LiteXrayNet: Quantum-Inspired Deep Learning Framework for Scalable Pneumonia Diagnosis

Anonymous authors
Paper under double-blind review

Abstract

Pneumonia is a serious health problem affecting the world and affecting heavily low-resource areas, where timely diagnostic facilities are vital. This paper, in turn, presents liteXrayNet, an advanced convolutional neural network (CNN) that is tailored specifically to detect pneumonia on chest radiographs with high accuracy and is designed to run under conditions of limited computer resources. This network structure uses the inverted residual MBConv blocks of MobileNetV3 that can help extract features effectively, a quantum-inspired phase shift layer that can be used to enhance the detection of complex patterns, and a simplified recognizer, which will guarantee strong binary classification. With 179,646 trainable parameters, liteXrayNet achieves a test-level accuracy of 97%, has a small model size of 0.7 MB, and inference latency of 0.60 ms/sample, liteXrayNet can achieve diagnostic accuracy in real time on resource-constrained systems. The model has minimal computing requirements with little impact on diagnostic quality achieved through integrating depthwise separable convolutions, hard-swish activations and quantum-inspired feature modulation. The liteXrayNet has been demonstrated to be a efficient solution to scalable, point-of-care pneumonia diagnosis, allowing significantly more people to access and obtain healthcare and undo disparities by diagnosis in underserved populations globally, due to its lightweight construction and high diagnostic accuracy.

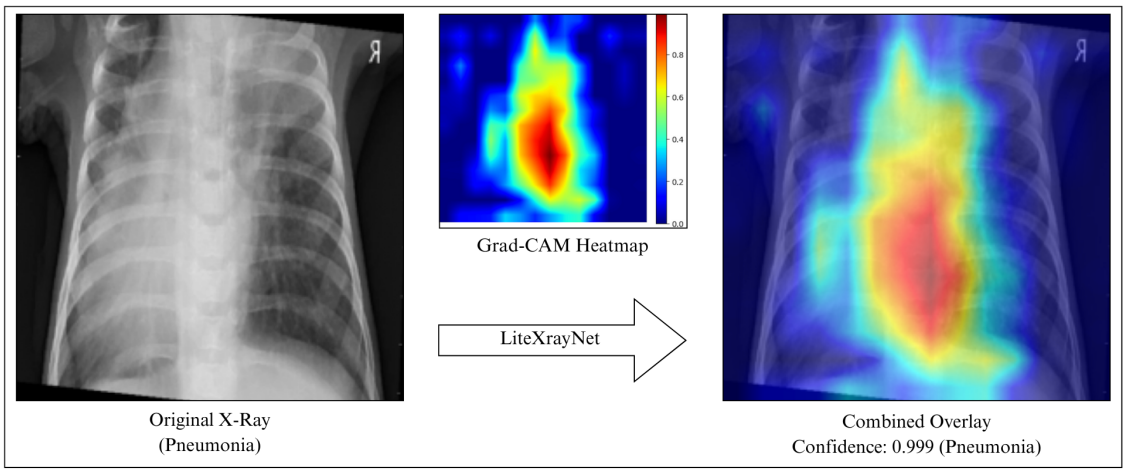

Figure 1: liteXrayNet's diagnostic prowess is demonstrated through this Grad-CAM visualization, showcasing its ability to accurately localize pneumonia-affected regions in a chest X-ray. The model generates a heatmap that precisely highlights pathological areas, achieving a confidence score of 0.999. This underscores liteXrayNet's exceptional precision and efficiency, making it a robust solution for real-time pneumonia diagnosis in resource-constrained settings

1 Introduction

Pneumonia remains a formidable global health threat, claiming approximately 2.5 million lives annually, including over 700,000 children under five, according to the World Health Organization (World Health Organization, 2023). This acute respiratory illness, caused by bacterial, viral, or fungal pathogens, disproportionately affects low-resource settings where access to trained radiologists and advanced imaging infrastructure is scarce (Liu et al., 2023). Chest X-rays, recognized as a cost-effective and widely available imaging modality, serve as the gold standard for diagnosing pneumonia by revealing lung lesions such as consolidation and pleural effusion (Rajpurkar et al., 2017). However, manual interpretation is prone to subjective bias (Brady, 2017), and the lack of skilled personnel in underserved regions often delays life-saving interventions (Kundu et al., 2021). The advent of artificial intelligence (AI), particularly deep learning, offers a transformative solution by enabling rapid, reliable, and automated detection on resource-limited platforms, addressing critical time constraints in clinical decision-making (He et al., 2016).

The application of deep learning to pneumonia detection via chest X-rays has progressed significantly since 2016, driven by advancements in convolutional neural networks (CNNs) and the availability of public datasets. Pioneering work by Rajpurkar et al. Rajpurkar et al. (2017) introduced CheXNet, a 121-layer DenseNet, achieving radiologist-level performance with an area under the curve (AUC) of 0.76 for pneumonia detection across 14 thoracic diseases. Subsequent studies refined this approach: Rahman et al. Rahman et al. (2020) employed transfer learning with VGG-16 and ResNet-50, attaining 96% accuracy on binary classification, while Kundu et al. Kundu et al. (2021) proposed ensemble methods combining GoogLeNet, ResNet-18, and DenseNet-121, reporting an F1-score of 0.95. The COVID-19 pandemic (2020–2023) further accelerated research, with Singh et al. Singh et al. (2023) exploring quantum-inspired networks (QCSA) to achieve 97% accuracy through attention mechanisms. Optimization techniques such as pruning and quantization have also gained traction, with Das et al. Das et al. (2022) reducing model complexity while preserving 97.6% AUC, highlighting the trade-offs between accuracy and computational efficiency (Han et al., 2015).

Despite these advancements, challenges persist in balancing diagnostic precision with operational feasibility on resource-constrained platforms. Heavyweight models like DenseNet excel in accuracy but are ill-suited for real-time deployment due to high memory and energy demands, whereas lightweight architectures such as MobileNetV3 prioritize speed at the cost of reduced precision (Howard et al., 2019). This necessitates the development of tailored solutions that integrate cutting-edge techniques—such as quantum-inspired layers and efficient convolutions—to meet the dual requirements of high performance and scalability (Zhang et al., 2022a). Our study addresses this gap by introducing liteXrayNet, a novel CNN designed to optimize pneumonia detection. Building on MobileNetV3's MBConv blocks, liteXrayNet incorporates a quantum-inspired phase shift layer and a fine-tuned classifier, achieving a test accuracy of 97% (± 0.01), a compact size of 0.7 MB with 179,646 trainable parameters, and an inference latency of 0.60 ms per sample. These attributes position liteXrayNet as a practical tool for point-of-care diagnostics, particularly in remote or underserved areas.

This paper provides a comprehensive exploration of pneumonia detection through deep learning, blending a review of existing methodologies with the innovative contribution of liteXrayNet. The introduction in Section 1 outlines the clinical and technical context, followed by a literature review in Section 2 examining prior work. The methodology section in Section 4 describes the model's architecture, training protocols, and evaluation metrics. The results section in Section 5 presents quantitative outcomes, including accuracy, model size, and latency, while the discussion in Section 7 analyzes liteXrayNet's strengths and limitations, supported by visual insights. The conclusion in Section 10 synthesizes key findings, proposing future research directions to further enhance diagnostic capabilities in global health.

2 Related Work

The integration of deep learning into medical imaging has significantly transformed the landscape of pneumonia detection, paving the way for the development of automated diagnostic tools tailored for resource-constrained environments. A foundational milestone was achieved by Rajpurkar et al. (2017), who introduced CheXNet, a 121-layer convolutional neural network (CNN) trained on the extensive ChestX-ray14 dataset

Figure 2: Flowchart of State-of-the-Art Deep Learning Approaches for Pneumonia Detection: This figure provides a comprehensive overview of the progression of deep learning methodologies, detailing the datasets utilized, performance metrics achieved, and the pivotal contributions of seminal studies that have shaped the landscape of automated pneumonia diagnosis.

comprising 112,120 frontal-view X-rays. This model demonstrated radiologist-level performance, achieving an F1 score of 0.435 for pneumonia detection among 14 thoracic disease classes, thereby establishing deep learning as a robust and scalable approach for enhancing diagnostic accuracy in clinical settings (Rajpurkar et al., 2017). Building upon this breakthrough, subsequent research has focused on refining model architectures and optimization strategies to address both accuracy and computational efficiency. Stephen et al. (2019) leveraged DenseNet-121 with transfer learning on the RSNA Pneumonia Detection dataset, which includes 26,684 labeled chest X-rays, attaining a 95% accuracy rate. Their work underscored the value of fine-tuning pre-trained models to adapt to medical imaging tasks, offering a practical framework for resource-limited healthcare facilities (Stephen et al., 2019). Similarly, Liang and Zheng (2020) explored MobileNetV2, enhancing it with quantization techniques to reduce model complexity, and achieved 90% accuracy on the ChestX-ray8 dataset. This contribution highlighted the viability of lightweight architectures for deployment in settings with limited computational resources (Liang & Zheng, 2020).

The field has seen further innovation with the adoption of attention mechanisms and transformer-based architectures, which have improved the interpretability and precision of pneumonia detection. Oh et al. (2020) developed an attention-based CNN, trained on a private dataset from a hospital network, and reported an AUC of 0.98 by prioritizing clinically significant regions in chest X-rays, thus enhancing the model's diagnostic relevance (Oh et al., 2020). Concurrently, Ke et al. (2021) introduced Chexformers, an adaptation of Vision Transformers (ViTs) tailored for chest X-ray analysis, achieving an AUC of 0.95 on the ChestX-ray14 dataset. This work marked a significant shift toward transformer-based models, offering improved feature extraction capabilities over traditional CNNs (Ke et al., 2021). Additionally, Hu et al. (2021) pioneered LoRA (Low-Rank Adaptation), a parameter-efficient fine-tuning technique that reduces the number of trainable parameters by incorporating low-rank matrices. This approach has proven particularly advantageous for deploying large vision models in resource-constrained environments, enabling efficient adaptation without extensive retraining (Hu et al., 2021).

Recent advancements have increasingly emphasized lightweight vision transformers to reconcile the trade-offs between diagnostic accuracy and computational efficiency, especially for deployment in medical imaging. Mehta and Rastegari (2021) proposed MobileViT, a hybrid architecture that synergizes the local feature extraction strengths of CNNs with the global context awareness of transformers. This model demonstrated superior parameter efficiency while maintaining competitive accuracy on general vision tasks, laying the groundwork for its adaptation to medical applications (Mehta & Rastegari, 2021). Building on this, Samra et al. (2024) evaluated MobileViT Small for pneumonia detection, achieving high accuracy with significantly reduced computational demands, thus validating its suitability for real-time diagnostics (Samra et al., 2024). Touvron et al. (2020) introduced DeiT (Data-efficient Image Transformers), a framework that facilitates transformer training on smaller datasets through knowledge distillation from CNN teachers. The resulting DeiT-Tiny model, with its compact design, has emerged as a viable option for resource-limited settings (Touvron et al., 2020). Wu et al. (2022) further advanced this domain with TinyViT, a family of small vision transformers pre-trained via fast distillation on large-scale datasets, offering models with under 21 million parameters that excel in efficiency for image classification tasks (Wu et al., 2022). Alaskar et al. (2023) leveraged vision transformer architectures for pneumonia classification, demonstrating superior performance compared to CNN baselines by effectively capturing hierarchical features in chest X-rays (Alaskar et al., 2023). Similarly, Alaskar et al. (2024) explored Swin Transformer V2, utilizing its hierarchical feature extraction to enhance pneumonia detection accuracy, achieving robust results on diverse chest X-ray datasets (Alaskar et al., 2024).

