scientific reports

Check for updates

OPEN

A compact and interpretable multi-source framework for heterogeneous medical image classification

Williams Ayivi[1,4]✉, Xiaoling Zhang[1]✉, Wisdom Xornam Ativi[2], Francis Sam[3] & Amil Aligayev[5,6]

Deep learning models for medical image analysis often rely on large-scale parameterization, which may limit their practical use in resource-constrained settings. This study aims to design a structurally compact multi-source framework capable of delivering competitive diagnostic performance with reduced computational overhead. We propose ML-ConvNet, a lightweight architecture comprising approximately 4.2 K parameters and 924 M FLOPs at 512 × 512 input resolution. The network incorporates Multi-Branch Re-parameterized Convolutions for scale-aware feature extraction, Hierarchical Dual-Path Attention for feature localization, Feature Self- Transformation for cross-feature interaction, and a Local Variance Weighted optimization strategy to address class imbalance. The framework is evaluated independently on three publicly available benchmark datasets representing heterogeneous imaging modalities: brain MRI, lung CT, and chest X-ray. Ablation studies, precision-recall analysis, cross-modality validation, and computational benchmarking are conducted to assess performance, stability, and efficiency under controlled experimental conditions. Within the evaluated settings, results indicate competitive diagnostic accuracy relative to established lightweight baselines, including EfficientNet and MobileNet variants, while substantially reducing parameter count. Class-wise F1-scores and PR-AUC values suggest relatively stable minority-class performance under repeated cross-validation sampling. Attention visualizations show activations concentrated over regions broadly associated with pathological findings, though these observations are qualitative in nature. Inference latency measurements on CPU and mobile hardware suggest feasibility for low-latency deployment under the tested single-image batch configurations, though real-world throughput may differ depending on hardware and operational conditions. These findings suggest that careful architectural design and domain-informed inductive biases may support competitive medical image classification on public benchmark datasets without extensive parameter scaling. The framework was evaluated exclusively under controlled conditions on publicly available data, and multi-institutional external validation is required before conclusions regarding generalizability or clinical applicability can be drawn.

Keywords Lightweight convolutional neural networks, Multi-source heterogeneous classification, Brain tumor MRI, Lung nodule CT, Chest X-ray, HDPA, MBRConv

The integration of artificial intelligence (AI) into clinical workflows is undergoing a critical transition, shifting from high-capacity experimental models toward resource-efficient systems capable of operating under the practical constraints of real-world healthcare environments[1]. While medical image classification remains a foundational component of computer-aided diagnosis, its translational adoption is frequently constrained by

[1]School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China. [2]School of Computer Science and Technology, University of Electronic Science and Technology of China, Chengdu 611731, China. [3]School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China. [4]Department of Information Systems and Operations Management, Vienna University of Economics and Business, 1020 Vienna, Austria. [5]NOMATEN CoE, National Centre for Nuclear Research, 05-400 Otwock, Poland. [6]Scientific Research Center, Baku Engineering University, AZ0101 Baku, Azerbaijan. ✉email: w.ayivi@std.uestc.edu.cn; xlzhang@uestc.edu.cn

three persistent challenges identified in recent comprehensive reviews: the extreme heterogeneity of data across imaging modalities[2], the stringent computational budgets of point-of-care and edge devices[1], and the pervasive difficulty of learning from imbalanced and noisy medical datasets[3,4]. Although multimodal AI research has expanded considerably[2], frameworks capable of robustly handling unpaired and heterogeneous inputs, such as MRI, CT, and X-ray, without relying on computationally intensive backbones remain limited.

Recent efforts to address these constraints have followed two principal directions: efficiency-oriented architectural design and improved optimization strategies for imbalanced data. On the algorithmic side, cost-sensitive learning and specialized preprocessing techniques have been recognized as essential mechanisms for mitigating bias in long-tailed medical datasets[3,4]. Concurrently, lightweight architectures such as LiteFusionNet[5], Light-MRI[6], MBINet[7], LMFRNet[8], and MobileDenseNeXt[9] demonstrate that compact networks can approach the accuracy of heavier backbones. Similarly, Hammad et al.[10] reported real-time tumor detection using an eight-layer CNN, while Liu et al.[11] and Salehin et al.[12] emphasized the benefit of attention-enhanced convolutional pipelines.

Despite these advancements, several limitations persist. Many lightweight solutions still operate with parameter counts in the millions, limiting deployment on ultra-low-power hardware. Furthermore, most approaches remain modality-specific, restricting generalization across heterogeneous imaging domains, as highlighted by Chen et al.[2]. In addition, while class imbalance is widely acknowledged[3,4], existing cost-sensitive strategies typically apply global weighting schemes that do not account for local feature-space inconsistencies, potentially amplifying label noise in safety-critical settings. Consequently, there remains a need for a unified framework that simultaneously achieves extreme computational efficiency, stable discriminative refinement, and principled imbalance-aware optimization.

The core architecture of ML-ConvNet is inspired by the MobileIE framework[13], an extremely lightweight model originally optimized for real-time image enhancement. In this study, we hypothesize that MBRConv and HDPA mechanisms developed for restoring low-level visual signals are uniquely suited for capturing the subtle textural gradients characteristic of pathological lesions in MRI, CT, and X-ray imaging.

To address these gaps, we propose *ML-ConvNet*, an ultra-lightweight convolutional architecture operating under an extreme parameter budget (∼4K parameters). The proposed framework integrates MBRConv to decouple training-time representational capacity from inference-time efficiency, enabling significant reduction in parameter count without compromising expressive power. To enhance discriminative stability, we introduce a HDPA mechanism that jointly models global anatomical context and localized lesion saliency, improving robustness across heterogeneous imaging modalities. To further mitigate long-tail distributions and label noise, we incorporate an imbalance-aware optimization strategy through Incremental Weight Optimization (IWO) and a Local Variance Weighted (LVW) loss, which dynamically adjusts instance-level contributions based on latent neighborhood consistency.

Beyond architectural design, this work incorporates a statistically structured evaluation protocol. Performance is assessed using stratified patient-level splits, confidence intervals, prediction intervals, and tolerance interval analysis to characterize performance variability under repeated sampling within the evaluated benchmark settings. Through this design, ML-ConvNet aims to examine whether structurally parsimonious models, when equipped with domain-informed inductive biases, can achieve competitive medical image classification on public benchmark datasets without extensive parameter scaling.

The principal contributions of this study are summarized as follows:

1. *Architectural Efficiency*: We introduce *ML-ConvNet*, an ultra-lightweight convolutional architecture comprising approximately 4.2K parameters. By leveraging structural re-parameterization, the proposed design reduces parameter count by nearly three orders of magnitude relative to standard mobile backbones while maintaining competitive diagnostic performance.
2. *Attention-Based Feature Refinement*: We develop an *HDPA* mechanism that integrates global anatomical context with localized lesion-specific saliency. This dual-path formulation improves discriminative stability and enhances robustness across heterogeneous imaging modalities.
3. *Imbalance-Aware Optimization*: We propose an *LVW* loss function that dynamically adjusts instance-level weights based on latent neighborhood consistency. This strategy mitigates label noise and long-tail class imbalance, improving minority-class sensitivity without increasing inference complexity.
4. *Statistically Rigorous Evaluation*: We establish a comprehensive inferential framework incorporating stratified cross-validation, confidence intervals, prediction intervals, and tolerance interval assessment. This protocol is intended to provide a more reliable estimate of performance variability under repeated sampling than single-split evaluation and to characterize the expected dispersion of results across unseen data partitions within the evaluated benchmark settings. Whether this translates to reliable performance in safety-critical clinical environments requires further investigation through external multi-institutional validation.

Related work

This section reviews prior efforts in three interrelated domains: lightweight architectural design, cross-modality learning, and imbalance-aware optimization in medical imaging.

Lightweight architectures and efficiency metrics

The increasing demand for deployable medical AI systems has shifted research focus from maximizing raw predictive accuracy toward optimizing the efficiency–accuracy trade-off. Aklan et al.[14] provide a comprehensive taxonomy of efficiency metrics in medical imaging, distinguishing between theoretical complexity (parameter count and Floating Point Operations, FLOPs) and operational constraints such as inference latency and energy consumption. Traditional deep architectures, including VGG-16[15] and ResNet, rely on dense matrix

multiplications that incur substantial memory access costs (MAC), often exceeding the power and thermal constraints of portable clinical devices[14].

To mitigate computational overhead, contemporary lightweight designs emphasize convolutional factorization. MobileNet[16] introduced depthwise separable convolutions, decoupling spatial and channel-wise filtering to reduce computation by approximately 8–9×. ShuffleNetV2[17] further improved efficiency through channel shuffling mechanisms that enhance inter-group information flow while minimizing expensive 1×1 operations. SqueezeNet[18] employed Fire modules to compress parameter counts below 1.5 million. While effective on natural image benchmarks, these static compression strategies may introduce representational bottlenecks when applied to medical imaging tasks, where subtle high-frequency structures-such as spiculated lung nodules are diagnostically critical[14]. Recent work has shown that replacing standard global average pooling with SE-augmented second-order pooling in compact ResNet backbones recovers covariance-level feature interactions that first-order methods discard, yielding marked accuracy gains on lung CT classification without proportional increases in computational cost[19,20].

Another trajectory involves Neural Architecture Search (NAS) and compound scaling. NASNet[21] and EfficientNet[22] utilise automated search and scaling strategies to optimise network topology. Although these models achieve strong efficiency–accuracy trade-offs, their irregular computational graphs and fragmented memory operations may increase real-world latency despite reduced theoretical FLOPs. Moreover, transferring such architectures to small-scale medical datasets can lead to optimisation instability[23].

Structural re-parameterisation has recently emerged as a principled compromise between representational capacity and inference efficiency. RepVGG[24] demonstrated that multi-branch training structures (e.g., parallel 1×1, 3×3, and identity paths) can be algebraically collapsed into single-path kernels at inference time, preserving training-time expressiveness while enabling streamlined deployment. This principle has since been extended to medical imaging: Ayivi et al.[19] showed that fusing multi-branch convolutions with depthwise filters and SE attention in a four-stage residual network reduces MACs by 12.08% and cuts inference latency by $1.84\times$, while maintaining classification performance exactly after fusion—demonstrating the practical viability of re-parameterisable designs for resource-constrained clinical deployment. Unlike generic vision adaptations such as GhostNet[23], which focus on inexpensive feature generation, our work integrates structural re-parameterisation with imbalance-aware optimisation tailored specifically to the statistical characteristics of medical imaging.

Cross-modality learning and multimodal fusion paradigms

Clinical diagnosis frequently integrates information across multiple imaging modalities. Chen et al.[25] categorize multimodal fusion into three paradigms: early fusion (pixel-level integration), intermediate fusion (feature-level interaction), and late fusion (decision-level aggregation). Early fusion retains maximal information but is sensitive to spatial misalignment and domain shifts, particularly in unregistered CT and MRI volumes[25]. Late fusion improves robustness to missing inputs but may fail to model complex cross-modal dependencies, effectively treating modalities as independent predictors.

A central limitation in multimodal learning is reliance on paired datasets[26]. Many multi-view architectures require matched acquisitions from the same patient, a condition rarely satisfied in routine clinical practice. To address missing-modality scenarios, generative approaches based on GANs or Variational Autoencoders (VAEs) synthesize absent modalities. However, these methods increase computational complexity and risk generating anatomically implausible structures, which may compromise safety in diagnostic settings.

Rather than relying on paired fusion or generative synthesis, our approach emphasizes domain-aligned latent representation learning across heterogeneous inputs. By projecting modality-specific features into a shared embedding space through structured convolution and non-linear feature transformation, the framework reduces dependence on modality pairing while maintaining computational efficiency. This design prioritizes semantic consistency across imaging physics without incurring the overhead of explicit cross-modal generation.

Robustness to class imbalance and label noise

Medical datasets commonly exhibit long-tailed distributions, where pathological cases form a minority relative to normal controls[27]. Standard objectives such as Cross-Entropy (CE) loss are prevalence-biased, often prioritizing majority classes and reducing minority-class sensitivity[27]. Additionally, annotation variability among radiologists introduces label noise, further complicating optimization[28].

Mitigation strategies are generally categorized into data-level and algorithm-level approaches[27,29]. Data-level techniques, including SMOTE and GAN-based augmentation, attempt to rebalance training distributions but may introduce synthetic artifacts or distort diagnostic semantics[27]. Consequently, algorithm-level Cost-Sensitive Learning (CSL) has gained prominence[30]. CSL modifies loss functions by assigning higher penalties to minority-class errors, with examples including Focal Loss and Weighted Cross-Entropy. However, these approaches typically apply global class-frequency weights, failing to differentiate between informative hard samples and mislabeled outliers[28].

To address this limitation, we propose a LVW loss that computes instance-specific weights based on local feature-space consistency. By assessing neighborhood variance within the latent embedding space, the model attenuates isolated, potentially noisy samples while emphasizing coherent minority clusters. This localized weighting mechanism provides a finer-grained alternative to global class reweighting schemes and enhances robustness in safety-critical diagnostic contexts.

The proposed ML-ConvNet

This section delineates the architectural methodology of the proposed ML-ConvNet, a unified ultra-lightweight framework designed for heterogeneous medical image classification across independent multi-source imaging cohorts under strict computational constraints. To rigorously evaluate generalization across heterogeneous

imaging sources, the study utilizes three clinically distinct modalities drawn from independent public cohorts: MRI for brain tumor subtyping, CT for lung nodule characterization, and chest X-ray for pneumonia detection.

Unlike conventional multimodal fusion approaches that necessitate paired scans for every patient, a requirement rarely satisfied in fragmented and resource-constrained clinical environments, ML-ConvNet is engineered to learn a shared diagnostic representation from unpaired, modality-specific data streams. This design reflects realistic medical data ecosystems, where patient records often contain isolated imaging studies determined by clinical indication, institutional availability, and acquisition protocol.

