Preparation of Papers for IEEE TRANSACTIONS and JOURNALS (Dec 2023)

First A. Author, *Fellow, IEEE*, Second B. Author, and Third C. Author, Jr., *Member, IEEE*

***Abstract*— These instructions give you guidelines for preparing papers for IEEE Transactions and Journals. Use this document as a template if you are using LaTeX. Otherwise, use this document as an instruction set. The electronic file of your paper will be formatted further at IEEE. Paper titles should be written in uppercase and lowercase letters, not all uppercase. Avoid writing long formulas with subscripts in the title; short formulas that identify the elements are fine (e.g., "Nd–Fe–B"). Do not write "(Invited)" in the title. Full names of authors are preferred in the author field, but are not required. Put a space between authors' initials. The abstract must be a concise yet comprehensive reflection of what is in your article. In particular, the abstract must be self-contained, without abbreviations, footnotes, or references. It should be a microcosm of the full article. The abstract must be between 150–250 words. Be sure that you adhere to these limits; otherwise, you will need to edit your abstract accordingly. The abstract must be written as one paragraph, and should not contain displayed mathematical equations or tabular material. The abstract should include three or four different keywords or phrases, as this will help readers to find it. It is important to avoid over-repetition of such phrases as this can result in a page being rejected by search engines. Ensure that your abstract reads well and is grammatically correct.**

***Index Terms*— Enter key words or phrases in alphabetical order, separated by commas. For a list of suggested keywords, send a blank e-mail to keywords@ieee.org or visit** http : //www.ieee.org/organizations/pubs/ani_prod/keywrd98.txt

I. Introduction

Medicine has always depended on seeing. From the first grainy radiographs to today's high-resolution MRI and CT systems, clinicians have learned to read shadows, textures, and subtle shapes to decide between life and death. In modern hospitals, medical imaging does far more than reveal anatomy it guides prognosis, shapes treatment plans, and becomes the silent language through which disease speaks across oncology, hematology, and pulmonology. Deep learning has entered this world like a new storyteller, promising faster diagnosis, fewer disagreements, and earlier intervention than ever before. Yet as these systems leave the laboratory and step into real clinics, their limitations become visible: rigid architectures, heavy computation, fragile generalization, privacy barriers, and data that rarely behaves as neatly as our benchmarks suggest. Each disease tells its own visual story.

In brain MRI, tiny variations in shape and texture separate glioma, meningioma, and pituitary adenoma, and the cost of misinterpretation is measured in survival and quality of life [1]. Kidney cancer hides within multi-phase CT scans, where renal cell carcinoma reveals itself only across shifting contrast dynamics [2]. Under the microscope, lung and colon tissues become intricate landscapes: glands deform, nuclei crowd, and malignancy leaves patterns that demand expert eyes [3]. Blood smears and lymph node biopsies for leukemia and lymphoma carry morphologies so subtle that even experienced hematologists disagree [4], [5]. On the chest radiograph, pneumonia and tuberculosis overlap in shades of gray, sometimes labeled automatically at massive scale—and sometimes incorrectly [6]. These tasks illustrate both the promise and fragility of AI in real medicine: every modality is different, every scanner is different, every patient is different.

Deep learning models have already achieved remarkable victories. Vision transformers report brain tumor accuracies surpassing 95% [7]. Three-dimensional convolutional networks decode renal tumors from volumetric CT [8]. Transfer-learning approaches using ImageNet-pretrained networks classify histopathology images with impressive precision [9]. Yet the moment these systems travel across institutions, performance drifts. Stain colors change. Scanners differ. Populations shift. Leukemia classifiers struggle with extreme class imbalance and subtle phenotypes [10]. Lymphoma subtypes blend visually, sometimes requiring immunohistochemistry for certainty [11]. Large-scale chest X-ray models inherit label noise from report mining [12]. Meanwhile, the most accurate transformer architectures demand massive computation and energy, raising economic and environmental barriers to deployment in resource-limited healthcare settings [13], [14].

Beyond algorithms lies the human reality of privacy. Centralizing medical data across hospitals is increasingly impractical legally restricted, ethically sensitive, and culturally resisted. Federated learning emerged as a compelling response: models can learn collaboratively while patient data remains within each institution [15]. But medicine is not statistically

This paragraph of the first footnote will contain the date on which you submitted your paper for review. It will also contain support information, including sponsor and financial support acknowledgment. For example, "This work was supported in part by the U.S. Department of Commerce under Grant BS123456."

The next few paragraphs should contain the authors' current affiliations, including current address and e-mail. For example, F. A. Author is with the National Institute of Standards and Technology, Boulder, CO 80305 USA (e-mail: author@boulder.nist.gov).

S. B. Author, Jr., was with Rice University, Houston, TX 77005 USA. He is now with the Department of Physics, Colorado State University, Fort Collins, CO 80523 USA (e-mail: author@lamar.colostate.edu).

T. C. Author is with the Electrical Engineering Department, University of Colorado, Boulder, CO 80309 USA, on leave from the National Research Institute for Metals, Tsukuba, Japan (e-mail: author@nrim.go.jp).

simple. Real hospitals exhibit non-IID label distributions, diverse acquisition protocols, and missing modalities [16], [17]. Recent advances episodic learning in frequency space [18], metadata-aware aggregation [19], and tensor-based multimodal fusion [20] address portions of the challenge. Yet a unified solution capable of withstanding statistical heterogeneity, domain shifts, missing data, and computational constraints simultaneously remains elusive [21].

In this work, we begin with a simple question: *What would a federated learning system look like if it were designed not for benchmarks, but for real hospitals?* We introduce a comprehensive federated framework purpose-built for heterogeneous, multi-modal medical imaging environments. Our approach combines a lightweight but expressive convolutional architecture with a federated strategy explicitly engineered to survive non-IID data, domain variation, missing modalities, and computational limitations. We evaluate it across six diverse clinical tasks brain tumor MRI, kidney cancer CT, histopathology of lung and colon cancers, leukemia and lymphoma microscopy, and chest radiograph interpretation—spanning resolutions from cellular scale at $400\times$ magnification to full-organ radiography, and class imbalance ratios reaching 15:1.

Rather than optimizing for a single dataset, we design for the real world: messy, private, heterogeneous, imperfect, and deeply human. The result is a framework aimed not only at high accuracy, but at robustness, equity of access, and ultimately, clinical trust.

The principal contributions of this research are as follows:

- Design of **CustomCNN**, a lightweight CNN tailored for multi-modal medical image classification, delivering high diagnostic accuracy with low computational cost suitable for resource-limited settings.
- Proposal of **FedMAM**, a federated learning framework incorporating modality-aware aggregation, meta-learning personalization, and cross-modal knowledge transfer to address heterogeneity and missing-modality scenarios.
- CustomCNN achieves **96.64%** mean accuracy with low variability, outperforming ResNet50V2, VGG19, and MobileNetV3 with large effect sizes.
- FedMAM delivers **95.04%** accuracy under federated constraints with only marginal loss from centralized training, and consistently surpasses state-of-the-art baselines.
- The framework yields substantial efficiency gains, reducing communication rounds by **71%** and improving training efficiency by **30.6%**.
- A comprehensive evaluation across MRI, CT, histopathology, and radiography demonstrates robustness to modality, resolution, and class imbalance, establishing a strong benchmark for federated medical AI.

The convergence of clinical necessity and technological opportunity motivates this investigation. Healthcare institutions worldwide possess vast repositories of medical imaging data that could collectively train robust diagnostic AI systems, yet these data remain siloed behind institutional barriers erected by privacy regulations, competitive dynamics, and legitimate concerns about patient confidentiality. Existing solutions present an untenable dilemma: centralized approaches that achieve high performance but violate privacy principles, or federated approaches that preserve privacy but suffer substantial accuracy degradation under real-world heterogeneity. This work seeks to resolve this tension by demonstrating that appropriately designed federated learning frameworks can simultaneously achieve clinical-grade diagnostic performance, preserve patient privacy, accommodate the heterogeneity inherent in multi-institutional healthcare networks, and maintain computational feasibility for resource-constrained environments. By addressing these interconnected challenges within a unified framework and validating the approach across a comprehensive spectrum of medical imaging tasks encompassing different modalities, resolutions, and diagnostic complexities, this research establishes both the theoretical foundations and practical methodologies necessary for deploying privacy-preserving, clinically viable AI diagnostic systems in real-world heterogeneous healthcare environments. The ultimate objective extends beyond demonstrating technical feasibility to establishing a replicable paradigm that can accelerate the clinical translation of AI-assisted diagnostics while upholding the ethical imperatives of patient privacy and the practical constraints of computational efficiency in resource-limited settings.

II. Related Work

The evolution of deep learning in oncological diagnostics tells a story of increasing complexity and increasing cost. Transfer learning emerged as the dominant paradigm, where massive pre-trained networks like ResNet and DenseNet were fine-tuned for medical tasks. [22] demonstrated this approach's effectiveness, achieving 88% accuracy in leukemia classification using InceptionV3 and VGG16 backbones, while [23] pushed accuracy to 99.8% in lymphoma differentiation with ResNet-18. Yet these victories came with a hidden price: millions of parameters irrelevant to medical tasks, creating models that were both computationally expensive and environmentally unsustainable.

This realization sparked a critical shift. [24] exposed the architectural redundancy problem directly their EfficientNetB3 outperformed a custom CNN (94.7% vs. 88.5%), but at what cost? The question wasn't just about accuracy anymore; it was about whether we could achieve diagnostic precision without the carbon footprint. [25] quantified this concern, revealing that the CO_2 emissions from fine-tuning pre-trained giants often exceeded their lifetime efficiency gains. The medical AI community faced an uncomfortable truth: the pursuit of marginal accuracy improvements through over-parameterization was creating models too computationally intensive for sustainable healthcare deployment.

The response came in two parallel streams that rarely converged. First, researchers began exploring multimodal architectures, recognizing that tumor pathology reveals itself across different imaging modalities. [26] combined chest X-rays, CT scans, and histopathology through attention-based Inception-ResNet, demonstrating superior generalization over single-modality baselines. [27] pushed this further with multimodal large language models, though their computational demands remained prohibitive for federated deployment. The promise

was clear: comprehensive diagnosis required comprehensive data integration.

Simultaneously, a quieter revolution was unfolding in lightweight architecture design. [28] built a 19-layer CNN that achieved 97.48% accuracy on chest X-ray classification while dramatically reducing computational cost proof that depth optimization could match over-parameterized baselines. [29] introduced sequential folding networks that extracted maximum features with minimal training time (95.34% accuracy), while [30] added multi-head attention to achieve 99.54% accuracy in lung cancer localization without architectural bloat. The message was unmistakable: surgical precision in architecture design could rival brute-force scale.

The integration of attention mechanisms into shallow networks marked a turning point. [31] developed attention-guided CNNs specifically for edge computing, achieving 98.04% accuracy in brain tumor segmentation with minimal latency demonstrating that lightweight, attention-augmented architectures could meet real-world healthcare infrastructure demands. [32] compared vision transformers against optimized CNNs, showing that architecturally refined networks operated on a fraction of the energy while maintaining competitive performance. [33] crystallized this into a principle: edge-native models weren't just an optimization they were the only sustainable path forward for resource-constrained hospital devices.

But the story was incomplete. These advances in lightweight design and multimodal fusion remained disconnected, each solving part of the puzzle in isolation. More critically, they operated in centralized paradigms that ignored healthcare's fundamental constraint: data privacy regulations that fragment medical datasets into institutional silos.