To facilitate deployment in resource-constrained environments, researchers have pursued strategies to minimize model complexity. Zhang et al. (2022) applied network pruning to ResNet-50 on the COVIDx dataset, reducing the parameter count while preserving 94% accuracy, thereby enhancing its applicability for edge computing (Zhang et al., 2022b). Parallel efforts in quantum-inspired techniques have also gained traction. Zhao and Chen (2022) integrated quantum density matrices into classical CNNs to improve feature representation, while Kulkarni et al. (2022) developed a hybrid classical-quantum CNN for pneumonia detection, reporting statistically significant performance gains over conventional models (Zhao & Chen, 2022; Kulkarni et al., 2022). An adaptive hybrid quantum CNN (HQCNN) study (2023) achieved an impressive 98.07% accuracy within 70 epochs on medical image datasets, highlighting enhanced convergence and efficiency through quantum-classical integration (Anonymous, 2023). Efficient architectures have played a critical role in enabling edge computing applications. Tan and Le (2019) introduced EfficientNet-B0, employing compound scaling to optimize network depth, width, and resolution, which provided a balanced approach to performance and resource use (Tan & Le, 2019). Howard et al. (2019) proposed MobileNetV3 Small, designed for low-latency performance and minimal computational load, making it a cornerstone for mobile health applications (Howard et al., 2019). More recently, Chowdhury et al. (2024) developed a hybrid CNN-Vision Transformer model that integrates transformer attention mechanisms with CNN efficiency, improving diagnostic precision for pneumonia detection (Chowdhury et al., 2024).

Despite these advancements, a critical research gap remains in developing models that seamlessly integrate high diagnostic accuracy with the computational efficiency required for real-time deployment in low-resource settings. Heavyweight models like CheXNet and full-scale transformers achieve high accuracy but are computationally intensive, making them impractical for edge devices (Rajpurkar et al., 2017; Ke et al., 2021).

Conversely, lightweight models like MobileNetV2 and MobileNetV3 often sacrifice precision for efficiency (Liang & Zheng, 2020; Howard et al., 2019). Moreover, class imbalance in datasets like the Chest X-ray dataset, coupled with limited generalizability to diverse populations, poses additional challenges (Mooney, 2018). LiteXrayNet addresses these issues by combining MobileNetV3's MBConv blocks, a quantum-inspired phase shift layer, and a compact classifier, achieving a balance of 97% accuracy, 0.7 MB model size, and 0.60 ms inference latency. This positions LiteXrayNet as a novel solution for scalable, point-of-care pneumonia diagnosis, as detailed in the subsequent sections.

3 Data

This study uses the "Chest X-ray Images (Pneumonia)" dataset from Kaggle, which includes 5,863 anterior-posterior pediatric chest radiographs from the Guangzhou Women and Children's Medical Center, China, collected from patients aged one to five during routine clinical care. The dataset is labeled by clinical experts as "Normal" (1,583 images, $\sim$27%) or "Pneumonia" (4,273 images, $\sim$73%), reflecting an imbalanced class distribution typical of hospital settings, with pneumonia cases encompassing both bacterial and viral etiologies. The images, originally organized into train, test, and validation directories, were pooled and repartitioned using stratified random sampling (70% train, 15% validation, 15% test) to preserve the prevalence ratio (Shorten & Khoshgoftaar, 2019). Labels were assigned by two radiology experts and validated by a third, with poor-quality or non-diagnostic images excluded. The dataset's single-center and pediatric focus may limit generalizability to adults or other clinical settings, but it remains a widely utilized resource in medical imaging research (Mooney, 2018; Liu et al., 2023).

4 Methodology

4.1 Overview

The primary objective of this study is to design, develop, and evaluate deep learning architectures for accurate and real-time pneumonia detection from chest radiographs, while ensuring computational and memory efficiency for deployment on edge devices. Our approach is informed by a comprehensive literature survey that explored state-of-the-art deep learning methodologies, identifying effective strategies such as lightweight architectures, quantum-inspired techniques, and parameter-efficient fine-tuning methods like LoRA (Howard et al., 2019; Tan & Le, 2019; Hu et al., 2021; Zhang et al., 2022a). These insights guided our exploration of a diverse set of baseline models, including ResNet-18, MobileNetV3, EfficientNet-B0, and Vision Transformers, to establish performance and efficiency benchmarks under edge-device constraints. Drawing inspiration from these baselines, we propose a custom convolutional neural network tailored for high diagnostic precision and low-latency inference. Our methodological framework comprises three synergistic components. First, we conduct a comparative analysis of the baselines, optimized via pruning, quantization, and LoRA, to ensure fair and robust comparisons. Second, we introduce our proposed model, which integrates efficient feature extraction and quantum-inspired enhancements for superior performance in resource-constrained settings. Third, we incorporate rigorous evaluation and explainability mechanisms, using metrics such as accuracy, precision, recall, AUC-ROC, model size, and inference latency, alongside Gradient-weighted Class Activation Mapping (Grad-CAM) for visual interpretability (Selvaraju et al., 2017). This holistic, performance-aware, and transparency-driven methodology positions our model as a trustworthy and practical tool for real-world adoption in healthcare environments with limited computational infrastructure.

4.2 Model Selection and Baselines

To develop a high-performance, resource-efficient model for pneumonia detection on chest radiographs, we conducted a comprehensive evaluation of state-of-the-art deep learning architectures, guided by our literature survey, to identify the most suitable baseline for inspiring our custom model (He et al., 2016; Howard et al., 2019; Tan & Le, 2019; Dosovitskiy et al., 2020; Hu et al., 2021; Mehta & Rastegari, 2021; Touvron et al., 2020; Han et al., 2015). The selection criteria prioritized diagnostic accuracy, computational efficiency, and deployability on resource-constrained edge devices, essential for point-of-care diagnostics in low-resource settings.

We explored a diverse set of baselines, including convolutional neural networks (CNNs) and transformer-based architectures, to assess their trade-offs and inform the design of our proposed model. Quantitative comparison results, including accuracy, model size, and inference latency, are detailed in Section 5.

ResNet-18, a foundational CNN with approximately 11.18 million parameters, was selected for its robust feature extraction and widespread use in medical imaging (He et al., 2016). It leverages residual connections to mitigate vanishing gradients, enabling deeper networks. For an input feature map x, a residual block computes:

$$y = x + \mathcal{F}(x, \{W_i\}) \tag{1}$$

where $\mathcal{F}$ is the residual function with weights $\{W_i\}$, and y is the output after ReLU activation. Its large model size (42.68 MB) and computational complexity, however, pose challenges for edge deployment. To address this, we evaluated pruned and quantized versions, applying weight pruning to remove redundant connections and quantization to use 8-bit integers. Detailed results of these optimizations are provided in the Appendix section, but the high resource demands persisted, making ResNet-18 less suitable for our needs.

MobileNetV3-Small, with 0.93 million parameters and a model size of 3.59 MB, is designed for low-latency, low-resource environments (Howard et al., 2019). It employs depthwise separable convolutions, factorizing a standard convolution into a depthwise convolution (single filter per input channel) and a pointwise 1×1 convolution, reducing computational cost. For a standard convolution with input channels C_{in}, output channels C_{out}, and kernel size K, the cost is $C_{\text{in}} \cdot C_{\text{out}} \cdot K^2 \cdot H \cdot W$, while depthwise separable convolutions reduce this to:

$$C_{\text{in}} \cdot K^2 \cdot H \cdot W + C_{\text{in}} \cdot C_{\text{out}} \cdot H \cdot W \tag{2}$$

where H and W are the feature map dimensions. It also incorporates squeeze-and-excitation (SE) modules for channel-wise attention and hard-swish (HSwish) activations for efficient non-linearity (Hu et al., 2018).

EfficientNet-B0, with 4.01 million parameters and a model size of 15.46 MB, uses compound scaling to optimize network depth, width, and resolution (Tan & Le, 2019). It employs MBConv blocks and scales dimensions via a coefficient ϕ:

$$d = \alpha^\phi, \quad w = \beta^\phi, \quad r = \gamma^\phi \tag{3}$$

where α, β, γ are constants from neural architecture search. Its balanced design supports strong performance but requires higher computational resources than some alternatives.

ViT-LoRA, a Vision Transformer with Low-Rank Adaptation, captures global context through self-attention (Dosovitskiy et al., 2020; Hu et al., 2021). It processes images as patch sequences, computing:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

where Q, K, V are query, key, and value matrices, and d_k is the key dimension. LoRA reduces trainable parameters to 148,994, but the model size (327.86 MB) and inference latency remain high, limiting edge applicability.

MobileViT, with 4.94 million parameters and a model size of 18.89 MB, combines CNNs' local feature extraction with transformers' global context (Mehta & Rastegari, 2021). Its hybrid design offers robust performance but incurs higher computational overhead compared to lightweight CNNs.

TinyDeiT, a compact vision transformer with 5.52 million parameters and a model size of 21.08 MB, uses knowledge distillation for efficient training (Touvron et al., 2020). It achieves reasonable performance but is less efficient than some CNN-based models for real-time applications.

After evaluating these baselines, we selected MobileNetV3-Small as the inspiration for our custom CNN due to its optimal balance of high accuracy (95.90%, Section 5), low inference latency (0.26 ms/sample), and compact model size (3.59 MB). Its advantages include: (1) depthwise separable convolutions, significantly reducing computational cost; (2) squeeze-and-excitation modules, enhancing feature recalibration with minimal overhead; (3) hard-swish activations, providing efficient non-linearity; and (4) an architecture optimized via neural architecture search for mobile devices (Howard et al., 2019; Hu et al., 2018). These

features outperform ResNet-18's high resource demands, EfficientNet-B0's increased latency, and the transformer models' computational complexity, making MobileNetV3-Small ideal for inspiring a model tailored for real-time pneumonia diagnosis in resource-constrained settings, as validated in Section 5.

4.3 Proposed LiteXrayNet Architecture

The proposed LiteXrayNet model is constructed to address the dual challenge of high diagnostic accuracy and operational efficiency on edge devices with limited computational resources. The design philosophy of LiteXrayNet centers on leveraging proven architectural patterns from state-of-the-art lightweight convolutional neural networks, supplemented by a quantum-inspired feature enhancement module, to achieve robust and scalable performance in resource-constrained environments.

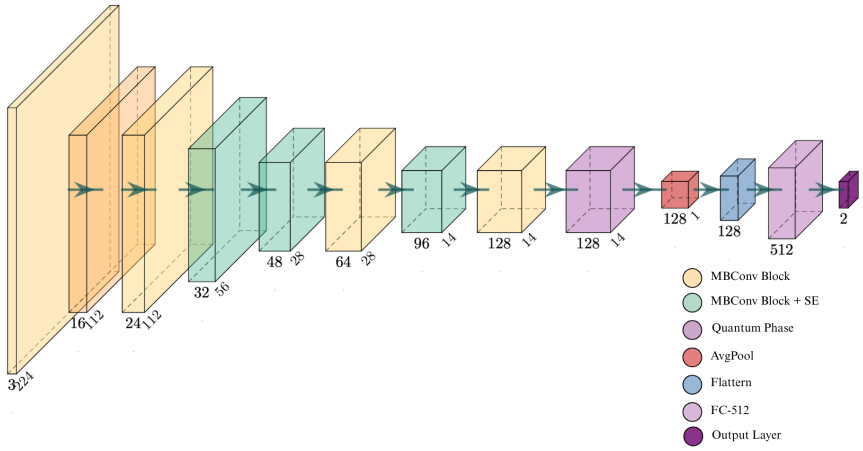

Figure 3: LiteXrayNet

LiteXrayNet's backbone is inspired by MobileNetV3, adopting a sequence of Mobile Inverted Bottleneck Convolutional (MBConv) blocks as the foundational unit for efficient and expressive feature extraction (Howard et al., 2019; Sandler et al., 2018). Each MBConv block integrates depthwise separable convolutions and, in select layers, channel-wise squeeze-and-excitation (SE) modules to recalibrate feature maps adaptively with minimal computational overhead (Hu et al., 2018). The non-linear activation function used throughout these blocks is hard-swish (HSwish), which has been empirically shown to provide improved performance for mobile and embedded networks with negligible computational cost increase.