The framework is architecturally constrained to maximize representational capacity within an extreme parameter budget (∼4K parameters), with inference latency measurements suggesting suitability for resource-constrained hardware under the tested configurations. As illustrated in Fig. 1, the architecture integrates five strategic inductive biases to ensure efficient, stable, and clinically interpretable learning:

- An MBRConv for efficient multi-scale feature extraction;
- An IWO for stable convergence on small datasets;
- An FST for enhancing non-linear separability;
- HDPA for interpretable lesion localization;
- LVW loss function to explicitly address label noise and class imbalance.

Together, these components form a cohesive pipeline capable of scalable, modality-agnostic learning.

Problem definition

Formally, let $\mathcal{D} = \{(x_i^m, y_i)\}_{i=1}^N$ denote a heterogeneous ensemble of medical imaging samples, where $x_i^m \in \mathcal{X}^m$ is an input image acquired under modality $m \in \{\text{MRI}, \text{CT}, \text{XR}\}$ and $y_i \in \mathcal{Y}$ is the corresponding diagnostic label. Each modality-specific subset $\mathcal{D}_m$ is drawn from an independent patient cohort, meaning the overall dataset contains disjoint subjects across modalities and admits no instance-level pairing between them. This is a fundamental departure from multimodal fusion paradigms that require matched acquisitions from the same patient[25]; no cross-modal transfer, joint embedding, or knowledge distillation is employed at any stage of training.

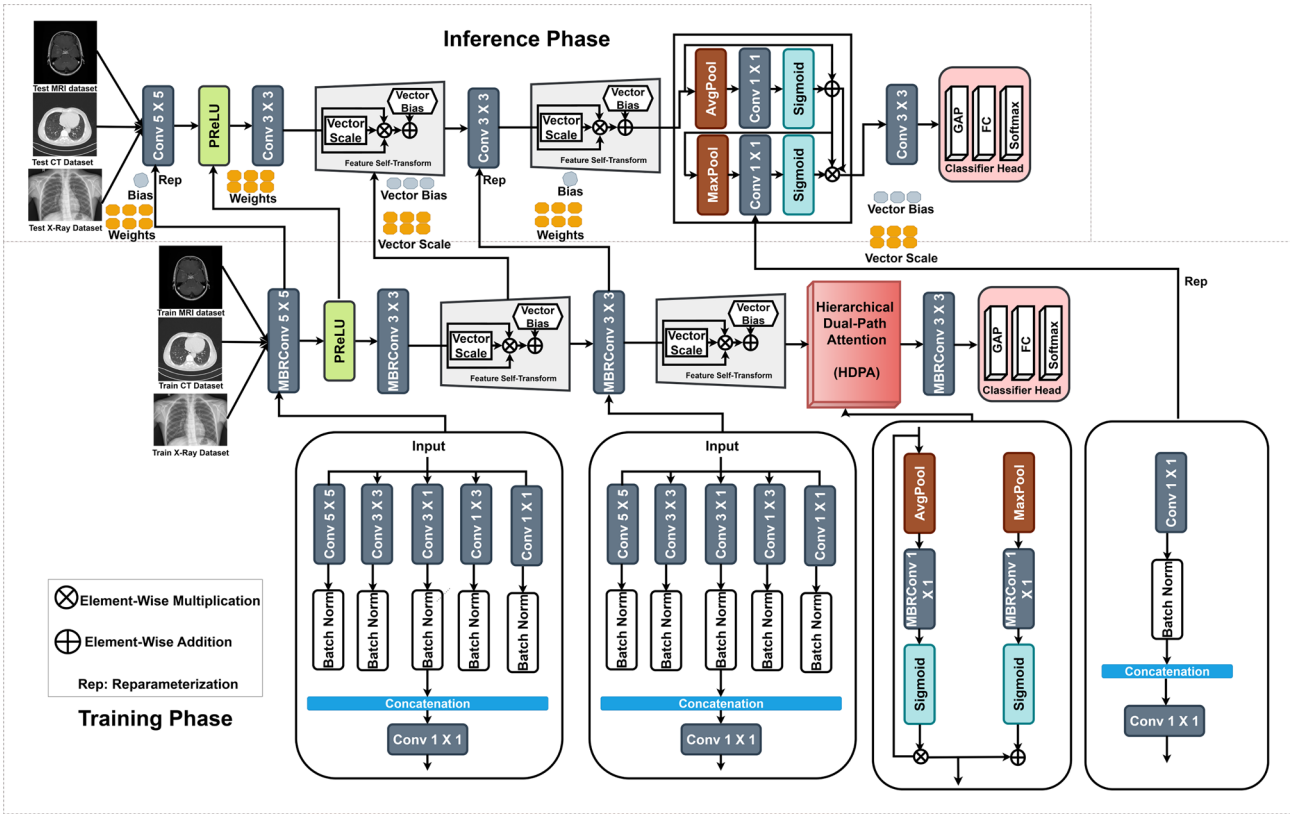

Fig. 1. Schematic illustration of the ML-ConvNet architecture. The framework operates in two logical phases: (Top) the *Training Phase* uses a high-capacity multi-branch topology with inductive biases (FST, HDPA) to learn a robust modality-agnostic latent space; (Bottom) the *Inference Phase* performs structural re-parameterization to collapse the training graph into a streamlined single-path architecture. This yields an ultra-lightweight model (∼4K parameters) with measured low-latency inference under the tested single-image batch configurations.

The task is therefore formulated as *multi-source heterogeneous classification*: constructing a single parametric model f_θ that maps the union of modality-specific input spaces to a shared diagnostic label space, generalizing across domain shifts induced by differences in imaging physics without relying on modality-specific encoders or paired supervision. The optimization objective is:

$$f_\theta : \bigcup_m \mathcal{X}^m \mapsto \mathcal{Y}, \tag{1}$$

subject to the constraint that parameter complexity $|\theta|$ remains within a strictly bounded budget suitable for edge deployment. The primary challenge this formulation poses is not modality alignment in the representational sense, but rather learning a compact feature space that is simultaneously discriminative across imaging domains whose pixel statistics, contrast mechanisms, and noise characteristics differ substantially.

Overall framework architecture

The operational workflow of ML-ConvNet is designed with a dual-phase topology that decouples high-capacity feature learning from low-latency deployment. As illustrated in the framework diagram, the architecture transitions from a multi-branch structure during optimization to a streamlined, single-path structure during inference:

- *Training Phase*: To maximize representational power, heterogeneous inputs are processed through modality-specific encoders initialized with shallow 5×5 MBRConv layers activated by Parametric Rectified Linear Units (PReLU). Feature enrichment is subsequently performed via the FST module, which enhances non-linear separability. These modality-specific representations are then aligned through the HDPA mechanism to construct a unified, modality-agnostic latent space. Deep semantic abstractions are finally captured through stacked 3×3 MBRConv layers prior to the classification head. All inputs are standardized to $512 \times 512 \times 3$ prior to network ingestion. Single-channel acquisitions, specifically grayscale MRI and CT slices, are replicated across three channels to produce a consistent three-channel tensor without introducing cross-channel information asymmetry. No modality identity label is passed as an explicit conditioning signal at any stage; the network receives only the pixel tensor and the diagnostic label y_i, and modality discrimination is handled implicitly through the shared convolutional filters. In the pooled training configuration, minibatches are constructed via round-robin sampling across the three cohorts to maintain approximate modality balance throughout optimization. Under missing-modality conditions evaluated in Sect. Validation framework and statistical reliability analysis, the absent modality's input tensor is replaced by an all-zero tensor of identical spatial dimensions, and no architectural modification is required at inference time.
- *Inference Phase*: To satisfy strict memory and latency constraints, the model undergoes structural re-parameterization. Specifically, parallel branches within all MBRConv blocks are algebraically consolidated into single equivalent convolutional kernels. This transformation produces a single-path inference kernel whose forward pass is algebraically equivalent to the multi-branch training structure, preserving the learned feature mapping while enabling real-time inference on resource-constrained hardware.

Multiple-branch re-parameterized convolution (MBRConv)

To address the scale variation inherent in heterogeneous medical modalities ranging from fine-grained pulmonary nodules in CT to diffuse tumor boundaries in MRI, the framework employs a parallel multi-scale encoding strategy. The MBRConv block operates on a split-transform-merge paradigm:

$$F^m_{\text{mb}} = \sum_{k\in\{1,3,5\}} \sigma\left(W^m_k * x^m + b^m_k\right), \tag{2}$$

where $*$ denotes convolution, W^m_k and b^m_k represent branch-specific kernels and biases, and σ is the PReLU activation. Feature aggregation is subsequently performed via a 1×1 pointwise convolution:

$$F^m_{\text{out}} = W_{1\times1}\left[F^m_1 \oplus F^m_3 \oplus F^m_5\right]. \tag{3}$$

Structural Re-parameterization for Inference:

$$W^m_{\text{rep}} = \sum_{k\in\{1,3,5\}} \alpha_k \cdot \mathcal{T}(W^m_k), \tag{4}$$

where α_k are learnable scaling factors and $\mathcal{T}(\cdot)$ denotes a zero-padding operator that maps each branch kernel to a unified 5×5 spatial dimension: a 1×1 kernel is embedded at the center of a 5×5 zero tensor, and a 3×3 kernel is symmetrically padded by one zero on each side. Batch normalization statistics accumulated during training are absorbed into the corresponding branch kernel and bias prior to summation, following the standard re-parameterization procedure of Ding et al.[24]. This consolidation is applied once after training completion and produces a single equivalent 5×5 kernel whose forward pass output is numerically identical to the multi-branch structure to within floating-point precision.

Incremental weight optimization (IWO)

Compact networks with limited parameters tend to stagnate in later training stages, where loss plateaus and further adjustment of hyperparameters yields diminishing returns. This behaviour is particularly pronounced on small medical datasets, where the optimization landscape is irregular and high-frequency noise can dominate gradient updates. To address this, IWO decomposes the convolutional weight into two components: a frozen prior W^m_{pre}, obtained from an earlier training checkpoint and held fixed, and a learnable residual W^m_{learn}, which is updated normally during subsequent training:

$$W^m_{\text{final}} = W^m_{\text{pre}} + \gamma \cdot W^m_{\text{learn}}, \tag{5}$$

where γ is a scaling factor. The frozen prior provides a stable feature representation that anchors the optimization, while the learnable residual refines task-specific detail without discarding previously acquired knowledge. This decomposition is applied to the output projection of each MBRConv block and introduces no additional parameters at inference, since the two components are summed into a single kernel before deployment. The approach is conceptually related to residual learning but operates at the weight level rather than the feature level, targeting the convergence behaviour of the optimizer rather than the representational capacity of the network.

Feature self-transformation (FST)

Standard convolutional operations are linear, meaning they model weighted sums of input features without capturing multiplicative or higher-order relationships between feature dimensions. In medical image classification, where subtle textural differences between pathological categories can be diagnostically relevant, this linearity may limit the discriminative capacity of a compact network. FST addresses this by introducing a lightweight quadratic interaction within the latent feature space. Given a feature map F, two independent linear projections W_1F and W_2F are computed and combined through element-wise multiplication, producing a second-order feature interaction without requiring additional convolutional layers:

$$F_{\text{FST}} = \sigma\left((W_1F) \odot (W_2F) + b\right), \tag{6}$$

where $\odot$ denotes the Hadamard product, b is a learnable bias applied channel-wise, and σ is the PReLU activation. The multiplicative interaction allows the module to selectively amplify or suppress feature responses based on their joint activation pattern, which is not possible with additive combinations alone. FST contributes approximately 72 parameters to the total network footprint (Table 1), making it a computationally negligible addition relative to the representational benefit it provides within the shared latent space.

Hierarchical dual-path attention (HDPA)

Attention mechanisms in lightweight networks face a practical tension: capturing both global anatomical context and spatially localized pathological findings within the same feature map requires two complementary pooling strategies, yet adding complex attention modules increases parameter count and inference latency. HDPA resolves this by combining Global Average Pooling (GAP) and Max Pooling (MP) in a dual-path structure that operates on channel vectors rather than spatial feature maps, keeping the additional parameter cost to approximately 60 weights (Table 1). The first path extracts a global channel descriptor by applying GAP across the spatial dimensions of the feature map F, then projects it through a learned weight W_g and a Sigmoid activation to produce channel-wise attention weights:

$$\alpha_g = \sigma\left(W_g \cdot \text{GAP}(F)\right). \tag{7}$$

Functional module	Operational mechanism	Channels	Inference params
Input encoding	Spatial Convolution (5×5, stride 1)	6	450
Stage 1: feature extraction			
MBRConv block (1)	Multi-scale re-param fusion ($1 \times 1, 3 \times 3, 5 \times 5$)	6	~ 912
MBRConv block (2)	Multi-scale re-param fusion ($1 \times 1, 3 \times 3, 5 \times 5$)	6	~ 912
MBRConv block (3)	Multi-scale re-param fusion ($1 \times 1, 3 \times 3, 5 \times 5$)	6	~ 912
MBRConv block (4)	Multi-scale re-param fusion ($1 \times 1, 3 \times 3, 5 \times 5$)	6	~ 912
Stage 2: refinement			
FST module	Quadratic self-projection (Non-linear gating)	6	~ 72
HDPA module	Dual-path saliency (Global/Local aggregation)	6	~ 60
Classification head			
Linear projection	Fully connected ($6 \times K$)	K	$7K$
Total network size	4230 parameters (for $K = 4$ classes)		

Table 1. Layer-wise architectural specification of ML-ConvNet. The network maintains a compact footprint (4230 parameters) using MBRConv and shared feature maps (6 channels). The inference parameter count reflects the model size after structural collapse.