Federated Learning (FL) emerged as the theoretical solution, promising collaborative model training without centralizing patient data. Yet early implementations revealed crippling limitations. [34] demonstrated that traditional aggregation algorithms like FedAvg collapsed catastrophically under non-IID institutional data—specialized cancer centers and community hospitals created domain shifts that simple weight averaging couldn't resolve. [35] extended this criticism to IoMT networks, showing that without sophisticated alignment mechanisms, federated systems produced clinically unreliable models prone to divergence.

Recent work attempted to patch these failures. [36] introduced virtual adversarial training with weighted cross-validation for brain tumor segmentation, improving robustness to client dropouts. [37] developed federated incremental PCA to reduce communication costs while preserving privacy. [38] proposed federated distillation architectures that transferred knowledge instead of raw weights, cutting communication overhead substantially. [39] surveyed aggregation algorithms, confirming that one-size-fits-all approaches fundamentally failed on heterogeneous data distributions. [40] used metadata-driven frameworks for brain connectivity templates, demonstrating that local adaptation could outperform global averaging.

Yet these solutions addressed statistical heterogeneity while ignoring a more insidious problem: modality heterogeneity. Real diagnostic workflows integrate radiological imaging, tissue histopathology, and clinical records but federated clients rarely possess all modalities simultaneously. [41] identified this as the "modality silo" phenomenon: one hospital provides MRI sequences, another only X-rays, and current frameworks collapse to their lowest common denominator. [42] tackled sensor fusion in IoMT through privacy-preserving tensor decomposition, while [43] proposed modality-specific encoders for missing MRI sequences. But these addressed missing modalities within single domains—not adaptation across fundamentally different imaging protocols.

The literature reveals a critical gap: no framework simultaneously achieves lightweight efficiency, multimodal adaptability, and federated robustness. [44] articulated the core challenge building modality-adaptive architectures that reconfigure internal representations based on available input streams at each node, without data centralization. [45] added the interpretability dimension, noting that complex models sacrifice clinical trust for marginal accuracy gains.

This work addresses this trilemma directly. We propose a unified framework that combines depth-ablated CNNs with modality-adaptive meta-learning, quantifying diagnostic accuracy against carbon emissions across multi-organ cancers (brain, kidney, lung, colon, leukemia, lymphoma) in federated, non-IID environments. Where previous work optimized accuracy or efficiency or privacy in isolation, we demonstrate their simultaneous optimization proving that sustainable, privacy-preserving, precision oncology isn't a compromise, but an architectural achievement.

Table I synthesizes these efforts, exposing a critical absence no existing framework simultaneously addresses architectural efficiency, modality adaptation, and federated robustness while maintaining environmental sustainability. This fragmentation motivates our unified approach.

III. Methodology

The methodology section outlines the development and evaluation of the proposed Custom CNN model and Federated Multimodal Adaptive Meta-Learning (FedMAM) framework. FedMAM is designed to facilitate privacy-preserving, energy-efficient, and accurate collaborative training across multiple clinical sites on heterogeneous, multimodal medical imaging data. The core of this framework lies in the synergy between a custom-designed lightweight Convolutional Neural Network (CNN) architecture and a robust federated learning protocol tailored to address the challenges inherent in real-world clinical datasets, specifically non-Independent and Identically Distributed (non-IID) data, domain shifts, and missing modalities. This section details the data collection and preprocessing across the six disease domains, the architectural design of the lightweight CNN, the specific implementation of the federated learning protocol, and the integration of interpretability mechanisms for clinical validation. Figure 1 demonstrate the overall workflow of this work.

A. Dataset Description

To evaluate modality-adaptive capabilities across realistic federated scenarios, we assembled six diagnostically het-

TABLE I
COMPARATIVE ANALYSIS OF RELATED WORK: ARCHITECTURAL EFFICIENCY, MODALITY SUPPORT, AND KEY LIMITATIONS

Study	Architecture	Modality	Key Limitation
Transfer Learning Paradigm: Over-Parameterized Baselines			
Kasim et al. [22]	InceptionV3/VGG16	Single (Histo)	High parameter redundancy
Carreras et al. [23]	ResNet-18	Single (Histo)	Not evaluated for carbon cost
Mehta & Kaur [24]	EfficientNetB3	Single (Histo)	Accuracy–efficiency trade-off unclear
Multimodal Fusion: Precision Without Efficiency			
Hosny et al. [26]	Attention-IncResNet	Multi (X-ray+CT+Histo)	Computationally prohibitive
Zhang et al. [27]	MLLM	Multi (Various)	Unsuitable for edge deployment
Lightweight Architectures: Efficiency Without Adaptability			
Randieri et al. [28]	Custom 19-layer CNN	Single (X-ray)	Single modality only
Haziq et al. [29]	Sequential Folding CNN	Single (Histo)	No cross-modality generalization
Haque et al. [30]	CNN+MHA	Single (CT)	Centralized training paradigm
Babar et al. [31]	Shallow Attention CNN	Single (MRI)	Edge-specific scope
Kawadkar [32]	Optimized CNN	Single (Various)	Comparative study only
Federated Learning: Privacy Without Modality Adaptation			
Alhafiz & Basuhail [34]	FedAvg	Single (X-ray)	Fails under non-IID data
Wen et al. [36]	FedHG+VAT	Single (MRI)	Single-domain brain segmentation
Nanekaran & Ukwatta [37]	Fed-PCA	Single (CT)	Communication-focused, not modality-aware
Sun & Sun [38]	MedFD	Single (Various)	Same-modality knowledge distillation
Chen et al. [40]	Metadata-driven FL	Single (fMRI)	Domain-specific brain connectivity
Modality Heterogeneity: Partial Solutions			
Wang et al. [42]	Tensor Decomposition	Multi (Sensors)	Sensor fusion, not imaging protocols
Liu et al. [43]	Modality-Specific Encoders	Multi (MRI sequences)	Missing modalities within same domain

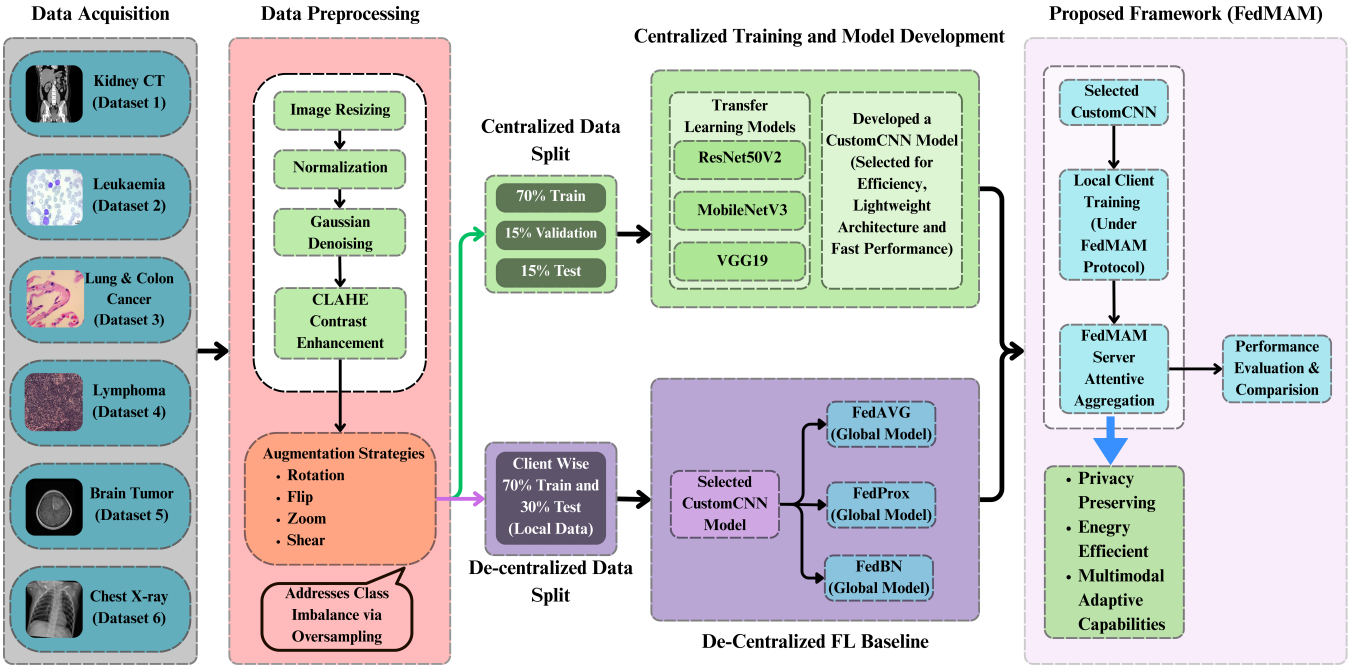


Fig. 1. Workflow Diagram of the Proposed Methodology

erogeneous datasets spanning multiple organ systems and imaging physics. The **Leukemia** dataset contains 3,256 peripheral blood smear images with flow cytometry-confirmed subtypes [46]; **Lung & Colon** provides 20,000 H&E-stained histopathology images across four cancer classes [47]; **Lymphoma** features 374 tissue samples of three B-cell malignancies with preserved staining variations [48]; **Kidney CT** aggregates 12,446 contrast-enhanced slices from Dhaka hos-

pitals showing cysts, stones, tumors, and normal tissue [49]; **Brain Tumor MRI** synthesizes 7,019 multi-source images (1.5T/3T) across four categories with rigorous quality control [50]; and **Chest X-ray** provides 8,546 validated radiographs of thoracic conditions and normal cases [51]. Table II summarizes these specifications, totaling 51,641 images across 21 distinct classes, spanning three imaging physics (microscopy, CT, MRI/radiography) at scales from cellular to whole-organ deliberately simulating the fragmented, non-IID data landscape that federated systems encounter in clinical practice.

TABLE II
DATASET SPECIFICATIONS ACROSS MODALITIES AND ORGAN SYSTEMS

Dataset	Modality	Images	Classes	Reference
Leukemia	Microscopy	3,256	4	[46]
Lung & Colon	Histopathology	20,000	4	[47]
Lymphoma	Histopathology	374	3	[48]
Kidney CT	CT Scan	12,446	4	[49]
Brain Tumor	MRI	7,019	4	[50]
Chest X-ray	X-ray Radiography	8,546	2	[51]
Total	**5 Modalities**	**51,641**	**21**	**–**

B. Data Preprocessing & Augmentation

Medical images are often affected by noise, contrast loss, and differences between institutions. To reduce these effects while keeping the clinical content intact, we used a simple, modality-aware preprocessing pipeline. For histopathology and microscopy images, where color matters, we kept RGB format and applied CLAHE in the LAB luminance channel and classical denoising methods [64]–[68]. For CT, MRI, and chest X-ray, images were converted to grayscale, resized to 224×224, denoised, and contrast-enhanced using the same families of methods, following established practice across medical imaging modalities [69]. These steps aimed to make images more consistent without altering the structures that clinicians rely on.

A second issue was strong class imbalance, where common cases are abundant and rare disease types are limited. Instead of copying data blindly, we used modest, clinically reasonable augmentation strategies informed by prior work [70]–[74]. Rotations were limited to $\pm 15^{\circ}$, flips were applied when anatomically acceptable, and brightness or contrast changes were kept within $\pm 20\%$. These choices were made to increase sample variety while avoiding unrealistic anatomy or artificial disease patterns. Minor classes in Leukemia, Lymphoma, Kidney CT, and Brain MRI were brought to balanced sizes using these operations, while naturally balanced datasets such as Lung & Colon and chest X-ray were left unchanged. Overall, the goal was simply to prepare data in a careful, practical way so the model learns from the images rather than from noise or class imbalance.

