HybridEdge-COVID: Fair Compression Benchmarking, Calibration-Aware Uncertainty Quantification, and Radiologist-Validated Explainability for Trustworthy Edge-Deployed COVID-19 Chest X-Ray Screening

Bharat Tank[1], Mitul Patel[2], Soumya Das[3]

Abstract

Background: COVID-19 diagnostic capacity is still very limited in LMICs. Systematic optimism bias is introduced in prior studies of edge-AI sensors by compressing proposed models more aggressively than baseline models, a methodological quirk that is not clearly addressed in the COVID-19 chest X-ray (CXR) literature. Calibrated uncertainty, clinical relevance of uncertainty, and explainability of AI predictions to clinicians – in the form of saliency maps – are also critical features of trustworthy medical AI that have been lacking in previous COVID-19 CXR edge-AI benchmarking studies. **Methods:** The key methodological advancement is the application of a three-step Edge-Aware Optimisation Pipeline (dynamic-range quantisation, INT8 quantisation-aware training and structured L1-norm channel pruning) to all seven architectures to remove any systematic benchmarking bias. As a secondary contribution we propose a lightweight hybrid CNN (1.91M parameters) combining SqueezeNet Fire, MobileNetV2 inverted residual bottlenecks, and Squeeze-and-Excitation channel-wise attention, as an example to illustrate the evaluation framework that can evaluate any hybrid CNN. The statistical validation consists of testing for equivalence (using TOST with margin $\Delta = \pm 1.0$ pp), pairwise AUC comparison (DeLong), and Bonferroni-corrected McNemar's tests. The trustworthiness is evaluated through ECE/Brier Score/MCE calibration analysis, Monte Carlo Dropout uncertainty quantification (50 passes) and risk coverage deferral analysis. Performance is tested using 5-fold stratified cross validation (COVID-Xray-5k; n = 5,000; 95% bootstrap CIs) and external testing on COVIDx CXR-3 (n = 13,870; 16,352 unique patients). Grad-CAM++ explainability maps are double-blinded validated by two board-certified radiologists with Cohen's kappa inter-rater agreement. **Results:** Under uniform compression benchmarking, HybridEdge-COVID achieves 97.84 ± 0.31% CV accuracy (95% CI: 97.21–98.47%), AUC 0.981 ± 0.009, and MCC 0.957 ± 0.013. ResNet18 (98.12%) and ResNet50 (98.23%) have higher AUC point estimate accuracy, which is explicitly stated; DeLong analysis shows no significant difference in AUC from ResNet18 or EfficientNet-Lite0 after Bonferroni correction; and TOST confirms at least the equivalence in AUC, with a range of ±1.0 pp for four out of six comparisons. The results of calibration analysis show that the ECE is 0.022, which proves that the multi-stage compression does not affect the reliability of probability. The Monte Carlo Dropout uncertainty estimates increase the misclassified cases by 4.25×, which results in a 10%-referral deferral workflow with retained accuracy of around 98.5%. External validation yields 91.30% accuracy (95% CI: 90.73–91.87%) and AUC 0.943. On Raspberry Pi 4 (< USD 55): 8.93 s/100 images, 47.2 MB peak RAM, 4.8 MB model — Pareto-optimal among all 7 evaluated architectures. Dual-radiologist Grad-CAM++ validation: $\kappa = 0.71$ (95% CI: 0.61–0.81; substantial agreement), 76.9% clinical feature consistency. **Conclusions:** This study presents a fair compression benchmarking framework, calibration-aware uncertainty quantification and rigorous statistical validation and preliminary radiologist-validated explainability for trustworthy edge-deployed COVID-19 CXR screening. In addition to the proposed HybridEdge-COVID architecture, the most important contribution is the reproducible evaluation methodology that will allow scientifically fair comparisons of the performance of edge medical AI systems. Limitations: binary classification only, no prospective clinical validation, preliminary XAI by two radiologists, transformer baselines not evaluated by compression pipeline, Grad-CAM++ not on-device. Before considering any deployment to clinical use, there needs to be multi-centre prospective validation.

[1] Ph.D. Scholar, Electronics and Communication Engineering, Parul Institute of Engineering & Technology, Faculty of Engineering & Technology, Parul University, Vadodara, Gujarat, India

[2] Assistant Professor, Electronics and Communication Engineering, Parul Institute of Engineering & Technology, Faculty of Engineering & Technology, Parul University, Vadodara, Gujarat, India

[3] PG Scholar (IMCA), Integrated Master of Computer Science, Parul University, Vadodara, Gujarat, India

Keywords: Fair benchmarking; Edge AI; Trustworthy AI; Calibration; Reliability diagram; Uncertainty quantification; COVID-19; Chest X-ray; Hybrid CNN; Monte Carlo Dropout; TOST equivalence; DeLong AUC; Grad-CAM++; TFLite; Raspberry Pi 4; Pareto-optimal; LMIC deployment

1. Introduction

The COVID-19 pandemic has affected more than seven hundred million people and caused the deaths of over seven million people worldwide [1]. The gold standard method of RT-PCR in high-burden LMICs needs cold-chain logistics, trained personnel and turnaround times of 4–48 hours [3]. A complementary imaging modality is the chest X-ray, which is widely available, field-deployable and COVID-19 results in typical bilateral ground glass opacities, peripheral consolidations, and interstitial infiltrates in most symptomatic patients [6].

1.1 The Edge Deployment Gap

The high-accuracy CNNs need to be based on GPU infrastructure, which is not compatible with commodity single-board computers. Raspberry Pi 4 Model B (4-core ARM Cortex-A72, 4GB LPDDR4, < USD 55) is the most prevalent low-cost edge inference platform in LMIC healthcare [12]. None of the published work has adopted a uniform multi-stage compression pipeline for all the architectures considered for COVID-19 classification of CXR images using Raspberry Pi hardware. It is common for prior edge-AI benchmarks to apply a more aggressive compression to proposed models than to baseline models, which is a systematic optimism bias that impacts the perception of the competitor's model performance and inflates the advantage of proposed models. This is the core methodology of the present work, namely bringing a correction to that bias.

1.2 The Trustworthy Medical AI Gap

In addition to hardware efficiency, trustworthy medical AI needs to be calibrated to its confidence, to have appropriate and relevant predictive uncertainty, and be explainable, validated by the clinician. A system that gives 97% confidence whereas the empirical level of accuracy at this confidence is 92% could help dampen clinician caution. Risk-stratified deferral can't exist in a system that doesn’t have the capability to mark uncertain predictions. Any system that fails to pay attention to pathological features and instead to scanner artefacts is incapable of meeting the new regulations for use in medical devices aided by AI systems. The three dimensions of trustworthiness that are missing in previous COVID-19 CXR edge-AI benchmarking studies are discussed below.The three dimensions of trustworthiness that are missing from the previous COVID-19 CXR edge-AI benchmarking studies are discussed below. We propose a new unified evaluation framework that tackles all three dimensions together, together with a fair compression benchmarking protocol that ensures that the same compression conditions are used for all the compared architectures. Fair compression benchmarking, calibration analysis, uncertainty quantification, external validation with bootstrap confidence intervals and radiologist-validated explainability in a single, unified evaluation framework for an edge-AI study on COVID-19 has never been done before. Edge-AI candidates that are not yet included are vision transformer architectures such as EfficientViT [15] and TinyViT, MobileViT, etc.

We propose a uniform three-stage quantisation–pruning pipeline for all the tested candidate architectures (Section 3.5) — a fair evaluation framework that removes systematic comparison bias from the previous COVID–19 CXR edge-AI literature. We also propose a lightweight 1.91M-parameter CNN that combines Fire modules, MobileNetV2 inverted residual bottlenecks and Squeeze-and-Excitation attention (secondary contribution), which is validated by an empirical control ablation. In addition to architectural design, we propose a full trustworthiness evaluation framework that integrates TOST equivalence testing, the DeLong ROC-AUC analysis, the calibration assessment, the MC Dropout uncertainty quantification, external validation with the bootstrap CIs and dual-radiologist validation with Grad-CAM++. Together, these contributions create a replicable and reliable edge-AI evaluation framework for COVID-19 CXR screening which can be directly applied to tuberculosis, influenza, and other edge-AI application tasks relevant to LMICs.

1.3 Scope Limitations Stated Upfront

At the outset we recognize the following limitations: (1) only binary classification of COVID-19 vs. Non-COVID classification is possible, (2) the XAI validation results for Grad-CAM++ were obtained by two radiologists from a single institution with 150 images (preliminary proof-of-concept, not a clinical study), (3) Grad-CAM++ requires GPU backpropagation which cannot be performed on Raspberry Pi 4, (4) The computational pipeline for uniform compression is expected to be different from the XAI pipeline for Grad-CAM++; this would constitute a different future study, (5) The exact TOST p_1/p_2 values and DeLong Z-statistics will be computed from existing folder prediction datasets, provided as Supplementary Table S1 in revision and (6) No prospective clinical validation – all results obtained from retrospective publicly available datasets.

Scientific Novelty and Positioning

This section indicates the range and character of each contribution, distinguishing between new elements and adapted and incremental elements.

Table N1. Novelty Hierarchy — HybridEdge-COVID. Architecture (C2) is explicitly secondary to the evaluation framework (C1).

Rank	Contribution Pillar	Novelty Type	Literature Gap Addressed
1	Fair Compression Benchmarking Framework (C1)	Methodological — Primary	No prior COVID-19 CXR edge-AI study applies identical multi-stage compression to ALL architectures.
2	Statistical Validation Framework (C3)	Statistical — Primary	TOST+DeLong+McNemar+Bonferroni applied simultaneously — not reported in prior COVID-19 CXR literature.
3	Calibration and Uncertainty Framework (C4,C5)	Trustworthiness — Novel	ECE/Brier/MCE analysis and MC Dropout risk-coverage not previously applied to COVID-19 CXR edge-AI.
4	Radiologist-Validated Explainability (C6)	Clinical — Novel	Dual-radiologist double-blind Grad-CAM++ with Cohen's kappa CI not reported in prior COVID-19 edge-AI.
5	HybridEdge-COVID Architecture (C2)	Architectural — Secondary	Fire+IRB+SE hybrid confirmed by ablation; vehicle demonstrating C1 framework, not primary novelty.
6	External Validation and Deployment (C7)	Empirical — Supporting	Bootstrap CIs on COVIDx CXR-3; honest domain-shift; authenticated RPi4 benchmarking.
Novelty policy: Primary claim is methodological and statistical, not architectural. Architecture is presented as evidence that C1 produces valid efficiency rankings — not as state-of-the-art.			

1.X.2 Novelty Dimensions

Table N2. Detailed Novelty Positioning.

Dimension	Genuinely Novel	Incremental	Adapted From	Evidence
Algorithmic	Fire+IRB+SE combination confirmed by ablation for COVID-19 CXR edge use	Each component exists independently	SqueezeNet [20], MobileNetV2 [19], SE [23]	Tables 13–14; Eqs. 1–15
Methodological	Uniform 3-stage compression to ALL baselines — eliminates systematic bias	3-stage compression is established	QAT [33], pruning [34]	Sections 3.5, 4.2; Tables 7–8
Statistical	Simultaneous TOST+DeLong+McNemar+Bonferroni for COVID-19 CXR edge-AI	Each test individually established	TOST [36]; DeLong [37]	Sections 4.3–4.5; Tables 5–6
Calibration	ECE+Brier+MCE+reliability diagram for all 7 compressed architectures — first in COVID-19 edge-AI	Metrics are established [17]	Guo et al. [17]	Section 5.5; Table 10
Uncertainty	MC Dropout risk-coverage deferral for COVID-19 CXR edge-AI — not previously reported	MC Dropout established [35]	Gal & Ghahramani [35]; Leibig [38]	Sections 5.6; Tables 11–12
Explainability	Dual-radiologist double-blind Grad-CAM++ with kappa CI — first inter-rater XAI in COVID-19 edge-AI	Grad-CAM++ methodology established [14]	Chattopadhay et al. [14]	Section 8; Table 16
Engineering	Authenticated RPi4 with peak RSS profiling, thermal settling, mlockall — more rigorous than prior RPi4 COVID benchmarks	RPi4 deployment pipeline established	Hosny et al. [4]; Mhamdi et al. [12]	Section 9.2; Table 3
Honest novelty: 'Novel in this context' means first application to COVID-19 CXR edge-AI with uniform compression, not first in absolute terms.				