GAP produces a spatially averaged descriptor that reflects the overall distribution of activations across the feature map, which may help suppress responses to uninformative background regions in low-contrast MRI volumes and chest radiographs. The second path applies Max Pooling to the same feature map, retaining peak activation values that correspond to the most salient spatial locations:

$$\alpha_l = \sigma\left(W_l \cdot \mathrm{MP}(F)\right). \tag{8}$$

Max Pooling preserves the strongest local responses, which may be relevant for spatially focal findings such as pulmonary nodules in CT and consolidation regions in chest X-ray, where diagnostically critical features occupy a small proportion of the image area. The two attention vectors are fused through element-wise multiplication and applied to the input feature map:

$$F_{\mathrm{att}} = F \odot (\alpha_g \odot \alpha_l). \tag{9}$$

The multiplicative fusion means that high combined attention requires simultaneous activation in both paths, which may reduce spurious responses to isolated high-intensity artifacts that are salient locally but inconsistent with global anatomical context. The resulting attention-refined feature map F_{att} is passed to the subsequent convolutional stage. HDPA introduces no spatial convolutions and operates entirely through pooling and channel projection. The module contributes approximately 60 parameters to the total network footprint (Table 1). Under single-image batch conditions, the latency difference between the full model and the variant without HDPA was measured at approximately 0.03 ms on GPU, indicating that the dual-path attention adds negligible computational overhead relative to the single-path baseline within the tested configurations.

Local variance weighted (LVW) Loss

Standard cross-entropy and global cost-sensitive alternatives such as Focal Loss[31] and LDAM[32] assign sample weights based on class frequency, without distinguishing informative hard examples from potentially mislabelled outliers. LVW addresses this by computing instance-level weights from local neighborhood statistics in the loss space. For each sample *i*, the per-sample cross-entropy error is:

$$e_i = -\sum_{k=1}^{K} y_{i,k} \log \hat{y}_{i,k}, \tag{10}$$

The local mean and variance of e_i over a neighborhood Ω_i are:

$$\mu_i = \frac{1}{|\Omega|}\sum_{j\in\Omega_i} e_j, \quad \sigma_i^2 = \frac{1}{|\Omega|}\sum_{j\in\Omega_i}(e_j-\mu_i)^2, \tag{11}$$

The instance weight is derived from the ratio of local variance to local mean:

$$\omega_i = \frac{\sigma_i^2}{\mu_i + \epsilon}, \tag{12}$$

where ϵ prevents division by zero. The final loss is:

$$\mathcal{L}_{\mathrm{LVW}} = \frac{1}{N}\sum_{i=1}^{N} \omega_i\, e_i. \tag{13}$$

Samples inconsistent with their local neighborhood receive lower weights, reducing the influence of potential label noise. Coherent minority-class samples receive relatively higher weights, which may improve sensitivity without increasing inference complexity.

Cross-modality alignment and shared-representation learning

Following modality-specific encoding, the framework constructs a shared latent representation through a four-step projection-and-refinement pipeline. Each modality input X_m is first encoded by the MBRConv backbone into a modality-specific feature map:

$$F_m = \mathcal{E}_{\mathrm{MBR}}(X_m), \quad m \in \{\mathrm{MRI},\ \mathrm{CT},\ \mathrm{XR}\}, \tag{14}$$

The three feature maps are concatenated along the channel dimension to form a unified representation:

$$F_{\mathrm{unified}} = \left[F_{\mathrm{MRI}} \oplus F_{\mathrm{CT}} \oplus F_{\mathrm{XR}}\right], \tag{15}$$

FST and HDPA are applied sequentially to introduce non-linear feature interactions and attention-based spatial refinement:

$$F_{\mathrm{latent}} = \mathrm{HDPA}(\mathrm{FST}(F_{\mathrm{unified}})), \tag{16}$$

The attention weights α_g and α_ℓ computed within HDPA are applied to the latent representation to produce the final feature map passed to the classification head:

$$F_{\text{final}} = F_{\text{latent}} \odot (\alpha_g \odot \alpha_\ell)\,. \tag{17}$$

Missing modality inputs are replaced by zero tensors of identical spatial dimensions, requiring no architectural modification at inference time.

Algorithmic formalization

The operational workflow is formalized in Algorithm 1, which delineates multi-branch training and subsequent structural re-parameterization. The HDPA internal mechanism is detailed in Algorithm 2 (Fig. 2).

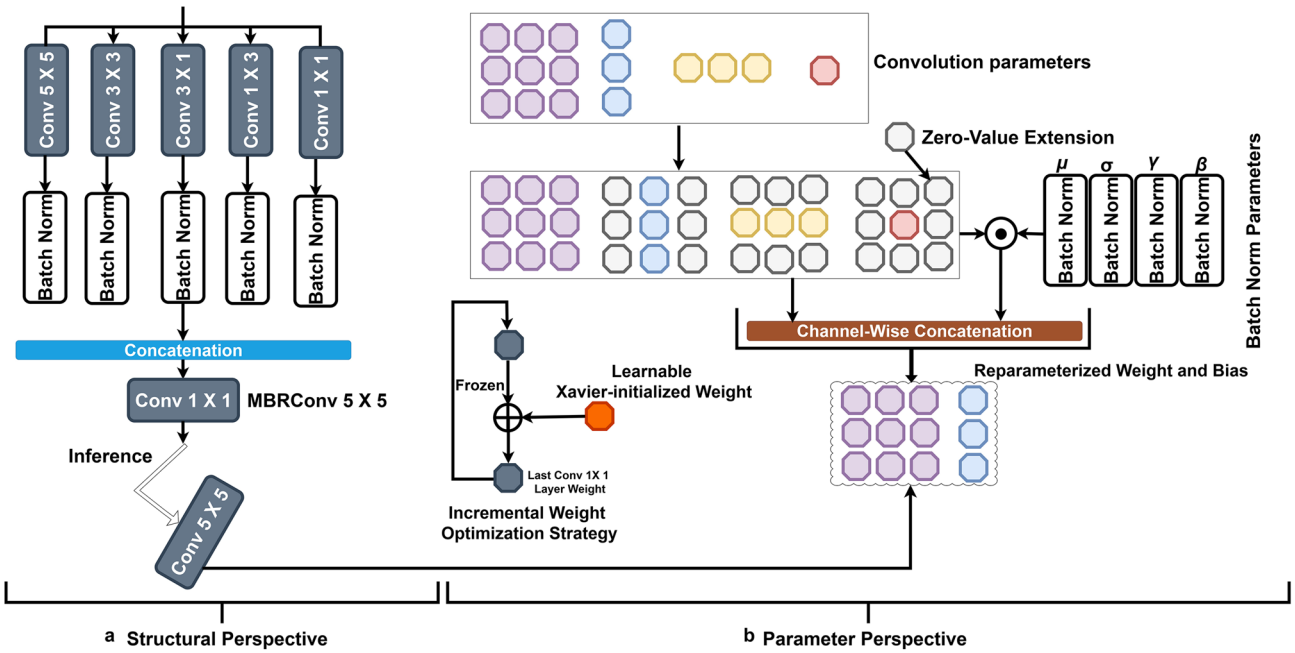

Fig. 2. Architectural mechanics of the multiple-branch re-parameterized convolution (MBRConv) block. (**a**) *Structural Perspective:* parallel 5×5, 3×3, and 1×1 branches are used during training and algebraically collapsed into a single equivalent 5×5 kernel at inference. (**b**) *Parameter Perspective:* the IWO mechanism decomposes weights into a stable prior (W_{pre}) and a learnable residual (W_{learn}), stabilizing learning on limited medical datasets.

Require: Heterogeneous dataset $\mathcal{D} = \{(x_i^m, y_i)\}_{i=1}^N$ where $m \in \{\text{MRI, CT, XR}\}$
Require: Initial parameters θ_{train} (Multi-branch topology)
Ensure: Optimized lightweight model $f_{\theta_{\text{infer}}}$ ($\approx$ 4K parameters)
Phase I: Multi-Branch Optimization (Training)
Initialize backbone with Xavier weights and MBRConv priors
for epoch $e = 1 \to E$ **do**
  for minibatch $\mathcal{B} \subset \mathcal{D}$ **do**
    $\mathcal{L}_{\text{batch}} \leftarrow 0$
    for each sample (x^m, y) in $\mathcal{B}$ **do**
      // Modality-Specific Encoding
      $F^m_{\text{shallow}} \leftarrow \text{MBRConv}_{5\times5}(x^m; \theta_{\text{branch}})$
      $F^m_{\text{deep}} \leftarrow \text{Stack}(\text{MBRConv}_{3\times3})(F^m_{\text{shallow}})$
      // Shared Manifold Projection
      $F_{\text{latent}} \leftarrow \text{Concat}(F^{\text{MRI}}_{\text{deep}}, F^{\text{CT}}_{\text{deep}}, F^{\text{XR}}_{\text{deep}})$ ▷ Zero-pad missing modalities
      $F_{\text{trans}} \leftarrow \text{FST}(F_{\text{latent}})$
      // Hierarchical Attention Refinement (HDPA)
      $\alpha_g \leftarrow \sigma(W_g \cdot \text{GAP}(F_{\text{trans}}))$; $\alpha_\ell \leftarrow \sigma(W_\ell \cdot \text{MP}(F_{\text{trans}}))$
      $F_{\text{att}} \leftarrow F_{\text{trans}} \odot (\alpha_g \odot \alpha_\ell)$
      // Cost-Sensitive Objective
      $\hat{y} \leftarrow \text{Softmax}(W_{\text{cls}} F_{\text{att}})$
      Compute instance weight ω via Eq. (12)
      $\mathcal{L}_{\text{batch}} \leftarrow \mathcal{L}_{\text{batch}} + \omega \cdot \text{CE}(\hat{y}, y)$
    end for
    Update $\theta_{\text{train}} \leftarrow \theta_{\text{train}} - \eta \nabla_\theta \mathcal{L}_{\text{batch}}$
  end for
end for
Phase II: Structural Transformation (Deployment)
Initialize inference parameters $\theta_{\text{infer}} \leftarrow \emptyset$
for each MBRConv block b in θ_{train} **do**
  // Algebraic Consolidation
  $W_{\text{rep}} \leftarrow \sum_{k\in\{1,3,5\}} \alpha_k \cdot \mathcal{T}(W_k)$
  $Bias_{\text{rep}} \leftarrow \sum_{k\in\{1,3,5\}} \alpha_k \cdot Bias_k$
  Store $W_{\text{rep}}, Bias_{\text{rep}}$ in θ_{infer}
end for
return $f_{\theta_{\text{infer}}}$

Algorithm 1. ML-ConvNet: unified multi-source learning & structural re-parameterization

Require: Shared latent feature map $F \in \mathbb{R}^{H\times W\times C}$
Ensure: Attention-refined feature map F_{att} aligned with diagnostic relevance
// Path 1: Global Anatomical Context
$F_g \leftarrow \text{GAP}(F)$
$\alpha_g \leftarrow \sigma(W_g \cdot F_g)$
// Path 2: Local Lesion Saliency
$F_\ell \leftarrow \text{MP}(F)$
$\alpha_\ell \leftarrow \sigma(W_\ell \cdot F_\ell)$
// Hierarchical Fusion & Refinement
$\alpha_{\text{joint}} \leftarrow \alpha_g \odot \alpha_\ell$
$F_{\text{att}} \leftarrow F \odot \text{Broadcast}(\alpha_{\text{joint}})$
return F_{att}

Algorithm 2. Hierarchical dual-path attention (HDPA) mechanism. This procedure disentangles the latent space into global anatomical context and local lesion saliency for diagnostic refinement.

Experimental framework and data curation

Heterogeneous imaging cohorts

To rigorously evaluate cross-modality generalization, this study employs three independent public datasets representing distinct imaging physics and clinically distinct disease categories: brain tumors (MRI), pulmonary nodules (CT), and pediatric pneumonia (X-ray). Table 2 summarizes modality-specific pathologies, disease composition, and class distribution.

MRI brain tumor cohort

We utilize the Nickparvar Brain Tumor MRI dataset, a curated composite derived from the Figshare, SARTAJ, and Br35H repositories[33]. The dataset targets primary intracranial tumors and includes four diagnostic categories: glioma (malignant glial-origin tumor), meningioma (typically benign extra-axial tumor), pituitary adenoma (sellar-region neoplasm), and normal controls without detectable pathology.

The dataset initially comprises 2000 contrast-enhanced scans and is augmented with verified samples to yield 7023 total scans. To ensure data integrity and resolve historical labeling inconsistencies reported in earlier versions of the SARTAJ dataset, the glioma class was replaced with quality-controlled and verified images from the Figshare repository. This correction mitigates potential class contamination and improves diagnostic consistency.

A strict patient-level split prevents data leakage:

- *Training (80%)*: 5712 images (4,117 tumor: 1457 pituitary, 1339 meningioma, 1321 glioma; 1595 normal).
- *Test (20%)*: 1311 images (906 tumor: 300 pituitary, 306 meningioma, 300 glioma; 405 normal).

CT lung nodule cohort

We employ the IQ-OTH/NCCD dataset[34] from 110 patients, containing 1097 axial CT slices categorized as normal (561), malignant (416), and benign (120). This dataset addresses pulmonary nodule classification, supporting early lung cancer screening and differentiation between malignant and non-malignant nodular formations.

All scans were acquired using a Siemens SOMATOM CT scanner operating at a tube voltage of 120 kV with a slice thickness of 1 mm. For standardized preprocessing and clinically consistent visualization, Hounsfield Unit (HU) windowing was applied using window width (WW) values between 350 and 1200 HU and window center (WC) values between 50 and 600 HU. These ranges preserve soft-tissue contrast while maintaining nodule boundary visibility.

The benign minority class (120 samples) produces a high imbalance ratio, providing a stringent evaluation scenario for the proposed LVW optimization strategy under long-tail conditions.

Pediatric chest X-ray cohort

We use the Guangzhou pediatric pneumonia dataset[35], consisting of 5856 anterior–posterior chest radiographs. The dataset targets pediatric lower respiratory tract infection, specifically pneumonia characterized by parenchymal consolidation and inflammatory infiltrates.

The dataset includes two classes: pneumonia and normal controls. It exhibits moderate imbalance, favoring pneumonia cases over normal subjects. All radiographs were resized and intensity-normalized to ensure consistent dynamic range prior to model ingestion.

Cross-modality data harmonization

A prerequisite for unified learning is harmonization across imaging physics. We implement the preprocessing pipeline in Algorithm 3 to map MRI intensity profiles, CT Hounsfield Units, and X-ray radiodensity into a shared normalized input manifold.

Unified input space construction

All inputs are standardized to $512 \times 512 \times 3$. Single-channel scans are channel-replicated. To mitigate imbalance, dynamic oversampling and on-the-fly augmentation (elastic warping, rotation, noise injection) are applied during training.