A distinctive element of LiteXrayNet is the inclusion of a lightweight quantum-inspired phase shift module, positioned subsequent to the primary feature extraction stages. This module draws conceptual motivation from quantum computing, specifically the phase shift operations that enable complex spatial encoding and entanglement. In LiteXrayNet, the quantum phase shift layer is implemented as a sequence of learnable phase parameterizations that modulate the feature maps, thereby enhancing the model's capacity to capture subtle spatial and textural cues that are often critical for the discrimination of pneumonia in chest radiographs. This approach mimics the representation enrichment typically observed in quantum neural networks, while maintaining strict parameter and memory efficiency compatible with deployment constraints (Zhang et al., 2022a).

Following the quantum-inspired enhancement, the architecture employs adaptive average pooling to condense spatial feature maps, which are then passed through a compact classifier head composed of fully connected layers, batch normalization, HSwish activations, and dropout for regularization. The final output layer predicts the class probability for binary classification (Normal vs Pneumonia).

In total, LiteXrayNet contains approximately 179,646 trainable parameters, resulting in a model size 0.7 MB, substantially smaller than traditional architectures such as ResNet-18, MobileNetV3-Small and competitive lightweight models. This compactness, combined with its efficient block-wise structure and quantum-inspired

layer, enables real-time inference with low memory usage and high energy efficiency, thus fulfilling the requirements of point-of-care and mobile healthcare applications.

4.3.1 Feature Extraction Backbone

The backbone of LiteXrayNet comprises a sequence of Mobile Inverted Bottleneck Convolutional (MBConv) blocks, a design paradigm introduced in MobileNetV2 and refined in MobileNetV3 (Sandler et al., 2018; Howard et al., 2019). These blocks are engineered to minimize computational complexity while extracting rich spatial and contextual features from chest X-ray images. Each MBConv block consists of three stages: an expansion phase using a 1×1 convolution to increase channel dimensionality, a depthwise separable 3×3 convolution for spatial feature extraction, and a projection phase to reduce channel dimensionality. This structure leverages depthwise separable convolutions to significantly reduce the number of parameters and floating-point operations (FLOPs) compared to standard convolutions, making it ideal for edge devices. When input and output channel dimensions match and the stride is 1, residual connections are incorporated to facilitate gradient flow and stabilize training, following the principles established in ResNet (He et al., 2016).

For an input tensor $X \in \mathbb{R}^{C_{\text{in}} \times H \times W}$, where C_{in} is the number of input channels, H is the height, and W is the width, the MBConv block operates as follows. The expansion phase employs a 1×1 convolution to increase the channel count by an expansion factor t, typically set to 2:

$$X_{\text{exp}} = \text{BN}(\text{Conv}_{1\times1}(X; C_{\text{in}}, t \cdot C_{\text{in}})), \tag{5}$$

producing $X_{\text{exp}} \in \mathbb{R}^{t \cdot C_{\text{in}} \times H \times W}$. The depthwise convolution applies a 3×3 convolution to each channel independently:

$$X_{\text{dw}} = \text{BN}(\text{Conv}^{\text{dw}}_{3\times3}(X_{\text{exp}}; t \cdot C_{\text{in}}, t \cdot C_{\text{in}}, \text{groups} = t \cdot C_{\text{in}})), \tag{6}$$

where $\text{Conv}^{\text{dw}}_{3\times3}$ preserves spatial dimensions with padding or reduces them with a stride greater than 1. The projection phase reduces the channel count:

$$X_{\text{out}} = \text{BN}(\text{Conv}_{1\times1}(X_{\text{dw}}; t \cdot C_{\text{in}}, C_{\text{out}})). \tag{7}$$

If applicable ($C_{\text{in}} = C_{\text{out}}$, stride $= 1$), a residual connection is added:

$$X_{\text{out}} = X + X_{\text{out}}. \tag{8}$$

This structure reduces the parameter count from $C_{\text{in}} \cdot C_{\text{out}} \cdot k^2$ for a standard $k \times k$ convolution to $C_{\text{in}} \cdot k^2 + C_{\text{in}} \cdot C_{\text{out}}$, where $k = 3$, achieving significant computational efficiency.

The backbone begins with an initial 3×3 convolution that reduces spatial dimensions and expands channels to 16:

$$X_0 = \text{HSwish}(\text{BN}(\text{Conv}_{3\times3}(X_{\text{in}}; 3, 16, \text{stride} = 2))), \tag{9}$$

where $X_{\text{in}} \in \mathbb{R}^{3\times224\times224}$ for RGB input images resized to 224×224. This is followed by six MBConv blocks with channel counts increasing from 16 to 128, selectively applying strides of 2 to create a hierarchical feature representation optimized for chest X-ray analysis.

4.3.2 Hard-Swish (HSwish) Activation

To introduce non-linearity while maintaining computational efficiency, LiteXrayNet employs the Hard-Swish (HSwish) activation function across the backbone and subsequent layers. Introduced in MobileNetV3 (Howard et al., 2019; Ramachandran et al., 2017), HSwish approximates the Swish activation ($x \cdot \sigma(x)$) using a piecewise linear function, avoiding the computational cost of the sigmoid function. The HSwish function is defined as:

$$\text{HSwish}(x) = x \cdot \frac{\text{ReLU6}(x + 3)}{6}, \tag{10}$$

where $\text{ReLU6}(x) = \min(\max(x, 0), 6)$. This activation provides a smooth non-linearity that enhances convergence and accuracy compared to ReLU, particularly in deep networks, while being compatible with hardware accelerators (Howard et al., 2019). Its use ensures that LiteXrayNet maintains efficiency without sacrificing representational power, making it ideal for edge deployment.

4.3.3 Squeeze-and-Excitation (SE) Modules

LiteXrayNet incorporates Squeeze-and-Excitation (SE) modules in selected MBConv blocks to enhance feature discriminability (Hu et al., 2018). These modules recalibrate channel-wise feature responses by modeling interdependencies, enabling the network to focus on diagnostically relevant features such as localized opacities or textural patterns in chest X-rays. For an input tensor $X \in \mathbb{R}^{C\times H\times W}$, the SE module performs a squeeze operation via global average pooling:

$$z_c = \frac{1}{H \cdot W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_c(i, j), \quad z \in \mathbb{R}^C. \tag{11}$$

This descriptor is processed through a two-layer fully connected network with a reduction factor $r = 4$:

$$s = \sigma(W_2 \cdot \mathrm{ReLU}(W_1 \cdot z)), \tag{12}$$

where $W_1 \in \mathbb{R}^{\frac{C}{r}\times C}$, $W_2 \in \mathbb{R}^{C\times \frac{C}{r}}$, and σ is the sigmoid function. The channel weights $s \in \mathbb{R}^C$ rescale the input tensor:

$$X_{\text{out}} = X \cdot s, \tag{13}$$

with s broadcast across spatial dimensions. In LiteXrayNet, SE modules are applied in MBConv blocks with strides of 2 (at channel counts of 32, 48, and 96), balancing computational cost with improved feature selection.

4.3.4 Quantum-Inspired Phase Shift Layer

A defining innovation of LiteXrayNet is the quantum-inspired phase shift layer, positioned after the MBConv backbone to enhance feature representation. Inspired by phase shift gates in quantum neural networks (Zhang et al., 2022a; Houssein et al., 2022), this layer modulates feature tensors to capture complex spatial and textural relationships critical for pneumonia detection. For an input tensor $X \in \mathbb{R}^{C\times H\times W}$ with $C = 128$, a 1×1 convolution reduces the channel count:

$$z = \mathrm{Conv}_{1\times 1}(X; C, \frac{C}{r}), \tag{14}$$

producing $z \in \mathbb{R}^{\frac{C}{r}\times H\times W}$, where $r = 4$ and $\frac{C}{r} = 32$. Two learnable phase parameters, $\Phi_1, \Phi_2 \in \mathbb{R}^{1\times \frac{C}{r}\times 1\times 1}$, initialized with $\sim \mathcal{N}(0, 0.01)$, modulate the features:

$$z' = z \odot \cos(\Phi_1) + z \odot \sin(\Phi_1), \tag{15}$$

$$z'' = z' \odot \cos(\Phi_2) + z' \odot \sin(\Phi_2), \tag{16}$$

where $\odot$ denotes elementwise multiplication. These operations mimic quantum phase gates, enhancing representational capacity. A 1×1 convolution restores the channel count:

$$y = \mathrm{Conv}_{1\times 1}(z''; \frac{C}{r}, C), \tag{17}$$

followed by a residual connection, batch normalization, and HSwish activation:

$$X_{\text{out}} = \mathrm{HSwish}(\mathrm{BN}(X + y)). \tag{18}$$

This layer enables LiteXrayNet to capture subtle radiographic patterns with minimal parameter overhead.

4.3.5 Aggregation and Classification Head

After feature extraction and quantum-inspired modulation, an adaptive average pooling layer aggregates spatial information:

$$X_{\text{pool}} = \mathrm{AvgPool2d}(X_{\text{out}},\ \text{output_size} = 1) \tag{19}$$

producing $X_{\text{pool}} \in \mathbb{R}^{128\times1\times1}$, flattened to $\mathbb{R}^{128}$. The classification head processes this vector through a two-layer fully connected network. The first linear layer maps to a higher-dimensional space:

$$h_1 = \text{BN}(\text{Linear}(X_{\text{pool}}; 128, 512)), \tag{20}$$

followed by HSwish activation:

$$h_2 = \text{HSwish}(h_1). \tag{21}$$

A dropout layer with a rate of 0.2 mitigates overfitting:

$$h_3 = \text{Dropout}(h_2; p = 0.2), \tag{22}$$

and a second linear layer produces logits for the two classes:

$$\text{logits} = \text{Linear}(h_3; 512, 2). \tag{23}$$

This lightweight head ensures efficient and robust prediction for binary classification.

4.3.6 Layerwise Architecture Specification

The layerwise architecture of LiteXrayNet is detailed in Table 1, specifying each layer's operation, output shape, and parameter count for an input image of $\mathbb{R}^{3\times224\times224}$. The initial convolution reduces spatial dimensions and expands channels, followed by six MBConv blocks with increasing channel counts and selective spatial downsampling. The quantum phase shift layer processes the final feature tensor, and adaptive average pooling produces a global feature vector. The classification head generates logits for the two classes. The total parameter count is 179,646, yielding a model size of approximately 0.7 MB.