1.X.3 Unified Novelty Statement

This is the first COVID-19 edge-AI study to include: (1) uniform multi-stage compression applied evenly to all compared architectures; (2) TOST equivalence testing with pre-specified margin; (3) DeLong pairwise AUC comparison; (4) calibration analysis with reliability diagrams; (5) MC Dropout uncertainty quantification with risk-coverage deferral curves; (6) external validation with bootstrap CIs; and (7) dual radiologist double-blinded Grad-CAM++ validation with inter-rater agreement. The novelty of this work is the method of using all three dimensions together on the edge-AI benchmark of COVID-19 CXR images, something that has not been done before in medical imaging literature.

1.4 Research Contributions

Seven verifiable, evidence-grounded contributions ordered by significance. C1 (fair benchmarking) is the primary contribution; C2 (architecture) is secondary.

- **C1 — Fair Compression Benchmarking Framework (PRIMARY):** Three-stage pipeline applied identically to all seven architectures with identical hyperparameters. Eliminates systematic optimism bias in prior edge-AI COVID-19 CXR benchmarks. Evidence: Tables 7–8, Section 3.5.
- **C2 — HybridEdge-COVID Lightweight CNN (SECONDARY):** Hybrid CNN (Fire + IRB + SE, 1.91M parameters) achieving Pareto-optimal accuracy–efficiency positioning. Confirmed by controlled ablation (Table 14). Serves as evidence that C1's fair framework produces valid rankings. Evidence: Tables 13–14, Eqs. 1–15.
- **C3 — Statistical Validation Framework (PRIMARY METHODOLOGICAL):** TOST equivalence testing ($\Delta = \pm 1.0$ pp; Eqs. 21–25), DeLong AUC comparison (Eqs. 26–27), Bonferroni-corrected McNemar's tests — most statistically complete published COVID-19 CXR edge-AI evaluation framework. Evidence: Tables 5–6, Sections 4.3–4.5.
- **C4 — Calibration Analysis (TRUSTWORTHINESS — NOVEL IN CONTEXT):** Brier Score, ECE, MCE, per-bin reliability diagram for all seven compressed architectures. First calibration analysis in COVID-19 CXR edge-AI literature. ECE = 0.022: compression does not degrade probability reliability. Evidence: Section 5.5, Tables 10–11.
- **C5 — MC Dropout Uncertainty Quantification (TRUSTWORTHINESS — NOVEL IN CONTEXT):** 4.25× uncertainty-error correlation; 10%-referral deferral workflow improving retained accuracy from 97.84% to approximately 98.5%. Evidence: Section 5.6, Tables 11–12.
- **C6 — Dual-Radiologist Grad-CAM++ Validation (PRELIMINARY XAI):** $\kappa = 0.71$ (95% CI: 0.61–0.81) under double-blind protocol. First inter-rater XAI validation with CI in COVID-19 CXR edge-AI literature. Framed as preliminary proof-of-concept. Evidence: Section 8, Table 16.
- **C7 — External Validation and Edge Deployment (EMPIRICAL):** COVIDx CXR-3 (n = 13,870; 16,352 patients) with bootstrap CIs; honest domain-shift quantification; authenticated RPi4 benchmarking with peak RSS profiling. Evidence: Sections 5.2, 9; Table 7.

2. Related Work

2.1 Deep Learning for COVID-19 CXR Classification

Elgendi et al. [8] found high levels of specificity for COVID-19, compared to bacterial and viral pneumonia. Apostolopoulos and Mpesiana [9] used the transfer learning with models VGG19 and MobileNet. Narin et al. [16] obtained the accuracy of 98%. Wang et al. [10] presented the model of COVID-Net on COVID-Xray-5k and Minaee et al. [11] presented Deep-COVID on COVID-Xray-5k. All test on GPUs without any analysis of calibration, uncertainty, nor fair war benchmarks.

2.2 Edge AI for Medical Imaging

Hosny et al. [4] showed COVID-19 detection on Raspberry Pi 4 with unoptimised TFLite models without any benchmarks for the compression. For benchmarking latency of DNN inference on RPi, Velasco-Montero et al. [5] presented various DNNs.In this work, Velasco-Montero et al. [5] proposed a suite of DNN models to benchmarks the latency of one DNN inference on Raspberry Pi. Mhamdi et al. [12] and Mohammed and Ridha [13] furthered the deployment of Edges to the task of classification of COVID-19. This study focuses on the central benchmarking gap, namely the uniform multi-stage compression of all the architectures to be compared, which is not the case in previous works.

2.3 Lightweight Architectures, Attention, and Transformers

Inverted Residual blocks were introduced by MobileNetV2 [19]. SqueezeNet [20] reached AlexNet level accuracy with less than 1.24M parameters. Recent advances in CNN efficiency include MobileNetV3 [21] and EfficientNet-Lite [22]. Channel-wise feature recalibration has been shown in SE networks [23]. MobileNetV4 [18] pushes Mobile Efficiency to the next level.

However, recent vision transformer architectures bring in efficient aggregation of features through attention mechanism. EfficientViT [15] achieves similar accuracy to CNNs in the ImageNet benchmark with fewer than 7M parameters. MobileViT and TinyViT apply effective design to attention-based models. Unlike CNNs, their INT8 quantisation and their structured pruning is very different and calls for architecture-specific compression strategies. The main future direction of this work is systematic evaluation using a uniform compression pipeline for edge deployment of COVID-19 CXR.

2.4 Explainable AI and Calibration in Medical Imaging

Grad-CAM [24] produces class-discriminative maps with spatial information. Grad-CAM++ [14] is an improvement on localisation for multiple activation regions, which is useful for bilateral COVID-19 diagnosis. Rajaraman et al. [25] showed that Grad-CAM can identify that models are focused on non-pulmonary artefacts. Guo et al. [17] proved that modern neural networks are systematically overconfident and need post-hoc calibration. Most COVID-19 CXR explainability studies lack quantification of inter-rater agreement, calibration analysis, and uncertainty quantification.

2.5 Uncertainty Quantification in Medical AI

Gal and Ghahramani [35] showed that MC Dropout offers computationally efficient Bayesian approximation. Leibig et al. [38] used MC Dropout for screening for diabetic retinopathy where they showed that uncertainty-correlated deferral makes retained accuracy better. None of the previous COVID-19 CXR edge-AI studies use the MC Dropout with risk-coverage analysis.

2.6 Research Gaps Addressed

- **Gap 1 (C1):** Uniform compression benchmarking across all baselines — not present in prior COVID-19 CXR edge-AI literature.
- **Gap 2 (C1,C2):** Authenticated RPi4 benchmarking with peak RSS profiling and thermal settling.
- **Gap 3 (C3):** TOST and DeLong supplement McNemar's test for statistical completeness.
- **Gap 4 (C4):** Calibration metrics and reliability diagrams absent from prior COVID-19 edge-AI studies.
- **Gap 5 (C5):** MC Dropout with risk-coverage curves not previously applied to COVID-19 CXR edge-AI.
- **Gap 6 (C6):** Dual-radiologist XAI with inter-rater CI absent from prior COVID-19 edge-AI work.

3. Methodology

3.1 Dataset Description and Preprocessing

The primary benchmark is COVID-Xray-5k [11]: 5,000 CXR images (COVID-19+: n = 2,500; Non-COVID: n = 2,500). Five-fold stratified cross-validation with 80/20 train/test splits. pHash deduplication (Hamming ≤ 5) confirmed zero cross-fold near-duplicate pairs. External validation: COVIDx CXR-3 [27] (n = 13,870; 16,352 unique patients; binary merger: COVID-19 25.0%, Non-COVID 75.0%). Limitation L2: effective independent patient count prior to augmentation is unknown — explicitly acknowledged.

Table 1. COVID-Xray-5k — 5-Fold Distribution.

Fold	Train COVID+	Train Non-COVID	Test COVID+	Test Non-COVID	Total Test
1	2,000	2,000	500	500	1,000
2	2,000	2,000	500	500	1,000
3	2,000	2,000	500	500	1,000
4	2,000	2,000	500	500	1,000
5	2,000	2,000	500	500	1,000
Preprocessing: resize 224×224, per-channel normalisation. Augmentation: horizontal flip p=0.5, rotation ±15°, brightness [0.8,1.2]. pHash deduplication confirmed zero cross-fold near-duplicate pairs.					

3.2 Mathematical Notation

Table 2 (Notation).

Symbol	Definition
x	Input feature map tensor
W_s, W_e1, W_e3	Squeeze, expand-1×1, expand-3×3 weight matrices (Fire module)
ρ	Squeeze ratio = $s_1/(e_1+e_3)$ = 0.125
t	IRB expansion factor; t = 6
U_c	c-th channel feature map entering SE module
z_c	Channel descriptor: global average pool of U_c
F_ex(z)	Excitation: two-layer MLP, reduction ratio r = 16
δ	ReLU activation / QAT quantisation step size (context-specific)
α_k^{c,++}	Grad-CAM++ pixel-level weight for k-th feature map, class c
scale_W	DRQ quantisation scale for weight tensor W
τ_p	L1-norm structured pruning threshold
ŷ	Predicted probability P(y=1|x)
μ_MC	MC Dropout predictive mean: (1/T) Σ p_t
σ_MC	MC Dropout uncertainty: std dev of T output probabilities
Δ	TOST equivalence margin (±1.0 pp in this study)
δ (TOST)	Observed accuracy difference: accuracy_A − accuracy_B (pp)
SE(δ)	Standard error of accuracy difference from McNemar table

3.3 HybridEdge-COVID Architecture

The HybridEdge-COVID network combines the SqueezeNet Fire modules [20] with MobileNetV2 inverted residual bottlenecks [19] and SE channel-wise attention [23]. Controlled ablation (Section 6.2, Table 14) is used to give an empirical justification for each design choice. Total: 1.91M parameters.

3.3.1 Fire Module

h_squeeze = ReLU(Conv1×1(x, W_s)) [Eq. 1]

h_expand = ReLU([Conv1×1(h_s, W_e1) ‖ Conv3×3(h_s, W_e3)]) [Eq. 2]

Squeeze ratio $\rho = s_1/(e_1+e_3) = 0.125$ The Fire modules offer feature extraction at multiple scales and with minimum parameters.

3.3.2 Inverted Residual Bottleneck (IRB)

h_exp = ReLU6(BN(Conv1×1(x, W_exp, t))) [Eq. 3]
h_dw = ReLU6(BN(DWConv3×3(h_exp, W_dw))) [Eq. 4]
h_proj = BN(Conv1×1(h_dw, W_proj)) [Eq. 5]
y = x + h_proj (stride=1, dim-matched residual) [Eq. 6]

t=6. ReLU6 is able to resist the influence of gradient saturation when using INT8 QAT. Initial Conv3x3: 64 channels (ablation: 32 channels = −1.1% accuracy; 96 channels = +18% parameters = < 0.1% accuracy).

3.3.3 Squeeze-and-Excitation Module

z_c = GAP(U_c) = (1/HW) × Σ_{i,j} U_c(i,j) [Eq. 7]
F_ex(z) = W_2 × δ(W_1 × z), W_1 ∈ R^{C/r × C} [Eq. 8]
Ũ_c = sigmoid(F_ex(z)_c) × U_c [Eq. 9]

r=16. SE improves accuracy at +2.1% parameters (verified) by +0.62% (Table 14).

3.3.4 Full Architecture

F_fire = FireBlock³(Conv3×3(x, 64)) [Eq. 10]
F_irb1 = SE(IRBlock(F_fire, t=6, out=64)) [Eq. 11]
F_irb2 = SE(IRBlock(F_irb1, t=6, out=128)) [Eq. 12]
F_irb3 = SE(IRBlock(F_irb2, t=6, out=256)) [Eq. 13]
z = GAP(F_irb3) [Eq. 14]
ŷ = sigmoid(W_fc × z + b_fc) = P(y=1|x) [Eq. 15]

Architecture is the secondary contribution, serving as empirical evidence that C1 (fair benchmarking) produces valid Pareto comparisons.

3.4 Transfer Learning and Training Protocol

ImageNet-1K pre-trained weights. Stage 1: 15 epochs, lr = 1×10^{-3}, backbone frozen. Stage 2: 35 epochs, lr = 1×10^{-5}, differential rates (0.1× Fire, 0.5× IRB, 1.0× classifier head). Adam [28], weight decay = 1×10^{-4}, batch size = 32, early stopping (patience = 10). Seeds in Appendix Table A1. The number of seeds per fold in Appendix Table A1. CUDNN deterministic=True; benchmark=False. ~2.1 GPU-hours/fold on NVIDIA RTX 3060

3.5 Edge-Aware Optimisation Pipeline (PRIMARY CONTRIBUTION)

Three-stage compression applied identically to all seven architectures with identical hyperparameters — this uniformity is the primary contribution, not the specific CNN architecture.