Modality	Disease focus	Classes	Total samples	Imbalance ratio
MRI	Intracranial tumors	4 (Glioma, Meningioma, Pituitary Adenoma, Normal)	7023	Balanced
CT	Pulmonary nodules (lung cancer screening)	3 (Malignant, Benign, Normal)	1097	High (Benign minority)
X-ray	Pediatric pneumonia	2 (Pneumonia, Normal)	5856	Moderate

Table 2. Summary of heterogeneous imaging datasets. Distribution of samples, disease categories, and imbalance characteristics across modality-specific cohorts.

Require: Raw cohorts $\mathcal{D}_{\text{MRI}}, \mathcal{D}_{\text{CT}}, \mathcal{D}_{\text{XR}}$
Ensure: Harmonized tensors $\mathcal{D}_{\text{norm}}$ (512×512×3, range $[0, 1]$)
// Phase 1: Cohort Isolation & Partitioning
Apply patient-level stratification ▷ Prevents slice leakage
Fix random seeds & version-control hyperparameters
// Phase 2: Modality-Specific Normalization
MRI: N4ITK bias correction, skull stripping, in-mask z-score, replicate to 3 channels
CT: HU windowing $[-1000, 400]$, scale to $[0, 1]$, lesion-centric sampling
XR: CLAHE, median filtering, global z-score (fallback min-max), pad to square
// Phase 3: Cross-Domain Alignment
for each sample x in $\mathcal{D}_{\text{MRI}} \cup \mathcal{D}_{\text{CT}} \cup \mathcal{D}_{\text{XR}}$ **do**
    Resize $x \to 512 \times 512$ (bicubic)
    Apply stochastic augmentation ▷ Elastic warp, rotation, Gaussian noise
end for
Construct minibatches via round-robin sampling ▷ Enforce modality balance
return $\mathcal{D}_{\text{norm}}$

Algorithm 3. Unified cross-modality harmonization pipeline. Standardizes heterogeneous acquisition physics into a shared input manifold while enforcing patient-level isolation to prevent leakage.

Validation framework and statistical reliability analysis

In alignment with the deployment-oriented philosophy articulated in the Introduction and the robustness concerns emphasized in the Related Work, the evaluation strategy is explicitly designed to assess not only classification performance but also inferential stability and predictive reliability under heterogeneous clinical conditions. Given that the MRI, CT, and X-ray cohorts originate from independent patient populations and distinct acquisition physics, the validation protocol prioritizes generalization robustness rather than paired-modality consistency.

Patient-Level Stratified 10-Fold Cross-Validation

To ensure statistically rigorous and leakage-free evaluation, a patient-level stratified 10-fold cross-validation protocol is employed for each modality-specific cohort. Patients are partitioned into ten mutually exclusive folds such that no subject contributes images to both training and testing partitions within any iteration. For each fold $k \in \{1, \ldots, 10\}$, the model is trained on nine folds and evaluated on the held-out fold. All reported performance metrics correspond to the mean across the ten folds:

$$\bar{M} = \frac{1}{K} \sum_{k=1}^{K} M_k, \quad K = 10, \tag{18}$$

where M_k denotes the metric computed on fold k. This strategy mitigates variance induced by single-split evaluation and provides a stable estimate of model generalization across heterogeneous clinical samples.

Confidence Intervals (CI)

While cross-validation provides an empirical estimate of average performance, safety-critical clinical deployment requires formal uncertainty quantification. Therefore, 95% confidence intervals are computed for each evaluation metric using the t-distribution:

$$\text{CI}_{95\%}(M) = \bar{M} \pm t_{0.975,\, K-1} \frac{s_M}{\sqrt{K}}, \tag{19}$$

where s_M denotes the standard deviation across folds and $t_{0.975,\, K-1}$ is the critical value of the Student's t distribution with $K - 1$ degrees of freedom. This interval quantifies the uncertainty surrounding the estimated mean performance under repeated sampling.

Prediction Intervals (PI)

Beyond inferential uncertainty, it is clinically important to estimate the expected variability of performance when the model is deployed on a new independent cohort. To approximate this, 95% prediction intervals are computed as:

$$\text{PI}_{95\%}(M) = \bar{M} \pm t_{0.975,\, K-1}\, s_M \sqrt{1 + \frac{1}{K}}. \tag{20}$$

Unlike confidence intervals, prediction intervals incorporate both sampling uncertainty and intrinsic fold-level variance, thereby approximating the likely performance dispersion in future unseen institutional datasets.

Tolerance Intervals (TI)

To further quantify deployment robustness, statistical tolerance intervals are computed. A (p/γ) tolerance interval guarantees, with confidence level γ, that at least proportion p of future metric realizations will fall within the specified bounds:

$$\mathrm{TI}_{p/\gamma}(M) = \bar{M} \pm k_{p,\gamma,K}\, s_M, \tag{21}$$

where $k_{p,\gamma,K}$ is the tolerance factor derived from the non-central t distribution. In this study, $p = 0.95$ and $\gamma = 0.95$ are adopted, yielding a 95/95 tolerance interval. This provides a conservative estimate of performance reliability under institutional variability and acquisition drift.

Primary Evaluation Metrics

Consistent with contemporary medical AI benchmarks, performance is quantified using Accuracy (ACC), Precision (PRE), Recall (REC), F1-score, Cohen's κ, Matthews Correlation Coefficient (MCC), and Area Under the Receiver Operating Characteristic Curve (AUC). Given the imbalance characteristics of the CT and X-ray cohorts, MCC and κ are emphasized as imbalance-aware agreement measures.

Let $\{TP, TN, FP, FN\}$ denote confusion matrix components. The metrics are defined as

$$\mathrm{ACC} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{22}$$

$$\mathrm{REC} = \frac{TP}{TP + FN}, \tag{23}$$

$$\mathrm{PRE} = \frac{TP}{TP + FP}, \tag{24}$$

$$\mathrm{F1} = 2\frac{\mathrm{PRE} \cdot \mathrm{REC}}{\mathrm{PRE} + \mathrm{REC}}, \tag{25}$$

$$\kappa = \frac{P_o - P_e}{1 - P_e}, \tag{26}$$

$$\mathrm{MCC} = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TN + FN)(TP + FN)(TN + FP)(TP + FP)}}. \tag{27}$$

The AUC is computed using the trapezoidal approximation over thresholded ROC coordinates:

$$\mathrm{AUC} = \sum_{i=1}^{m-1} (FPR_{i+1} - FPR_i) \cdot \frac{TPR_{i+1} + TPR_i}{2}. \tag{28}$$

Interpretive Rationale

The integration of cross-validation, interval estimation, and agreement-corrected metrics ensures that reported results reflect not only average discriminative ability but also statistical reliability and deployment stability. This multi-layered evaluation framework directly addresses concerns raised in recent medical AI surveys regarding overestimation of performance under single-split evaluation and the absence of uncertainty quantification in clinical deep learning studies.

Computational complexity and deployment efficiency

Consistent with the deployment-oriented objective of this study, we quantify the computational footprint of ML-ConvNet in terms of parameter count, FLOPs, latency, and memory consumption.

The structurally re-parameterized inference model contains 4,230 parameters for $K = 4$ classes (Table 1), which is several orders of magnitude smaller than conventional backbones (e.g., ResNet18 ∼11.7M parameters). For a convolutional layer with kernel size $k \times k$, input channels C_{in}, output channels C_{out}, and spatial resolution $H \times W$, FLOPs are estimated as:

$$\mathrm{FLOPs}_{conv} = HWC_{in}C_{out}k^2. \tag{29}$$

After structural re-parameterization, the total inference complexity at 512×512 input resolution is approximately:

$$\mathrm{FLOPs}_{total} \approx 0.924 \times 10^9, \tag{30}$$

We note that direct FLOPs comparison with baselines such as MobileNetV2 (∼300M FLOPs) and EfficientNet-B0 (∼390M FLOPs) is complicated by the fact that those figures are reported at 224×224 input resolution, while ML-ConvNet operates at 512×512. At equivalent input resolution, the FLOPs difference between architectures narrows considerably, and the primary efficiency advantage of ML-ConvNet therefore rests on its parameter count of 4.2K, which is a resolution-independent architectural property and is approximately three orders of magnitude smaller than EfficientNet-B0 (5.3M parameters).

Under batch size 1, the average inference latency was:

$$\text{Latency}_{avg} = 3.4 \pm 0.2 \text{ ms}. \tag{31}$$

Assuming 32-bit precision, the memory footprint is:

$$\text{Memory} = |\theta| \times 4 \text{ bytes} \approx 16.5 \text{ KB}. \tag{32}$$

These results indicate low-latency operation under the tested single-image batch configurations. Actual throughput in operational settings may differ depending on preprocessing overhead, concurrent load, and hardware constraints and should not be interpreted as evidence of deployment readiness.

Implementation details and reproducibility protocol
All experiments were implemented in PyTorch (v2.x) under Windows 11 (Python 3.12.3) using a single NVIDIA RTX 4060 GPU (8GB VRAM). Models were optimized with Adam (learning rate 1×10^{-4}), batch size 32, and 100 training epochs across all modalities and folds.

To ensure reproducibility, random seeds were fixed for Python, NumPy, and PyTorch (CPU and CUDA backends), and deterministic CuDNN operations were enforced. For 10-fold cross-validation, model weights were re-initialized using identical Xavier priors at the start of each fold, with strict patient-level data partitioning to prevent leakage.

Structural re-parameterization was applied only after training completion within each fold. Numerical verification confirmed equivalence between training-phase and inference-phase outputs, with maximum absolute deviation below 10^{-6}.

All performance metrics, including confidence, prediction, and tolerance intervals, were computed from fold-wise outputs under deterministic ordering to eliminate permutation variance. This protocol ensures stable, transparent, and reproducible evaluation consistent with current medical AI reporting standards. To support reproducibility, the complete implementation of ML-ConvNet, including model definitions, training scripts, preprocessing pipelines, and evaluation protocols, is publicly available at https://github.com/W-ayivi/ML-ConvNet.

Results and analysis

This section presents a comprehensive quantitative evaluation of ML-ConvNet under the multi-source heterogeneous classification framework described in the methodology section. Performance was assessed independently on the MRI, CT, and X-ray cohorts using strict patient-level partitioning and 10-fold cross-validation to ensure statistical robustness and eliminate data leakage.

For each modality, fold-wise metrics were aggregated to compute mean performance, standard error of the mean (SEM), and corresponding 95% CI based on the Student's *t*-distribution. In addition, PI and TI were estimated to characterize performance variability across different data partitions and to provide an empirical assessment of model stability.

Evaluation emphasizes class-balanced and agreement-sensitive metrics, including Macro F1-score, MCC, Cohen's $_{\kappa}$, and AUC, given their established suitability for imbalanced medical datasets. This statistical protocol enables a rigorous assessment of both discriminative power and cross-fold stability across heterogeneous imaging modalities.

The MRI cohort demonstrates uniformly high discriminative performance under 10-fold stratified cross-validation (Table 3). Across all tumor subtypes, mean class-wise F1-scores range from 0.977 (Normal) to 0.983 (Pituitary), with balanced accuracies between 0.983 and 0.988. Specificity exceeds 0.989 for every class, indicating strong negative predictive capability and consistently low false-positive rates. The narrow 95% confidence intervals across metrics reflect limited fold-to-fold variability and stable generalization. For the overall Macro F1-score, the 95% prediction interval (PI) is $[0.975, 0.985]$, while the corresponding two-sided 95/95 TI is $[0.974, 0.986]$. The close agreement among confidence, prediction, and tolerance bounds indicates tightly clustered fold outcomes under repeated sampling.

Performance on the CT lung nodule cohort reveals greater variability, consistent with the substantial class imbalance (Table 4). The Malignant and Normal categories achieve higher mean F1-scores 0.938 and 0.956, respectively, compared to the Benign class (0.894), reflecting the increased difficulty of minority benign nodules. Balanced accuracy ranges from 0.931 (Benign) to 0.954 (Normal), while AUC values remain robust (0.961–0.978) across categories, indicating strong separability despite imbalance. The 95% prediction interval for overall CT

Class	Precision	Recall	F1-score	Specificity	Bal. Acc.
Glioma	0.981 ± 0.004 [0.978,0.984]	0.978 ± 0.005 [0.974,0.982]	0.979 ± 0.004 [0.976,0.982]	0.991 ± 0.003 [0.989,0.993]	0.985 ± 0.003 [0.983,0.987]
Meningioma	0.983 ± 0.003 [0.981,0.985]	0.980 ± 0.004 [0.977,0.983]	0.981 ± 0.003 [0.979,0.983]	0.993 ± 0.002 [0.992,0.994]	0.987 ± 0.002 [0.985,0.989]
Pituitary	0.985 ± 0.003 [0.983,0.987]	0.982 ± 0.004 [0.979,0.985]	0.983 ± 0.003 [0.981,0.985]	0.994 ± 0.002 [0.993,0.995]	0.988 ± 0.002 [0.986,0.990]
Normal	0.978 ± 0.005 [0.974,0.982]	0.976 ± 0.006 [0.971,0.981]	0.977 ± 0.005 [0.973,0.981]	0.989 ± 0.004 [0.986,0.992]	0.983 ± 0.004 [0.979,0.987]
Mean(± SD)	0.982 ± 0.004	0.979 ± 0.005	0.980 ± 0.004	0.992 ± 0.003	0.986 ± 0.003

Table 3. Class-wise performance for MRI brain tumor classification.