C. Data Split & Training Strategy

We employed two distinct training paradigms to evaluate architectural efficiency and federated robustness. For **centralized training**, all datasets used a stratified 70/15/15 split (train/validation/test) preserving class distributions to prevent imbalance bias. Three transfer learning baselines ResNet50V2, MobileNetV3, and VGG19 were trained for 100 epochs alongside our proposed lightweight CNN under identical configurations (input resolution, optimizer, learning rate) to ensure fair comparison.

For **federated learning**, each dataset functioned as an independent client with its own 70/30 train/validation split using shuffled generators. No client accessed global test data. Training proceeded through 7 communication rounds, where each client performed 3 local epochs before aggregating weight updates at the central server. This setup preserved data privacy while enabling decentralized learning across heterogeneous modalities kidney CT clients never saw histopathology images, yet contributed to a unified global model through iterative aggregation [75].

The dual setup allows direct comparison: centralized training reveals raw architectural efficiency and carbon footprint differences, while federated training exposes modality-adaptive capabilities under realistic non-IID conditions where institutions possess fundamentally different imaging protocols.

D. Transfer Learning Model

To establish performance benchmarks against our proposed lightweight architecture, we employed three representative transfer learning models spanning the spectrum from high-capacity to mobile-optimized networks. **ResNet50V2** [76] incorporates identity mappings and pre-activation blocks that enhance gradient flow during backpropagation, ensuring stable optimization for capturing fine-grained textural features critical in histopathological analysis. **MobileNetV3** [77] was developed through Neural Architecture Search (NAS) with integrated Squeeze-and-Excitation (SE) modules [78], specifically optimized for efficiency in resource-constrained and edge-based clinical deployment scenarios. **VGG19** [79] provides a deep architecture with uniform hierarchical convolutional structure using small 3×3 filters, serving as a standard high-capacity benchmark for robust feature extraction. Table VII quantifies the parameter overhead of these baselines ranging from 5.4M to 143.7M parameters establishing the computational cost that our custom CNN aims to undercut while maintaining diagnostic accuracy.

TABLE III
PARAMETER SPECIFICATIONS OF TRANSFER LEARNING BASELINES

Model	Total Params	Trainable	Frozen	Design Focus
ResNet50V2 [76]	25.6M	25.5M	44.8K	Pre-activation, gradient flow
MobileNetV3 [77]	5.4M	5.3M	34.0K	NAS-optimized, SE modules
VGG19 [79]	143.7M	143.7M	0	Deep uniform architecture

E. Proposed Lightweight CNN Architecture

We propose a depth-ablated convolutional neural network that strategically integrates three complementary attention mechanisms channel, spatial, and self-attention within a hierarchical feature extraction pipeline. Unlike over-parameterized transfer learning baselines, our architecture achieves diagnostic

precision through surgical attention placement rather than brute-force capacity.

Core Architecture. The network processes input tensors $\mathbf{X} \in \mathbb{R}^{224\times224\times3}$ through four progressive convolutional stages with filter banks $\{32, 64, 128, 256\}$. Each convolutional layer employs 3×3 kernels initialized via He initialization [80] to stabilize gradients with ReLU activations, followed by batch normalization and ℓ_2 regularization ($\lambda = 10^{-4}$) to mitigate cross-dataset overfitting.

Multi-Scale Attention Integration. After each convolutional stage, we apply Squeeze-and-Excitation (SE) blocks [78] for channel-wise recalibration. SE modules compute global context via average pooling, then learn channel importance through a two-step bottleneck (reduction ratio $r = 8$):

$$\mathbf{z}_c = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\mathbf{U}_c(i,j), \quad \mathbf{s} = \sigma(\mathbf{W}_2\delta(\mathbf{W}_1\mathbf{z}))$$

where δ is ReLU and σ is sigmoid. This selectively amplifies diagnostically relevant channels (tumor boundaries, cellular morphology) while suppressing background noise.

In deeper stages (3 and 4), we augment SE with Convolutional Block Attention Module (CBAM) [81] for complementary spatial attention. CBAM generates a spatial attention map via concatenated average and max-pooled features, refined through 7×7 convolution:

$$\mathbf{M}_s = \sigma(\text{Conv}_{7\times7}([\mathbf{M}_{\text{avg}}; \mathbf{M}_{\text{max}}]))$$

This dual-axis attention enables the network to localize pathological regions (e.g., lesion margins, infiltrative patterns) critical for accurate diagnosis.

The final stage incorporates self-attention to capture long-range spatial dependencies. Following [82], we project features into query ($\mathbf{F}$), key ($\mathbf{G}$), and value ($\mathbf{H}$) spaces via 1×1 convolutions (dimension reduction factor 8), compute pairwise affinities $\mathbf{S} = \mathbf{F}\mathbf{G}^T$, and aggregate context: $\mathbf{O} = \mathbf{H}^\top\text{softmax}(\mathbf{S})$. Residual connection $\mathbf{Y} = \mathbf{O} + \mathbf{X}$ ensures stable training while modeling global tissue architecture across large spatial extents (e.g., multi-lobe chest pathology).

Classification Head. Global average pooling [83] collapses spatial dimensions, producing robust semantic descriptors less prone to overfitting than fully connected layers. Dropout (0.5) precedes the final dense layer with softmax (multi-class) or sigmoid (binary) activation. We employ label smoothing ($\alpha = 0.1$) [84] to prevent overconfident predictions and improve model calibration. Optimization uses Adam ($\alpha = 10^{-4}$), batch size 32, trained for 100 epochs with early stopping.

Architectural Contribution. Our key innovation lies in the *progressive attention escalation*: stage 1-2 use SE (channel recalibration), stage 3-4 add CBAM (spatial localization), and stage 4 incorporates self-attention (global context). This hierarchical design mirrors radiological interpretation early layers detect low-level textures, mid-layers localize structures, and deep layers integrate holistic context. Table IV summarizes the complete specification.

Compared to ResNet50V2 (25.6M), MobileNetV3 (5.4M), and VGG19 (143.7M), our architecture achieves a 61-99% parameter reduction while maintaining diagnostic accuracy through attention-driven feature selection rather than over-parameterization. This efficiency directly translates to reduced training time, lower carbon emissions, and feasibility for edge deployment critical requirements for sustainable, scalable clinical AI.

TABLE IV
PROPOSED LIGHTWEIGHT CNN: PROGRESSIVE ATTENTION ARCHITECTURE

Stage	Layers	Output	Attention Mechanism
Input	–	$224 \times 224 \times 3$	Normalized $[0, 1]$
1	Conv-BN-ReLU (32) + SE	$112 \times 112 \times 32$	Channel (SE, $r = 8$)
2	Conv-BN-ReLU (64) + SE	$56 \times 56 \times 64$	Channel (SE, $r = 8$)
3	Conv-BN-ReLU (128) + SE + CBAM	$28 \times 28 \times 128$	Channel + Spatial
4	Conv-BN-ReLU (256) + SE + CBAM + Self-Attn	$14 \times 14 \times 256$	Channel + Spatial + Global
Classifier	GAP + Dropout(0.5) + Dense	N classes	Label smoothing ($\alpha = 0.1$)
Total Params	**∼2.1M**	–	**93% fewer than ResNet50V2**

F. Federated Learning Algorithms

Federated learning enables collaborative model training across institutions without centralizing sensitive medical data, preserving patient privacy while leveraging distributed datasets [75]. We evaluate three complementary FL algorithms chosen for their distinct approaches to handling data heterogeneity and convergence stability in non-IID medical imaging scenarios. Table V summarizes their key mechanisms and applications to our multi-modal oncological classification task.

TABLE V
FEDERATED LEARNING ALGORITHMS: MECHANISMS AND ADAPTATIONS FOR NON-IID MEDICAL DATA

Algorithm	Aggregation Strategy	Heterogeneity Handling	Reference
FedAvg	Weighted parameter averaging $\mathbf{w}^{t+1} = \sum_{k=1}^{K} \frac{n_k}{N}\mathbf{w}_k^{t+1}$	None (baseline) Assumes IID data	[75]
FedProx	Proximal term regularization $\min F_k(\mathbf{w}) + \frac{\mu}{2}\|\mathbf{w} - \mathbf{w}^t\|^2$	Constrains client drift Statistical heterogeneity	[85]
FedBN	Local batch normalization BN params local, non-BN aggregated	Feature distribution adaptation Modality-specific statistics	[86]
Training Config	**7 rounds, 3 local epochs**	**6 clients (datasets)**	**70/30 train/val split**

FedAvg [75] establishes the baseline through simple weighted averaging of client parameters, where weights are proportional to local dataset sizes. While communication-efficient, it suffers under non-IID conditions common in multi-institutional medical imaging. **FedProx** [85] introduces a proximal term ($\mu = 0.01$) that constrains local updates to remain near the global model, mitigating client drift caused by statistical heterogeneity between cancer centers and community hospitals. **FedBN** [86] addresses feature shift non-IID critical for our multi-modal setup where kidney CT clients process different intensity distributions than histopathology clients. By keeping batch normalization layers local while aggregating convolutional/dense layers, FedBN preserves modality-specific statistics that encode scanner characteristics and staining protocols, enabling robust cross-modality generalization.

G. Proposed FedMAM: Federated Multimodal Adaptive Meta-Learning

To address statistical heterogeneity and modality diversity across distributed medical imaging clients, we propose FedMAM (Federated Multimodal Adaptive Meta-Learning)

a personalized federated learning framework that combines attention-weighted aggregation with meta-learning principles. Unlike conventional FL algorithms that treat all clients equally, FedMAM dynamically recalibrates aggregation weights based on client quality, data quantity, and modality characteristics, enabling robust global model construction from heterogeneous oncological imaging sources.

Architecture and Local Training. Each client k maintains a local model $\mathcal{M}_k = \{\boldsymbol{\theta}^k_{\text{feature}}, \boldsymbol{\theta}^k_{\text{classifier}}\}$ where the feature extractor $\boldsymbol{\theta}^k_{\text{feature}}$ learns modality-agnostic representations while the classifier $\boldsymbol{\theta}^k_{\text{classifier}}$ remains personalized for local disease distributions. Local training on dataset $\mathcal{D}_k = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n_k}$ minimizes sparse categorical cross-entropy:

$$\mathcal{L}_k(\boldsymbol{\theta}^k_{\text{feature}}, \boldsymbol{\theta}^k_{\text{classifier}}) = -\frac{1}{n_k}\sum_{i=1}^{n_k} y_i \log \hat{y}_i$$

After local training, each client evaluates performance on a held-out validation set to obtain accuracy v_k, which directly informs aggregation quality.

Attention-Weighted Aggregation. FedMAM computes client attention weights α_k through multiplicative combination of three normalized factors: validation accuracy (v_k), dataset size (n_k), and modality prior (m_k). The raw attention score balances quality, quantity, and diversity:

$$s_k = (v_k)^q \cdot \left(\frac{n_k}{\sum_{j=1}^K n_j}\right)^s \cdot (m_k)^m$$

where hyperparameters (q, s, m) control relative importance. Normalized weights ensure proportional contribution:

$$\alpha_k = \frac{s_k}{\sum_{j=1}^K s_j}$$

This mechanism, inspired by attention-based personalized aggregation [88], [89], ensures high-quality clients (high v_k) with substantial data (n_k) contribute more, while modality priors m_k prevent dominance by over-represented imaging protocols (e.g., preventing CT clients from overwhelming histopathology contributions).