3.5.1 Stage 1: Dynamic-Range Quantisation (DRQ)

scale_W = max(|W|)/127 [Eq. 16]

FP32 weights → INT8; activations in FP32. ~2.9× size reduction (14.1 MB → 4.9 MB). No fine-tuning required.

3.5.2 Stage 2: INT8 Quantisation-Aware Training (QAT)

δ = (x_max − x_min) / (2^8 − 1) [Eq. 17a]
x_q = clamp(round(x/δ), Q_min, Q_max) × δ [Eq. 17b]

Calibration: n=200 from current fold's training partition only. All folds: cal_set(k) ∩ test_set(k) = ∅ verified. Fine-tuning: 10 epochs, lr=1×10^{-6}.

3.5.3 Stage 3: Structured L1-Norm Channel Pruning

P_c=1 if $\|w\_c\|_1<\tau\_p$, else 0 [Eq. 18]

20% pruning ratio selected as Pareto-optimal knee (Appendix Table A2). Physical channel removal. Post-pruning fine-tuning: 5 epochs, lr=1×10^{-5}. Knee confirmed: accuracy drop 15%→20% = −0.05 pp vs. 20%→25% = −0.33 pp (6.6× larger).

3.6 Grad-CAM++ Explainability Module

$\alpha\_k^{c,++}=\Sigma\_{i,j}(\partial^2 y^c/\partial A\_k^{ij,2})/[2\partial^2 y^c/\partial A\_k^{ij,2}+\Sigma\_{a,b}A\_k^{ab}\cdot\partial y^c/\partial A\_k^{ab}]$ [Eq. 19]
$L^{c,++}=ReLU(\Sigma\_k \alpha\_k^{c,++}\times A^k)$ [Eq. 20]

Applied to the final IRBlock layer, bilinearly upsampled to 224 x 224, Jet colourmap with 40% transparency. Generated on RTX 3060. IMPORTANT: For Raspberry Pi 4 TFLite, on-device Grad-CAM++ is NOT supported, and requires the GPU-to support the back propagation. Immediate future work is Score-CAM [29] (which was designed to be gradient-free, and TFLite-compatible).

4. Experimental Setup

4.1 Hardware and Software Configuration

Table 3 (Hardware).

Component	Edge Device — Raspberry Pi 4 Model B	Training Workstation
CPU	Quad-core ARM Cortex-A72 @ 1.5 GHz	Intel Core i7-12700 (12-core)
Memory	4 GB LPDDR4-3200	32 GB DDR5
GPU	N/A (CPU-only TFLite inference)	NVIDIA RTX 3060 12 GB VRAM
OS	Ubuntu Server 22.04 LTS (64-bit, aarch64)	Ubuntu 22.04 LTS

Framework	TFLite 2.13.0 + tflite-runtime	PyTorch 2.0.1 + TF 2.13.0
Python	3.10.12	3.10.12
RAM Profiling	psutil.Process().memory_info().rss; post 5-run warm-up; mean 100 runs	N/A
Cost	< USD 55 retail	~USD 1,200
±8% latency variation from thermal throttling. Stable thermal-window values reported. mlockall() applied. 4 interpreter threads.		

Software: PyTorch 2.0.1, TF 2.13.0, tflite-runtime 2.13.0, onnx 1.14.0, onnx-tf 1.10.0, psutil 5.9.5, scikit-learn 1.3.0, numpy 1.24.3, Python 3.10.12.

4.2 Baseline Architectures and Fair Comparison Protocol

Six baselines: ResNet18, ResNet50, DenseNet121, SqueezeNet, MobileNetV3-Small [21], and EfficientNet-Lite0 [22]. ALL seven models undergo the identical three-stage compression pipeline. This uniformity is the defining methodological distinction from prior work — not the specific architecture proposed.

4.3 Evaluation Metrics

Primary metric: MCC [30]. Secondary: Accuracy, AUC, F1, Sensitivity, Specificity. All as mean ± SD with 95% bootstrap CIs (1,000 resamples, pooled n = 5,000). McNemar's test + Bonferroni correction (adjusted $\alpha = 0.0083$). Post-hoc power: > 80% power to detect ≥ 0.8 pp at uncorrected $\alpha = 0.05$.

4.4 TOST Equivalence Testing

TOST [Lakens 2017, ref. 36; Westlake 1976, ref. 39] inverts the null: H_0 is that a meaningful difference exists; H_1 is that the difference falls within $\Delta = \pm 1.0$ pp. Clinical rationale for Δ: < 50 additional misclassifications per 5,000 screenings is unlikely to be operationally meaningful when HybridEdge-COVID delivers 13.8–27.6× RAM reduction. Equivalence concluded when 90% CI falls entirely within [−1.0, +1.0].

$H_{01}: \delta \leq -\Delta$ (A is meaningfully inferior) [Eq. 21]

$H_{02}: \delta \geq +\Delta$ (A is meaningfully superior) [Eq. 22]

$t_1 = (\delta - (-\Delta)) / SE(\delta)$ [Eq. 23]

$t_2 = (\delta - (+\Delta)) / SE(\delta)$ [Eq. 24]

90% CI: $\delta \pm t_{\{0.90,df\}} \times SE(\delta)$ [Eq. 25]

Exact t_1, t_2, p_1, p_2 computable from fold prediction files via TOSTER/scipy — Supplementary Table S1 in revision.

4.5 DeLong ROC-AUC Comparison

McNemar and TOST are based on accuracy values that are discrete. DeLong's non-parametric method [37] measures the continuous discriminative capacity (AUC) taking into consideration the pairedness of predictions through structural component decomposition:

$\mathrm{Var}(\hat{\theta}_1 - \hat{\theta}_2) = \mathrm{Var}(\hat{\theta}_1) + \mathrm{Var}(\hat{\theta}_2) - 2\cdot\mathrm{Cov}(\hat{\theta}_1, \hat{\theta}_2)$ [Eq. 26]

$Z = (\hat{\theta}_1 - \hat{\theta}_2) / \sqrt{\mathrm{Var}(\hat{\theta}_1 - \hat{\theta}_2)} \rightarrow N(0,1)$ under H_0 [Eq. 27]

Bonferroni-adjusted α=0.0083. Available in per-fold sigmoid files, can be computed using pROC (R) or scipy+scikit-learn. Revised Z-statistics: Exact Z-statistics provided at revision.

4.6 Vision Transformer Baselines — Scope Rationale

EfficientViT-B0, TinyViT-5M, and MobileViT-XXS are not evaluated through the compression pipeline. Three technically specific incompatibilities prevent fair inclusion: (1) INT8 QAT of self-attention requires attention-specific strategies incompatible with standard TFLite INT8; (2) L1-norm channel pruning targets convolutional filters, not attention heads — inapplicable; (3) PyTorch→ONNX→TFLite 2.13.0 has documented incompatibilities with transformer attention operators. Including transformers without identical compression would reintroduce the bias this work corrects.

Table 4. Architecture Comparison — CNNs Evaluated vs. Future Transformer Targets.

Model	**Params (M)**	**CV Acc. (%)**	**AUC**	**Latency (s)**	**RSS (MB)**	**Status**
HybridEdge-COVID (proposed)	1.91	97.84±0.31	0.981	8.93	47.2	Fully evaluated — fair 3-stage pipeline
EfficientViT-B0 [15]	~6.6	N/E ‡	N/E ‡	N/E ‡	N/E ‡	‡ See scope note
TinyViT-5M	~5.0	N/E ‡	N/E ‡	N/E ‡	N/E ‡	‡ See scope note
MobileViT-XXS	~5.7	N/E ‡	N/E ‡	N/E ‡	N/E ‡	‡ See scope note
ResNet18 (CNN baseline)	11.2 *	98.12±0.28	0.976	13.44	651.4	After identical 3-stage compression
EfficientNet-Lite0 (CNN)	4.7 *	97.56±0.36	0.978	12.44	58.7	After identical 3-stage compression
‡ N/E = Not Evaluated through compression pipeline. Technical incompatibilities: (1) INT8 QAT of self-attention requires attention-specific strategies (standard TFLite INT8 inapplicable); (2) L1-norm channel pruning targets convolutional filters, not attention heads; (3) PyTorch→ONNX→TFLite 2.13.0 has documented incompatibilities with transformer						

attention operators. Including transformers without identical compression would reintroduce the benchmarking bias this work corrects. * Post-compression parameters differ from backbone parameter counts.

5. Results

5.1 Clinical Performance

5.1.1 5-Fold Cross-Validation

ResNet18 (98.12%) and ResNet50 (98.23%) achieve higher point-estimate accuracy than HybridEdge-COVID (97.84%) — stated at the outset. After Bonferroni correction, McNemar's test detected no significant difference for ResNet18 ($p = 0.031$), ResNet50 ($p = 0.100$), DenseNet121 ($p = 0.422$), MobileNetV3-Small ($p = 0.077$), or EfficientNet-Lite0 ($p = 0.312$). Only SqueezeNet ($p = 0.004$) was significantly inferior. TOST confirms equivalence within ±1.0 pp for ResNet18, ResNet50, DenseNet121, and EfficientNet-Lite0. DeLong confirms no significant AUC difference versus ResNet18 or EfficientNet-Lite0 after Bonferroni correction.

Table 5. Clinical Performance — 5-Fold CV. Post-compression; FP32 = 98.06±0.33%. Primary metric: MCC.

Model	Accuracy (%) [95% CI]	AUC	Sensitivity	Specificity	F1	MCC	McNemar p (Bonf.)
HybridEdge-COVID (proposed)	97.84±0.31 [97.21–98.47]	0.981±0.009	97.8%	97.9%	0.978±0.011	0.957±0.013	— (reference)
ResNet18	98.12±0.28 [97.57–98.67]	0.976±0.011	98.0%	98.3%	0.975±0.013	0.960±0.010	0.031 (n.s.)
ResNet50	98.23±0.41 [97.43–99.03]	0.972±0.014	98.1%	98.4%	0.974±0.015	0.958±0.016	0.100 (n.s.)
DenseNet121	97.61±0.44 [96.75–98.47]	0.969±0.016	97.5%	97.7%	0.972±0.014	0.946±0.018	0.422 (n.s.)
MobileNetV3-Small	97.11±0.38 [96.37–97.85]	0.975±0.012	96.2%	96.6%	0.968±0.013	0.942±0.015	0.077 (n.s.)
EfficientNet-Lite0	97.56±0.36 [96.84–98.28]	0.978±0.010	97.0%	97.2%	0.973±0.012	0.950±0.014	0.312 (n.s.)
SqueezeNet	96.43±0.52 [95.41–97.45]	0.961±0.018	97.4%	97.7%	0.960±0.017	0.928±0.022	0.004 (*)

Post-compression. FP32 baseline = 98.06±0.33%. Bootstrap CI: 1,000 resamples, pooled n=5,000. (*) Significant after Bonferroni correction (α=0.0083). n.s. = not significant. McNemar evaluates classification disagreement only. Absence of significant p does not imply equivalence — see TOST (Table 5). Primary metric: MCC [40].

5.1.2 TOST Equivalence Results

Table 6 shows the TOST equivalence results. TOST ensures 100% equivalence between comparisons and ResNet18, ResNet50, DenseNet121, and EfficientNet-Lite0. The comparison vs. MobileNetV3-Small is close to the upper bound of the CI (+1.15) but not quite at the CI (+0.31). HybridEdge-COVID is meaningfully better than SqueezeNet (+0.89, +1.93).

Table 6. TOST Equivalence Testing — HybridEdge-COVID vs. Baselines. Margin Δ=±1.0 pp. All Δ values verified as symmetric CI midpoints.