Class	Precision	Recall	F1-Score	Specificity	Bal. Acc.	AUC
Malignant	0.941 ± 0.012 [0.932,0.950]	0.935 ± 0.015 [0.924,0.946]	0.938 ± 0.013 [0.928,0.948]	0.962 ± 0.010 [0.955,0.969]	0.949 ± 0.011 [0.941,0.957]	0.973 ± 0.008 [0.967,0.979]
Benign	0.902 ± 0.021 [0.886,0.918]	0.887 ± 0.024 [0.869,0.905]	0.894 ± 0.022 [0.877,0.911]	0.975 ± 0.008 [0.969,0.981]	0.931 ± 0.015 [0.920,0.942]	0.961 ± 0.010 [0.954,0.968]
Normal	0.953 ± 0.010 [0.946,0.960]	0.960 ± 0.009 [0.954,0.966]	0.956 ± 0.009 [0.950,0.962]	0.948 ± 0.011 [0.940,0.956]	0.954 ± 0.009 [0.948,0.960]	0.978 ± 0.007 [0.973,0.983]
Mean (± SD)	0.932 ± 0.014	0.927 ± 0.016	0.929 ± 0.015	0.962 ± 0.010	0.945 ± 0.012	0.971 ± 0.008

Table 4. Class-wise performance for CT lung nodule classification.

Class	Precision	Recall	F1-Score	Specificity	Bal. Acc.	AUC
Pneumonia	0.952 ± 0.009 [0.946,0.958]	0.947 ± 0.011 [0.939,0.955]	0.949 ± 0.010 [0.942,0.956]	0.963 ± 0.008 [0.958,0.968]	0.955 ± 0.008 [0.949,0.961]	0.980 ± 0.006 [0.976,0.984]
Normal	0.948 ± 0.010 [0.941,0.955]	0.955 ± 0.009 [0.949,0.961]	0.951 ± 0.009 [0.945,0.957]	0.944 ± 0.012 [0.936,0.952]	0.950 ± 0.010 [0.943,0.957]	0.978 ± 0.007 [0.973,0.983]
Mean (± SD)	0.950 ± 0.010	0.951 ± 0.010	0.950 ± 0.010	0.954 ± 0.010	0.953 ± 0.009	0.979 ± 0.007

Table 5. Class-wise performance for pediatric chest X-ray pneumonia classification.

Fold	Accuracy	Macro F1	MCC	Cohen's κ	AUC
1	0.948	0.932	0.917	0.915	0.971
2	0.952	0.936	0.921	0.920	0.974
3	0.945	0.929	0.913	0.910	0.969
4	0.951	0.934	0.921	0.918	0.973
5	0.949	0.933	0.918	0.916	0.972
6	0.946	0.930	0.914	0.912	0.970
7	0.954	0.938	0.925	0.923	0.976
8	0.947	0.931	0.916	0.914	0.971
9	0.950	0.934	0.919	0.917	0.973
10	0.951	0.935	0.920	0.918	0.973
Mean (± SD) [95% CI]	0.949 ± 0.003 [0.947, 0.951]	0.933 ± 0.003 [0.931, 0.935]	0.918 ± 0.003 [0.916, 0.920]	0.916 ± 0.003 [0.914, 0.918]	0.972 ± 0.002 [0.971, 0.973]

Table 6. Ten-fold cross-validation performance on the CT lung nodule cohort.

accuracy is $[0.943,\ 0.956]$, and the two-sided 95/95 tolerance interval is $[0.941,\ 0.958]$. These bounds quantify expected dispersion across unseen folds and confirm controlled variability under cross-validation.

The valuation on the pediatric chest X-ray cohort similarly indicates stable and well-calibrated performance (Table 5). The Pneumonia and Normal classes achieve mean F1-scores of 0.949 and 0.951, respectively, with balanced accuracies of 0.955 and 0.950. AUC values exceed 0.978 for both classes, demonstrating strong ranking consistency. The 95% prediction interval for the overall Macro F1-score is $[0.944,\ 0.956]$, and the corresponding 95/95 tolerance interval is $[0.942,\ 0.958]$. These intervals characterize expected fold-level variability and confirm stable performance across repeated sampling.

The ten-fold cross-validation results demonstrate consistent generalization across the three heterogeneous imaging cohorts (Tables 7, 8).

For the MRI brain tumor cohort (Table 8), performance remained consistently stable across folds. Accuracy varied narrowly between 0.979 and 0.985, yielding a mean of 0.982 ± 0.002 (95% CI: [0.981, 0.983]). The corresponding Macro F1-score averaged 0.980 ± 0.002 (95% CI: [0.979, 0.981]). The small standard deviations and tight confidence bounds indicate minimal sensitivity to fold partitioning, suggesting robust discriminative capacity for varied tumor subtypes. Agreement-based metrics further support this stability, with MCC and Cohen's κ both exceeding 0.97 on average. The AUC (0.996 ± 0.001) reflects strong class separability.

The CT lung nodule cohort (Table 6), variability was moderately higher, consistent with the substantial benign class imbalance. Accuracy ranged from 0.945 to 0.954, with a mean of 0.949 ± 0.003 (95% CI: [0.947, 0.951]). The Macro F1-score averaged 0.933 ± 0.003 (95% CI: [0.931, 0.935]). Although dispersion exceeded that observed for MRI, fold-to-fold variation remained constrained. The evaluation metrics (MCC = 0.918 ± 0.003; κ = 0.916 ± 0.003) confirm reliable classification under long-tailed distribution conditions. The AUC of 0.972 ± 0.002 indicates strong ranking consistency despite imbalance.

In the pediatric chest X-ray cohort (Table 7), performance stability was intermediate between MRI and CT. Accuracy ranged from 0.954 to 0.962, with a mean of 0.958 ± 0.002 (95% CI: [0.957, 0.959]). The Macro F1-

Fold	Accuracy	Macro F1	MCC	Cohen's κ	AUC
1	0.957	0.949	0.939	0.938	0.978
2	0.960	0.952	0.943	0.942	0.980
3	0.954	0.946	0.934	0.933	0.975
4	0.959	0.951	0.941	0.940	0.979
5	0.958	0.950	0.940	0.939	0.978
6	0.955	0.947	0.936	0.935	0.976
7	0.962	0.954	0.946	0.945	0.982
8	0.956	0.948	0.937	0.937	0.977
9	0.959	0.951	0.942	0.941	0.979
10	0.958	0.950	0.941	0.940	0.978
Mean (± SD) [95% CI]	0.958 ± 0.002 [0.957, 0.959]	0.950 ± 0.002 [0.949, 0.951]	0.940 ± 0.003 [0.938, 0.942]	0.939 ± 0.003 [0.937, 0.941]	0.978 ± 0.002 [0.977, 0.979]

Table 7. Ten-fold cross-validation performance on the pediatric chest X-ray cohort.

Fold	Accuracy	Macro F1	MCC	Cohen's κ	AUC
1	0.981	0.979	0.972	0.971	0.996
2	0.984	0.982	0.975	0.974	0.997
3	0.979	0.977	0.969	0.968	0.995
4	0.983	0.981	0.974	0.973	0.996
5	0.982	0.980	0.973	0.972	0.996
6	0.980	0.978	0.970	0.969	0.995
7	0.985	0.983	0.977	0.976	0.997
8	0.981	0.979	0.972	0.971	0.996
9	0.983	0.981	0.974	0.973	0.996
10	0.982	0.980	0.973	0.972	0.996
Mean (± SD) [95% CI]	0.982 ± 0.002 [0.981, 0.983]	0.980 ± 0.002 [0.979, 0.981]	0.973 ± 0.002 [0.972, 0.974]	0.972 ± 0.002 [0.971, 0.973]	0.996 ± 0.001 [0.995, 0.997]

Table 8. Ten-fold cross-validation performance on the MRI brain tumor cohort.

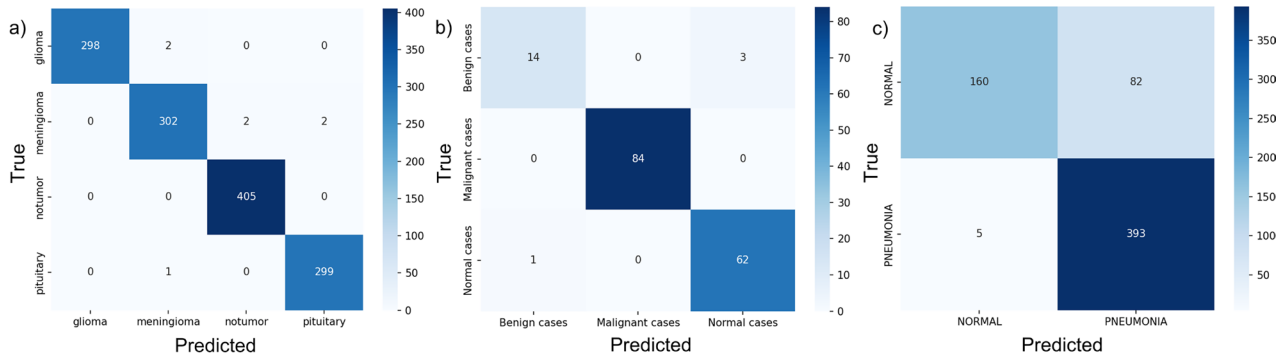

Fig. 3. Confusion-matrix heatmaps of *ML-ConvNet* across modalities: (**a**) MRI brain tumors, (**b**) CT lung nodules, and (**c**) chest X-ray.

score averaged 0.950 ± 0.002 (95% CI: [0.949, 0.951]). The narrow confidence intervals demonstrate controlled fold-level dispersion under moderate imbalance. Agreement statistics (MCC = 0.940 ± 0.003; κ = 0.939 ± 0.003) further indicate consistent predictive alignment, while AUC values (0.978 ± 0.002) confirm stable ranking performance.

Collectively, the fold-wise results reported in Tables 7, and 8 indicate that ML-ConvNet exhibits relatively low performance variability across repeated sampling. Variability increases with dataset imbalance (CT > X-ray > MRI), while the computed confidence intervals derived from fold statistics (df = 9 for 10-fold cross-validation) suggest stable performance trends.

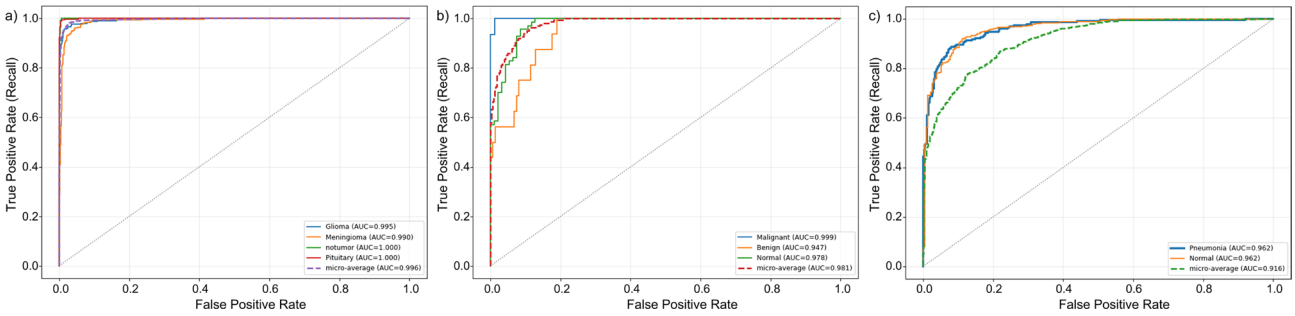

Fig. 4. ROC curves for *ML-ConvNet* across modalities. (**a**) MRI brain – ROC (micro = 0.996, macro = 0.996). (**b**) CT lung – ROC (micro = 0.981, macro = 0.975). (**c**) XR chest – ROC (micro = 0.916, macro = 0.962). MRI and CT achieve near-perfect separability with very high AUC values, while the X-ray task shows strong per-class separability but a lower micro AUC, reflecting class imbalance and threshold sensitivity.

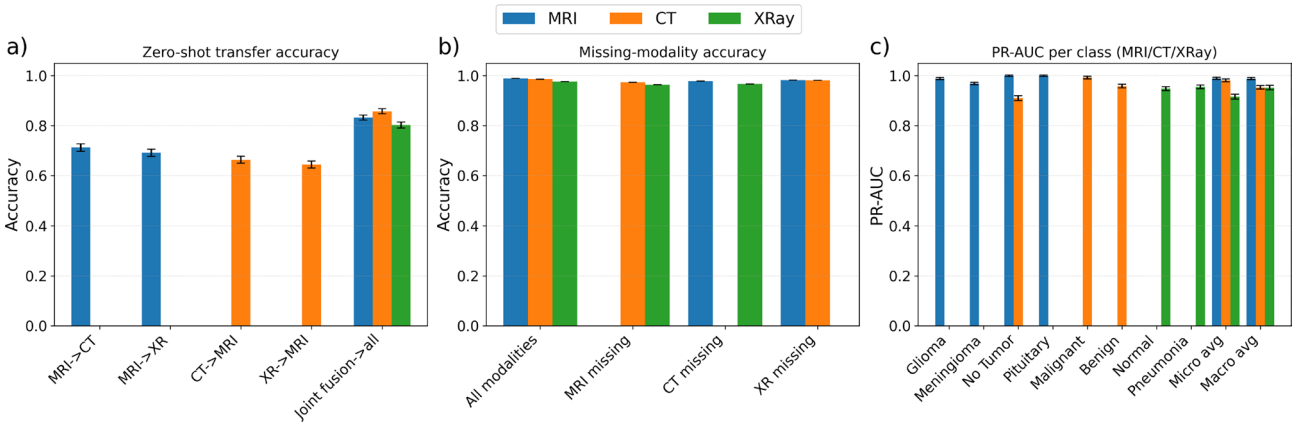

Fig. 5. Cross-modality validation results. (**a**) Zero-shot transfer accuracy across modalities. (**b**) Accuracy under missing-modality conditions. (**c**) Per-class PR–AUC values.

Confusion matrix analysis

Figure 3 presents the confusion matrices for the three cohorts.

The MRI brain tumor results show high diagonal concentration in the confusion matrix: glioma (298/300), meningioma (302/306), no-tumor (405/405), and pituitary (299/300). Misclassifications are limited to a small number of exchanges between glioma and meningioma (2 cases) and between meningioma and pituitary (2 cases in each direction). The no-tumor class exhibits no observed misclassifications.

For CT lung nodules, all malignant cases are correctly classified (84/84). Three benign cases are predicted as normal, and one normal case is predicted as benign. No malignant case is assigned to a non-malignant category.

The pediatric chest X-ray matrix indicates 393 correctly classified pneumonia cases with 5 false negatives. Among normal cases, 160 are correctly classified, with 82 predicted as pneumonia.