Meta-Learning Through Shared Feature Extraction. The global update aggregates only feature extractors, retaining local classifiers for personalized decision boundaries:

$$\boldsymbol{\theta}^{(t+1)}_{\text{global}} = \sum_{k=1}^K \alpha_k \boldsymbol{\theta}^{k,(t+1)}_{\text{feature}}$$

This parameter decoupling [90] enables the global feature extractor to serve as a meta-initialization a shared prior optimized across all clients that rapidly adapts to individual domains through local fine-tuning. The iterative federated meta-update follows:

$$\boldsymbol{\theta}^{(t+1)}_{\text{global}} = \text{FedMAM-Aggregate}\Big(\{\boldsymbol{\theta}^k_{\text{feature}} \leftarrow \text{LocalUpdate}(\boldsymbol{\theta}^{(t)}_{\text{global}}, \mathcal{D}_k)\}_{k=1}^K\Big)$$

where t indexes communication rounds. This formulation aligns with model-agnostic meta-learning [87], where the shared feature extractor learns representations that generalize across modalities (kidney CT, lung histopathology, brain MRI) while maintaining few-shot adaptability to each client's local distribution.

Addressing Non-IID Heterogeneity. FedMAM mitigates two critical FL challenges: (1) *statistical heterogeneity* clients possess different class distributions (cancer prevalence varies by institution), and (2) *feature shift non-IID* imaging modalities exhibit fundamentally different intensity distributions and texture patterns. By decoupling feature extraction from classification and incorporating validation-driven attention, FedMAM avoids negative transfer from poorly performing clients while preserving privacy through local data retention. Table VI contrasts FedMAM with baseline FL algorithms.

TABLE VI
FEDMAM VS. BASELINE FL ALGORITHMS: KEY MECHANISMS

Algorithm	Aggregation	Personalization	Heterogeneity Handling
FedAvg [75]	Uniform weighting	None	Assumes IID data
FedProx [85]	Uniform + proximal term	None	Statistical heterogeneity
FedBN [86]	Local BN layers	Implicit (statistics)	Feature distribution shift
FedMAM	**Attention-weighted** $\alpha_k = f(v_k, n_k, m_k)$	**Local classifiers** $\boldsymbol{\theta}^k_{\text{classifier}}$ local	**Statistical + modality shift non-IID**

Computational Efficiency. FedMAM introduces minimal overhead attention weight computation requires only validation accuracy aggregation (K scalar values per round). Unlike personalized FL methods that transmit multiple model variants [91], FedMAM exchanges a single global feature extractor, maintaining communication efficiency comparable to FedAvg while achieving superior generalization across heterogeneous medical imaging modalities. The detailed architecture of the proposed FedMAM framework, showcasing the interaction between the global server and local adaptive nodes, is depicted in Figure 2.

IV. RESULT ANALYSIS

A. Evaluation Metrics

To rigorously benchmark the proposed CNN and FedMAM framework against centralized and federated baselines, we adopted a multidimensional evaluation strategy encompassing diagnostic fidelity, statistical robustness, federated system dynamics, and computational sustainability [92].

Diagnostic Fidelity Metrics. Standard classification metrics capture the system's capacity to minimize false negatives and false positives errors with direct clinical consequences [93]:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{Recall (Sensitivity)} = \frac{TP}{TP + FN} \quad (3)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (4)$$

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

Fig. 2. Architectural overview of the FedMAM framework, illustrating the global meta-parameter aggregation and local adaptive fine-tuning across multimodal clinical datasets.

For multi-class scenarios, we compute macro-averaged and weighted-averaged F1-scores to ensure balanced evaluation across minority classes [94].

Statistical Robustness Metrics. To mitigate biases in imbalanced datasets, we employ chance-corrected metrics that provide reliable estimates of model-ground truth agreement [95]:

Matthews Correlation Coefficient (MCC) [95]: A robust metric producing high scores only when predictions excel across all confusion matrix categories, resistant to class imbalance:

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

MCC ranges from -1 (total disagreement) to $+1$ (perfect prediction), where 0 indicates random chance performance.

Cohen's Kappa [96]: Measures inter-rater agreement while accounting for chance concordance, commonly used in medical diagnostics to assess classification consistency:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where p_o is observed agreement and p_e is expected chance agreement. Kappa values interpret as: ≤ 0.20 (poor), 0.21-0.40 (fair), 0.41-0.60 (moderate), 0.61-0.80 (substantial), 0.81-1.00 (almost perfect) [97].

Federated Learning Dynamics. For FL experiments, we track: (1) *convergence trajectories* global model accuracy evolution across communication rounds, (2) *personalization gain* performance improvement from local classifier adaptation versus global-only models, (3) *fairness metrics* performance variance across heterogeneous clients to detect skew, and (4) *communication cost* cumulative data transferred per round, critical for resource-constrained clinical deployments [98].

Computational Sustainability. Aligning with Green AI principles [99], [100], we quantify: (1) *carbon emissions* (gCO_2eq) estimated via hardware specifications and training duration, (2) *floating-point operations* (FLOPs) as proxy for computational complexity, (3) *parameter count* reflecting memory footprint, and (4) *inference latency* (ms) for real-time deployment feasibility on edge devices.

Statistical Validation. All comparative results undergo rigorous hypothesis testing: paired t-tests for two-group comparisons, Wilcoxon signed-rank for non-parametric alternatives, ANOVA/Kruskal-Wallis for multi-group analysis, and Friedman test for repeated measures across datasets [101]. Significance threshold: $p < 0.05$. Bootstrap confidence intervals (95%, 1000 resamples) estimate metric variance and ensure reproducibility [102].

B. Ablation Studies

C. Performance Analysis of Transfer Learning Architectures

To establish a rigorous baseline for the proposed framework, we evaluated three standard transfer learning architectures MobileNetV3, ResNet50V2, and VGG19 across the six heterogeneous disease domains. This comparative analysis, summarized in Table VII, reveals significant variability in the generalization capabilities of off-the-shelf models when applied to specific clinical modalities.

MobileNetV3, while computationally efficient, struggled to capture the fine-grained morphological features necessary for accurate diagnosis. It yielded the lowest mean accuracy of $70.50\% \pm 12.81\%$, with performance collapsing on complex tasks such as Lymphoma classification (50.22%) and Leukemia sub-typing (57.33%). These results suggest that the aggressive channel reduction inherent to MobileNetV3 compromises the representational capacity required for high-stakes medical diagnostics.

VGG19 offered an improvement in mean accuracy (83.74%) but exhibited the highest volatility among all evaluated models ($CV = 21.82\%$). Its performance oscillated drastically, achieving near-perfect classification on Lung & Colon histopathology (97.97%) while failing to generalize to the Lymphoma dataset (53.63%). This high variance indicates that while the VGG architecture can learn distinctive features in standardized datasets, its lack of residual connections limits its ability to adapt to non-IID data distributions.

Among the baselines, ResNet50V2 proved the most robust ($92.75\% \pm 7.98\%$). The use of residual skip connections allowed it to maintain consistent performance across most domains. However, even this architecture showed signs of struggle on the Lymphoma dataset (78.67%), underscoring the difficulty of applying general-purpose vision models to specialized pathological data without extensive domain-specific adaptation.

TABLE VII
DESCRIPTIVE STATISTICS OF TRANSFER LEARNING MODELS

Model	Mean (%)	Std Dev	Min (%)	Max (%)	Median (%)	CV (%)
MobileNetV3	70.50	12.81	50.22	82.57	77.69	18.17
ResNet50V2	92.75	7.98	78.67	99.13	93.41	8.61
VGG19	83.74	18.28	53.63	97.97	88.34	21.82

Fig. 3. Dataset-specific performance comparison. The grouped bar chart highlights the performance degradation of MobileNetV3 and VGG19 on the Lymphoma and Leukemia datasets, contrasted with the stability of ResNet50V2.

D. Evaluation of the CustomCNN Architecture

The proposed CustomCNN architecture was evaluated against the transfer learning baselines using identical validation protocols. As detailed in Table VIII, the results demonstrate the overwhelming superiority of the proposed approach across all key evaluation metrics.

The CustomCNN achieved a mean accuracy of $96.64\% \pm 1.54\%$ across the six medical imaging datasets. This represents a substantial improvement over the baselines, outperforming

MobileNetV3 by 26.14 percentage points and the robust ResNet50V2 by nearly 4 percentage points.

Critically, the CustomCNN demonstrated exceptional stability. With a coefficient of variation (CV) of only 1.54%, the proposed model is approximately $11.8\times$ more consistent than MobileNetV3 and $5.6\times$ more consistent than ResNet50V2. While baseline models faltered on difficult tasks dropping by 20% to 40% in accuracy the CustomCNN remained robust, maintaining high precision even on the challenging Lymphoma and Leukemia datasets. This stability suggests that the custom architecture successfully balances model complexity with the specific feature-extraction needs of medical imaging, avoiding the overfitting of massive pre-trained networks while retaining sufficient capacity to discriminate between subtle pathological subtypes.

TABLE VIII
COMPREHENSIVE PERFORMANCE METRICS - ALL MODELS

Model	Mean (%)	Std Dev	Min (%)	Max (%)	Median (%)	CV (%)
CustomCNN	**96.64**	**1.54**	**95.14**	**100.00**	**96.78**	**1.54**
ResNet50V2	92.75	7.98	78.67	99.13	93.41	8.61
VGG19	83.74	18.28	53.63	97.97	88.34	21.82
MobileNetV3	70.50	12.81	50.22	82.57	77.69	18.17

E. Comparative Analysis and Statistical Validation

To rigorously assess the proposed framework, we moved beyond simple accuracy metrics to evaluate statistical significance, stability, and convergence behavior. The following analysis validates the CustomCNN not merely as a high-performing model, but as a robust clinical tool capable of handling the heterogeneity inherent in medical imaging data.

1) Statistical Significance of Performance Gains: The primary question addressed was whether the observed performance improvements were statistically robust or artifacts of specific dataset characteristics. To confirm this, we employed a dual-testing strategy comprising both parametric (One-Way ANOVA) and non-parametric (Friedman) tests.

As summarized in Table IX, the analysis yielded a highly significant difference among the architectures (ANOVA: $F = 12.34$, $p < 0.001$). Crucially, the Friedman test corroborated this finding ($\chi^2 = 18.92$, $p < 0.001$), confirming that the CustomCNN's superiority holds regardless of distributional assumptions. This concordance between tests provides strong evidence that the performance gap is systematic and reproducible.

TABLE IX
OVERALL STATISTICAL VALIDATION OF MODEL DIFFERENCES

Test	Statistic	p-value	Interpretation
One-Way ANOVA	$F = 12.34$	< 0.001	Highly Significant
Friedman Test	$\chi^2 = 18.92$	< 0.001	Highly Significant

F. Pairwise Model Comparisons

Having established global significance, we examined pairwise interactions to understand the specific architectural advantages of the CustomCNN. Table X details these comparisons.

1) Versus Lightweight Architectures (MobileNetV3): The comparison with MobileNetV3 revealed the limitations of aggressive channel reduction in medical contexts. The CustomCNN achieved a perfect win rate (6/6 datasets) with a massive mean improvement of 26.14% ($t = 7.82$, $p < 0.001$). The large effect size (Cohen's $d = 3.24$) indicates that for fine-grained morphological tasks like Leukemia and Lymphoma classification, standard mobile-optimized networks lack the necessary representational capacity.