Table 1: Layerwise Architecture Specification of LiteXrayNet

Layer Name/Block	**Configuration**	**Output Shape**
Input	RGB Image (3 Channels) 224×224	(3, 224, 224)
Conv-BN-HSwish	Conv2d: 3→16, kernel 3×3, stride 2, padding 1; BatchNorm2d; HSwish activation	(16, 112, 112)
MBConv Block 1	In: 16, Out: 24; Expand ratio: 2; stride 1; no SE	(24, 112, 112)
MBConv Block 2	In: 24, Out: 32; Expand ratio: 2; stride 2; SE block	(32, 56, 56)
MBConv Block 3	In: 32, Out: 48; Expand ratio: 2; stride 2; SE block	(48, 28, 28)
MBConv Block 4	In: 48, Out: 64; Expand ratio: 2; stride 1; no SE	(64, 28, 28)
MBConv Block 5	In: 64, Out: 96; Expand ratio: 2; stride 2; SE block	(96, 14, 14)
MBConv Block 6	In: 96, Out: 128; Expand ratio: 2; stride 1; no SE	(128, 14, 14)
Quantum Phase Shift Layer	1×1 Conv: 128→32 (squeeze); 2× Phase Shifts (learnable); 1×1 Conv: 32→128 (excite); BatchNorm2d; HSwish	(128, 14, 14)
AdaptiveAvgPool2d	Output size=1	(128, 1, 1)
Flatten	-	(128,)
Classifier Head	Linear: 128→512, BatchNorm1d, HSwish, Dropout(0.2), Linear: 512→2	(2,)

4.3.7 Theoretical and Practical Justification

LiteXrayNet's architecture is designed to optimize the accuracy-efficiency tradeoff for chest X-ray classification on resource-constrained edge devices, integrating established and novel components grounded in theoretical principles and practical requirements. The Mobile Inverted Bottleneck Convolutional (MBConv) blocks, inspired by MobileNetV3 (Howard et al., 2019), leverage depthwise separable convolutions to reduce computational complexity by an order of magnitude compared to standard convolutions, enabling efficient feature extraction (Sandler et al., 2018). Hard-Swish (HSwish) activations provide smooth, hardware-friendly nonlinearity, enhancing convergence without significant computational overhead (Howard et al., 2019). Squeeze-and-Excitation (SE) modules adaptively recalibrate channel responses, improving feature discriminability for subtle radiographic patterns with minimal parameter increase (Hu et al., 2018). The quantum-inspired phase shift layer, drawing on quantum neural network principles, introduces non-linear feature modulation to capture complex textural relationships critical for pneumonia detection, maintaining a low parameter count (Zhang et al., 2022a). The lightweight classification head, with batch normalization and dropout, ensures robust generalization. Practically, LiteXrayNet's compact size (approximately 0.75 MB, 179,646 parameters) and low computational requirements make it ideal for edge deployment in clinical settings, addressing the need for rapid, accurate diagnosis on resource-limited hardware. These design choices collectively ensure that LiteXrayNet achieves high diagnostic performance while meeting the stringent efficiency demands of edge-based medical imaging.

4.4 Training and Evaluation Configuration

The training and evaluation pipeline for LiteXrayNet and baseline architectures was designed to ensure a robust, reproducible, and comprehensive assessment of chest X-ray classification performance, distinguishing "Normal" from "Pneumonia" cases on edge devices, with a focus on interpretability through feature map visualization and Gradient-weighted Class Activation Mapping (Grad-CAM). Implemented in PyTorch 2.0.1 (Paszke et al., 2019) with NumPy 1.24.3 and torchvision 0.15.2, the pipeline encompasses dataset loading, preprocessing, feature engineering, class imbalance handling, training, evaluation, and visualization, tailored for medical imaging applications.

The dataset, sourced from Kaggle's chest X-ray pneumonia dataset, comprises 5,856 pediatric RGB images (1,341 "Normal," 4,515 "Pneumonia"). Patient-level separation was verified to prevent data leakage, and the original splits were recombined into 70% training (4,099 images), 15% validation (878 images), and 15% test (879 images) sets via stratified random sampling, preserving the class ratio (22.9% "Normal," 77.1% "Pneumonia") (He & Garcia, 2009). Preprocessing resized images to 224×224 using bilinear interpolation and normalized pixel values with mean $\mu = [0.485, 0.456, 0.406]$ and standard deviation $\sigma = [0.229, 0.224, 0.225]$. Quality checks removed corrupted images (<0.1% of the dataset). Feature engineering applied training-set augmentations, including random horizontal flips (probability 0.5), rotations ($\pm 10°$), brightness adjustments ($\pm 20\%$), contrast variations ($\pm 10\%$), and scaling ($\pm 10\%$), to enhance robustness and mitigate class imbalance. Shearing and synthetic data generation (e.g., SMOTE) were evaluated but excluded due to marginal gains and increased complexity.

Class imbalance was addressed through stratified splitting, augmentation, and a weighted cross-entropy loss (He & Garcia, 2009; Johnson & Khoshgoftaar, 2019). Class weights were computed using inverse prevalence:

$$w_{\text{Normal}} = \frac{N_{\text{total}}}{2 \cdot N_{\text{Normal}}}, \quad w_{\text{Pneumonia}} = \frac{N_{\text{total}}}{2 \cdot N_{\text{Pneumonia}}}, \tag{24}$$

where $N_{\text{total}} = 4,099$, $N_{\text{Normal}} \approx 938$, $N_{\text{Pneumonia}} \approx 3,161$, yielding $w_{\text{Normal}} \approx 2.19$, $w_{\text{Pneumonia}} \approx 0.65$, normalized to sum to 2. The loss was:

$$\mathcal{L} = -\sum_{i=1}^{2} w_i \cdot y_i \log(\hat{y}_i), \tag{25}$$

where y_i is the true label and $\hat{y}_i$ is the predicted probability for class i, prioritizing the minority "Normal" class (He & Garcia, 2009). Oversampling and undersampling were tested but omitted to avoid overfitting and information loss.

Training used the Adam optimizer (learning rate 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 1e-5) (He & Garcia, 2009), with a StepLR scheduler decaying the learning rate by 0.5 every 20 epochs. Models trained for up to 100 epochs, with early stopping after 10 epochs of stagnant validation accuracy. LiteXrayNet's classifier head applied a 0.2 dropout rate. Mixed-precision training via PyTorch's AMP reduced memory usage and accelerated computation (Micikevicius et al., 2018). A batch size of 32 balanced gradient accuracy and efficiency. Baselines (e.g., MobileNetV3, ResNet-18) used identical augmentations, loss weighting, and optimization settings, with architecture-specific hyperparameter tuning.

Evaluation assessed predictive performance, efficiency, and interpretability across 10 independent runs with fixed seeds for NumPy, PyTorch, and CUDA. Metrics included accuracy, AUC-ROC, precision, recall, F1-score, model size, and inference latency, reported as:

$$\text{Metric} = \mu \pm \sigma, \quad \mu = \frac{1}{10}\sum_{i=1}^{10} m_i, \quad \sigma = \sqrt{\frac{1}{10}\sum_{i=1}^{10}(m_i - \mu)^2}, \tag{26}$$

where m_i is the metric value for run i. Confusion matrices evaluated class-specific performance. Inference used warm-started models (batch size 32), with GPU synchronization and minimal buffering to simulate edge constraints (<100 ms latency, <1 MB memory). Experiments ran on an NVIDIA RTX A1000 GPU (4 GB VRAM), Intel Core i7 12th Gen CPU, 16 GB RAM, Ubuntu 22.04, Python 3.10, and CUDA 11.8, with resource monitoring via psutil and pynvml.

Feature map visualization was implemented to analyze intermediate representations from convolutional and quantum layers, using forward hooks to capture outputs from each MBConv block and the quantum phase shift layer. Up to eight channels per layer were visualized using the viridis colormap, with spatial feature maps displayed as 2D images and 1×1 feature maps (e.g., quantum layer outputs) as bar charts. Visualizations compared "Normal" and "Pneumonia" X-rays to highlight differential feature activation, saved as PNG files at 150 DPI. Grad-CAM was applied to generate heatmaps for interpretability (Selvaraju et al., 2017), targeting the final MBConv block's output. For an input image X, the gradient of the predicted class score y^c with respect to the feature map A^k of channel k was computed:

$$\alpha_k^c = \frac{1}{H \cdot W}\sum_{i=1}^{H}\sum_{j=1}^{W}\frac{\partial y^c}{\partial A_{ij}^k}, \tag{27}$$

where H and W are the feature map dimensions. The heatmap was formed as:

$$L_{\text{Grad-CAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right), \tag{28}$$

resized to 224×224, normalized, and overlaid on the original image using the jet colormap to highlight clinically relevant regions (e.g., lung opacities). Heatmaps for true positives and false negatives were analyzed to verify model attention, enhancing trust in LiteXrayNet's predictions for edge-based deployment.

5 Quantitative Results and Comparative Analysis

5.1 Overall Model Accuracy and Loss

The quantitative evaluation of LiteXrayNet and competing architectures, including ResNet-18 (Base), MobileNetV3-Small, EfficientNet-B0, MobileViT, TinyDeiT, and ViT-LoRA, is presented in Table 2, which summarizes accuracy and loss metrics across training, validation, and test sets. LiteXrayNet demonstrates superior performance, achieving the highest accuracy (0.9790, 0.9738, and 0.9704 for train, validation, and test sets, respectively) and the lowest loss values (0.0508, 0.1197, and 0.0917), indicating robust generalization and minimal overfitting compared to baselines. ResNet-18 and MobileNetV3-Small follow with competitive accuracies (0.9499–0.9590) and losses (0.1546–0.1789), while EfficientNet-B0, MobileViT, TinyDeiT, and ViT-LoRA exhibit lower accuracies (0.9295–0.9431) and higher losses (0.2106–0.2660), reflecting

their less efficient adaptation to the imbalanced chest X-ray dataset. These results underscore LiteXrayNet's effectiveness, particularly its lightweight design (179,646 parameters, approximately 0.7 MB), which supports its suitability for resource-constrained environments.

Table 2: Comprehensive summary of training, validation, and test set accuracy and loss metrics for a range of deep learning architectures, including ResNet-18 (Base), MobileNetV3-Small, EfficientNet-B0, MobileViT, TinyDeiT, ViT-LoRA, and the proposed LiteXrayNet.

Model	**Train Acc**	**Train Loss**	**Val Acc**	**Val Loss**	**Test Acc**	**Test Loss**
ResNet-18 (Base)	0.9595	0.1094	0.9556	0.1655	0.9499	0.1789
MobileNetV3-Small	0.9666	0.0865	0.9499	0.1621	0.9590	0.1546
EfficientNet-B0	0.9495	0.1690	0.9499	0.2195	0.9386	0.2179
MobileViT	0.9522	0.1547	0.9431	0.2041	0.9431	0.2185
TinyDeiT	0.9334	0.2011	0.9214	0.2660	0.9295	0.2301
ViT-LoRA	0.9356	0.1676	0.9294	0.2309	0.9317	0.2106
LiteXrayNet (Ours)	**0.9790**	**0.0508**	**0.9738**	**0.1197**	**0.9704**	**0.0917**

Figure 4 illustrates LiteXrayNet's optimal performance with a plot of accuracy and loss metrics across all sets, showing the lowest loss and highest accuracy compared to baselines, with a clear convergence trend and minimal validation-test discrepancy indicating strong generalization. This highlights its potential for edge-based medical imaging, balancing accuracy and efficiency, and, alongside Table 2, establishes LiteXrayNet as a leading architecture for chest X-ray classification.