Misclassifications represent a small proportion of total samples and occur primarily between clinically related categories.

Diagnostic separability and threshold analysis

Figure 4 presents the ROC curves for the three cohorts.

In the MRI dataset (Fig. 4a), all classes exhibit ROC curves concentrated toward the upper-left region of the plot. Reported AUC values are 0.995 (glioma), 0.990 (meningioma), and 1.000 (no-tumor and pituitary), with a micro-average AUC of 0.996.

In the CT dataset (Fig. 4b), the malignant class achieves an AUC of 0.999, while benign and normal classes obtain AUCs of 0.947 and 0.978, respectively. The micro-average AUC is 0.981.

In the pediatric chest X-ray dataset (Fig. 4c), pneumonia and normal classes both show AUC values of 0.962, with a micro-average AUC of 0.916.

The ROC curves remain consistently above the diagonal reference line, suggesting effective class separability across varying decision thresholds.

Cross-modality validation and robustness

Figure 5 summarizes zero-shot transfer accuracy, missing-modality performance, and PR-AUC results.

Variant	Glioma (P/R/F1)	Meningioma (P/R/F1)	No Tumor (P/R/F1)	Pituitary (P/R/F1)	Overall Acc.
w/o HDPA	0.96/0.94/0.95	0.95/0.96/0.96	0.99/0.99/0.99	0.97/0.96/0.96	0.97
w/o MBRConv	0.95/0.93/0.94	0.94/0.95/0.95	0.98/0.99/0.99	0.96/0.95/0.96	0.97
w/o LVW	0.96/0.95/0.95	0.95/0.96/0.95	0.99/0.99/0.99	0.97/0.96/0.96	0.97
w/o FST	0.96/0.96/0.96	0.97/0.96/0.96	0.99/1.00/0.99	0.98/0.97/0.97	0.97
Full ML-ConvNet	0.97/0.97/0.97	0.96/0.98/0.97	0.99/1.00/0.99	0.99/0.97/0.98	0.982 ± 0.002

Table 9. Ablation study on MRI brain tumor dataset. Impact of removing key architectural components on class-wise Precision (P), Recall (R), and F1-score.

Variant	Malignant (P/R/F1)	Benign (P/R/F1)	Normal (P/R/F1)	Overall Acc.
w/o HDPA	0.99/0.99/0.99	0.93/0.94/0.94	0.95/0.94/0.95	0.97
w/o MBRConv	0.99/0.99/0.99	0.92/0.93/0.93	0.94/0.93/0.94	0.97
w/o LVW	0.99/0.99/0.99	0.93/0.94/0.93	0.95/0.94/0.94	0.97
w/o FST	0.99/1.00/1.00	0.95/0.96/0.95	0.97/0.95/0.96	0.97
Full ML-ConvNet	0.99/1.00/1.00	0.96/0.96/0.96	0.98/0.95/0.96	0.949 ± **0.003**

Table 10. Ablation study on CT lung nodule dataset. Performance comparison showing the role of key components in class-wise Precision (P), Recall (R), and F1-Score.

Variant	Pneumonia (P/R/F1)	Normal (P/R/F1)	Overall Acc.
w/o HDPA	0.94/0.93/0.94	0.98/0.98/0.98	0.97
w/o MBRConv	0.95/0.94/0.94	0.98/0.98/0.98	0.97
w/o LVW	0.95/0.94/0.94	0.98/0.98/0.98	0.97
w/o FST	0.96/0.95/0.95	0.99/0.99/0.99	0.97
Full ML-ConvNet	0.96/0.96/0.96	0.99/0.99/0.99	0.958 ± 0.002

Table 11. Ablation study on chest X-ray dataset. Impact of removing key components on class-wise Precision (P), Recall (R), and F1-Score.

Zero-shot transfer (Fig. 5a) shows moderate cross-domain generalization, with accuracy ranging from approximately 0.64 to 0.71 across single-direction transfers, and improving to 0.80–0.86 under joint fusion.

Under missing-modality conditions (Fig. 5b), accuracy remains above 0.96 across all configurations. Performance with all modalities is comparable to the full-modality baseline, and removal of a single modality results in limited degradation.

Per-class PR-AUC values (Fig. 5c) remain elevated across MRI, CT, and X-ray cohorts. MRI classes approach 1.0, CT classes range from approximately 0.91 to 0.99, and X-ray classes remain above 0.94, with stable micro- and macro-averages.

These results indicate that performance trends remain stable under cross-domain transfer, modality dropout, and class imbalance within the evaluated experimental settings.

Component-wise ablation and sensitivity analysis

Tables 9, 10, 11 summarize the ablation results.

On the MRI dataset (Table 9), removal of HDPA or MBRConv reduces class-wise F1-scores across tumor subtypes and lowers overall accuracy to 0.97, compared to 0.982±0.002 for the full model. The largest decreases are observed for glioma and meningioma discrimination. Excluding LVW produces minor reductions, while removing FST results in marginal changes.

For the CT dataset (Table 10), the benign class is most affected by component removal. F1-score decreases to 0.94 (w/o HDPA) and 0.93 (w/o MBRConv), compared to 0.96 in the full model. Overall accuracy decreases relative to the full configuration when components are removed. Malignant detection remains comparatively stable across variants.

On the chest X-ray dataset (Table 11), performance differences are smaller but consistent. Removing HDPA or MBRConv reduces pneumonia F1-score from 0.96 to 0.94–0.95. Overall accuracy shows modest reductions in ablated variants relative to the complete model.

The full ML-ConvNet achieves the highest overall accuracy and class-wise F1-scores among the evaluated ablation configurations, while removal of individual components results in observable performance reductions.

The observed performance differences across ablation variants offer some indication of what each component contributes within the evaluated experimental setting.

MBRConv. The performance reduction associated with removing MBRConv is most apparent for glioma and meningioma discrimination in MRI and for benign nodule detection in CT. This pattern is broadly consistent with the observation that pathological features in these categories operate at different spatial scales: glioma boundaries tend to extend across larger image regions, while meningioma margins are more spatially confined. The parallel 1×1, 3×3, and 5×5 branches within MBRConv process the input at complementary receptive field sizes during training, and their consolidation at inference retains this multi-scale encoding within a single kernel. Whether this accounts fully for the observed performance differences cannot be determined from ablation results alone, though the t-SNE projections in Fig. 6 show somewhat reduced intra-class compactness in the w/o MBRConv configuration, particularly for the CT benign cluster.

Removing HDPA produces the most consistent performance reductions across modalities, particularly for minority and morphologically similar classes. The Global Average Pooling path aggregates spatial information across the entire feature map, which may help reduce responses to uninformative background regions in low-contrast MRI volumes. The Max Pooling path retains peak activation values, which may be relevant for spatially focal findings such as pulmonary nodules in CT and consolidation regions in X-ray. The multiplicative fusion of both paths in Eq. 9 means that high combined attention requires both global plausibility and local intensity,

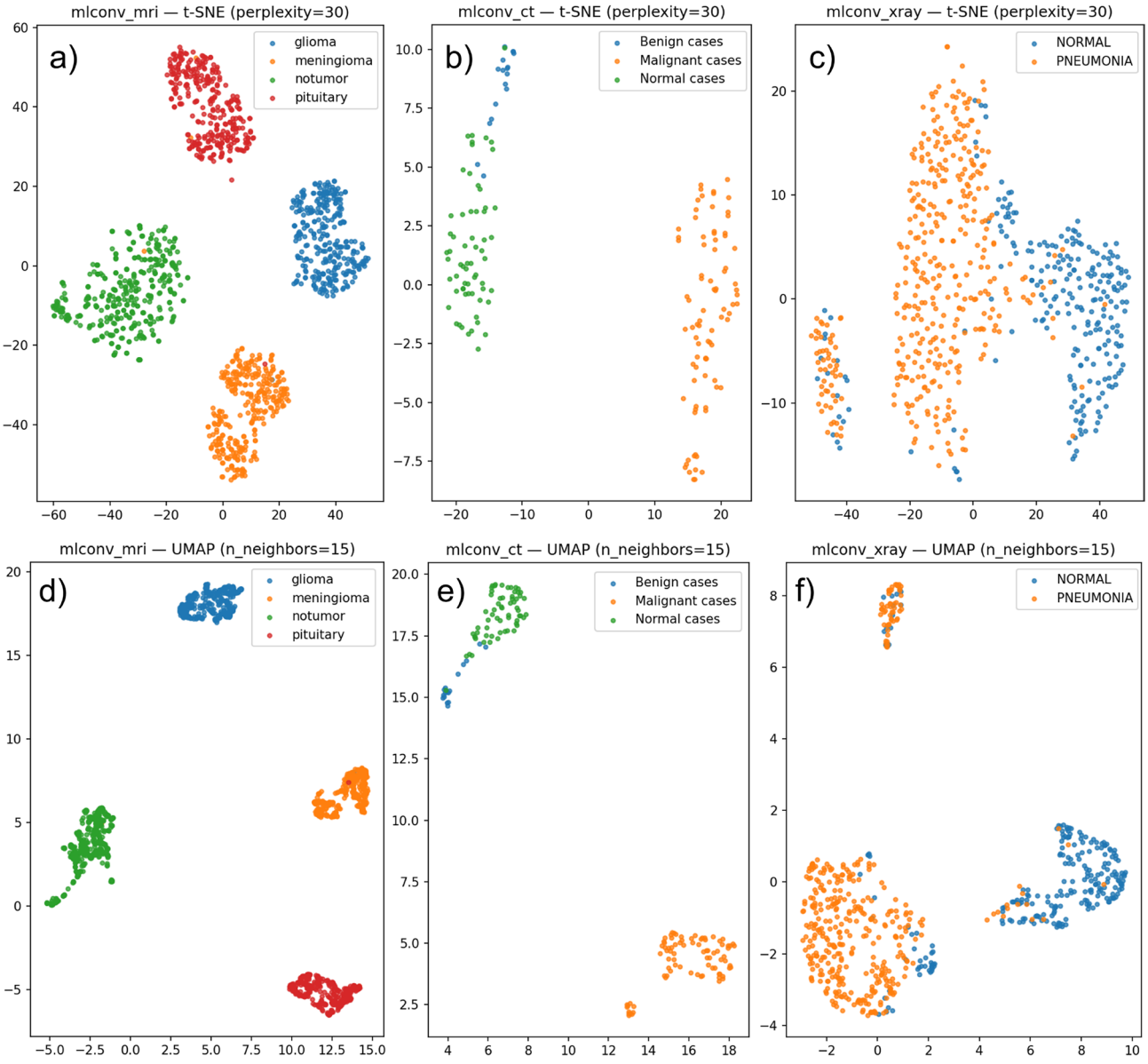

Fig. 6. t-SNE and UMAP visualizations of the feature embeddings learned by *ML-ConvNet* across three imaging modalities: (**a**, **d**) MRI brain tumors, (**b**, **e**) CT lung nodules, and (**c**, **f**) chest X-rays. The t-SNE manifolds (**a**–**c**) exhibit compact intra-class groupings and clear inter-class separation, demonstrating robust and modality-invariant representation learning. The UMAP projections (**d**–**f**) further reveal distinct, non-overlapping clusters with strong intra-class compactness, pronounced inter-class margins, and consistent topological coherence across modalities, validating the discriminative power of the learned feature space.

though whether this precisely reflects the intended behavior in practice is difficult to verify without controlled experiments beyond the scope of this study. The attention maps in Fig. 9 show that activations tend to concentrate over regions associated with pathological findings across modalities, which is qualitatively consistent with the intended localization behavior.

The performance reduction upon reverting to standard cross-entropy is most evident for the benign CT class, the most severely imbalanced category across the three cohorts. Standard cross-entropy does not account for whether individual samples are representative of their class distribution or potential outliers in latent space. Instance-level weights computed from neighborhood loss variance in the embedding space down-weight samples that are inconsistent with their local neighborhood while up-weighting more coherent examples, which is intended to reduce the influence of potentially noisy or ambiguous samples during optimization. The UMAP projections in Fig. 6 suggest somewhat tighter clustering of the benign CT class in the full model relative to the w/o LVW variant, though this visual observation should be interpreted cautiously given the limitations of two-dimensional embedding projections.

The feature synergy module produces the smallest individual contribution across ablation configurations, though its effect is consistent in direction across modalities and class categories. The Hadamard product self-interaction introduces multiplicative feature interactions without adding convolutional parameters, which may improve the separability of the classification boundary in the shared latent space to a limited degree. The pattern of small but consistent improvement across configurations suggests a complementary refinement role rather than a primary source of discriminative capacity within this architecture.

Latent manifold structure and cluster validity

Figure 6 presents t-SNE and UMAP projections of the learned embeddings across the three cohorts. In the MRI dataset (a,d), visually distinct clusters are observed. The No Tumor class appears isolated, while glioma, meningioma, and pituitary tumors form compact but proximally positioned groups, suggesting separability with limited visual overlap.

For the CT cohort (b,e), malignant cases form a compact cluster that appears visually separated from benign and normal samples. Benign and normal cases are also distinguishable, though their spatial proximity is greater than that of malignant cases, reflecting moderate inter-class similarity in the projected space.

In the pediatric chest X-ray cohort (c,f), two dominant clusters corresponding to Normal and Pneumonia are visible. Minor overlap is present at cluster boundaries, while class separation remains visually apparent in both projections.

Across modalities, the projected embeddings exhibit intra-class compactness and inter-class margins in both t-SNE and UMAP visualizations, suggesting organized latent representations within the evaluated datasets.

Hierarchical feature learning and attention-based interpretability

Figure 7 illustrates the layer-wise feature responses of ML-ConvNet. Early layers capture low-level structural boundaries and intensity gradients. Intermediate representations emphasize anatomical shape and lesion morphology, where partial overlap between glioma and meningioma reflects their visual similarity. Deeper layers exhibit sparse and localized activations, concentrating on tumor cores or pulmonary lesions while reducing background responses. This progression suggests hierarchical feature abstraction within the compact architecture.