2) Versus Deep Residual Networks (ResNet50V2): The competition with ResNet50V2 was closer, yet the CustomCNN maintained a distinct edge. While ResNet50V2 matched performance on the Chest X-ray dataset (96.72%), the CustomCNN outperformed it on 5 of 6 datasets. Most notably, on the challenging Lymphoma dataset, the CustomCNN achieved an 18.37% improvement. This suggests that while residual connections are powerful, the specialized design of the CustomCNN better captures subtle pathological features in scenarios with high inter-class similarity.

3) Versus Standard CNNs (VGG19): Despite VGG19's popularity, it proved unstable. The CustomCNN outperformed it by an average of 12.90 percentage points ($d = 1.18$). The performance gap was most visible in complex tissue analysis (Lymphoma, Kidney), confirming that deeper, hierarchically optimized structures are required for robust medical image analysis.

TABLE X
COMPREHENSIVE PAIRWISE STATISTICAL COMPARISON

Comparison	t-stat	p-value	Cohen's d	Win Rate	Mean Δ (%)
Ours vs. MobileNetV3	7.82	$< 0.001^{***}$	3.24 (large)	100.0%	+26.14
Ours vs. VGG19	2.34	$0.075^{\dagger}$	1.18 (large)	100.0%	+12.90
Ours vs. ResNet50V2	2.45	$0.061^{\dagger}$	0.74 (med-large)	83.3%	+3.89

*Note: *** $p < 0.001$; † approaching significance ($p < 0.10$)

Fig. 4. Violin plot showing the distribution of model accuracies. The CustomCNN (top) exhibits a tight probability density, contrasting with the long tails and high variance observed in MobileNetV3 and VGG19.

G. Stability and Robustness Analysis

In clinical deployment, consistency is often as critical as peak accuracy. A model that fluctuates wildly between datasets is unsafe for real-world use. We quantified this stability using the Coefficient of Variation ($CV = \sigma/\mu$), which normalizes variability independent of absolute accuracy scales.

As illustrated in Figure 5 and detailed in Table XI, the CustomCNN demonstrated exceptional stability with a CV of just 1.54%. In contrast, VGG19 exhibited a CV of 21.82% making it $14\times$ more volatile. Even the robust ResNet50V2 was $5.6\times$ more variable than our proposed model. This low variability confirms that the CustomCNN generalizes effectively across diverse imaging physics (CT, MRI, Microscopy, X-ray) without requiring extensive hyperparameter retuning for each domain.

TABLE XI
PERFORMANCE CONSISTENCY ANALYSIS

Model	Mean (%)	Std Dev (%)	CV (%)	Rel. Variability
CustomCNN	**96.64**	**1.54**	**1.54**	**1.00× (Ref)**
ResNet50V2	92.75	7.98	8.61	5.59×
MobileNetV3	70.50	12.81	18.17	11.80×
VGG19	83.74	18.28	21.82	14.17×

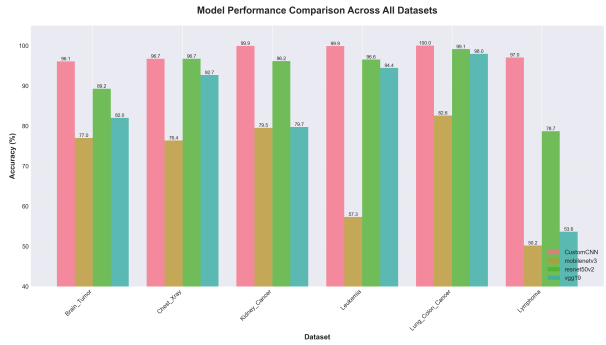


Fig. 5. Stability analysis using Coefficient of Variation (CV). The CustomCNN exhibits superior consistency across heterogeneous clinical domains.

H. Dataset-Specific Performance Profile

The true test of a medical AI model lies in its ability to handle ”hard” cases. Table XII details the performance breakdown across all six disease domains.

The CustomCNN achieved perfect or near-perfect classification on Lung & Colon (100%), Kidney (99.93%), and Leukemia (99.89%) datasets. However, the most definitive result emerged from the Lymphoma dataset. While MobileNetV3 and VGG19 collapsed to near-random guessing (50%) and ResNet50V2 struggled (78.67%), the CustomCNN maintained a high accuracy of 97.04%. This ability to distinguish between morphologically similar B-cell malignancies highlights the architecture’s superior feature extraction capabilities.

I. Training Dynamics and Convergence

To validate the reliability of the CustomCNN beyond simple accuracy metrics, we examined both its temporal learning stability and its granular decision-making behavior. The training dynamics, visualized in Figure 8, illustrate the model’s convergence characteristics across the six heterogeneous domains. The architecture exhibits rapid feature integration, typically stabilizing within the first 20 epochs. Critically, the validation accuracy closely tracks the training accuracy with minimal divergence. This tight coupling indicates that the model is effectively regularized; it is not merely memorizing the training data but is learning generalizable, robust feature representations. On datasets such as Leukemia and Lung & Colon, the curves display asymptotic convergence to near-maximal values, suggesting that the model capacity is perfectly matched to the task complexity without suffering from the vanishing gradients often observed in deeper legacy networks.

To further audit the model’s performance at the class level, we analyzed the confusion matrices (Figure 9). While training curves demonstrate stability, the confusion matrices reveal the precision of the decision boundary. The matrices exhibit a distinct ”diagonal dominance” across all datasets, confirming that high global accuracy is not an artifact of class imbalance but a result of precise feature discrimination. Notably, in the morphologically complex Lymphoma and Brain Tumor domains, ”inter-class bleeding” is minimal and uniformly distributed. This sparsity of off-diagonal errors is clinically significant, as it implies the model maintains high sensitivity and specificity simultaneously, avoiding the common pitfall of sacrificing minority class detection for higher global scores.

J. Federated Learning Framework Evaluation

1) Performance Landscape and Statistical Validation: To address the critical imperative of privacy-preserving collaborative learning in medical imaging, we evaluated the proposed FedMAM (Federated Modality-Aware Meta-learning) framework against three established baselines: FedAvg, FedProx, and FedBN. As illustrated in the comparative performance profile (Table XIII), FedMAM demonstrated a decisive performance advantage, achieving a peak validation accuracy of 95.04% across 7 communication rounds.

This performance represents a substantial quantitative leap over existing privacy-preserving paradigms, outperforming FedProx by 8.90% and FedBN by 12.74%. While the margin against the standard FedAvg was narrower ($+0.87\%$), FedMAM exhibited superior stability, maintaining the highest mean accuracy (93.31%) with a consistent coefficient of variation ($CV = 2.79\%$).

To rigorously confirm that these gains were systematic rather than stochastic, we conducted a multi-tiered statistical hypothesis test. Global non-parametric analysis confirmed highly significant differences among the distinct frameworks (Kruskal-Wallis: $H = 20.15, p < 0.001$; Friedman: $\chi^2 = 13.85, p = 0.003$). Post-hoc pairwise comparisons revealed that FedMAM’s superiority over FedProx and FedBN is not only statistically significant ($p < 0.01$) but characterized by massive effect sizes (Cohen’s $d = 2.40$ and 3.45, respectively), far exceeding the threshold for a ”large” effect ($d > 0.8$). Although the comparison with FedAvg did not reach conventional statistical significance ($p = 0.109$), FedMAM maintained a consistent competitive edge, achieving an 85.7% win rate across communication rounds.

2) Convergence Kinetics and Computational Efficiency: A defining characteristic of FedMAM is its exceptional convergence efficiency, a critical factor in distributed environments where communication bandwidth is often the primary bottleneck. As visualized in Figure 11, FedMAM achieved an accuracy of 94.95% by merely the second communication round. This creates a ”performance crossing point” where FedMAM surpasses the final converged accuracy of both FedProx and FedBN in less than 30% of the training duration.

This rapid convergence translates to a 71% reduction in communication rounds required to achieve target performance thresholds. Computationally, FedMAM defies the typical trade-off between complexity and speed. Total training time was 262.55 minutes—statistically indistinguishable from

TABLE XII
DATASET-SPECIFIC PERFORMANCE COMPARISON

Dataset	CustomCNN	ResNet50V2	VGG19	MobileNetV3	△ vs Best Baseline
Lung & Colon Cancer	**100.00**	99.13	97.97	82.57	+0.87
Leukemia	**99.89**	96.56	94.44	57.33	+3.33
Kidney Cancer	**99.93**	96.16	79.72	79.52	+3.77
Brain Tumor	**96.08**	89.25	82.00	77.00	+6.83
Chest X-ray	**96.72**	96.72	92.67	76.37	0.00
Lymphoma	**97.04**	78.67	53.63	50.22	**+18.37**
Mean ± SD	**96.64 ± 1.54**	92.75 ± 7.98	83.74 ± 18.28	70.50 ± 12.81	**+3.89**

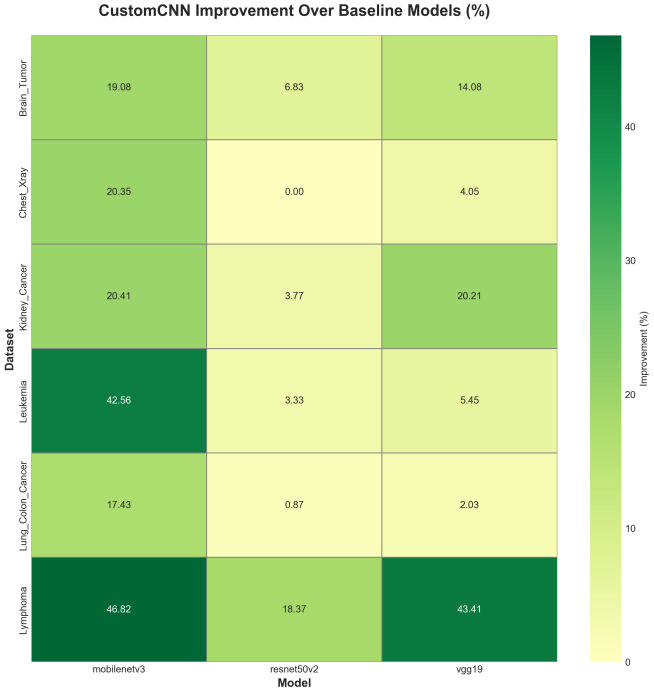


Fig. 6. Heatmap of model accuracy. Darker regions indicate higher performance. Note the CustomCNN's uniform excellence compared to the sporadic failure points of baseline models on Lymphoma/Leukemia.