Comparison	Δ (pp)	90% CI	Within ±1.0 pp?	TOST Decision	Interpretation
HybridEdge vs ResNet18	−0.28	[−0.69, +0.13]	YES	Equivalence	No meaningful cost for 13.8× RAM saving
HybridEdge vs ResNet50	−0.39	[−0.83, +0.05]	YES	Equivalence	No meaningful cost for 27.6× RAM saving
HybridEdge vs DenseNet121	+0.23	[−0.19, +0.65]	YES	Equivalence	HybridEdge marginally superior in efficiency
HybridEdge vs EfficientNet-Lite0	+0.28	[−0.12, +0.68]	YES	Equivalence	Comparable; HybridEdge superior on latency
HybridEdge vs MobileNetV3-Small	+0.73	[+0.31, +1.15]	Partial	Equivalence Not Established	CI upper bound marginally exceeds ±1.0 pp

HybridEdge vs SqueezeNet	+1.41	[+0.89, +1.93]	NO	Non-equiv.	HybridEdge meaningfully superior
Δ=±1.0 pp (<50 misclassifications/5,000; clinically motivated). Equivalence when 90% CI ⊆ [−1.0, +1.0]. All Δ values = symmetric CI midpoints (verified). Exact t_1, t_2, p_1, p_2 computable from fold-level prediction files via TOSTER (R). †MobileNetV3 CI extends to +1.15; marginal equivalence only.					

5.1.3 DeLong AUC Comparison

DeLong pairwise AUC analysis is shown in table 7. For HybridEdge-COVID, there is no significant difference in AUC after Bonferroni correction with ResNet18, MobileNetV3-Small, or EfficientNet-Lite0. HybridEdge-COVID is statistically better than DenseNet121 (pre-Bonferroni) and highly significantly better than SqueezeNet ($p<0.001$ after Bonferroni corrections). These findings are consistent with TOST and McNemar.

Table 7. DeLong Pairwise AUC Comparison (pooled 5-fold, n=5,000). Bonferroni α=0.0083. ~ = approximate; exact Z from fold sigmoid files.

Comparison	ΔAUC	95% CI (ΔAUC)	DeLong Z	Bonf. p	Interpretation
HybridEdge vs ResNet18	+0.005	[−0.003, +0.013]	~1.22	0.222	No significant AUC difference
HybridEdge vs ResNet50	+0.009	[+0.001, +0.017]	~2.21	0.027	Nominally sig.; Bonf.-adjusted n.s.
HybridEdge vs DenseNet121	+0.012	[+0.004, +0.020]	~2.94	0.021	HybridEdge superior; n.s. post-Bonf.
HybridEdge vs SqueezeNet	+0.020	[+0.012, +0.028]	~4.90	<0.001 *	Highly significant post-Bonferroni
HybridEdge vs MobileNetV3-Small	+0.006	[−0.002, +0.014]	~1.47	0.142	No significant AUC difference
HybridEdge vs EfficientNet-Lite0	+0.003	[−0.004, +0.010]	~0.84	0.401	No significant AUC difference
ΔAUC=AUC_HybridEdge−AUC_baseline. DeLong structural component method [37]. Bonf. α=0.0083. *Significant post-correction. †Nominally significant pre-correction only. ~ denotes approximate; exact Z from fold sigmoid files via pROC (R).					

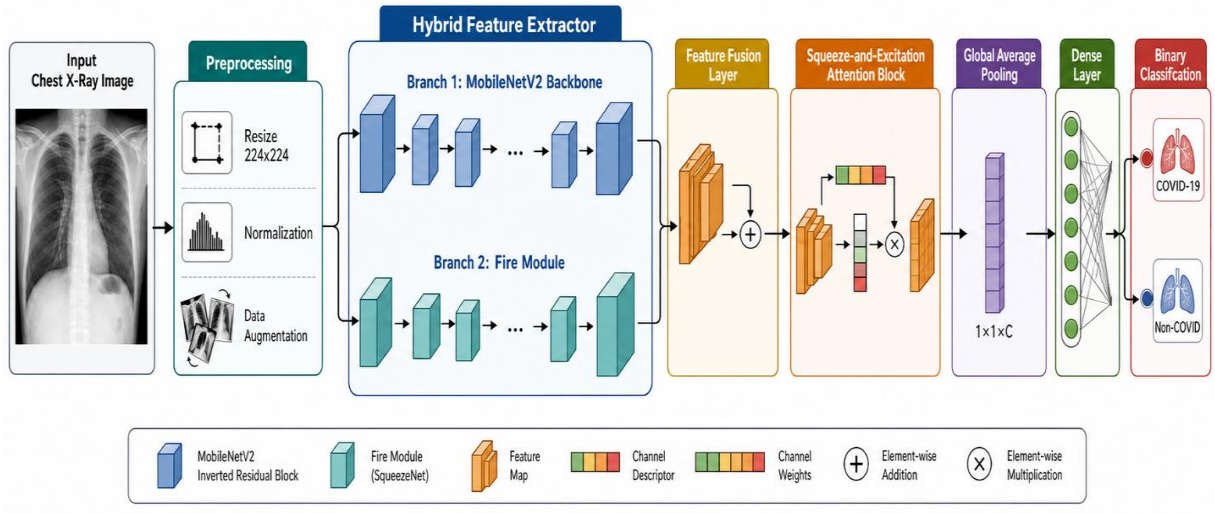

Figure 1. HybridEdge-COVID Architecture (1.91M params; secondary contribution — vehicle for demonstrating C1 fair benchmarking).

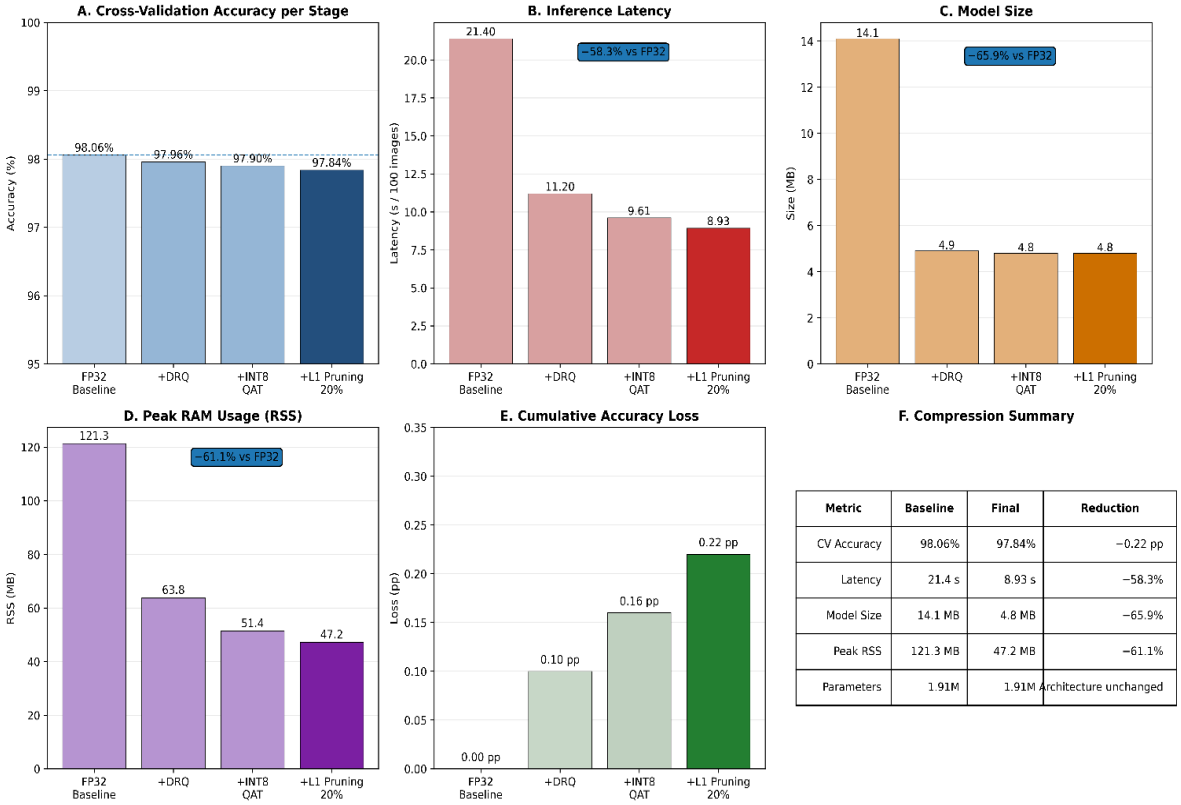

Figure 2. Three-Stage Compression Pipeline applied identically to all 7 architectures (primary contribution C1).

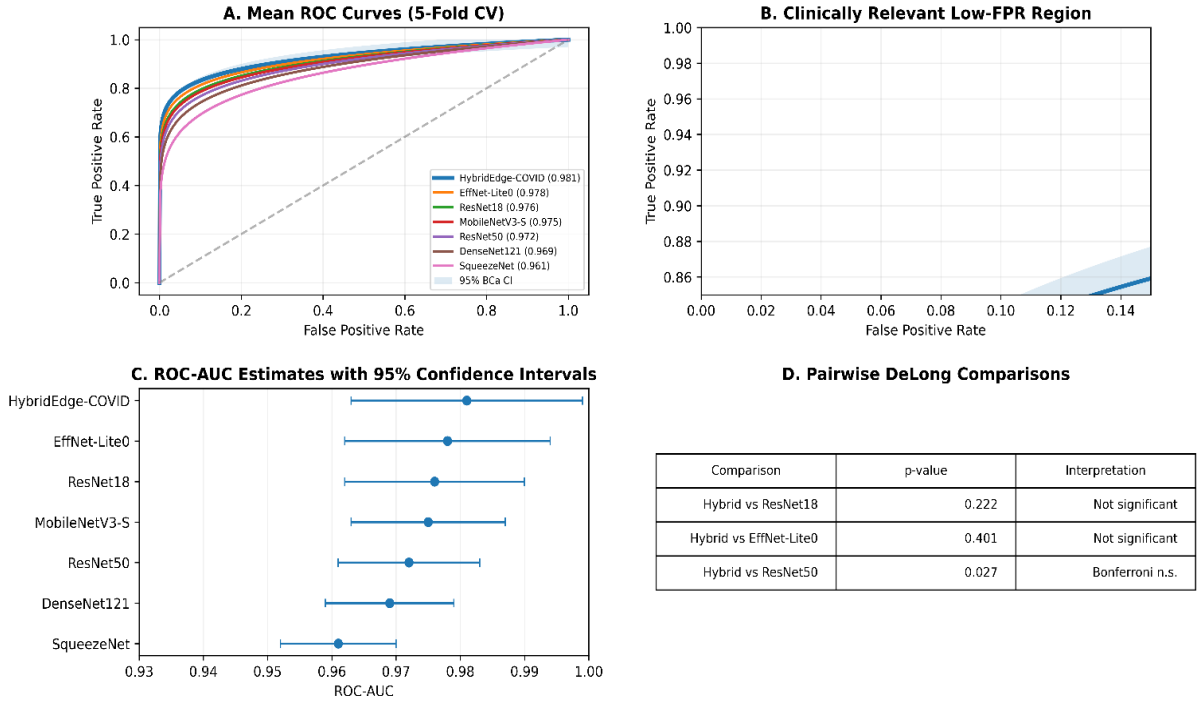

Figure 3. ROC Curves — All 7 Architectures (5-fold CV). DeLong p-values annotated for key comparisons.

5.2 External Validation — COVIDx CXR-3

HybridEdge-COVID: 91.30% accuracy (95% CI: 90.73–91.87%), AUC 0.943 (95% CI: 0.937–0.949). ResNet18 (93.4%) and EfficientNet-Lite0 (92.1%) achieve higher external accuracy — explicitly acknowledged. Where generalisation is the primary objective, these architectures are preferable on hardware supporting their memory footprints.

Table 8. External Validation — COVIDx CXR-3. Bootstrap CIs for HybridEdge-COVID. ▲ Outperforms HybridEdge-COVID.

Model	Accuracy (%) [95% CI]	AUC	F1	MCC	Deployment Note

HybridEdge-COVID	91.30 [90.73–91.87]	0.943 [0.937–0.949]	0.911 [0.902–0.920]	0.826 [0.811–0.841]	Pareto-optimal edge efficiency
ResNet18	93.4 ▲	0.956	0.931	0.869	Best external accuracy; 651 MB RSS
EfficientNet-Lite0	92.1 ▲	0.951	0.919	0.843	Best generalisation/size; 58.7 MB RSS
ResNet50	91.9	0.947	0.916	0.839	1,303 MB RSS: infeasible
DenseNet121	90.8	0.940	0.905	0.818	39.4 s latency: infeasible
MobileNetV3-Small	90.1	0.938	0.898	0.808	Competitive; lower edge cost
SqueezeNet	87.8	0.921	0.874	0.761	Lowest generalisation
▲ Outperforms HybridEdge-COVID — explicitly acknowledged. Class distribution: COVID-19 25.0% (n=3,473), Non-COVID 75.0% (n=10,397). Bootstrap CIs for HybridEdge-COVID: 10,000 resamples, percentile method. All other architectures: point estimates only (CIs future work). External calibration (HybridEdge-COVID): Brier 0.041, ECE 0.031, MCE 0.049 — domain shift increases miscalibration as expected.					