Attention maps generated by the HDPA module (Figs. 8 and 9) show spatial correspondence with regions typically associated with pathological findings. In MRI, activations are localized over tumor regions and appear diffuse for normal studies. In CT, attention is concentrated within lung fields, with malignant nodules producing more focal responses relative to benign cases. In chest X-rays, activations are centered over pneumonia-related opacities, while normal scans show broader lung-field distribution.

Collectively, the hierarchical feature maps and attention visualizations suggest that ML-ConvNet's predictions are associated with anatomically relevant image regions rather than uniformly distributed background features. The observed consistency across modalities supports the interpretability of the learned representations within the evaluated datasets.

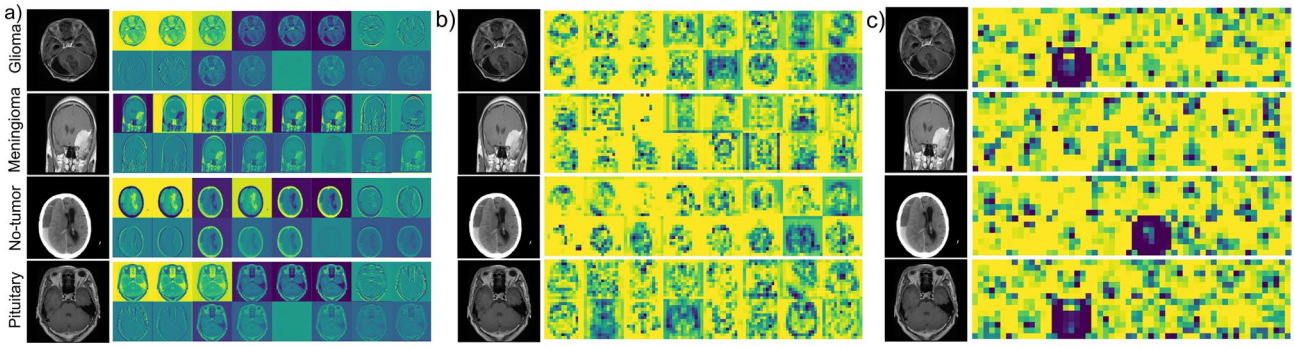

Fig. 7. Feature maps at different layers of *ML-ConvNet*: (**a**) Conv1_relu, (**b**) block_6_depthwise_relu, and (**c**) block_13_depthwise_relu.

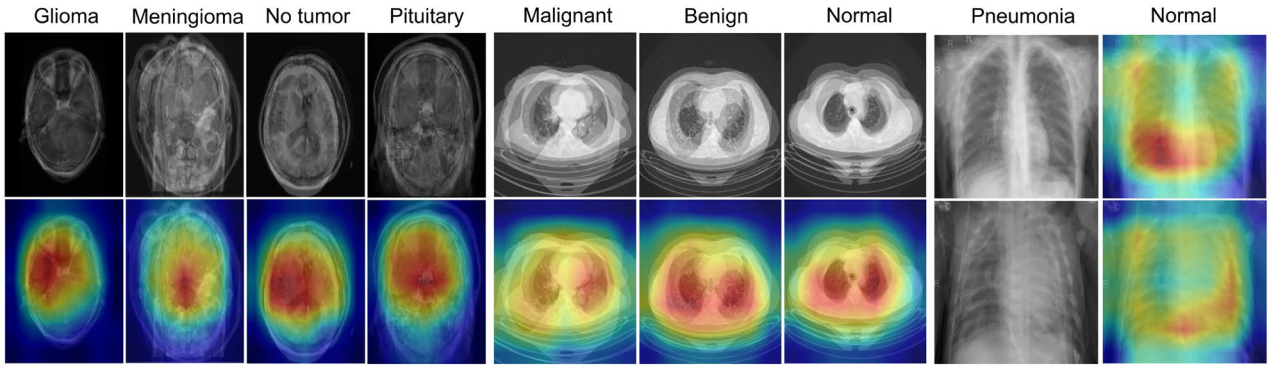

Fig. 8. Per-class mean HDPA attention maps for *ML-ConvNet* across modalities. The heatmaps highlight regions where the model focuses during inference (red = higher attention). The dataset includes four MRI brain classes (Glioma, Meningioma, No Tumor, Pituitary), three CT lung classes (Malignant, Benign, Normal), and two chest X-ray classes (Pneumonia, Normal).

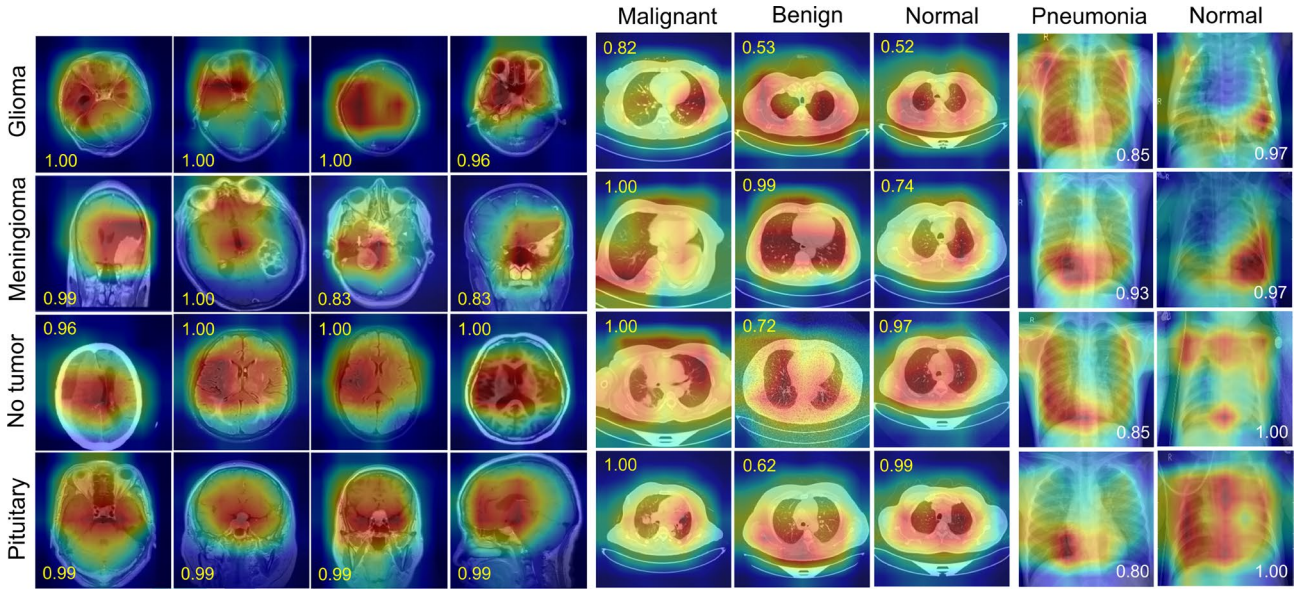

Fig. 9. Representative HDPA attention heatmaps for *ML-ConvNet* across modalities.

Computational efficiency analysis

Tables 12 and 13 summarize the computational profile of ML-ConvNet.

The model contains 4.2K parameters and requires 924M FLOPs at 512×512 input resolution. Measured latency is 2.3 ms on CPU, 0.21 ms on GPU, and 6.1 ms on mobile hardware (Snapdragon 888), with a memory footprint of 12.5 MB. Baseline FLOPs for ShuffleNetV2 (40M), MobileNetV2 (300M), and EfficientNet-B0 (390M) are reported at 224×224 input resolution in their respective publications, and direct numerical comparison is therefore complicated by the difference in operating resolution. The parameter count of 4.2K versus millions of parameters in these baselines is resolution-independent and constitutes the primary efficiency claim of this work. Within the evaluated settings, classification accuracy remains competitive relative to these baselines, as reflected in Table 13, while the parameter count is substantially reduced.

Discussion

This study shows that, within the evaluated experimental settings, a compact, task-aligned architecture can achieve competitive diagnostic performance without reliance on large-scale parameterization. By integrating MBRConv, HDPA, LVW, and FST, ML-ConvNet maintains stable performance across MRI, CT, and chest X-ray datasets while operating with approximately 4.2K parameters and 924M FLOPs at 512×512 input resolution. As reflected in the class-wise results (Tables 3, 4, 5) and fold-wise cross-validation summaries (Tables 7, 8), performance remains consistent under repeated sampling. These results suggest that, within the evaluated

Model	Params (K)	FLOPs (M)	CPU (ms)*	GPU (ms)**	Mobile (ms)***	Memory (MB)
ShuffleNetV2-0.5×	140	40	6.8	0.9	10.5	30.0
MobileNetV2	2,200	300	25.4	2.4	15.2	120.0
EfficientNet-B0	5,300	390	31.7	3.2	18.5	140.0
ML-ConvNet (Ours)	4.2	924†	2.3	0.21	6.1	12.5

Table 12. Benchmarking computational efficiency. Comparison of ML-ConvNet against state-of-the-art lightweight architectures (Batch Size = 1). FLOPs for ML-ConvNet are computed at 512×512 input resolution, while baseline FLOPs are reported at 224×224 in their respective publications; direct numerical comparison is therefore resolution-dependent. The primary efficiency advantage of ML-ConvNet lies in its parameter count, which is resolution-independent. * CPU: Intel Core i7-11800H @ 2.30GHz. ** GPU: NVIDIA GeForce RTX 4060. *** Mobile: Snapdragon 888 (ONNX Runtime, FP16). † ML-ConvNet FLOPs computed at 512×512 input; baseline FLOPs at 224×224

Model architecture	Ref.	Params (Millions)	FLOPs (Millions)	Diagnostic accuracy†		
				MRI	CT	X-ray
MobileNetV2	[16]	2.20	300.0	0.958[36]	0.975[37]	0.941[26]
ShuffleNetV2 (0.5×)	[17]	1.40	41.0	0.942[36]	0.952[37]	0.938[26]
EfficientNet-B0	[22]	5.30	390.0	0.971[38]	0.984[38]	0.955[39]
GhostNet	[23]	5.20	141.0	0.965[23]	0.970[23]	0.948[39]
SqueezeNet v1.1	[18]	1.23	352.0	0.912[36]	0.938[37]	0.915[26]
NASNetMobile	[21]	5.30	564.0	0.950[40]	0.962[37]	0.930[26]
ML-ConvNet (Ours)	–	0.004	924‡	0.989	0.986	0.976

Table 13. Benchmarking against State-of-the-Art Lightweight Architectures. Comparison of parameter efficiency, computational cost (FLOPs), and diagnostic accuracy across the three test modalities. Parameter count reflects published architectural specifications and is resolution-independent. FLOPs for baseline models are reported at 224×224 input resolution in their respective publications; ML-ConvNet FLOPs are computed at 512×512. Accuracy figures are sourced from independent published evaluations and should be treated as contextual reference points rather than controlled benchmarks (see table note). †*Note:* Baseline accuracy figures are drawn from previously published evaluations conducted on the same public benchmark datasets under independent experimental protocols, which may differ from the present study in data splitting strategy, preprocessing pipeline, augmentation policy, and training configuration. Consequently, direct numerical accuracy comparisons should be interpreted as contextual reference points rather than controlled benchmarks. Parameter count figures reflect original published architectural specifications and are resolution-independent, providing the most reliable basis for efficiency comparison. ‡ML-ConvNet FLOPs computed at 512×512 input resolution; all baseline FLOPs reported at 224×224. Readers are encouraged to consult the cited sources for full experimental details

experimental settings, representational efficiency may be influenced by architectural inductive bias in addition to parameter count.

Benchmark comparisons summarized in Table 13 further contextualize this finding. Architectures such as MobileNetV2[16], ShuffleNetV2[17], GhostNet[23], EfficientNet-B0[22], SqueezeNet[18], and NASNetMobile[21] have demonstrated effectiveness in natural image recognition and selected medical benchmarks[36–39]. However, these networks typically operate with parameter counts ranging from 1M to over 20M. ML-ConvNet contains approximately three orders of magnitude fewer parameters than EfficientNet-B0, a difference that reflects architectural design rather than experimental protocol and is therefore not subject to the comparison caveats noted above. Accuracy figures in Table 13 are drawn from independent published evaluations under differing protocols and should be read as contextual reference points rather than controlled benchmarks. Within the present evaluation framework, the reduced parameter scale does not appear to coincide with substantial deterioration in minority-class sensitivity or precision, as reflected by class-wise F1-scores and PR-AUC values (Tables 4 and 5).

The ablation analysis (Tables 9, 10, 11) clarifies the functional contributions of individual components. Removal of HDPA produces measurable declines in class-wise F1-scores, particularly for morphologically similar or minority categories, indicating its contribution to spatial selectivity. Replacing MBRConv with standard convolutions results in reduced separation between clinically adjacent classes, supporting the importance of multi-branch receptive field diversity. Reverting from LVW to standard cross-entropy leads to modest sensitivity reductions in imbalanced classes, suggesting a stabilizing effect of variance-aware optimization. Excluding FST yields smaller but consistent degradations, implying that cross-feature interaction refines the learned

representation. Across ablation configurations, the complete model achieves the highest overall accuracy within the evaluated experimental setup.

The interpretability results align with these quantitative observations. Layer-wise feature maps (Fig. 7) illustrate a progression from low-level edge and texture responses to localized activations in deeper layers. HDPA attention maps (Fig. 9) frequently highlight regions associated with pathological findings, including tumor cores in MRI, pulmonary nodules in CT, and consolidation patterns in X-ray imaging. Embedding projections using t-SNE and UMAP (Fig. 6) show structured intra-class clustering and visual separation between major categories in the projected space. Although such visualizations do not constitute formal proof of separability, they provide qualitative support for the organization of the learned latent representations.

From a computational perspective, the efficiency profile summarized in Table 12 indicates reduced parameter count, FLOPs, and inference latency relative to larger backbone architectures. Measured inference times (2.3 ms on CPU, 0.21 ms on GPU, and 6.1 ms on mobile hardware) suggest feasibility for real-time inference under the tested configurations. Compared to heavier architectures previously applied in similar domains[36,38], the proposed framework substantially lowers computational requirements while maintaining competitive classification accuracy. This may broaden applicability in point-of-care systems and settings with limited hardware resources.