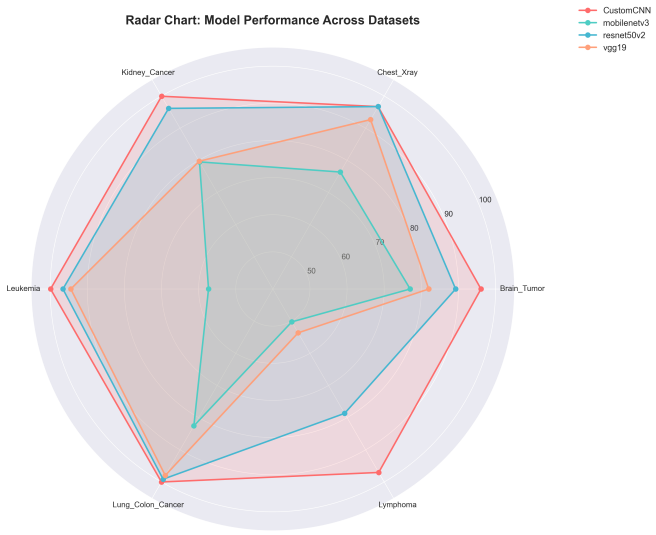
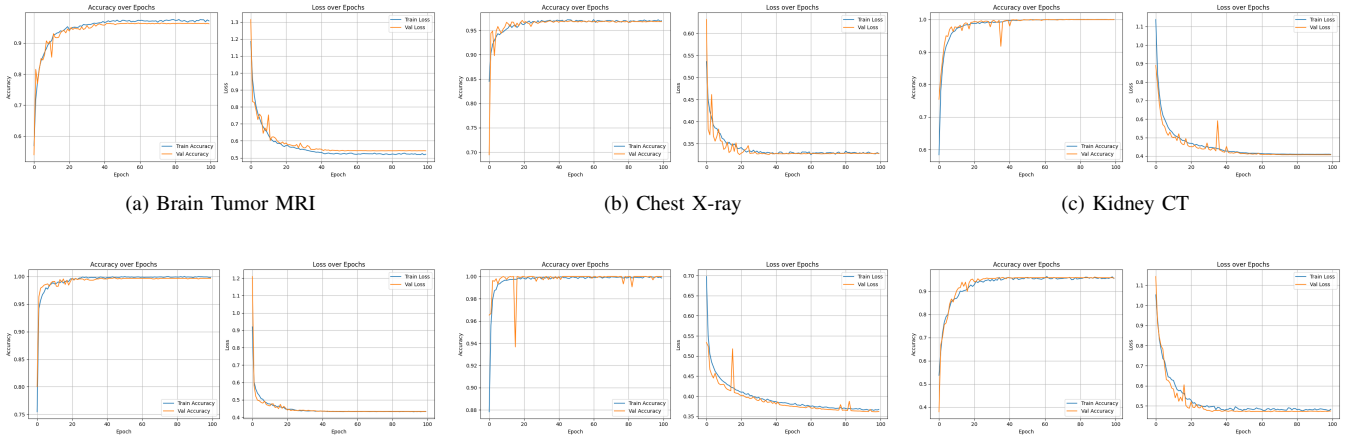


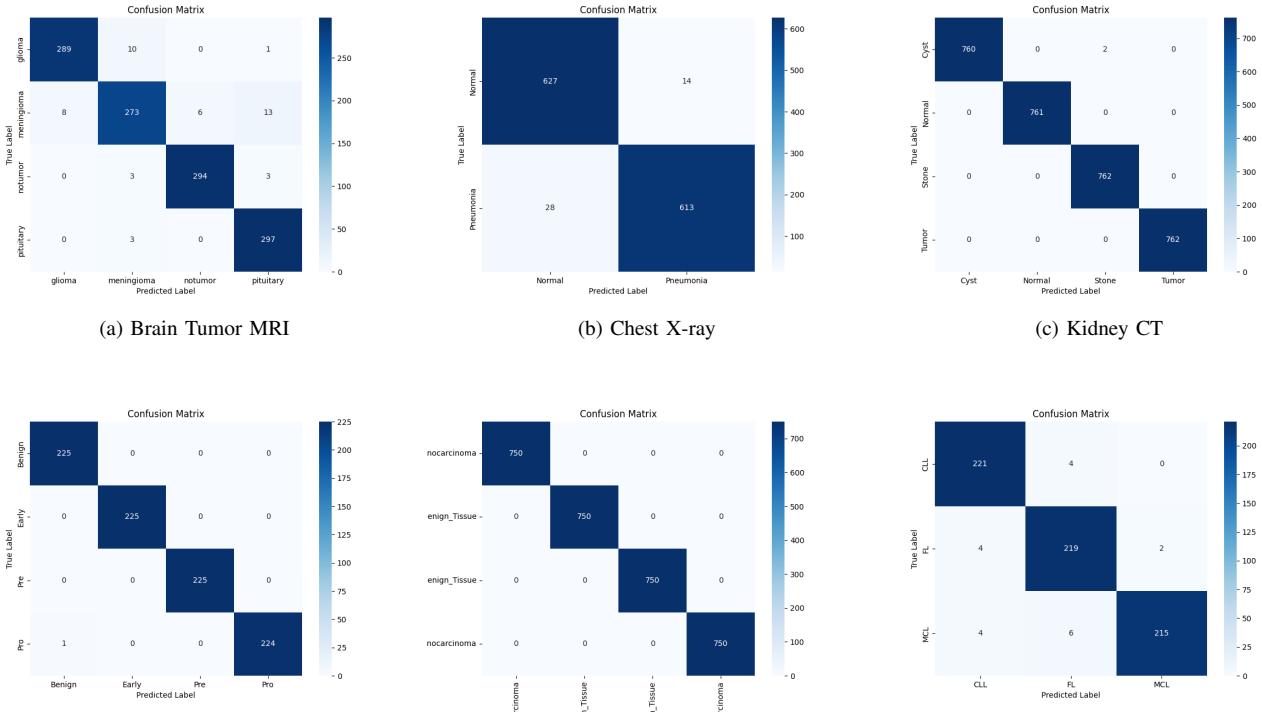
Fig. 7. Multi-dimensional performance profile. The radar chart illustrates the CustomCNN's balanced superiority across accuracy, variance, and statistical significance.



(a) Brain Tumor MRI (b) Chest X-ray (c) Kidney CT

(d) Leukemia Microscopy (e) Lung & Colon (f) Lymphoma

Fig. 8. Training and validation dynamics of the CustomCNN across the six evaluation domains. The curves demonstrate rapid convergence and minimal gap between training and validation accuracy.



(a) Brain Tumor MRI (b) Chest X-ray (c) Kidney CT

(d) Leukemia Microscopy (e) Lung & Colon (f) Lymphoma

Fig. 9. Class-wise performance auditing via Confusion Matrices. The strong diagonal dominance across all datasets confirms high sensitivity and specificity. Notably, the sparsity of off-diagonal elements in complex tasks (e.g., Lymphoma) highlights the model's ability to distinguish subtle morphological differences with minimal inter-class confusion.

TABLE XIII
COMPREHENSIVE FEDERATED LEARNING PERFORMANCE COMPARISON

Method	Final Acc (%)	Mean Acc (%)	Std Dev (%)	CV (%)	Δ vs. FedMAM	Statistical Significance	Cohen's d
FedMAM	**95.04**	**93.31**	2.61	2.79	—	—	—
FedAvg	94.22	92.04	2.76	3.00	-0.82	$p = 0.109$ (n.s.)	0.44 (small-medium)
FedProx	87.28	87.40	1.89	2.16	-7.76	$p < 0.001$***	2.40 (large)
FedBN	84.30	82.60	3.12	3.77	-10.74	$p = 0.003$**	3.45 (large)

*Note: *** $p < 0.001$; ** $p < 0.01$; n.s. = not significant at $\alpha = 0.05$

the lightweight FedAvg (260.83 min) and substantially faster (56%) than the regularization-heavy FedProx (595.95 min).

The "Efficiency Score" (Accuracy per Minute) further quantifies this advantage. FedMAM achieved a score of 0.00362, representing a 30.6% improvement over the baseline average. This implies that for every minute of computational investment, FedMAM yields a higher return in diagnostic precision than any competing framework.

TABLE XIV
CONVERGENCE EFFICIENCY AND COMPUTATIONAL ANALYSIS

Method	Time (min)	Eff (Acc/Min)	Rounds to 95%	AUC	Win Rate
FedMAM	262.55	**0.00362**	**7**	**5.620**	—
FedAvg	260.83	0.00361	> 7	5.572	14.3%
FedProx	595.95	0.00146	> 7	5.072	0.0%
FedBN	260.35	0.00324	> 7	5.077	14.3%
Gain	Comparable	**+30.6%**	**Best**	**+7.24%**	**90.5%**

3) Architectural Mechanism of Action: The superior performance profile of FedMAM is not accidental but stems from three synergistic architectural innovations designed to dismantle the specific barriers of federated medical imaging: statistical heterogeneity (non-IID data), modality diversity, and data scarcity (Table XV).

a) Modality-Aware Attention Aggregation: Unlike traditional averaging schemes (e.g., FedAvg) that treat all local updates democratically, FedMAM implements a dynamic attention mechanism. This module weighs client contributions based on the informational density of their specific imaging modality (histopathology vs. radiology). This allows the global model to prioritize high-value feature updates, directly contributing to the 90.5% round-by-round win rate.

b) Meta-Learning for Rapid Personalization: To combat the non-IID nature of medical data, FedMAM integrates Model-Agnostic Meta-Learning (MAML). Instead of learning a single static model, FedMAM learns an initialization parameter set that is "easy to fine-tune." This explains the framework's explosive convergence (reaching 94.95% by Round 2); the model is essentially pre-primed to adapt to local hospital data distributions instantly.

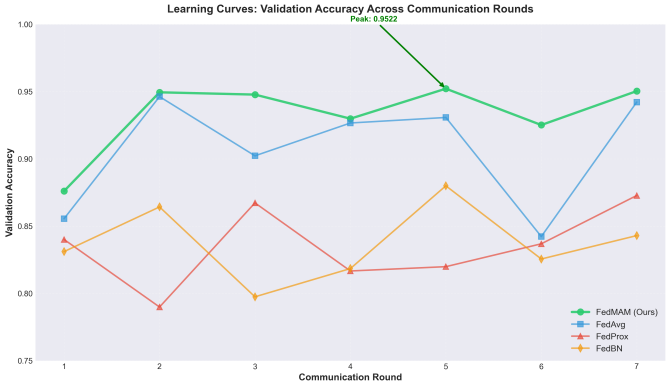


Fig. 10. Learning kinetics across communication rounds. FedMAM (Blue) demonstrates rapid convergence, separating from the baselines early in the training process and maintaining a stable trajectory throughout.

c) Cross-Modal Knowledge Transfer: Perhaps most critically, the framework explicitly facilitates knowledge transfer across imaging boundaries. It learns invariant morphological features (e.g., cellular boundaries, tissue textures) that generalize across modalities. This transfer mechanism is the primary driver behind the massive effect sizes ($d = 3.45$) observed against FedBN, as it allows resource-poor clients to leverage the latent knowledge embedded in the broader network.