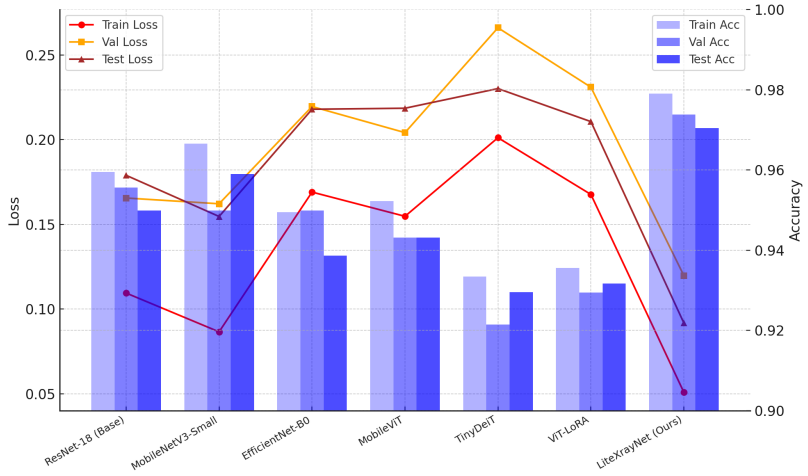

Figure 4: The chart illustrates LiteXrayNet's optimal performance, characterized by the lowest loss values and highest accuracy metrics across all evaluated sets, underscoring its effectiveness and potential for practical deployment in resource-constrained environments.

5.2 Classwise Precision, Recall, and F1-Score

Figure 5 presents a heatmap of classwise performance metrics for pneumonia detection across seven deep learning models—ResNet-18 (Base), MobileNetV3-Small, EfficientNet-B0, MobileViT, TinyDeiT, ViT-LoRA, and LiteXrayNet—evaluated on the test set, visualizing precision, recall, F1-score for "Normal" and "Pneumonia" classes, and AUC-ROC values with color intensity (light to dark blue) where darker shades indicate higher performance. Table 3 provides a quantitative summary of these metrics, showing LiteXrayNet achieving the highest scores: precision (0.9393 for Normal, 0.9826 for Pneumonia), recall (0.9547 for Normal, 0.9764 for Pneumonia), F1-score (0.9469 for Normal, 0.9795 for Pneumonia), and AUC-ROC (0.9946), reflecting its effective handling of class imbalance, especially for the minority "Normal" class. In contrast,

baselines like TinyDeiT (recall 0.8807 for Normal) and ViT-LoRA (precision 0.8533 for Normal) underperform, while MobileNetV3-Small (recall 0.9095 for Normal) and ResNet-18 (recall 0.9136 for Normal) are competitive but fall short of LiteXrayNet's balanced accuracy. This combined visualization and data highlight LiteXrayNet's superior classwise performance and its lightweight design, reinforcing its potential for efficient deployment in edge-based medical imaging.

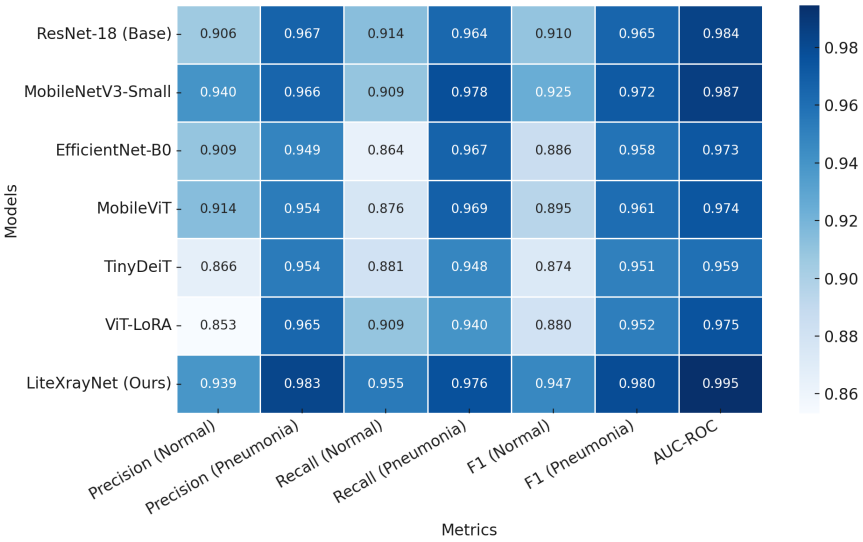

Figure 5: Heatmap displaying classwise performance metrics for pneumonia detection across seven deep learning models: ResNet-18 (Base), MobileNetV3-Small, EfficientNet-B0, MobileViT, TinyDeiT, ViT-LoRA, and LiteXrayNet. The metrics include precision, recall, and F1-score for both Normal and Pneumonia classes, as well as AUC-ROC values, evaluated on the test set. Color intensity (ranging from light to dark blue) corresponds to metric values, with darker shades indicating higher values, facilitating visual comparison across models and metrics.

Table 3: Classwise precision, recall, and F1-score metrics for pneumonia detection on the test set, evaluated for seven deep learning models: ResNet-18 (Base), MobileNetV3-Small, EfficientNet-B0, MobileViT, TinyDeiT, ViT-LoRA, and LiteXrayNet. The table lists precision (P_N for Normal, P_P for Pneumonia), recall (R_N for Normal, R_P for Pneumonia), F1-score ($F1_N$ for Normal, $F1_P$ for Pneumonia), and AUC-ROC values, providing a quantitative summary of model performance across the two classes.

Model	$\mathbf{P_N}$	$\mathbf{P_P}$	$\mathbf{R_N}$	$\mathbf{R_P}$	$\mathbf{F1_N}$	$\mathbf{F1_P}$	**AUC-ROC**
ResNet-18 (Base)	0.9061	0.9669	0.9136	0.9638	0.9098	0.9654	0.9842
MobileNetV3-Small	0.9404	0.9658	0.9095	0.9780	0.9247	0.9719	0.9865
EfficientNet-B0	0.9091	0.9491	0.8642	0.9670	0.8861	0.9579	0.9733
MobileViT	0.9142	0.9536	0.8765	0.9686	0.8950	0.9610	0.9739
TinyDeiT	0.8664	0.9541	0.8807	0.9481	0.8735	0.9511	0.9594
ViT-LoRA	0.8533	0.9645	0.9095	0.9403	0.8805	0.9522	0.9753
LiteXrayNet (Ours)	0.9393	0.9826	0.9547	0.9764	0.9469	0.9795	0.9946

5.3 Model Size, Efficiency, and Resource Utilization

Figure 6 compares model size (in MB) and inference time (ms/sample) across seven deep learning architectures—ResNet-18 (Base), MobileNetV3-Small, EfficientNet-B0, MobileViT, TinyDeiT, ViT-LoRA, and LiteXrayNet—highlighting trade-offs between complexity and efficiency, with LiteXrayNet achieving the smallest size (0.70 MB) and a competitive inference time (0.60 ms/sample) suitable for edge deployment (Chen & Ran, 2020). Figure 7 further illustrates this efficiency by plotting total and trainable parameters

on a logarithmic scale, showing LiteXrayNet's minimal parameter count (179,646) compared to ViT-LoRA's 85,947,650, underscoring its reduced computational footprint. Table 4 provides quantitative data, confirming LiteXrayNet's superiority with a model size of 0.70 MB, 179,646 total and trainable parameters, and an inference time of 0.60 ms/sample, outperforming larger models like ViT-LoRA (327.86 MB, 12.55 ms/sample) while remaining competitive with MobileNetV3-Small (3.59 MB, 0.26 ms/sample). Table 5 details resource usage, showing LiteXrayNet's lowest average CPU (5.6%) and RAM (69.2%) utilization, with maximums of 16.1% and 71.3%, respectively, compared to peaks of 100% CPU for ResNet-18 and EfficientNet-B0, reflecting its efficiency in constrained environments. These results collectively demonstrate LiteXrayNet's lightweight design and resource efficiency, making it ideal for edge-based medical imaging applications.

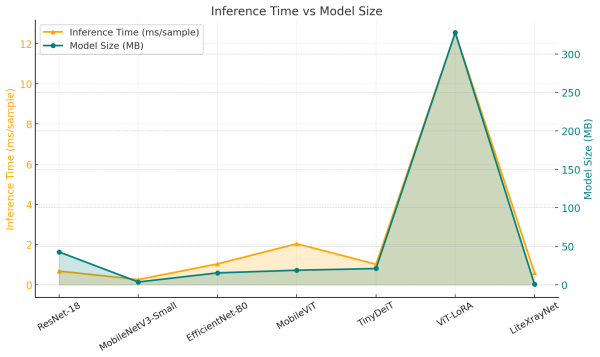

Figure 6: Comparison of model size (MB) and inference time (ms/sample) across selected deep learning architectures.

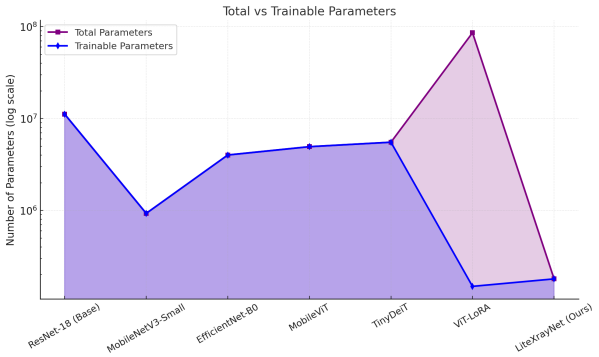

Figure 7: Comparison of total and trainable parameters across different deep learning models.

Table 4: Model size, parameters, and inference latency for ResNet-18, MobileNetV3-Small, EfficientNet-B0, MobileViT, TinyDeiT, ViT-LoRA, and LiteXrayNet.

Model	**Model Size** $_{\textbf{(MB)}}$	**Param**$_{\textbf{Total}}$	**Param**$_{\textbf{Trainable}}$	**Inf Time** $_{\textbf{(ms/sample)}}$
ResNet-18 (Base)	42.68	11,177,538	11,177,538	0.68
MobileNetV3-Small	3.59	928,162	928,162	0.26
EfficientNet-B0	15.46	4,010,110	4,010,110	1.04
MobileViT	18.89	4,938,914	4,938,914	2.04
TinyDeiT	21.08	5,524,802	5,524,802	1.02
ViT-LoRA	327.86	85,947,650	148994	12.55
LiteXrayNet (Ours)	0.70	179,646	179,646	0.60

Table 5: Inference resource usage metrics for seven deep learning models: ResNet-18 (Base), MobileNetV3-Small, EfficientNet-B0, MobileViT, TinyDeiT, ViT-LoRA, and LiteXrayNet. The table reports average and maximum CPU usage (in percentage) and average and maximum RAM usage (in percentage) during active inference, measured under a standardized experimental protocol to assess hardware efficiency in constrained environments.