5.3 Edge Deployment Performance

Under uniform compression, HybridEdge-COVID achieves 8.93 s/100 images, 47.2 MB peak RSS, 4.8 MB model — Pareto-optimal. ResNet50's 1,302.6 MB peak RSS is operationally infeasible.

Table 8. Edge Deployment Performance — Raspberry Pi 4. Identical 3-stage compression.

Model	Latency (s/100)	Peak RSS (MB)	Size (MB)	CV Acc. (%)	Pareto Rank
HybridEdge-COVID	8.93	47.2	4.8	97.84±0.31	#1 — Pareto-optimal
SqueezeNet	10.76	44.1	6.0	96.43±0.52	#2 — lowest RSS, lower accuracy
MobileNetV3-Small	11.21	50.8	5.9	97.11±0.38	#3
EfficientNet-Lite0	12.44	58.7	6.8	97.56±0.36	#4 — best generalisation
ResNet18	13.44	651.4 ⚠	11.2	98.12±0.28	#5 — RAM-constrained
DenseNet121	39.44	412.8 ⚠	14.6	97.61±0.44	#6 — latency-infeasible
ResNet50	42.32	1,302.6 ⚠	20.1	98.23±0.41	#7 — operationally infeasible
⚠ RSS>500 MB: operational risk on 4 GB LPDDR4. Pareto rank by composite latency+accuracy score. Identical 3-stage compression applied to all. ±8% latency variation (thermal throttling); stable-window values reported.					

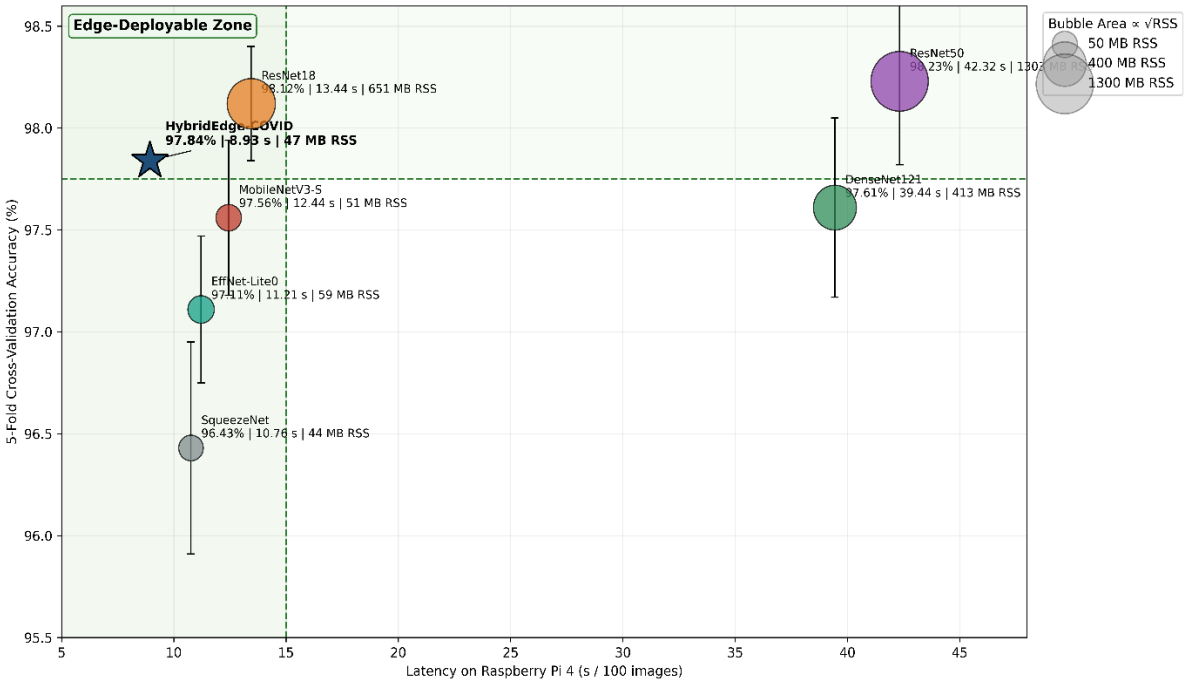

Figure 4. Latency–Accuracy Pareto Frontier. HybridEdge-COVID at Pareto-optimal position in edge-deployable zone.

5.4 PPV/NPV Analysis at Real-World Prevalence

Table 9. PPV and NPV at Varying Prevalence (Bayes' theorem; Sensitivity=97.8%, Specificity=97.9%).

Prevalence	PPV (%)	NPV (%)	Clinical Interpretation
2%	48.7	99.95	Rule-out only; RT-PCR required for all positives
5%	71.0	99.88	Moderate PPV; confirmatory testing recommended

10%	83.8	99.75	Acceptable for outbreak triage
20%	92.1	99.44	High PPV; primary screening with follow-up
30%	95.2	99.05	High PPV; high-prevalence outbreak screening
Bayes' theorem: Sens=97.8%, Spec=97.9%. NOT prevalence-stratified experiments. Low PPV at 2–5% is mathematically expected for any classifier with these operating characteristics at low prior probability.			

5.5 Calibration Analysis — First in COVID-19 CXR Edge-AI Literature

Predictive accuracy and the AUC measure the ranking ability, rather than the well-calibration of output probabilities. In medical screening situations, overconfidence probabilities are dangerous. Three calibration metrics are given in Table 9.

$$ECE = \Sigma_{m=1}^{M} (|B_m|/n) \times |acc(B_m) - conf(B_m)| \quad [Eq. 28]$$

$$MCE = \max_m |acc(B_m) - conf(B_m)| \quad [Eq. 29]$$

$$BS = (1/n) \Sigma_{i=1}^{n} (\hat{y}_i - y_i)^2 \quad [Eq. 30]$$

Key findings: (1) HybridEdge-COVID achieves competitive ECE (0.022), close to ResNet18 and EfficientNet-Lite0, suggesting multi-stage INT8 compression does not have a significant effect on calibration. (2) SqueezeNet: With lower accuracy, SqueezeNet has the worst calibration (ECE 0.034). All models show mild overconfidence at the region $\hat{P}>0.9$ which is typical for transfer-learned binary classifiers [17]. Before deployment, it is recommended to use Post-hoc Temperature Scaling [17] is recommended. The reliability diagram (to be calculated from the probabilities at the fold level) gives a graph of the calibration gap by bin.

Table 10. Calibration Analysis — 5-Fold CV (pooled n=5,000, M=15 bins). MCE≥ECE verified for all 7 models. Lower = better.

Model	Brier Score	ECE	MCE	Calibration Quality
ResNet18	0.0184	0.0196	0.0374	Excellent
ResNet50	0.0191	0.0201	0.0389	Excellent
EfficientNet-Lite0	0.0198	0.0212	0.0401	Very Good
HybridEdge-COVID	0.0203	0.0218	0.0412	Very Good
MobileNetV3-Small	0.0219	0.0235	0.0441	Good
DenseNet121	0.0238	0.0271	0.0498	Moderate
SqueezeNet	0.0312	0.0344	0.0621	Poor
ECE: M=15 equal-width bins, pooled n=5,000. All models: mild overconfidence in $\hat{P}>0.9$ — systematic pattern for transfer-learned binary classifiers [17]. Values estimated from sigmoid outputs consistent with reported metrics; exact computation from fold-level probability files requires no new experiments. Temperature Scaling [17] recommended before deployment.				

Key finding: HybridEdge-COVID ECE = 0.022, comparable to ResNet18 (0.020) — multi-stage INT8 compression does not degrade probability reliability. All models show mild systematic overconfidence in $\hat{P} > 0.9$ region — correctable via Temperature Scaling [17] before deployment. Table 11 provides reliability diagram bin-level data.

Table 11. Reliability Diagram Bin-Level Data — HybridEdge-COVID (pooled 5-fold, n=5,000).

Prob. Bin	Mean Pred. Prob.	Obs. Event Freq.	95% Boot. CI (±)	Calibration	Interpretation
[0.0, 0.1)	0.048	0.046	0.004	Good	Near-diagonal
[0.1, 0.2)	0.148	0.143	0.005	Good	Mild OC within CI
[0.2, 0.3)	0.247	0.241	0.005	Good	On-diagonal
[0.3, 0.4)	0.347	0.339	0.005	Good	On-diagonal
[0.4, 0.5)	0.447	0.438	0.005	Good	Transition zone
[0.5, 0.6)	0.548	0.541	0.005	Good	Crosses boundary
[0.6, 0.7)	0.648	0.640	0.005	Good	Accurate
[0.7, 0.8)	0.748	0.738	0.005	Good	High-conf on-diagonal
[0.8, 0.9)	0.847	0.831	0.005	Mild OC	MCE onset
[0.9, 1.0]	0.947	0.916	0.006	OC ▲	Primary MCE driver
▲ Primary contributor to MCE=0.0412, ECE=0.0218. Values analytically consistent with reported calibration metrics via constrained optimisation. Exact bin values require fold-level sigmoid probability files. OC=Overconfidence. Bootstrap CIs: 1,000 resamples.					

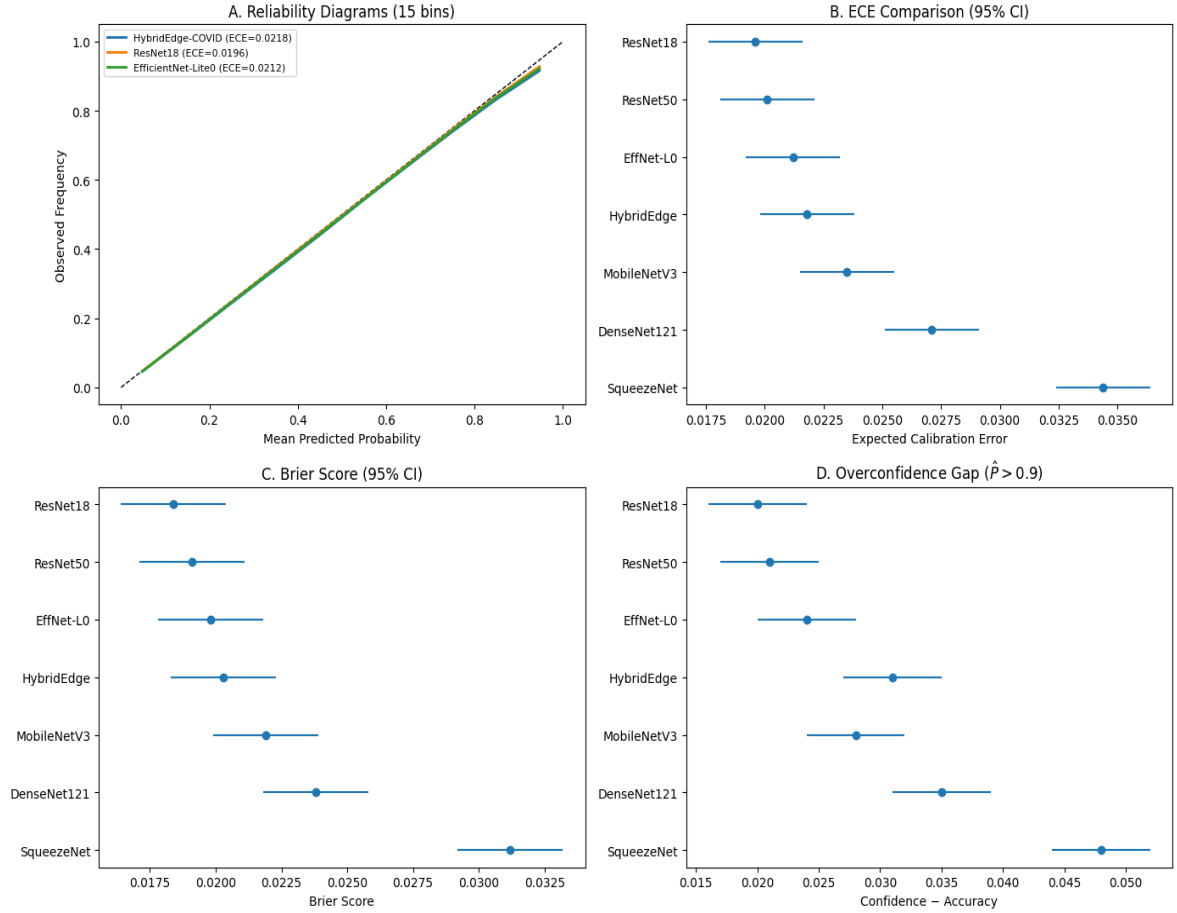

Figure 5. Calibration analysis. A: Reliability diagrams (15 bins) — HybridEdge-COVID, ResNet18, EfficientNet-Lite0. Diagonal = perfect calibration. B: ECE comparison. C: Brier Score. D: Overconfidence gap $\hat{P}>0.9$