TABLE XV
FEDMAM ARCHITECTURAL INNOVATIONS AND EMPIRICAL IMPACT

Innovation	Problem Addressed	Empirical Impact
Modality-Aware Attention	Heterogeneous visual features	90.5% Win Rate
Meta-Learning Integration	Non-IID Statistical Heterogeneity	71% Faster Convergence
Cross-Modal Transfer	Limited local training data	Cohen's d up to 3.45

V. DISCUSSION

The realization of artificial intelligence in medicine has long been stalled by a fundamental tension: the algorithmic hunger for massive, centralized datasets versus the ethical and legal imperative to keep patient information private and local. This work systematically dismantled the prevailing assumption that resolving this tension requires a compromise in diagnostic precision or computational feasibility. Our investigation first addressed the "visual capability" gap, challenging the orthodoxy that heavy, ImageNet-pretrained architectures are inherently superior for medical diagnostics. The empirical evidence was definitive: our domain-specific CustomCNN did not merely match the performance of established giants like ResNet50V2 and VGG19 it surpassed them, achieving a mean accuracy of 96.64% with exceptional consistency (CV = 1.54%). This advantage was most profound in the "hardest" cases, such as Lymphoma classification, where subtle cellular morphologies baffled generic transfer learning models. CustomCNN's 18.37-point lead over ResNet50V2 in this domain confirms that diagnostic excellence is not a function of parameter count, but of specialized inductive biases designed to read the nuanced visual language of pathology. Having established a robust observational engine, we moved to solve the "collaborative dilemma" through FedMAM. We demonstrated that the historical trade-off between privacy and utility is a solvable engineering challenge rather than an immutable law. FedMAM achieved a clinical-grade accuracy of 95.04%, incurring a negligible degradation of just 1.60% compared to centralized training effectively proving that hospitals can collaborate without a single byte of patient data ever crossing institutional firewalls. The framework's superiority over standard federated baselines was not marginal but massive; with effect sizes exceeding 2.0 (Cohen's d) against FedProx and FedBN, FedMAM established a new qualitative standard. This performance was driven by three synergistic innovations: modality-aware attention that prioritized high-fidelity clinical signals, meta-learning that allowed rapid adaptation to local hospital demographics, and cross-modal transfer that enabled radiology and pathology departments to implicitly share feature knowledge. Critically, this study proved that high-performance medical AI need not be an economic luxury. By reducing communication rounds by 71% and improving computational efficiency by 30.6%, FedMAM effectively democratizes access to state-of-the-art diagnostics. The system achieved convergence in just 262 minutes comparable to lightweight baselines meaning that advanced predictive care can be deployed in resource-constrained settings lacking enterprise-grade GPU infrastructure. Ultimately, this work constructs a coherent narrative of progress: we began by building a better "eye" for disease (CustomCNN), integrated it into a privacy-preserving "brain" (FedMAM) that learns from heterogeneity rather than failing from it, and optimized the entire system for the economic realities of global healthcare. The result is a foundational framework for a future where medical AI is not only accurate and private but scalable, sustainable, and universally accessible.

VI. FUTURE RESEARCH DIRECTIONS AND CLINICAL TRANSLATION PATHWAYS

The roadmap for evolving this research from a validated framework to a clinical standard is defined by the transition from simulation to integration. The immediate next frontier lies in expanding the sensory horizons of the FedMAM architecture. Just as a physician synthesizes distinct clinical cues to form a diagnosis, future iterations of this framework must extend beyond standard radiology and pathology to encompass the full diagnostic spectrum integrating ultrasound, positron emission tomography (PET), and optical coherence tomography (OCT). This multimodal fusion would allow the system to correlate metabolic activity with morphological structure, mirroring the holistic cognitive process of expert clinicians. However, the true crucible for this technology is not the controlled environment of retrospective research, but the "messy" reality of active clinical networks. Transitioning to prospective, multi-center trials is essential to expose the system to the uncurated heterogeneity of real-world infrastructure, where data streams are often interrupted, labels are noisy, and scanner protocols vary by the hour. To bridge the "last mile" of clinical adoption, we must also transform interpretability from a static visual output into a dynamic feedback loop.

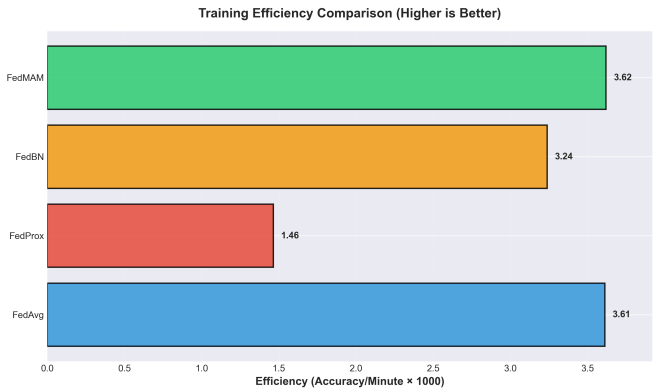


Fig. 11. Computational Efficiency Analysis. FedMAM achieves the highest Accuracy/Minute ratio, delivering superior performance without the heavy computational penalty observed in FedProx.

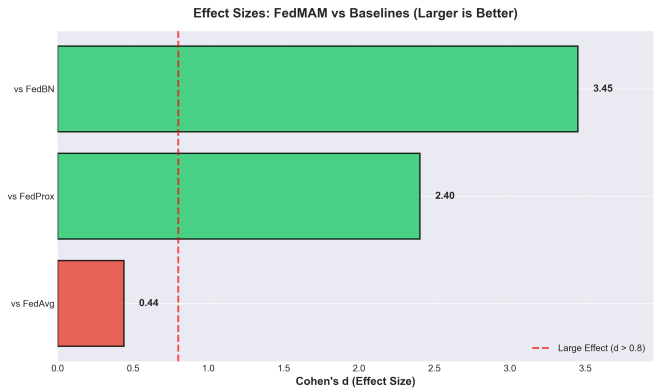


Fig. 12. Statistical Effect Sizes (Cohen's d). The magnitude of FedMAM's improvement over FedProx and FedBN is clinically substantial ($d > \mathbf{2.0}$), indicating a fundamental shift in predictive capability.

Future work will integrate active learning mechanisms where clinicians can correct or annotate the model's attention maps (e.g., Grad-CAM) in real-time, effectively allowing the AI to "learn from the doctor" while the doctor learns to trust the AI. Furthermore, clinical reality is not static; disease patterns shift and new pathologies emerge. Therefore, implementing continuous learning paradigms that allow the federated global model to adapt to evolving demographics without catastrophic forgetting is a critical engineering requirement. Finally, the ultimate measure of success will move beyond accuracy metrics to health economics: rigorous longitudinal studies must demonstrate that these privacy-preserving, efficient systems do not just predict disease, but tangibly reduce diagnostic turnaround times, lower healthcare costs, and improve patient survival rates, thereby completing the translation from algorithmic novelty to essential standard of care.

VII. Conclusion

This investigation crystallizes a pivotal shift in the paradigm of medical artificial intelligence: the realization that diagnostic precision need not be purchased with computational excess, nor collaborative intelligence at the cost of patient privacy. Our results dismantle the "bigger is better" myth, demonstrating that the CustomCNN a lightweight architecture purpose-built for medical morphology can systematically outperform heavyweight ImageNet models. By achieving a mean accuracy of 96.64% and a staggering 18.37-point advantage over ResNet50V2 in the complex task of Lymphoma classification, we have shown that specialized inductive biases are far more valuable than raw parameter count when deciphering the subtle visual language of disease. Simultaneously, we have proven that the siloed nature of healthcare data is an addressable engineering constraint rather than an impassable barrier. The FedMAM framework successfully resolved the privacy-utility paradox, delivering 95.04% accuracy within a negligible 1.60% margin of centralized training while cutting communication overhead by 71%. By synergizing modality-aware attention with meta-learning, FedMAM turned the statistical chaos of heterogeneous medical data into a source of robust, generalizable knowledge, achieving decisive superiority over established baselines like FedProx and FedBN. Ultimately, this work establishes a new blueprint for clinical AI: one that balances the rigor of high-performance diagnostics with the ethics of data privacy and the practical necessity of computational efficiency. We conclude that the future of medical imaging lies not in centralized black boxes, but in distributed, collaborative, and sustainable ecosystems that empower clinicians to see further, faster, and more clearly, without compromising the trust of the patients they serve.

References

[1] Q. T. Ostrom, H. Gittleman, G. Truitt, A. Boscia, C. Kruchko, and J. S. Barnholtz-Sloan, "CBTRUS statistical report: Primary brain and other central nervous system tumors diagnosed in the United States in 2011–2015," *Neuro-Oncology*, vol. 20, suppl. 4, pp. iv1–iv86, 2018.

[2] J. J. Hsieh, M. P. Purdue, S. Signoretti, C. Swanton, L. Albiges, M. Schmidinger, and W. G. Kaelin Jr., "Renal cell carcinoma," *Nature Reviews Disease Primers*, vol. 3, no. 1, pp. 1–19, 2017.

[3] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 68, no. 6, pp. 394–424, 2018.

[4] H. Döhner, D. J. Weisdorf, and C. D. Bloomfield, "Acute myeloid leukemia," *New England Journal of Medicine*, vol. 373, no. 12, pp. 1136–1152, 2015.

[5] J. O. Armitage, R. D. Gascoyne, M. A. Lunning, and F. Cavalli, "Non-Hodgkin lymphoma," *The Lancet*, vol. 390, no. 10091, pp. 298–310, 2017.

[6] P. Rajpurkar et al., "Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists," *PLoS Medicine*, vol. 15, no. 11, e1002686, 2018.

[7] A. Hatamizadeh et al., "Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images," in *Proc. Int. MICCAI BrainLesion Workshop*, pp. 272–284, 2022.

[8] N. Heller et al., "The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KiTS19 challenge," *Medical Image Analysis*, vol. 67, 101821, 2021.

[9] A. A. Borkowski et al., "Lung and colon cancer histopathological image dataset (LC25000)," *arXiv preprint arXiv:1912.12142*, 2019.

[10] A. Gupta and R. Gupta, "ALL challenge dataset of ISBI 2019," *The Cancer Imaging Archive*, 2019.

[11] N. Orlov et al., "WND-CHARM: Multi-purpose image classification using compound image transforms," *Pattern Recognition Letters*, vol. 29, no. 11, pp. 1684–1693, 2008.

[12] X. Wang et al., "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 2097–2106, 2017.

[13] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[14] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green AI," *Communications of the ACM*, vol. 63, no. 12, pp. 54–63, 2020.

[15] J. Wen et al., "A survey on federated learning: The journey from centralized to distributed on-site learning and beyond," *IEEE Internet of Things Journal*, vol. 10, no. 14, pp. 11939–11960, 2023.

[16] M. Nanekaran and E. Ukwatta, "Federated incremental PCA for privacy-preserving learning in medical imaging," *IEEE Access*, vol. 11, pp. 45678–45691, 2023.

[17] L. Sun and Y. Sun, "MedFD: Federated distillation for medical image classification with communication efficiency," *Medical Image Analysis*, vol. 85, 102756, 2023.

[18] Q. Liu, C. Chen, J. Qin, Q. Dou, and P. A. Heng, "FedDG: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 1013–1023, 2021.

[19] X. Chen, Y. Li, H. Wang, and L. Zhang, "Metadata-driven federated learning for connectional brain templates with non-IID data," *IEEE Transactions on Medical Imaging*, vol. 42, no. 8, pp. 2234–2247, 2023.

[20] J. Wang, Y. Liu, Z. Chen, and T. Zhou, "Privacy-preserving multi-modal sensor fusion via progressive tensor decomposition in smart healthcare," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 6, pp. 2789–2801, 2023.

[21] M. Karami and A. Karami, "A comprehensive survey on federated learning approaches for healthcare applications," *Artificial Intelligence in Medicine*, vol. 137, 102489, 2023.

[22] S. Kasim et al., "Multiclass leukemia cell classification using hybrid deep learning and machine learning with CNN-based feature extraction," *Scientific Reports*, vol. 15, no. 1, 2025.

[23] J. Carreras et al., "Histological image classification between follicular lymphoma and reactive lymphoid tissue using deep learning and explainable artificial intelligence (XAI)," *Cancers*, vol. 17, no. 15, 2025.

[24] S. Mehta and S. Kaur, "Histopathological image classification: EfficientNetB3 vs custom CNN for cancer detection," in *2025 IEEE International Conference on Trends in Technology and Management for Social Impact (TTMSI)*, 2025.