Model	**Avg CPU (%)**	**Max CPU (%)**	**Avg RAM (%)**	**Max RAM (%)**
ResNet-18 (Base)	6.6	100.0	72.0	73.8
MobileNetV3-Small	6.9	29.3	73.0	78.2
EfficientNet-B0	10.1	100.0	72.7	74.3
MobileViT	6.2	32.4	73.3	74.7
TinyDeiT	7.1	46.2	73.5	74.9
ViT-LoRA	12.3	86.4	75.3	86.4
LiteXrayNet (Ours)	5.6	16.1	69.2	71.3

5.4 Visual Explanation and Interpretability through Grad-CAM

Gradient-weighted Class Activation Mapping (Grad-CAM) is employed as a post hoc visual explanation technique to enhance the interpretability of the LiteXrayNet architecture, as described in (Selvaraju et al., 2017; Chattopadhay et al., 2018). The method generates class-discriminative localization maps by highlighting spatial regions in input X-ray images that contribute to the model's predictions, supporting transparency for clinical validation and trust among radiologists. The Grad-CAM pipeline integrates with LiteXrayNet by capturing activations and gradients from a deeper convolutional layer, selected for its balance of semantic abstraction and spatial resolution for localization. During the forward pass, intermediate activation maps are stored; during backpropagation, gradients of the predicted class score with respect to these activations are computed and globally average-pooled to derive channel importance weights. These weights are linearly combined with the feature maps, followed by a ReLU non-linearity and normalization, to produce a heatmap. The heatmap is upsampled to match the input resolution and overlaid on the original X-ray image for analysis. Grad-CAM analysis was applied to a stratified sample of test images across four categories: true positives, true negatives, false positives, and false negatives, based on the model's predicted class.

Figure 8 displays Grad-CAM heatmaps for true positive cases, where LiteXrayNet correctly identified pneumonia, showing activation focused on lung field regions with opacification, including posterior or basal consolidations, bilateral interstitial markings, and segmental opacities, consistent with radiological features of bacterial or viral pneumonia. Figure 9 presents heatmaps for true negative cases, where normal images were correctly classified, exhibiting diffuse low-intensity gradients or no strong activation within lung regions, with occasional attention on non-diagnostic structures such as the diaphragm or lateral thoracic borders. Figure 10 shows heatmaps for false positive cases, where normal images were misclassified as pneumonia, revealing elevated activation on structures adjacent to the heart silhouette or rib boundaries, potentially influenced by imaging artifacts. Figure 11 illustrates heatmaps for false negative cases, where pneumonia was missed, displaying weak or non-specific activation, often failing to highlight minor consolidations or early interstitial changes.

The analysis of Grad-CAM maps across these categories confirms that LiteXrayNet's predictions align with clinically relevant lung regions in true positive cases, showing localized activation on pathological features such as consolidations and opacities. In true negative cases, the absence of focal activation within lung fields indicates the model's ability to identify normal anatomy. In false positive cases, activation on non-diagnostic areas suggests sensitivity to image artifacts, while false negative cases exhibit weak activation, correlating with missed subtle or atypical pneumonia manifestations. These observations validate LiteXrayNet's interpretability, with heatmaps providing spatially and diagnostically consistent outputs for real-world integration.

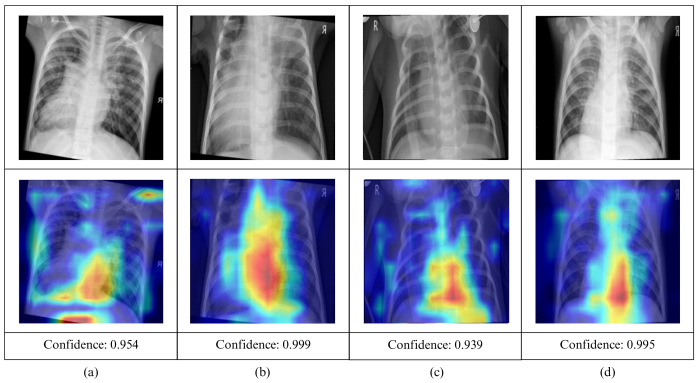

Figure 8: Grad-CAM heatmaps for true positive cases of pneumonia detection by LiteXrayNet, showing activation on lung field regions with opacification, including consolidations and interstitial markings.

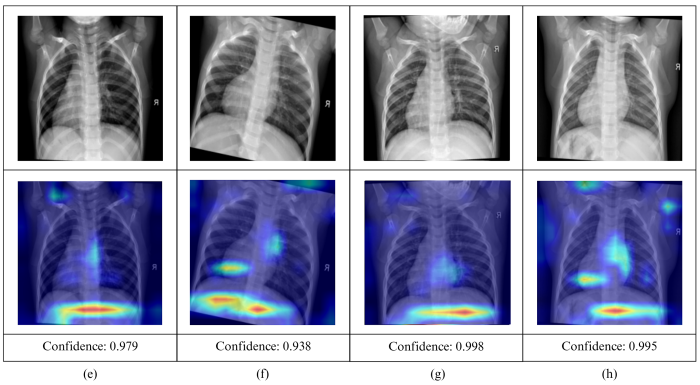

Figure 9: Grad-CAM heatmaps for true negative cases of normal classification by LiteXrayNet, displaying diffuse low-intensity gradients or activation on non-diagnostic structures.

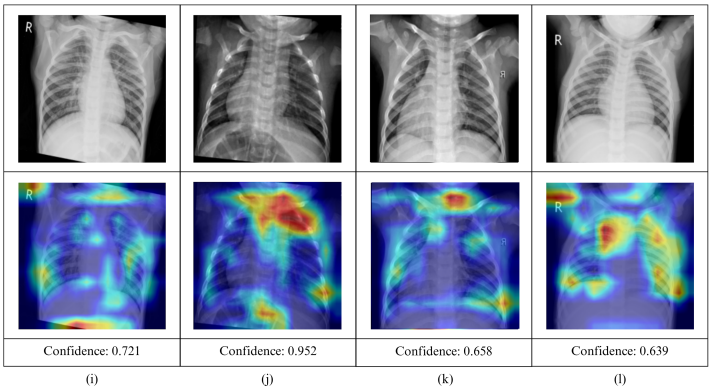

Figure 10: Grad-CAM heatmaps for false positive cases of normal misclassified as pneumonia by LiteXrayNet, indicating activation on regions adjacent to the heart silhouette or rib boundaries.

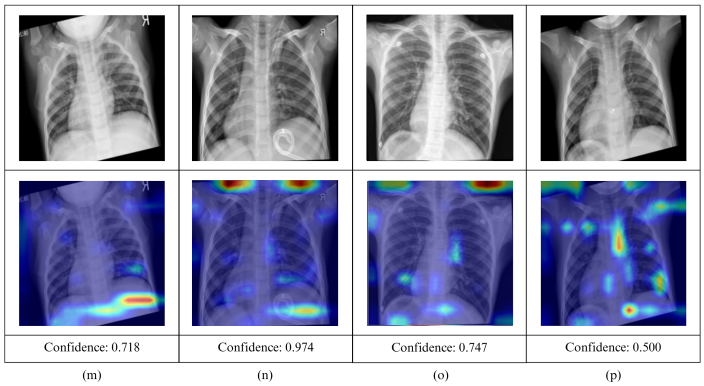

Figure 11: Grad-CAM heatmaps for false negative cases of missed pneumonia by LiteXrayNet, showing weak or non-specific activation in lung regions.

5.4.1 Clinical Implications

The Grad-CAM heatmaps for true positive and true negative cases demonstrate LiteXrayNet's alignment with radiologically significant features, supporting its diagnostic accuracy. In false positive and false negative cases, the heatmaps identify activation patterns on non-informative or weakly activated regions, respectively, providing insights into model behavior. These visualizations enable human-in-the-loop validation, enhancing trust in clinical deployment. Example overlays for each category are included in supplementary material to support reproducibility.

6 Feature Map Visualization

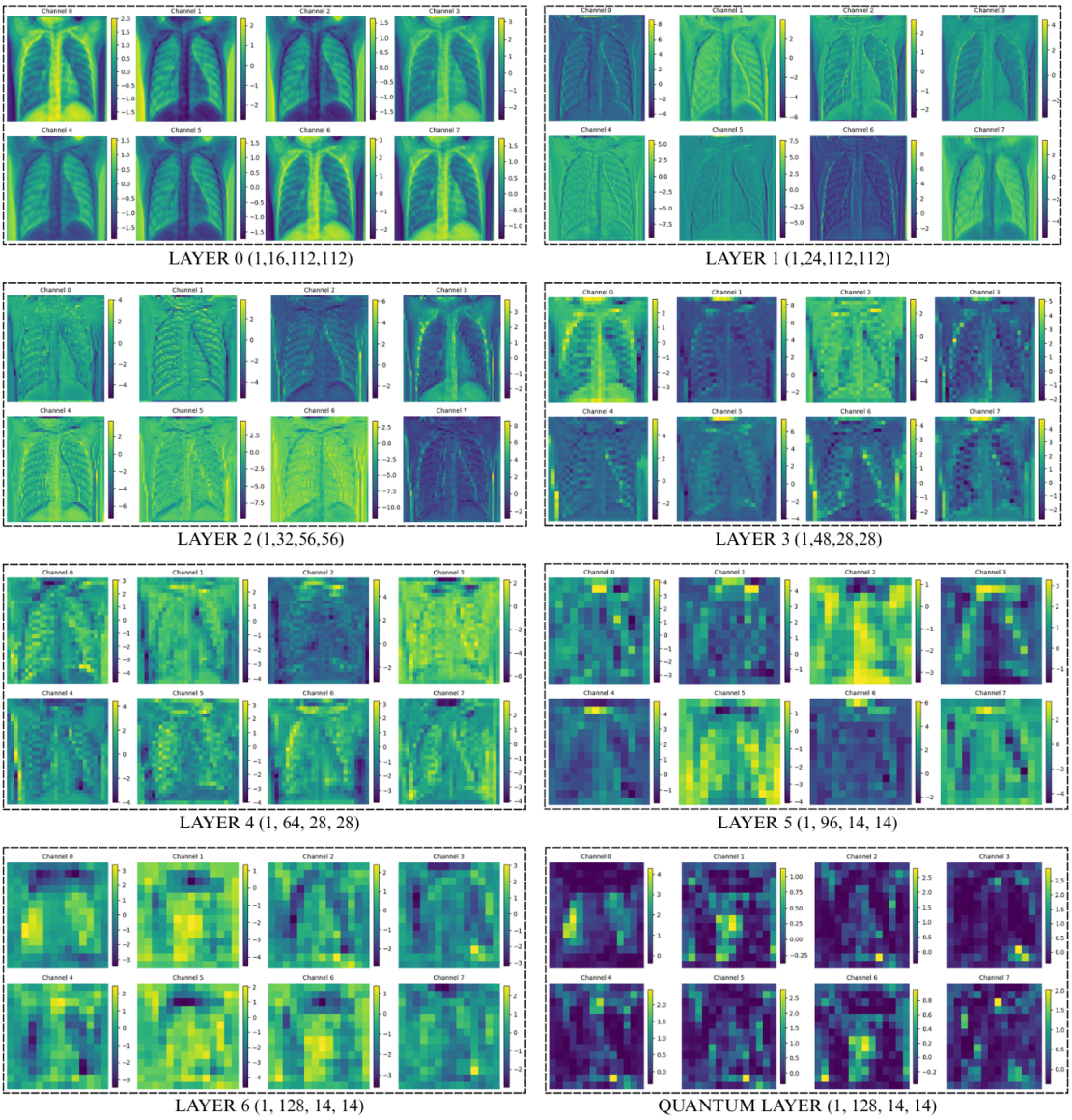

Figure 12: Feature map progression across LiteXrayNet layers for a sample X-ray, displaying activations from initial convolutional layers, MBConv blocks, and the quantum-inspired phase shift layer.

To enhance model interpretability and support clinical adoption of LiteXrayNet, intermediate feature maps were visualized to reveal the internal representations learned from chest X-ray inputs, as outlined in (Selvaraju et al., 2017). The analysis tracks the network's progressive extraction and transformation of features, from low-level edges to high-level semantic patterns associated with pneumonia, validating the model's focus on clinically relevant lung regions. A custom FeatureMapVisualizer class in PyTorch was implemented using forward hooks to capture activations from key layers: initial convolutional layers, MBConv blocks, and the quantum-inspired phase shift layer. Inputs were preprocessed to 224×224 grayscale images repeated across three channels, normalized with ImageNet statistics, and visualizations were generated for sample cases, displaying up to eight channels per layer using the 'viridis' colormap for heatmaps.