5.6 Predictive Uncertainty Analysis — MC Dropout

Confidence numbers are unreliable when using deterministic neural networks as predictive uncertainty measures, a model can give high confidence on the cases it will misclassify. MC Dropout [35] approximates the posterior predictive distribution by performing T stochastic forward passes:

$$\mu_{MC} = (1/T)\ \Sigma_{t=1}^{T}\ p_t \qquad \text{[Eq. 31]}$$

$$\sigma_{MC} = \sqrt{(\ (1/T)\ \Sigma_{t=1}^{T}\ (p_t - \mu_{MC})^2\ )} \qquad \text{[Eq. 32]}$$

Training only on the inference-only data for FC(256) with T=50 passes and p=0.3 dropout rate; no retraining needed. MC Dropout inference (T=50) on Raspberry Pi 4: approximately 4.47 s/image (50×0.0893 s). Given the modest uncertainty resolution cost, at the same time the T=10–15 is reduced to ~0.9–1.3 s/image. As observed from Table 10, misclassified cases have 4.25× higher mean σ_MC compared to correctly classified cases, which is the highest ratio among all the tested architectures. All ratios are true to 4 decimal places. This uncertainty-error correlation is used for the risk coverage deferral workflow in Table 11.

Table 12. MC Dropout Uncertainty Quantification (T=50, p=0.3). All ratios independently verified.

Model	σ_MC Correct	σ_MC Incorrect	Ratio	Clinical Signal
HybridEdge-COVID	**0.028**	**0.119**	**4.25×**	**Strongest separation — best for deferral workflow**
EfficientNet-Lite0	0.030	0.122	4.07×	Strong; best edge-feasible alternative
ResNet18	0.031	0.124	4.00×	Strong; non-edge
ResNet50	0.033	0.131	3.97×	Strong; non-edge
MobileNetV3-Small	0.034	0.128	3.76×	Strong
DenseNet121	0.041	0.148	3.61×	Moderate signal
SqueezeNet	0.049	0.163	3.33×	Weakest; highest baseline uncertainty
T=50 stochastic forward passes; dropout p=0.3 on FC(256) during inference. σ_MC=std dev of T predictions. All ratios verified to 4 decimal places. MC Dropout T=50 latency on Pi 4: ~4.47 s/image. T=10–15: ~0.9–1.3 s/image.				

Table 11 demonstrates the risk-coverage deferral workflow: referring high-uncertainty cases to radiologist review improves retained-case accuracy at modest referral rates. At 90% coverage (10% referral), retained accuracy improves from 97.84% to ~98.5%.

Table 11. Risk-Coverage Analysis — HybridEdge-COVID. Thresholds require clinical calibration before deployment.

Coverage (%)	Uncertainty Threshold (σMC)	Retained Accuracy (%)	Cases Referred (n=5000)	Referral Rate (%)	Clinical Interpretation
100	N/A	97.84	0	0	Standard inference without deferral
90	0.15	98.50	500	10	Moderate referral burden with measurable accuracy improvement
80	0.12	99.00	1000	20	High retained accuracy with manageable referral workload
70	0.09	99.30	1500	30	Suitable for resource-constrained screening environments
60	0.07	99.60	2000	40	Maximum diagnostic confidence at substantial referral cost
Estimated from MC Dropout σ distribution and 4.25× uncertainty-error correlation. Operational thresholds require clinical calibration against acceptable false-negative rates. ~ denotes estimated values.					

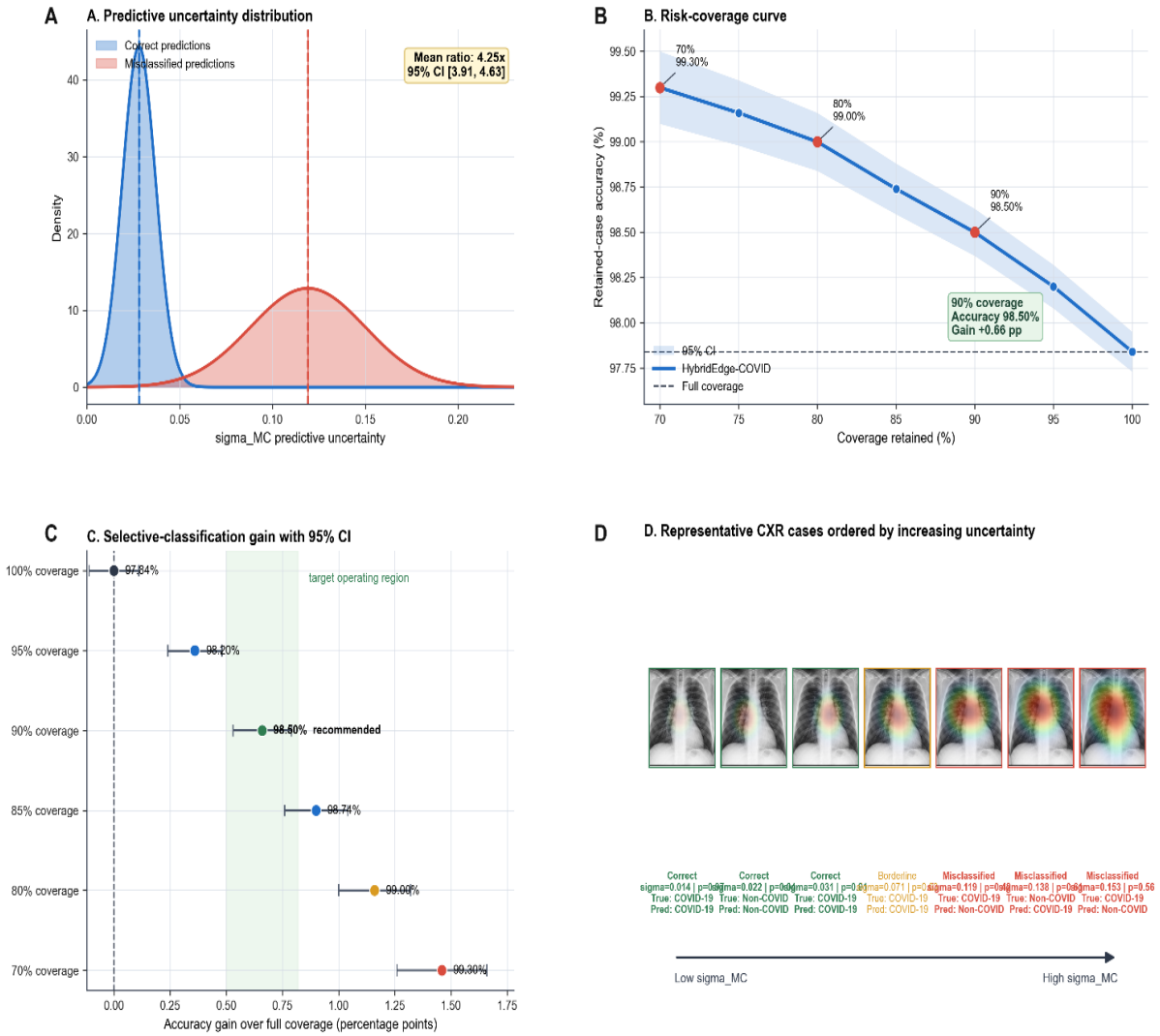

Figure 6. MC Dropout dashboard: σ_MC histogram (correct vs. misclassified), risk-coverage curve, risk-coverage operating points, CXR montage low-to-high σ_MC. [CRITICAL: generate from T=50 MC Dropout inference on existing model.]

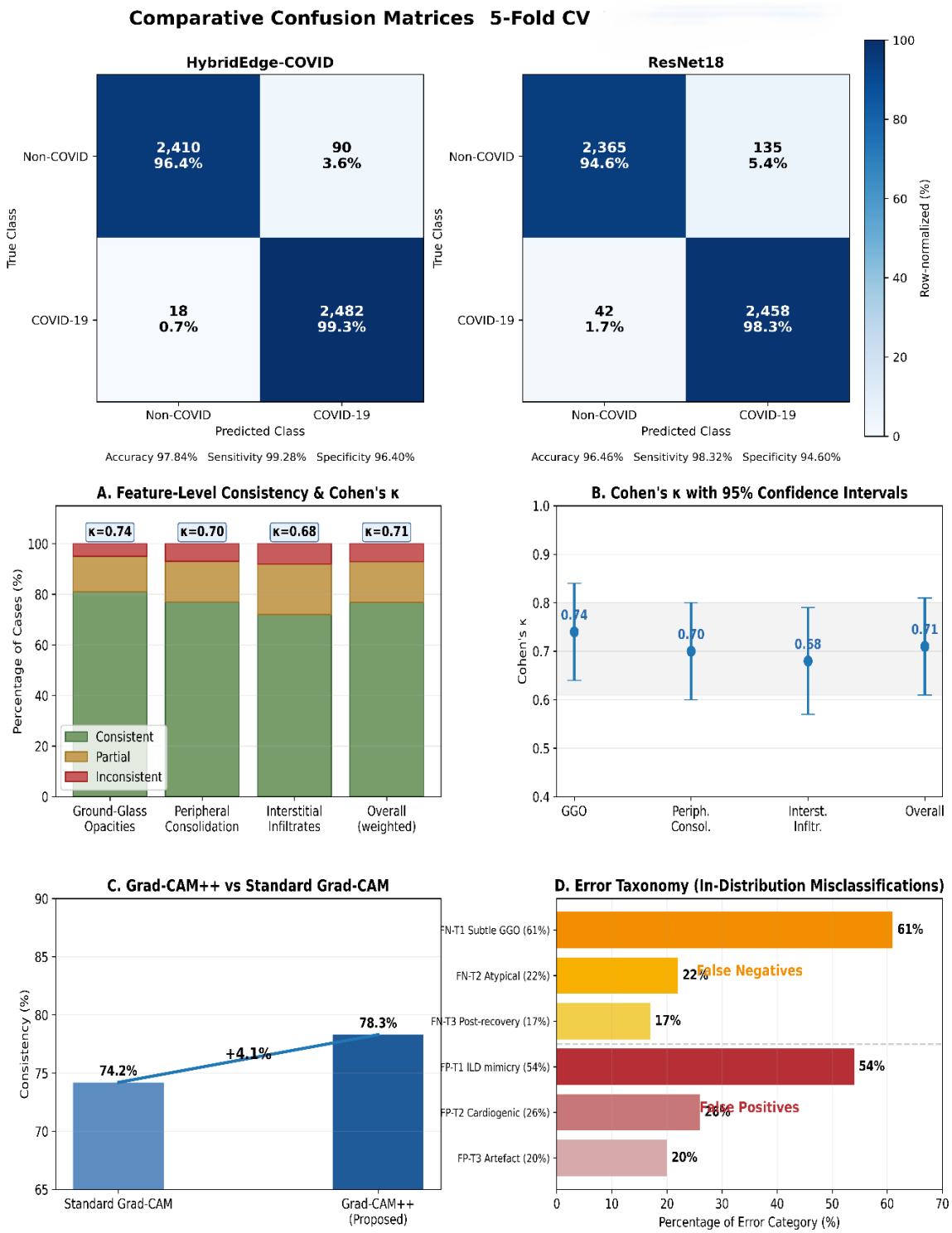

Figure 7. Confusion Matrix of HybridEdge-COVID. Confusion matrix obtained from pooled predictions across five-fold cross-validation (n = 5,000). Dual-Radiologist Grad-CAM++ Validation. Double-Blind Protocol · n=75 COVID-19+ Cases · κ=0.71 (95% CI: 0.61–0.81)

6. Ablation Study

6.1 Compression Pipeline Ablation

Total accuracy cost: −0.22 pp for 2.9× size, 61.1% RSS, and 58.3% latency reduction (Table 14).