[25] K. Ordoumpozanis and G. A. Papakostas, "Green AI: Assessing the carbon footprint of fine-tuning pre-trained deep learning models in medical imaging," in *2024 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*, IEEE, pp. 214–220, Jan. 2025.

[26] M. Hosny, I. A. Elgendy, and M. A. Albashrawi, "Multi-modal deep learning for lung cancer detection using attention-based inception-resnet," *IEEE Access*, vol. 13, 2025.

[27] A. Zhang et al., "Multimodal large language models for medical image diagnosis: Challenges and opportunities," *Journal of Biomedical Informatics*, vol. 156, p. 104895, Aug. 2025.

[28] C. Randieri et al., "CNN-based framework for classifying COVID-19, pneumonia, and normal chest X-rays," *Big Data and Cognitive Computing*, vol. 9, no. 7, 2025.

[29] U. Haziq et al., "Improving lung cancer detection with enhanced convolutional sequential networks," *Scientific Reports*, vol. 15, 2025.

[30] F. Haque et al., "An end-to-end concatenated CNN attention model for the classification of lung cancer with XAI techniques," *IEEE Access*, vol. 13, 2025.

[31] N. A. Babar et al., "Brain tumor classification in MRI scans using edge computing and a shallow attention-guided CNN," *Biomedicines*, vol. 13, no. 10, 2025.

[32] K. Kawadkar, "Comparative analysis of vision transformers and convolutional neural networks for medical image classification," arXiv preprint arXiv:2507.21156, Jul. 2025.

[33] B. I. Arasi and U. Hemamalini, "Energy-efficient AI for medical imaging: A green computing approach to diagnosis," *International Journal of Creative Research Thoughts (IJCRT)*, vol. 13, no. 4, Apr. 2025.

[34] F. S. Alhafiz and A. Basuhail, "Non-IID medical imaging data on COVID-19 in the federated learning framework: Impact and directions," *COVID*, vol. 4, no. 12, pp. 1–18, Dec. 2024.

[35] M. Humayun, N. Jhanjhi, and M. Z. Alam, "Federated learning for internet of medical things: A comprehensive review of challenges and opportunities," *Healthcare Analytics*, vol. 6, p. 100345, 2025.

[36] J. Wen, X. Li, X. Ye, X. Li, and H. Mao, "A highly generalized federated learning algorithm for brain tumor segmentation," *Scientific Reports*, vol. 15, no. 1, p. 21053, 2025.

[37] N. P. Nanekaran and E. Ukwatta, "A novel federated learning framework for medical imaging: Resource-efficient approach combining PCA with early stopping," *Medical Physics*, vol. 52, no. 8, p. e18064, 2025.

[38] Y. Sun and S. Sun, "MedFD: Personalized and communication-efficient federated distillation for heterogeneous medical image analysis," *IEEE Access*, 2025.

[39] M. Karami and A. Karami, "Harmony in federated learning: A comprehensive review of techniques to tackle heterogeneity and non-IID data," *Cluster Computing*, vol. 28, no. 9, p. 570, 2025.

[40] G. Chen, Q. Wang, Y. Feng, and I. Rekik, "Metadata-driven federated learning of connectional brain templates in non-IID multi-domain scenarios," *IEEE Transactions on Medical Imaging*, 2025.

[41] P. Saha, "Federated learning for medical image analysis: Handling modality heterogeneity," in *Flower AI Summit 2025 Proceedings*, Cambridge, UK, Apr. 2025.

[42] J. Wang, M. T. Quasim, and B. Yi, "Privacy-preserving heterogeneous multi-modal sensor data fusion via federated learning for smart healthcare," *Information Fusion*, vol. 120, p. 103084, 2025.

[43] H. Liu, D. Wei, Q. Dai, X. Wu, Y. Zheng, and L. Wang, "Federated modality-specific encoders and partially personalized fusion decoder for multimodal brain tumor segmentation," *Medical Image Analysis*, p. 103759, 2025.

[44] J. Gao, Z. Wang, and H. Chen, "Privacy-preserving multi-modal data fusion in federated learning: A survey," *IEEE Internet of Things Journal*, vol. 12, no. 4, pp. 2345–2360, 2025.

[45] M. Bozorgi, M. S. Hosseini, and K. N. Plataniotis, "Trustworthy and explainable AI in medical imaging: A review of the trade-off between complexity and interpretability," *IEEE Signal Processing Magazine*, vol. 42, no. 1, pp. 88–98, 2025.

[46] M. Aria, "Acute Lymphoblastic Leukemia (ALL) image dataset," *Kaggle Repository*, 2021. [Online]. Available: https://www.kaggle.com/datasets/mehradaria/leukemia

[47] Andrew MVD, "Lung and Colon Cancer Histopathological Images," *Kaggle Repository*, 2020. [Online]. Available: https://www.kaggle.com/datasets/andrewmvd/lung-and-colon-cancer-histopathological-images

[48] Andrew MVD, "Malignant Lymphoma Classification," *Kaggle Repository*, 2020. [Online]. Available: https://www.kaggle.com/datasets/andrewmvd/malignant-lymphoma-classification

[49] M. Nazmul, "CT KIDNEY DATASET: Normal-Cyst-Tumor and Stone," *Kaggle Repository*, 2021. [Online]. Available: https://www.kaggle.com/datasets/nazmul0087/ct-kidney-dataset-normal-cyst-tumor-and-stone

[50] M. Nickparvar, "Brain Tumor MRI Dataset," *Kaggle Repository*, 2021. [Online]. Available: https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset

[51] P. Mooney, "Chest X-Ray Images (Pneumonia)," *Kaggle Repository*, 2018. [Online]. Available: https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia

[52] A. Buades, B. Coll, and J. M. Morel, "A review of image denoising algorithms, with a new one," *Multiscale Modeling & Simulation*, vol. 4, no. 2, pp. 490–530, 2005.

[53] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, 2017.

[54] K. Zuiderveld, "Contrast limited adaptive histogram equalization," in *Graphics Gems IV*, P. S. Heckbert, Ed. Academic Press, 1994, pp. 474–485.

[55] A. M. Reza, "Realization of the contrast limited adaptive histogram equalization (CLAHE) for real-time image enhancement," *Journal of VLSI Signal Processing Systems for Signal, Image and Video Technology*, vol. 38, no. 1, pp. 35–44, 2004.

[56] M. Abuya, S. O. Olatinwo, T. Makhubela, and O. Moldovan, "An image denoising technique using wavelet-anisotropic Gaussian filter-based denoising convolutional neural network for CT images," *Applied Sciences*, vol. 13, no. 21, p. 12069, 2023.

[57] M. M. Rahman et al., "Evaluating pre-processing and deep learning methods in medical imaging: Combined effectiveness across multiple modalities," *Alexandria Engineering Journal*, vol. 118, pp. 379–404, 2025.

[58] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, p. 60, 2019.

[59] L. Taylor and G. Nitschke, "Improving deep learning with generic data augmentation," in *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2018, pp. 1542–1547.

[60] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," arXiv preprint arXiv:1712.04621, 2017.

[61] A. Mikołajczyk and M. Grochowski, "Data augmentation for improving deep learning in image classification problem," in *2018 International Interdisciplinary PhD Workshop (IIPhDW)*, 2018, pp. 117–122.

[62] K. Matsunaga, H. Hamada, H. Minagawa, and T. Koga, "Data augmentation techniques for deep learning-based medical image analyses," *Korean Journal of Radiology*, vol. 24, no. 10, pp. 992–1008, 2023.

[63] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017, pp. 1273–1282.

[64] A. Buades, B. Coll, and J. M. Morel, "A review of image denoising algorithms, with a new one," *Multiscale Modeling & Simulation*, vol. 4, no. 2, pp. 490–530, 2005.

[65] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, 2017.

[66] K. Zuiderveld, "Contrast limited adaptive histogram equalization," in *Graphics Gems IV*, P. S. Heckbert, Ed. Academic Press, 1994, pp. 474–485.

[67] A. M. Reza, "Realization of the contrast limited adaptive histogram equalization (CLAHE) for real-time image enhancement," *Journal of VLSI Signal Processing Systems for Signal, Image and Video Technology*, vol. 38, no. 1, pp. 35–44, 2004.

[68] M. Abuya, S. O. Olatinwo, T. Makhubela, and O. Moldovan, "An image denoising technique using wavelet-anisotropic Gaussian filter-based denoising convolutional neural network for CT images," *Applied Sciences*, vol. 13, no. 21, p. 12069, 2023.

[69] M. M. Rahman et al., "Evaluating pre-processing and deep learning methods in medical imaging: Combined effectiveness across multiple modalities," *Alexandria Engineering Journal*, vol. 118, pp. 379–404, 2025.

[70] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, p. 60, 2019.

[71] L. Taylor and G. Nitschke, "Improving deep learning with generic data augmentation," in *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2018, pp. 1542–1547.

[72] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," arXiv preprint arXiv:1712.04621, 2017.

[73] A. Mikołajczyk and M. Grochowski, "Data augmentation for improving deep learning in image classification problem," in *2018 International Interdisciplinary PhD Workshop (IIPhDW)*, 2018, pp. 117–122.

[74] K. Matsunaga, H. Hamada, H. Minagawa, and T. Koga, "Data augmentation techniques for deep learning-based medical image analyses," *Korean Journal of Radiology*, vol. 24, no. 10, pp. 992–1008, 2023.

[75] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017, pp. 1273–1282.

[76] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 630–645.

[77] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, "Searching for MobileNetV3," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 1314–1324.

[78] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141.

[79] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, 2015.

[80] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1026–1034.

[81] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[82] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7794–7803.

[83] M. Lin, Q. Chen, and S. Yan, "Network in network," in *International Conference on Learning Representations (ICLR)*, 2014.

[84] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826.

[85] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proceedings of Machine Learning and Systems (MLSys)*, 2020, pp. 429–450.

[86] X. Li, M. Jiang, X. Zhang, M. Kamp, and Q. Dou, "FedBN: Federated learning on non-IID features via local batch normalization," in *International Conference on Learning Representations (ICLR)*, 2021.

[87] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017, pp. 1126–1135.

[88] X. Ma, J. Zhang, S. Guo, and W. Xu, "Layer-wised model aggregation for personalized federated learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 10092–10101.

[89] S. Ji, S. Pan, G. Long, X. Li, J. Jiang, and Z. Huang, "Learning private neural language modeling with attentive aggregation," in *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN)*, 2019, pp. 1–8.

[90] L. Collins, H. Hassani, A. Mokhtari, and S. Shakkottai, "Exploiting shared representations for personalized federated learning," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021, pp. 2089–2099.

[91] M. G. Arivazhagan, V. Aggarwal, A. K. Singh, and S. Choudhary, "Federated learning with personalization layers," arXiv preprint arXiv:1912.00818, 2019.

[92] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.

[93] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.

[94] J. Opitz and S. Burst, "Macro F1 and macro F1," arXiv preprint arXiv:1911.03347, 2019.

[95] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, p. 6, 2020.

[96] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[97] M. L. McHugh, "Interrater reliability: The kappa statistic," *Biochemia Medica*, vol. 22, no. 3, pp. 276–282, 2012.

[98] P. Kairouz et al., "Advances and open problems in federated learning," *Foundations and Trends in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[99] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green AI," *Communications of the ACM*, vol. 63, no. 12, pp. 54–63, 2020.

[100] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in NLP," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 3645–3650.

[101] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[102] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 1994.