphasizing the power of conditional computation [12]. Explainable methods (e.g., Grad-CAM) can still be used even with compressed models, which can be trusted by clinicians. The real-world applications indicate an increasing possibility of edge-based clinical AI, as TinyML systems can run on low power (around 100 mW) and the smartphones can infer in 50-100 ms. Nevertheless, the literature does not provide standard evaluation frameworks that can be used to collectively evaluate baseline models, compression techniques, and explainability, which is an open research gap.

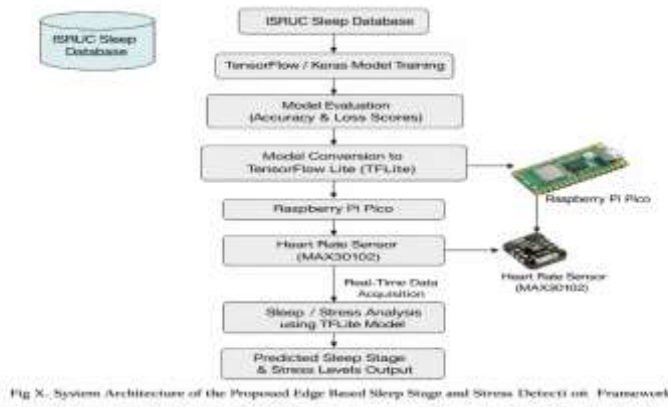


Fig. 2.2.1. TinyML pipeline including data collection, training, quantization, and deployment.

3. METHOD, EXPERIMENTS AND RESULTS

3.1. Methodology

The method includes three main steps: (1) literature search and classification, (2) comparative performance of the model optimization methods, and (3) assessment according to metrics related to deployment. To begin with, the systematic review included recent works published between 2021-2025, centering on the use of AI in the clinical imaging field, natural language processing, and physiological signal analysis. Second, a comparative framework was created to compare these approaches on their capacity to decrease the computational complexity and retain the diagnostic performance. Lastly, the techniques have been assessed on the basis of a common set of metrics representing practical deployment conditions, namely accuracy, model size, latency, energy usage and interpretability.

3.2. Evaluation Metrics

An effective assessment framework is needed to evaluate lightweight clinical AI systems.

In the paper, several performance dimensions are taken into account:

Accuracy / F1-score / AUC: Diagnostic and predictive performance.

Model Size (MB): Number of storage requirements of the edge deployment.

Latency (ms): Inference time, which is important in real-time applications.

Energy Consumption (mW): Of significance to wearable and battery-powered devices.

Compression Ratio: Level of downsizing of model over baseline.

Explainability Metrics: Incorporating IAUC, DAUC, and saliency consistency.

3.3. Results and Observations

The comparative analysis shows that there are a few significant trends. First, compression methods like pruning and quantization of models are highly effective in minimizing the computational cost at the expense of reasonable accuracy. Losses of up to 80-95% in model size with little performance loss is obtained in most cases. Second, the lightweight and sparse architecture is an alternative to using post-training compression, which includes efficiency into model design. MobileNet and EfficientNet architectures can be highly accurate with fewer parameters, whereas mixture-of-experts models can do conditional computation, achieving greater efficiency without compromising accuracy. Third, the methods of

explainability still work even when compressed models are used. Lastly, deployment-related studies show that lightweight models can be deployed to edge devices, including microcontrollers, smartphones, and embedded systems, and can provide real-time inference with a low power budget.

4. DISCUSSIONS

Although the development is fast, there are several issues that are not addressed:

- **Integration and Regulatory Compliance:** One of the gaps in the reviewed literature is the lack of standardized comparative reporting. A lot of research introduces compressed or lightweight models without a consistent comparison to strong baseline architectures or recent state-of-the-art algorithms as well as popular evaluation protocols. Subsequent work ought to put more focus on reproducible benchmark configurations, which collectively report on predictive, computational, and interpretability results. The connection between compressed models and hospital systems (PACS, EHR) has not been greatly discussed. This can be filled in the future, e.g. with designing interoperable pipelines and the requirements of FDA/CE. Improved XAI standards and user research are required clinical adoption requirements to support physicians in their trust and proper interpretation of model results.
- **Energy and Runtime Benchmarking:** It is not common to have overall benchmarks of a wide range of hardware. Community intervention is required to make comparisons between models on standard medical problems (e.g. chest X-ray classification, ECG analysis) on devices (microcontrollers, RPi, phones) based on latency, energy and memory. This would inform architecture/hardware co-design.