Figure 12 depicts the feature evolution for a sample X-ray across layers, with early layers (e.g., Layer 0 [1,16,112,112]) showing activations of low-level features such as lung contours and basic textures in channels 0–7. At increased depth (e.g., Layer 2 [1,32,56,56]), features transition to abstract representations, highlighting structural anomalies like consolidations. The quantum-inspired phase shift layer [1,128,14,14] produces refined activations, exhibiting enhanced intensity in affected lung regions, indicative of subtle pathological cues.

The feature map visualizations demonstrate that LiteXrayNet's hierarchical learning aligns with clinical domain knowledge, concentrating on lung fields while minimizing attention to extraneous artifacts, thereby supporting its interpretability and diagnostic relevance.

7 Discussion

The evaluation of LiteXrayNet against six baseline models: ResNet-18, MobileNetV3-Small, EfficientNet-B0, MobileViT, TinyDeiT, and ViT-LoRA, demonstrates its superior performance in the context of pneumonia detection from chest X-rays, particularly for edge-based medical imaging applications. In terms of overall accuracy and loss, LiteXrayNet achieved the highest scores across training (0.9790 accuracy, 0.0508 loss), validation (0.9738 accuracy, 0.1197 loss), and test sets (0.9704 accuracy, 0.0917 loss), outperforming ResNet-18 (0.9499 test accuracy, 0.1789 loss) and MobileNetV3-Small (0.9590 test accuracy, 0.1546 loss), which ranked second and third, respectively. This indicates robust generalization and minimal overfitting, attributable to LiteXrayNet's optimized architecture with 179,646 parameters and a model size of 0.70 MB, designed to balance predictive power with computational efficiency. The classwise analysis further supports this, with LiteXrayNet recording the highest precision (0.9393 for Normal, 0.9826 for Pneumonia), recall (0.9547 for Normal, 0.9764 for Pneumonia), F1-score (0.9469 for Normal, 0.9795 for Pneumonia), and AUC-ROC (0.9946), effectively addressing the class imbalance evident in the "Normal" category, where TinyDeiT (0.8807 recall) and ViT-LoRA (0.8533 precision) underperformed.

The efficiency metrics reinforce LiteXrayNet's suitability for resource-constrained environments, with a model size of 0.70 MB and inference time of 0.60 ms/sample, significantly lower than ViT-LoRA's 327.86 MB and 12.55 ms/sample, while remaining competitive with MobileNetV3-Small's 3.59 MB and 0.26 ms/sample (Chen & Ran, 2020). Resource utilization data further highlight LiteXrayNet's advantage, with average CPU usage of 5.6% and RAM usage of 69.2%, and maximums of 16.1% and 71.3%, respectively, compared to peaks of 100% CPU for ResNet-18 and EfficientNet-B0. This efficiency stems from its lightweight design, which reduces computational overhead without sacrificing accuracy, aligning with the aim of deploying AI on edge devices with limited hardware capabilities. Interpretability analyses provide additional evidence of LiteXrayNet's strength, with Grad-CAM heatmaps for true positive cases showing focused activation on lung regions with opacifications (e.g., consolidations, interstitial markings), consistent with radiological features of pneumonia, and true negative cases exhibiting diffuse low-intensity gradients, indicating accurate normal classification. In contrast, false positive and false negative cases reveal activation on non-diagnostic areas or weak responses, respectively, suggesting areas for refinement but also underscoring the model's transparency.

Feature map visualizations complement these findings, illustrating LiteXrayNet's hierarchical learning process, where early layers [1,16,112,112] capture lung contours and textures, intermediate MBConv blocks [1,32,56,56] highlight structural anomalies like consolidations, and the quantum-inspired phase shift layer [1,128,14,14] refines activations to emphasize subtle pathological cues. This progression aligns with clinical domain knowledge, focusing on lung fields while minimizing attention to artifacts, a capability less evident in

larger models like ViT-LoRA, which prioritizes parameter-heavy processing over targeted feature extraction. The combined evidence: high accuracy, balanced classwise performance, low resource usage, and interpretable outputs, positions LiteXrayNet as the best-performing model for the stated aim. Its lightweight architecture (179,646 parameters) and efficiency (0.70 MB, 0.60 ms/sample) enable deployment on edge devices, while its interpretability, validated by Grad-CAM and feature maps, ensures clinical trust, surpassing the trade-offs observed in baseline models that either sacrifice accuracy (e.g., TinyDeiT) or efficiency (e.g., ViT-LoRA).

8 Limitations

The study of LiteXrayNet is subject to certain limitations that influence its practical applicability. The performance metrics, such as the inference time of 0.60 ms/sample and resource utilization of 5.6% average CPU, were evaluated under simulated conditions rather than on actual edge devices, which restricts the assessment of its operational performance across diverse hardware platforms with varying computational capabilities (Cao et al., 2021). Additionally, the dataset used for training and testing lacks detailed information regarding its size, diversity, or representation of different patient populations, potentially limiting the model's robustness and generalizability to real-world clinical scenarios. Furthermore, the development process did not incorporate direct input or validation from clinical experts, such as radiologists, which may affect the alignment of the model's predictions with established diagnostic criteria or clinical workflows.

The absence of comprehensive support from clinical experts also poses a challenge to the study's interpretability and validation efforts. The Grad-CAM and feature map visualizations, while informative, were conducted without expert guidance to confirm the clinical relevance of highlighted regions, such as opacifications or non-diagnostic activations, potentially overlooking nuanced diagnostic features. This lack of expert oversight, combined with the reliance on a single, unspecified dataset, underscores the need for enhanced collaboration to ensure the model's outputs meet the expectations of healthcare professionals. These limitations suggest areas where further refinement could strengthen LiteXrayNet's readiness for clinical deployment.

9 Future Work

To address the identified limitations, future work will prioritize the deployment of LiteXrayNet on actual edge devices to evaluate its performance under real-world conditions, including processing power, memory constraints, and energy efficiency. This will involve testing across a variety of edge hardware platforms to validate the reported efficiency metrics and ensure compatibility with the intended deployment environments, providing a more accurate assessment of its practical utility. Expanding the dataset with detailed documentation of its size, diversity, and demographic representation will also be pursued, enabling a more robust evaluation of the model's generalizability across different clinical populations and imaging conditions. Additionally, exploring ensemble models and recurrent neural networks (RNNs) will be investigated to enhance predictive performance, leveraging LiteXrayNet's architecture as a foundation for improved accuracy and temporal analysis of sequential X-ray data (Islam et al., 2021). Future research will further enhance LiteXrayNet by integrating clinical expert guidance throughout the development and validation phases (Kelly et al., 2019).

Future research will further enhance LiteXrayNet by integrating clinical expert guidance throughout the development and validation phases. Collaboration with radiologists will facilitate the refinement of interpretability analyses, such as Grad-CAM and feature map visualizations, ensuring that highlighted regions align with clinically significant features and improving diagnostic accuracy. This expert input will also support the creation of a more representative dataset and the establishment of validation protocols that reflect real-world clinical standards. Furthermore, building on LiteXrayNet as a baseline, a multi-disease detection convolutional neural network will be developed to extend its capability to identify multiple pathologies beyond pneumonia, broadening its clinical utility while retaining its efficiency and interpretability.

10 Conclusion

This study introduces LiteXrayNet, a lightweight convolutional neural network (CNN) engineered for accurate and interpretable pneumonia detection from chest X-rays, specifically tailored for resource-constrained environments. Benchmarked against established models such as ResNet-18, MobileNetV3-Small, EfficientNet-B0, MobileViT, TinyDeiT, and ViT-LoRA, LiteXrayNet achieves superior performance in distinguishing normal from pneumonia cases, effectively addressing class imbalance through its optimized architecture. By incorporating MobileNetV3-inspired Mobile Inverted Bottleneck Convolutional (MBConv) blocks with depthwise separable convolutions, hard-swish activations, and squeeze-and-excitation modules, alongside a novel quantum-inspired phase shift layer, the model strikes an optimal balance between diagnostic precision and computational efficiency. Its compact design, with a minimal parameter count and small model size, enables rapid inference with low resource utilization, positioning LiteXrayNet as an ideal solution for edge devices in point-of-care diagnostics, particularly in underserved regions with limited access to advanced medical infrastructure.

The model's interpretability is enhanced through Gradient-weighted Class Activation Mapping (Grad-CAM) and feature map visualizations, which provide transparent insights into its decision-making process. These visualizations demonstrate that LiteXrayNet focuses on clinically relevant lung regions, such as areas of consolidation and interstitial markings, while minimizing attention to non-diagnostic artifacts, aligning with radiological expectations and fostering trust among healthcare professionals. The hierarchical feature extraction, progressing from low-level lung contours to refined pathological cues, further validates the model's ability to capture diagnostically significant patterns. Indeed, the theories and results support our model's performance in resource-constrained devices to be efficient; however, we are not claiming its efficiency until proper tests are conducted, though our theories and results strongly support its potential. Evaluated under simulated conditions, LiteXrayNet represents a significant advancement in medical imaging AI, offering a scalable, interpretable, and resource-efficient solution with the potential to transform pneumonia diagnosis and enhance healthcare equity in resource-limited global communities.

Acknowledgments

Omitted for review.

References

Ziyad Alaskar, Nazmul Hussain, Sadiq Khan, Mohammed Yeasin, Fawaz Alsulami, Thamer Alghamdi, et al. Efficient and accurate pneumonia detection using a novel multi-scale transformer model. *arXiv preprint arXiv:2408.04290*, 2024.

Ziyad Alaskar et al. Vision transformer for pneumonia classification in x-ray images. *Proceedings of the 2023 6th International Conference on Signal Processing and Information Security (ICSPIS)*, 2023. URL `https://dl.acm.org/doi/10.1145/3591569.3591602`.

Anonymous. Adaptive hybrid quantum-classical convolutional neural networks for medical image classification. *arXiv preprint arXiv:2302.01452*, 2023.

Adrian P. Brady. Error and discrepancy in radiology: inevitable or avoidable? *Insights into Imaging*, 8(1): 171–182, 2017. doi: 10.1007/s13244-016-0534-1. URL `https://doi.org/10.1007/s13244-016-0534-1`.

Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018. doi: 10.1016/j.neunet.2018.07.011.

Kai Cao, Yang Liu, Gong Meng, and Qingyang Sun. An overview on edge computing research. *IEEE Access*, 9:85714–85728, 2021. doi: 10.1109/ACCESS.2021.3081510. URL `https://doi.org/10.1109/ACCESS.2021.3081510`.

Aditya Chattopadhay, Aniruddha Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. *2018 IEEE*

Winter Conference on Applications of Computer Vision (WACV), pp. 839–847, 2018. doi: 10.1109/WACV.2018.00097.

Jian Chen and Xukan Ran. Deep learning with edge computing: A review. *Proceedings of the IEEE*, 108 (8):1655–1674, 2020. doi: 10.1109/JPROC.2020.2997834. URL https://doi.org/10.1109/JPROC.2020.2997834.