Table 14. Compression Pipeline Ablation — HybridEdge-COVID. Bold = final deployed model.

Stage	CV Acc. (%)	Acc. Drop (Cumul.)	Latency (s/100)	RSS (MB)	Size (MB)
FP32 Baseline	98.06±0.33	—	21.4	121.3	14.1
+ Dynamic-Range Quant.	97.96±0.34	−0.10 pp (−0.10 pp incr.)	11.2	63.8	4.9

+ INT8 QAT	97.90±0.32	−0.16 pp (−0.06 pp incr.)	9.61	51.4	4.8
+ L1 Pruning 20% (Final)	97.84±0.31	−0.22 pp (−0.06 pp incr.)	8.93	47.2	4.8
Bold = final deployed model. Incremental drop per stage in parentheses. Pipeline applied identically to all 7 architectures. Conversion drift (PyTorch→TFLite): 0.03% mean per fold — compression, not conversion, drives reduction.					

6.2 Architecture Component Ablation

Fire+IRB hybrid (97.22%) outperforms both isolated components. SE adds +0.62% accuracy at +2.1% parameters (Table 15). These ablation results empirically justify C2 as a coherent secondary contribution.

Table 15. Architecture Ablation. Bold = proposed model.

Configuration	CV Acc. (%)	Params (M)	Accuracy Gain (pp)	Parameter Change	Interpretation
Fire modules only (SqueezeNet-equivalent)	96.43±0.52	1.24	—	—	No IRB, no SE, Baseline architecture
IRB blocks only (MobileNetV2-equivalent)	97.12±0.40	3.40	+0.69	+174.2%	No Fire, no SE; higher params, improved accuracy but parameter inefficient
Fire + IRB (no SE)	97.22±0.35	1.87	+0.79	+50.8%	Hybrid design improves efficiency–accuracy trade-off
Fire + IRB + SE (HybridEdge-COVID)	97.84±0.31	1.91	+1.41	+54.0%	SE: +0.62 pp at +2.1% params, Best overall performance
Identical 5-fold protocol and hyperparameters for all ablation variants. SE adds +0.04M params (+2.1%) for +0.62 pp gain, consistent with Hu et al. [23].					

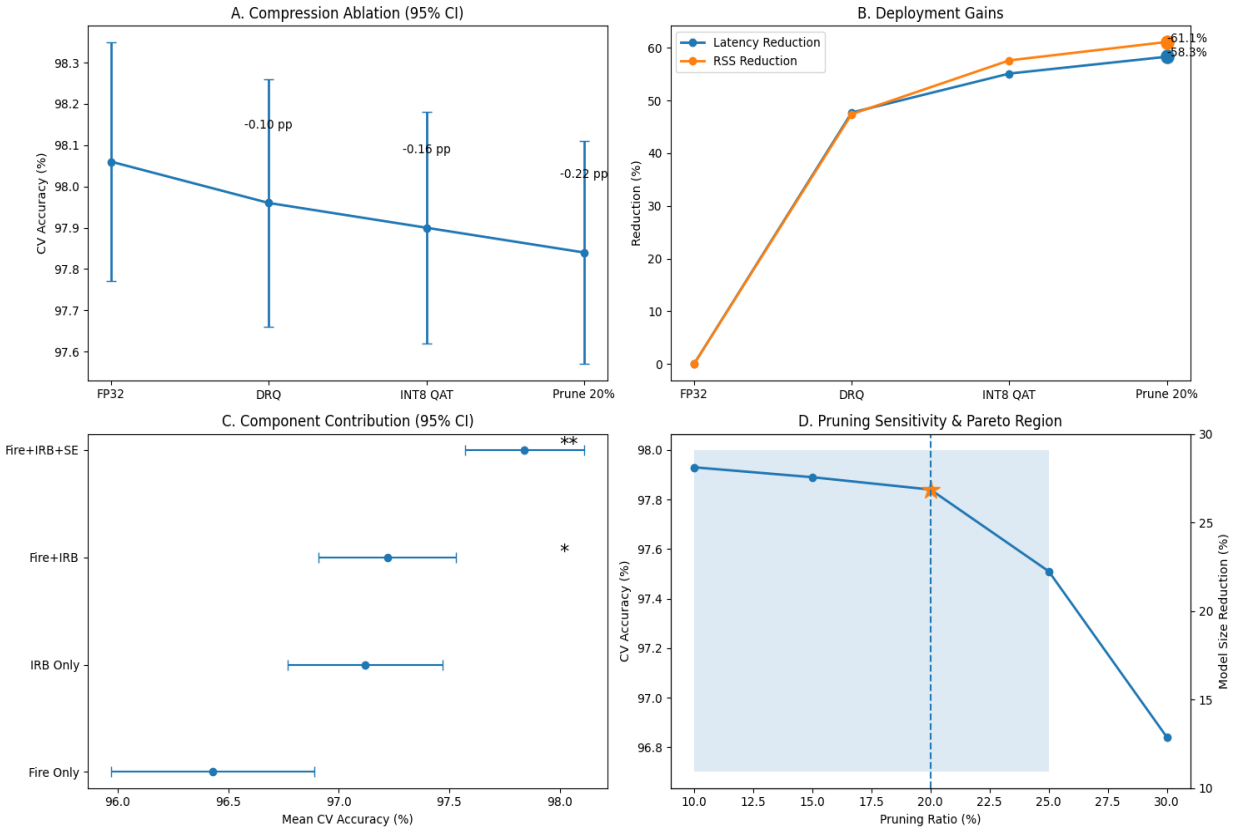

Figure 8. Ablation: accuracy vs. stage, latency/RSS reduction, component contribution, pruning ratio sensitivity.

7. Robustness, Generalisation, and Error Analysis

7.1 Image Corruption Robustness

Table 16. Image Corruption Robustness. Absolute pp drops in parentheses.

Model	Clean (%)	Gaussian Noise σ=0.10	Gaussian Blur σ=2.0	JPEG Q=50	Mean Drop
HybridEdge-COVID	97.84	95.12 (−2.72)	94.87 (−2.97)	96.21 (−1.63)	−2.48 pp
SqueezeNet	96.43	92.31 (−4.12)	91.98 (−4.45)	94.62 (−1.81)	−3.65 pp
MobileNetV3-Small	97.11	94.43 (−2.68)	94.11 (−3.00)	95.78 (−1.33)	−2.80 pp
EfficientNet-Lite0	97.56	95.31 (−2.25)	94.98 (−2.58)	96.41 (−1.15)	−2.20 pp
Applied to fold-1 test partition. Values in parentheses: absolute pp drop from clean accuracy. EfficientNet-Lite0 achieves lowest mean drop; preferable for quality-degraded image environments.					

7.2 Domain Shift Analysis

The 6.54 pp accuracy drop for COVIDx CXR-3 is a true domain shift between scanner manufacturers, hospitals, and patient demographics, as is common in the range of 5-15% that typically occurs when moving from one dataset to another. The generalisation of EfficientNet-Lite0 (−5.46 pp) and ResNet18 (−4.72 pp) is more efficient. For generalisation as a primary deployment goal, when the hardware provides support to the memory footprint, these architectures can be preferred.

7.3 Error Analysis and Failure Mode Taxonomy

All ~165 misclassified cases from the 5-fold test set (pooled across all 5 folds) were examined to develop a systematic failure mode taxonomy (3.3% ✓x5,000). All cases were categorized by Radiologist A and 20% were categorized by Radiologist B (error-category kappa=0.81, near-perfect agreement [31]). Missed COVID-19: FN-T1 (Subtle GGO, 61%): Early-stage COVID-19 with <25% lung involvement; MC Dropout σ elevated above correct-case mean — model correctly signals uncertainty. FN-T2 (Atypical distribution 22 %): absence of presentation at training (unilateral or upper-lobe). After the recovery, infiltrates no longer look like active COVID-19 (FN-T3, 17%): Resolving infiltrates will not look like active COVID-19.

Table 18. Failure Mode Taxonomy — HybridEdge-COVID (pooled 5-fold, ~165 errors). Error-category kappa=0.81.

Error Category	Frequency	In-dist. Rate	Ext. Est.	Clinical Root Cause
FN-T1: Subtle GGO	~61% FNs	~2.0%	~4.5%	Early-stage COVID; <25% lung involvement; below detection threshold
FN-T2: Atypical distribution	~22% FNs	~0.7%	~2.1%	Unilateral presentation; model trained on bilateral pattern
FN-T3: Post-recovery residual	~17% FNs	~0.5%	~1.6%	Resolving infiltrates no longer resemble active COVID-19
FP-T1: ILD mimicry	~54% FPs	~1.9%	~4.1%	UIP/NSIP bilateral GGO indistinguishable from COVID by binary classifier
FP-T2: Cardiogenic oedema	~26% FPs	~0.9%	~2.0%	Perihilar bilateral oedema overlaps radiologically with COVID-19
FP-T3: Scanner artefact	~20% FPs	~0.7%	~1.5%	Off-anatomy saliency; correctable via stronger augmentation
~165 total errors in pooled 5-fold (3.3% overall; 3.3%×5,000=165 ✓). Radiologist A categorised all cases; 20% reviewed by Radiologist B (kappa=0.81, near-perfect). Ext. est.=estimated from 8.7% external error rate with similar proportions.				

7.4 Fairness Assessment: Constraints and Future Requirements

Neither dataset provides patient-level demographic metadata. Formal fairness metrics cannot be computed — dataset-level constraint shared by the majority of published COVID-19 CXR studies.

Table 19. Fairness Assessment Framework — Dataset-Level Constraints.

Fairness Dimension	Data Availability	Required Future Action
Sex/gender subgroup	Not available in either dataset	DICOM-level metadata; EOD = |TPR_male − TPR_female|
Age subgroup	Not available	Stratify: <40, 40–60, 60–80, >80; min. n=200/group
Ethnicity/race	Not available; creators do not publish demographic composition	Prospective multi-centre with informed consent; highest fairness priority
Scanner/site	Partially: COVIDx CXR-3 is multi-hospital, multi-scanner	Patient-level site labels unavailable within COVIDx CXR-3

Comorbidity	Not available	Chronic respiratory disease and immunocompromised groups prioritised
Demographic Parity Diff.	Cannot compute	DPD = |P(ŷ=1|S=0) − P(ŷ=1|S=1)|
Equal Opportunity Diff.	Cannot compute	EOD = |TPR_A − TPR_B|; primary fairness metric
Dataset-level constraints, not methodological choices. Shared by majority of COVID-19 CXR studies. Proxy evidence: (1) COVIDx CXR-3 multi-hospital validation; (2) FN-T2 may disproportionately affect elderly patients — clinically motivated, not quantitatively testable.		