5. CONCLUSIONS

1. **Problem Statement / Motivation:** High computational demands and limited hardware resources and interpretability are the reasons why clinical AI models are not easily deployed.
2. **Method Used:** A systematic comparative study of lightweight AI methods such as pruning, quantization, knowledge distillation, sparse architectures and explainable AI methods was performed.
3. **Key Findings:** Lightweight models offer major advantages in reducing model size (80-95%), latency and competitive performance. The efficiency of sparse architecture is even more efficient and surpasses large architecture in some situations. Explainability methods enhance transparency and aid clinical adoption.
4. **Limitations and Future Work:** The study is also disadvantaged by the absence of standardized benchmarks and similar evaluation techniques. The future research efforts should be on standardized benchmarking systems, incorporation of privacy-preserving methods and clinical feasibility.

6. REFERENCES

- [1] C. M. Nguyen and T.-S. Hy, "Efficient Deep Learning for Medical Imaging: Bridging the Gap Between Research Environments and Clinical Deployment," arXiv preprint arXiv:2602.00910, 2026.

- [2] E. Paula et al., “Comparative analysis of model compression techniques for achieving carbon efficient AI,” *Sci. Rep.*, vol. 15, p. 23461, 2025.
- [3] C. Johnvictor et al., “TinyML-Based Lightweight AI Healthcare Mobile Chatbot Deployment,” *J. Multidiscip. Healthc.*, vol. 17, pp. 5091–5104, 2024.
- [4] P. Srivastava, N. Shah, and K. Jaiswal, “Microcontroller-Based EdgeML: Health Monitoring for Stress and Sleep via HRV,” *Engineering Proceedings*, vol. 78, no. 1, 2024, art. 3.
- [5] E. F. Rivera-Guzmán et al., “Quantization of Deep Neural Networks for Medical Image Analysis: A Systematic Review and Meta-Analysis,” *Technologies*, vol. 14, no. 1, art. 76, 2026.
- [6] F. M. Alabbasy et al., “Compressing medical deep neural network models for edge devices using knowledge distillation,” *J. King Saud Univ. – Comput. Inf. Sci.*, vol. 35, no. 7, 2023, art. 101616.
- [7] X. Sui et al., “A Hardware-Friendly High-Precision CNN Pruning Method and Its FPGA Implementation,” *Sensors*, vol. 23, no. 2, art. 824, 2023.
- [8] N. Malik, “Interpretability-Aware Pruning for Efficient Medical Image Analysis,” *arXiv*, 2025 (preprint).
- [9] M. S. Vani et al., “Personalized health monitoring using explainable AI: bridging trust in predictive healthcare,” *Sci. Rep.*, vol. 15, p. 31892, 2025.
- [10] T. Dutta et al., “Towards On-Device Explainable Artificial Intelligence for the Semantic Web of Everything,” *Future Gener. Comp. Syst.*, 2026 (preprint).
- [11] D. Shehada et al., “An Explainable Framework for Mental Health Monitoring Using Lightweight and Privacy-Preserving Federated Facial Emotion Recognition,” *Sensors*, vol. 25, no. 7, 2024.
- [12] S. Jiang et al., “MoE-TinyMed: Mixture-of-Experts for Tiny Medical Large Vision-Language Models,” *arXiv preprint arXiv:2404.10237*, 2024.

Biographies



Shital Ajit Dumbre is an Assistant Professor in the Department of Computer Science and Engineering at MIT Art, Design and Technology University, Pune, India. She holds a M.E. in Computer Science and Engineering Her research interests include Artificial Intelligence, Machine Learning, Data Science, Healthcare Analytics, and Explainable AI.



Arti Kale is an Assistant Professor in the Department of Computer Science and Engineering at MIT Art, Design and Technology University, Pune, India. She holds a Ph.D. in Computer Science and Engineering Her research interests include Artificial Intelligence, Machine Learning, Data Science, Healthcare Analytics, and Explainable AI.