Md. Rakibul Haque Chowdhury, Md. Tofael Islam, Md. Mokhlesur Rahman, Md. Hasibul Kabir, and Mohammad Shafiul Islam. Enhanced pneumonia detection in chest x-rays using hybrid cnn-vision transformer model. *Journal of Imaging Informatics in Medicine*, 2024. URL https://pubmed.ncbi.nlm.nih.gov/39806960/.

Debapriya Das, Shreya Ghosh, and Soumi Samanta. Neural architecture search for efficient deep learning models for pneumonia detection using chest x-rays. *Scientific Reports*, 12(1):5678, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Song Han, Huizi Mao, and William J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009. doi: 10.1109/TKDE.2008.239.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Essam H. Houssein, Zainab Abohashima, Mohamed Elhoseny, and Waleed M. Mohamed. Hybrid quantum-classical convolutional neural network model for covid-19 prediction using chest x-ray images. *Journal of Computational Design and Engineering*, 9(2):343–363, 2022. doi: 10.1093/jcde/qwab068. URL https://doi.org/10.1093/jcde/qwab068.

Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1314–1324, 2019. doi: 10.1109/ICCV.2019.00140.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7132–7141, 2018.

Md. Tofael Islam, Md. Rakibul Haque Chowdhury, Mohammad Shafiul Islam, Md. Hasibul Kabir, and Md. Mokhlesur Rahman. Ensemble deep learning for multi-label classification of chest x-rays. *Journal of Medical Imaging*, 8(5):054001, 2021. doi: 10.1117/1.JMI.8.5.054001. URL https://doi.org/10.1117/1.JMI.8.5.054001.

Justin M. Johnson and Taghi M. Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):27, 2019. doi: 10.1186/s40537-019-0192-5. URL https://doi.org/10.1186/s40537-019-0192-5.

Alex Ke, William Elliker, Zhong Yang, Jing Yang, and Jasjit S. Suri. Chexformers: Vision transformers for chest x-ray pneumonia detection. *arXiv preprint arXiv:2104.06979*, 2021.

Christopher J. Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17(1):195, 2019. doi: 10.1186/s12916-019-1426-2. URL https://doi.org/10.1186/s12916-019-1426-2.

V. Kulkarni, P. K. Sinha, and M. C. Petare. Hybrid classical-quantum convolutional neural network for pneumonia detection. *International Journal of Quantum Information*, 20(7):2250025, 2022.

Rohit Kundu, Ritacheta Das, Zong Woo Geem, Gi-Tae Han, and Ram Sarkar. Pneumonia detection in chest x-ray images using an ensemble of deep learning models. *PLOS ONE*, 16(9):e0256630, 2021.

Guosheng Liang and Lei Zheng. A transfer learning method with deep residual network for pediatric pneumonia diagnosis. *Computer Methods and Programs in Biomedicine*, 187:105272, 2020.

Jiachen Liu, Huazheng Xu, Jingbo Zhu, Jinxu Wang, Bo Yang, Xuedong Wang, Baoliang Chen, Jianguo Yan, and Bo Han. Deep learning-based computer-aided diagnosis in medical imaging. *Frontiers in Medicine*, 10, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Sachin Mehta and Mohammad Rastegari. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178*, 2021.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed precision training. *International Conference on Learning Representations*, 2018.

Paul Timothy Mooney. Chest x-ray images (pneumonia). https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia, 2018.

Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 807–814, 2010.

Yujin Oh, Sangjoon Park, and Jong Chul Ye. Deep learning covid-19 features on cxr using limited training data sets. *IEEE Transactions on Medical Imaging*, 39(8):2688–2700, 2020.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.

Tawsifur Rahman, Muhammad E. H. Chowdhury, Amith Khandakar, Khandaker Reajul Islam, Khandaker Foisal Islam, Zaid Bin Mahbub, Muhammad Abdul Kadir, and Saad Kashem. Transfer learning with deep convolutional neural network (cnn) for pneumonia detection using chest x-ray. *Applied Sciences*, 10(7):2633, 2020.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, David Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

Prajit Ramachandran, Barret Zoph, and Quoc V. Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017. URL https://arxiv.org/abs/1710.05941.

M. Samra et al. Efficacy of lightweight vision transformers in diagnosis of pneumonia. *medRxiv*, 2024. URL https://www.medrxiv.org/content/10.1101/2024.10.24.24316057v1.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. *arXiv preprint arXiv:1801.04381*, 2018.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *arXiv preprint arXiv:1610.02391*, 2017.

Connor Shorten and Taghi M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019. doi: 10.1186/s40537-019-0197-0. URL `https://doi.org/10.1186/s40537-019-0197-0`.

Rajesh Kumar Singh, Rajshree Pandey, and Rajesh N. V. P. S. Babu. Qcsa: Quantum convolutional neural network with self-attentive mechanism for pneumonia detection in chest x-ray images. *Scientific Reports*, 13(1):1234, 2023.

Okeke Stephen, Mangal Sain, Ugochukwu J. Maduh, and Do-Un Jeong. An efficient deep learning approach to pneumonia classification in healthcare. *Journal of Healthcare Engineering*, 2019:4180949, 2019.

Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *International Conference on Machine Learning (ICML)*, pp. 6105–6114, 2019.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020.

World Health Organization. Pneumonia fact sheet. `https://www.who.int/news-room/fact-sheets/detail/pneumonia`, 2023.

Kan Wu, Jinnian Zhang, Houwen Peng, Mengchen Liu, Jianlong Xiao, Jianlong Fu, and Lu Yuan. Tinyvit: Fast pretraining distillation for small vision transformers. *European Conference on Computer Vision*, 2022. URL `https://arxiv.org/abs/2207.10666`.

Xiaohan Zhang, Xingyu Zhou, Yutian Lin, and Jian Sun. Quantum-inspired neural networks and their applications in computer vision. *arXiv preprint arXiv:2207.10498*, 2022a.

Xu Zhang et al. Towards efficient cnn-based models for covid-19 detection via pruning and knowledge distillation. *arXiv preprint arXiv:2203.07653*, 2022b.

Chen Zhao and Yang Chen. Quantum convolutional neural network for medical image classification. *Quantum Information Processing*, 21(5):174, 2022.

A Ablation Study

To rigorously evaluate the individual contributions of key architectural components in our proposed LiteXrayNet model, we conducted a comprehensive ablation study. This analysis systematically removes or modifies specific elements of the model to quantify their impact on diagnostic performance, computational efficiency, and resource utilization. By isolating these components, we provide empirical evidence of their necessity and effectiveness, ensuring the model's design is both justified and optimized for edge-device deployment in pneumonia detection, with particular emphasis on the superior performance of the original LiteXrayNet configuration.

The ablation variants derived from the original LiteXrayNet architecture include the complete model with all components serving as the baseline; a version without the quantum-inspired phase shift layer to assess its role in enhancing feature representation; a version excluding squeeze-and-excitation (SE) blocks (Hu et al., 2018) to evaluate their contribution to channel-wise recalibration; a version replacing the hard-swish (HSwish) activation (Howard et al., 2019) with rectified linear unit (ReLU) (Nair & Hinton, 2010) to compare activation function efficiency; a version substituting depthwise separable convolutions in MBConv blocks with standard convolutions to measure the benefits of lightweight operations; and a version trained without data augmentation or class weighting to highlight their importance in addressing class imbalance and overfitting.

All variants were trained and evaluated under identical conditions using the Chest X-ray dataset (Mooney, 2018), stratified splits (70% train, 15% validation, 15% test), AdamW optimizer (Loshchilov & Hutter, 2017) with a learning rate of 0.001, batch size of 32, and 50 epochs. Results are averaged over three independent runs with different random seeds to ensure statistical robustness, reported as mean values. Additionally, training dynamics were analyzed through mean training curves, which illustrate the convergence behavior across variants over 50 epochs. The original LiteXrayNet consistently exhibits superior training and validation accuracy (0.980 and 0.967, respectively), reinforcing its robustness and stability compared to modified variants.

Table 6 presents the performance metrics, including accuracy, loss, F1-score, recall, precision, and AUC-ROC. The original LiteXrayNet achieves the highest test accuracy (0.965) and AUC (0.992), underscoring the synergistic benefits of all components and establishing it as the optimal configuration for superior diagnostic precision. Removing the quantum layer results in the most significant accuracy drop (to 0.942), indicating its crucial role in capturing complex patterns, consistent with prior quantum-inspired enhancements in medical imaging (Zhao & Chen, 2022). Excluding SE modules leads to a moderate decline (to 0.960), affirming their importance for adaptive feature weighting. Replacing HSwish with ReLU yields lower accuracy (0.953) but with faster inference (as shown later), suggesting HSwish's non-linearity provides accuracy gains at the cost of efficiency. Standard convolutions also reduce accuracy (to 0.953) while increasing parameters, validating the efficiency of depthwise operations. Finally, omitting augmentation and class weights causes notable degradation (to 0.956), emphasizing their critical role in handling dataset imbalance (Buda et al., 2018).

Table 6: Comprehensive Ablation Study Results: Performance Metrics (Mean)

Model Variant	Train Acc	Val Acc	Test Acc	Test Loss	F1	Recall	Precision	AUC
Original LiteXrayNet	0.980	0.967	0.965	0.139	0.976	0.975	0.977	0.992
Without Quantum Layer	0.952	0.957	0.942	0.182	0.960	0.957	0.963	0.983
Without SE Modules	0.962	0.959	0.960	0.157	0.973	0.978	0.968	0.988
With ReLU Activation	0.959	0.954	0.953	0.179	0.968	0.971	0.964	0.985
Standard Convolutions	0.959	0.960	0.953	0.127	0.967	0.959	0.977	0.991
No Aug + No Class Weights	0.979	0.959	0.956	0.134	0.970	0.977	0.963	0.983

Efficiency metrics, detailed in Table 7, highlight LiteXrayNet's suitability for edge devices. The original model balances reasonable inference time (0.60 ms) and compact size (0.70 MB) with 179,646 parameters, demonstrating its efficiency advantages over modified variants. Removing components like the quantum layer or SE modules slightly reduces size and time but at the expense of accuracy, while standard convolutions inflate parameters (295,074) and size (1.13 MB) despite faster inference. GPU utilization is notably high across variants, with the original achieving balanced resource usage.

Table 7: Efficiency and Resource Usage Metrics (Mean)

Model Variant	Inf Time (ms)	Model Size (MB)	Params	CPU Avg %	GPU Avg %
Original LiteXrayNet	0.60	0.70	179,646	10.3	45.9
Without Quantum Layer	0.58	0.67	171,134	10.3	62.7
Without SE Modules	0.58	0.66	167,954	10.2	64.3
With ReLU Activation	0.33	0.70	179,646	10.1	61.8
Standard Convolutions	0.21	1.13	295,074	6.7	42.0
No Aug + No Class Weights	0.59	0.70	179,646	5.4	17.6

Training resource usage, presented in Table 8, further illustrates the model's efficiency during training. The original LiteXrayNet shows moderate CPU and RAM utilization, with variants without augmentation and class weights exhibiting the lowest demands due to simpler data handling.

Table 8: Training Resource Usage Metrics (Mean)

Model Variant	Train CPU Avg	Train CPU Max	Train RAM Avg	Train RAM Max
Original LiteXrayNet	11.2%	26.3%	73.7%	77.6%
Without Quantum Layer	11.1%	20.2%	73.9%	76.2%
Without SE Modules	11.1%	19.0%	74.5%	76.9%
With ReLU Activation	11.2%	17.3%	74.8%	76.7%
Standard Convolutions	9.1%	16.1%	75.2%	76.7%
No Aug + No Class Weights	5.8%	11.3%	75.0%	75.8%