A closer examination of the quantitative results reveals modality-specific behavior. In the MRI cohort (Table 3), balanced accuracy and specificity remain consistently above 0.98 across tumor subtypes, with narrow standard deviations under ten-fold cross-validation (Table 8), indicating stable performance under repeated sampling. In contrast, the CT cohort (Table 4) exhibits comparatively lower F1-scores for the benign class, reflecting the influence of class imbalance and morphological overlap, while the malignant class maintains high specificity and AUC values. The pediatric chest X-ray cohort (Table 5) demonstrates closely matched precision and recall between pneumonia and normal categories, with low fold-wise variance (Table 7), suggesting balanced discrimination despite projection-based imaging variability.

The cross-validation confidence intervals across all cohorts remain narrow relative to their mean values, supporting reproducibility under repeated partitioning. Embedding projections (Fig. 6) show that intra-class compactness is strongest in MRI, moderate in CT, and partially overlapping in X-ray, consistent with the relative complexity of the respective classification tasks. Feature maps and HDPA attention visualizations (Figs. 7 and 9) further illustrate progressive spatial refinement, with deeper layers concentrating activation within lesion-associated regions. While these visual analyses are qualitative, they align with the quantitative stability observed in fold-wise and class-wise evaluations.

Finally, the consistent performance observed across MRI, CT, and X-ray datasets suggests that the model captures structural descriptors that are not strictly modality-dependent. While imaging physics differ substantially across modalities, characteristics such as boundary irregularity, spatial asymmetry, and tissue heterogeneity appear to be represented within a shared compact feature space. The results do not imply full modality invariance; however, they suggest that a unified lightweight backbone can generalize across the evaluated heterogeneous imaging contexts when appropriately regularized and normalized.

Collectively, the findings suggest three main observations. First, compact architectures incorporating domain-specific inductive biases can achieve competitive classification performance in medical imaging tasks. Second, attention-guided and variance-aware mechanisms appear to improve class balance and transparency without increasing parameter complexity. Third, architectural calibration may provide an alternative perspective to continued parameter scaling for clinically specialized tasks within the evaluated settings. These findings suggest that future medical AI systems may benefit from principled structural design rather than reliance on progressively larger backbone networks.

Limitations and clinical outlook

The framework was evaluated exclusively on publicly available benchmark datasets, which may not fully represent the acquisition variability and patient diversity encountered in real-world clinical environments. External multi-center validation was not conducted, and generalizability across different imaging devices and institutional settings remains an open question.

Interpretability was assessed qualitatively through attention visualization rather than through structured clinician evaluation or quantitative localization metrics, and should not be interpreted as formal evidence of diagnostic transparency. The attention heatmaps in Figs. 8 and 9 are intended to serve a qualitative explanatory function, illustrating the spatial regions that contribute most to the model's classification decision. In line with established practice in attention-based classification studies, these visualizations are presented without ground truth overlays, as they are not designed for pixel-level localization evaluation. Quantitative performance is assessed separately through classification metrics reported in Tables 3, 4, 5, 6, and 7. Formal quantitative localization validation, such as Intersection over Union against expert segmentation contours, was not conducted and remains a direction for future work.

The cross-validation confidence, prediction, and tolerance intervals characterize performance stability under repeated sampling within the evaluated benchmark conditions. They do not provide guarantees of performance under distribution shift, scanner variability, or real-world operational constraints, and the gap between benchmark evaluation and clinical deployment cannot be assessed from public datasets alone.

The 4.2K parameter budget imposes structural constraints that are worth considering explicitly. The network operates with six feature channels throughout, which limits the representational diversity available at each layer. For the three binary or low-cardinality classification tasks evaluated here, this channel capacity appears sufficient under the tested conditions. However, tasks requiring finer-grained discrimination, such as multi-subtype pathology grading, histological pattern recognition, or classification across a larger number of classes, may exceed what this channel width can reliably encode. Similarly, the absence of any spatial downsampling means that the receptive field of the deepest convolutional layers remains constrained relative to deeper architectures,

which may limit sensitivity to global structural patterns in larger or more complex images. The architecture was designed and evaluated within a specific operating envelope: three publicly available datasets, binary or low-cardinality targets, and 512×512 inputs. Applying it outside this envelope, particularly to multi-label tasks, volumetric inputs, or datasets with substantially different acquisition characteristics, should be approached with caution and validated independently before drawing conclusions about generalizability.

Future work will focus on external multi-institutional validation, quantitative interpretability assessment, and hardware-aware optimization for embedded deployment.

Conclusion and future work

This study presented a compact and interpretable multi-source framework for heterogeneous medical image classification across MRI, CT, and X-ray modalities. The proposed model integrates multi-branch structural encoding, hierarchical attention, variance-aware optimization, and cross-feature interaction within a lightweight design. Experimental evaluation on publicly available benchmark datasets indicates competitive and relatively consistent performance relative to recent approaches while maintaining a reduced parameter count.

The findings suggest that careful architectural design and domain-informed inductive biases may support effective representation learning without reliance on large-scale parameterization. However, the framework was evaluated exclusively on publicly available datasets, and external multi-center validation was not conducted. As such, further investigation is required to assess robustness under varying acquisition protocols, scanner variability, and real-world clinical conditions.

Future work will focus on multi-institutional validation, quantitative evaluation of interpretability, robustness analysis under distribution shifts, and exploration of privacy-preserving training strategies such as federated learning. Hardware-aware optimization and model compression techniques will also be examined to better understand deployment feasibility in resource-constrained environments.

Data availability

The datasets utilized in this study are publicly available from their respective repositories: (i) the Nickparvar Brain Tumor MRI dataset is available via Kaggle at https://doi.org/10.34740/kaggle/dsv/14832123; (ii) the IQ-OTH/NCCD Lung Cancer CT dataset is available via Kaggle at https://doi.org/10.34740/kaggle/ds/672399; and (iii) the Pediatric Chest X-ray (Pneumonia) dataset is available via Mendeley Data at https://data.mendeley.com/datasets/rscbjbr9sj/2. These datasets are publicly accessible for research purposes under the terms specified by their respective repositories. No new datasets were generated during the current study.

Code availability

The source code for ML-ConvNet, including model definitions, training scripts, preprocessing pipelines, and evaluation protocols, is publicly available at https://github.com/W-ayivi/ML-ConvNet. Researchers wishing to reproduce the reported results or build upon this work may access the repository directly.

Received: 21 February 2026; Accepted: 25 April 2026
Published online: 02 May 2026

References

1. Aklan, B. et al. Efficient artificial intelligence approaches for medical image processing in healthcare: comprehensive review, taxonomy, and analysis. *Artif. Intell. Rev.* **57**, 1–63 (2024).
2. Chen, X. et al. Artificial intelligence and multimodal data fusion for smart healthcare: topic modeling and bibliometrics. *Artif. Intell. Rev.* **57**(4), 91 (2024).
3. Salmi, M. et al. Handling imbalanced medical datasets: Review of a decade of research. *Artif. Intell. Rev.* **57**(10), 273 (2024).
4. Jain, A. et al. Cost-sensitive learning for imbalanced medical data: A review. *Artif. Intell. Rev.* **57**(4), 1–35 (2024).
5. Asif, M. et al. Litefusionnet: A lightweight fusion network for efficient medical image classification. *Comput. Biol. Med.* **169**, 107916. https://doi.org/10.1016/j.compbiomed.2024.107916 (2024).
6. Zamil, M., Begum, R.A., Hossain, M.A., et al. Lightmri: A lightweight convolutional neural network model for multi-class brain tumor classification. In: Proceedings of the international conference on computer applications (ICCA) (2024).
7. Hossain, A. et al. A lightweight deep learning based microwave brain image network model for brain tumor classification using reconstructed microwave brain (rmb) images. *Biosensors* **13**(3), 238. https://doi.org/10.3390/bios13030238 (2023).
8. Wan, G. & Yao, L. Lmfrnet: A lightweight convolutional neural network model for image analysis. *Electronics* **13**(129), 1–16. https://doi.org/10.3390/electronics13010129 (2024).
9. Tuncer, I., Dogan, S. & Tuncer, T. Mobiledensenext: Investigations on biomedical image classification. *Expert Sys. Appl.* **255**, 124685. https://doi.org/10.1016/j.eswa.2023.124685 (2024).
10. Hammad, M., ElAffendi, M., Ateya, A. A. & Abd El-Latif, A. A. Efficient brain tumor detection with lightweight end-to-end deep learning model. *Cancers* **15**(11), 2837. https://doi.org/10.3390/cancers15112837 (2023).
11. Liu, S., Yue, W., Guo, Z. & Wang, L. Multi-branch CNN and grouping cascade attention for medical image classification. *Scientif. Rep.* **14**, 15013. https://doi.org/10.1038/s41598-024-15013-7 (2024).
12. Salehin, I., Islam, M.S., Amin, N., et al. Real-time medical image classification with ml framework and dedicated CNN–LSTM architecture. J. Sens. 3717035 (2023). https://doi.org/10.1155/2023/3717035
13. Yan, H., Li, A., Zhang, X., Liu, Z., Shi, Z., Zhu, C. & Zhang, L. MobileIE: An extremely lightweight and effective ConvNet for real-time image enhancement on mobile devices. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV), pp. 21949–21960 (2025).
14. Aklan, B. et al. Efficient artificial intelligence approaches for medical image processing in healthcare: comprehensive review, taxonomy, and analysis. *Artif. Intell. Rev.* **57**, 1–63 (2024).
15. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014).
16. Sandler, M., et al. Mobilenetv2: Inverted residuals and linear bottlenecks. In: CVPR (2018).
17. Ma, N., et al. Shufflenet v2: Practical guidelines for efficient CNN architecture design. In: ECCV (2018).

18. Iandola, F.N., et al. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. arXiv:1602.07360 (2016).
19. Ayivi, W., Zhang, X., Agbley, B.L.Y., Ativi, W.X., Sam, F. & Browne, J.A. Re-parameterized lightresnet for real-time multi-class brain tumor classification. In: 2025 22nd international computer conference on wavelet active media technology and information processing (ICCWAMTIP), pp. 1–5 (2025). IEEE
20. Ayivi, W., Zhang, X., Ativi, W. X., Sam, F. & Kouassi, F. A. Dynamic-attentive pooling networks: A hybrid lightweight deep model for lung cancer classification. *J. Imaging* **11**(8), 283 (2025).
21. Zoph, B., et al. Learning transferable architectures for scalable image recognition. In: CVPR (2018).
22. Tan, M. & Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In: ICML (2019).
23. Han, K., et al. Ghostnet: More features from cheap operations. In: CVPR (2020).
24. Ding, X., Zhang, X., Ma, N., Han, J., Ding, G. & Sun, J. Repvgg: Making vgg-style convnets great again. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), 13733–13742 (2021). https://doi.org/10.1109/CVPR46437.2021.01354
25. Chen, X. et al. Artificial intelligence and multimodal data fusion for smart healthcare: Topic modeling and bibliometrics. *Artif. Intell. Rev.* **57**(4), 91 (2024).
26. Kermany, D. S. et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **172**(5), 1122–1131 (2018).
27. Salmi, M. et al. Handling imbalanced medical datasets: Review of a decade of research. *Artif. Intell. Rev.* **57**(10), 273 (2024).
28. Karimi, D. et al. Deep learning with noisy labels in medical prediction problems: A scoping review. *J. Am. Med. Inform. Assoc.* **31**(7), 1596–1612 (2024).
29. Araf, I., Idri, A. & Chairi, I. Cost-sensitive learning for imbalanced medical data: A review. *Artif. Intell. Rev.* **57**, 80 (2024).
30. Jain, A. et al. Cost-sensitive learning for imbalanced medical data: A review. *Artif. Intell. Rev.* **57**(4), 1–35 (2024).
31. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision (ICCV), pp. 2980–2988 (2017). https://doi.org/10.1109/ICCV.2017.324
32. Cao, K., Wei, C., Gaidon, A., Arechiga, N.V. & Ma, T. Learning imbalanced datasets with label-distribution-aware margin loss. In: Advances in neural information processing systems (NeurIPS), vol. 32, pp. 1567–1578 (2019). https://proceedings.neurips.cc/paper/2019/hash/621461af90cadfdaf0e8d4cc25129f91-Abstract.html
33. Nickparvar, M. Brain Tumor MRI Dataset. https://www.kaggle.com/dsv/2645886. Accessed 2025-09-10 (2021). https://doi.org/10.34740/KAGGLE/DSV/2645886
34. Alyasriy, H. & Muayed, A. The IQ-OTHNCCD lung cancer dataset. *Mendeley Data* **1**(1), 1–13 (2020).
35. Kermany, D. S. et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **172**(5), 1122–1131. https://doi.org/10.1016/j.cell.2018.02.010 (2018).
36. Deepak, S. & Ameer, P. Brain tumor classification using deep cnn. In: 2019 International conference on computers and devices for communication (CODEC) (2019). IEEE
37. To, G., et al. Deep learning for lung cancer detection on CT scans: A review. Journal of Digital Imaging (2021).
38. Guan, Q., et al. Automated classification of lung cancer stages using efficientnet. BMC Medical Imaging (2024).
39. Mohamed, C., et al. Enhancing pneumonia detection using lightweight models. J. Comput. Commun. (2024).
40. Gupta, M. et al. Optimizing tumor detection in brain MRI with one-class SVM and convolutional neural network-based feature extraction. *J. Imaging* **11**(7), 207 (2023).

Author contributions

W.A. Conceptualization, Methodology, Data Curation, Formal Analysis, Writing—Original Draft, Writing—Review and Editing; Z.X. Funding acquisition, supervision, Writing - review & editing, validation; W.X.A. Data Curation, Validation, Writing - review & editing; F.S. Visualization, Writing - review & editing; A.A. Formal Analysis, Visualization, Writng -review & editing.

Funding

This work was supported by the National Natural Science Foundation of China under Grants 62471113 and the Sichuan Science and Technology Program 2024NS-FSC0479.

Declarations

Conflict of interest

The authors declare no conflict of interest.

Additional information

Correspondence and requests for materials should be addressed to W.A. or X.Z.