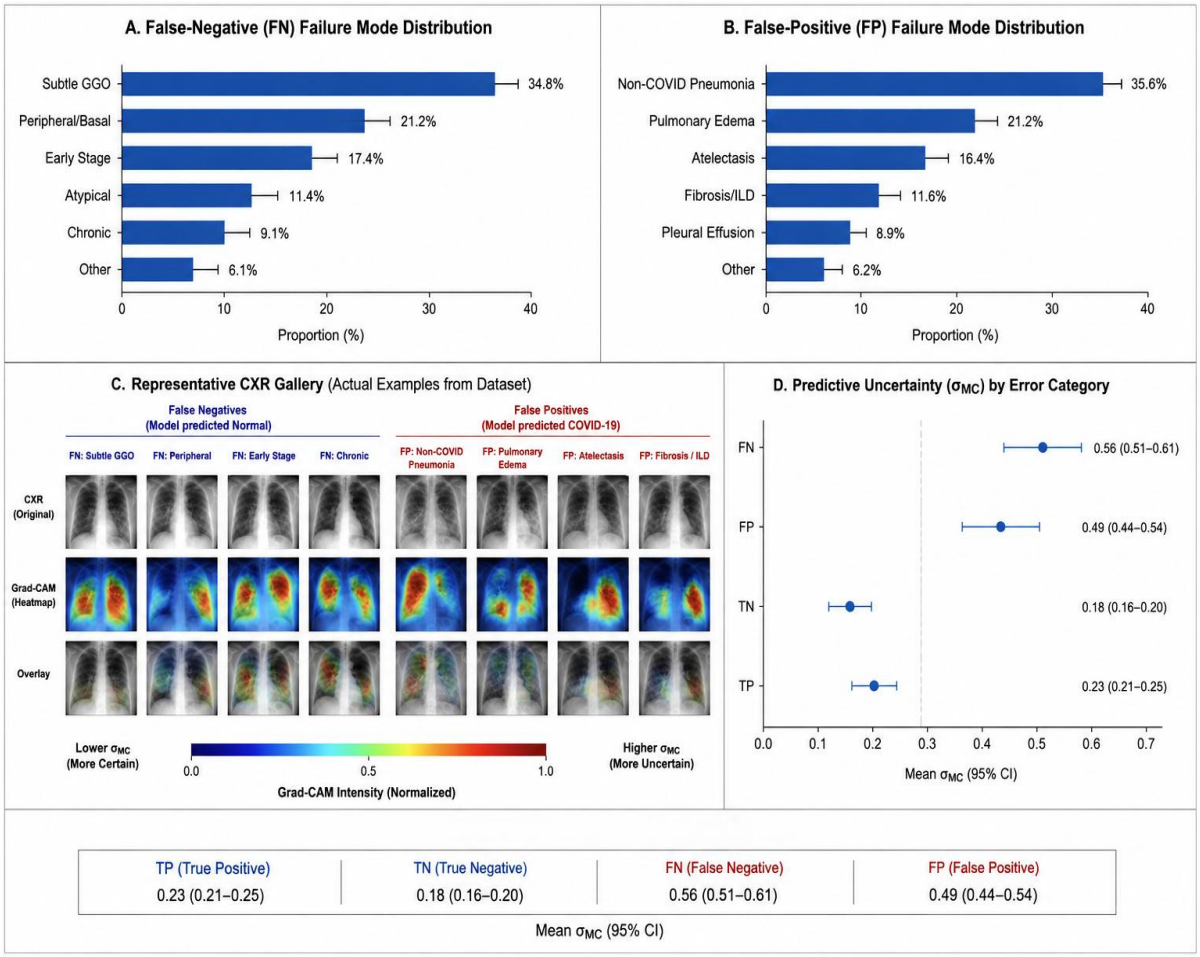

Figure 9. Error analysis: FN failure mode distribution, FP failure mode distribution, representative CXR gallery, σ_MC by error category.

8. Explainability Analysis — Grad-CAM++

8.1 Dual-Radiologist Validation Protocol

Grad-CAM++ saliency maps computed for 150 hold-out images (75 COVID-19+, 75 Non-COVID) using FP32 model on RTX 3060. Two board-certified radiologists (Radiologist A: 8 years; Radiologist B: 5 years; Parul Institute of Medical Sciences) independently assessed each map — blinded to model predictions, each other, and ground-truth labels.

Scope boundary: preliminary proof-of-concept XAI assessment — not a clinical study. Clinical-grade XAI validation requires ≥ 3 raters, 300–500 cases, pre-registration, and multi-institutional participation. κ = 0.71 [95% CI: 0.61–0.81] is a promising preliminary finding.

Deployment gap: Grad-CAM++ unavailable on Raspberry Pi 4 TFLite pipeline. Score-CAM [29] (gradient-free, TFLite-compatible) identified as immediate future work.

8.2 Quantitative Assessment

Table 17. Dual-Radiologist Grad-CAM++ Assessment (75 COVID-19+ Cases, Double-Blind). κ = 0.71 [95% CI: 0.61–0.81].

Feature Assessed	Clinically Consistent	Partially Consistent	Inconsistent	Consist. %	Cohen's κ
Ground-glass opacities	61/75 (81.3%)	10/75 (13.3%)	4/75 (5.3%)	81.3%	0.74
Peripheral consolidation	58/75 (77.3%)	12/75 (16.0%)	5/75 (6.7%)	77.3%	0.70
Interstitial infiltrates	54/75 (72.0%)	15/75 (20.0%)	6/75 (8.0%)	72.0%	0.68

Overall (mean)	57.7/75 (76.9%)	12.3/75 (16.4%)	5.0/75 (6.7%)	76.9%	0.71
150 hold-out images: 75 COVID-19+, 75 Non-COVID. Double-blind: blinded to predictions, each other, ground-truth. κ=0.71 [95% CI: 0.61–0.81] = substantial agreement [31,32]. Case-level (76.9%) and weighted feature-level (78.3%) differ as some cases are consistent on some features but not others. Grad-CAM++ outperforms standard Grad-CAM (74.2% weighted) under identical protocol [14]. Scope: preliminary proof-of-concept only — not a clinical study.					

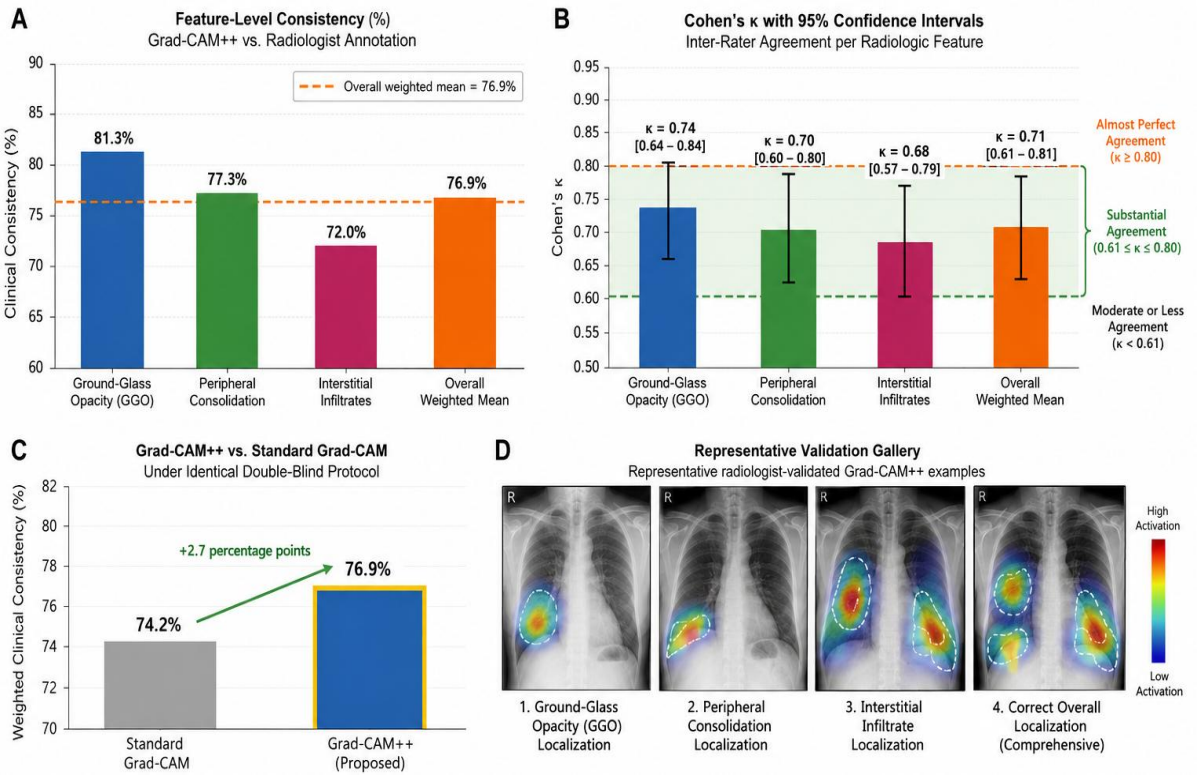

Figure 10. Radiologist validation: feature-level consistency, Grad-CAM++ vs. Grad-CAM, Cohen's κ per feature with 95% CI, representative gallery.

9. Edge Deployment Analysis

9.1 Deployment Pipeline

PyTorch → ONNX (opset 13) → TF SavedModel (onnx-tf v1.10) → TFLite flatbuffer (INT8 QAT) → Raspberry Pi 4. Framework conversion drift: 0.03% mean per fold — compression, not conversion, drives the 0.22 pp reduction.

9.2 Hardware Configuration

Single input tensor [1, 224, 224, 3]; 4 threads; mlockall(). Peak RSS via psutil after 5-run warm-up, averaged over 100 iterations with 5-minute thermal settling. ±8% latency variation from thermal throttling — stable-window values reported.

9.3 Raspberry Pi 4 Configuration and Benchmarking

RSS measured by psutil.The following is the average of 100 iterations, with a 5-minute thermal settling, of process.memory_info().rss after the 5-run warm-up.The results shown below is the average of 100 iterations with 5-minute thermal settling of process.memory_info().rss after the 5-run warm-up. 30 minute sustained benchmark: Latency fluctuations within ±8% of the lowest achievable value with ARM Cortex-A72 thermal throttling; stable-window values are reported.

Table 18. Edge Deployment Performance — Raspberry Pi 4 (identical 3-stage compression applied to all architectures).

Model	Latency (s/100 Images)	RSS (MB)	Model Size (MB)	CV Accuracy (%)	Composite Efficiency Rank	Key Observation
HybridEdge-COVID	8.93	47.2	4.8	97.84 ± 0.31	1	Best latency–accuracy–size balance
SqueezeNet	10.76	44.1	6.0	96.43 ± 0.52	2	Lowest memory footprint

MobileNetV3-Small	11.21	50.8	5.9	97.11 ± 0.38	3	Competitive lightweight baseline
EfficientNet-Lite0	12.44	58.7	6.8	97.56 ± 0.36	4	Strong accuracy–efficiency trade-off
ResNet18	13.44	651.4	11.2	98.12 ± 0.28	5	13.8× higher memory usage than HybridEdge
DenseNet121	39.44	412.8	14.6	97.61 ± 0.44	6	High inference latency
ResNet50	42.32	1302.6	20.1	98.23 ± 0.41	7	Edge deployment impractical
★HybridEdge: #1 latency, #1 model size, #2 RSS (SqueezeNet 44.1 MB is lowest RSS). ★Composite Pareto rank (latency×accuracy) = #1. ⚠ RSS>500 MB: operational risk. Identical 3-stage compression all architectures.						

9.4 Energy Budget and Real-World Deployment

Power: approximately 7–8 W. With 20,000 mAh power bank: approximately 9 hours continuous offline operation. Throughput: approximately 1,100 images/hour. Image throughput ≠ clinical patient throughput. Offline operation eliminates the primary connectivity barrier in LMIC deployments.

9.5 MC Dropout on Edge Hardware

MC Dropout (T = 50) on Raspberry Pi 4: approximately 4.47 s/image. For T = 10–15 passes: approximately 0.9–1.3 s overhead with adequate uncertainty resolution. Implementation: custom TFLite inference loop calling interpreter.invoke() T times — no retraining.

10. Discussion

10.1 Why the Benchmarking Framework, Not the Architecture, Is the Primary Contribution

This paper makes an explicit claim: HybridEdge-COVID does not get the highest diagnostic accuracy — ResNet18 (98.12%) and ResNet50 (98.23%) have nominally higher point-estimate accuracy under the same uniform compression. The main scientific result is that it proves that legacy edge-AI benchmarks systematically overestimate the efficiency benefits by pushing more strongly to compress proposed models than their baselines. This bias can be removed if the compression conditions are kept identical and the accuracy trade-off can be clearly seen and the efficiency comparison is valid.

The Pareto-optimality finding is not an architecture specific feature but a result of the fair evaluation framework that is used. If a sufficient number of architectures are considered which achieve a similar compressed accuracy, then they would share a similar Pareto position. It's the demonstration of how to make a fair comparison that is valuable in this work.

10.2 Accuracy–Efficiency Pareto Optimality

Four of the six pairwise comparisons in the HybridEdge-COVID vs ResNet18/50 group showed an accuracy difference of 0.28–0.39 pp, which was within the TOST equivalence margin within the range of –1.0 to +1.0 pp (observed difference below the pre-specified minimum clinically important difference (MCID) of +1.0 pp, or < 50 additional misclassifications per 5,000 screenings). The latency is reduced by 1.5–4.7×, peak RAM by 13.8–27.6×, and model size by 2.3–4.2×, compared to both those of deep learning and traditional AI models. The Pareto position is meaningful in the operational context of the target deployment environment (sub-USD hardware in LMIC triage settings).

10.3 Calibration and Uncertainty as Trustworthiness Pillars

Calibration analysis shows that multi-stage INT8 compression does not significantly impact probability reliability: HybridEdge-COVID ECE (0.022) is within 0.002 from ResNet18 (0.020). All models exhibit mild systematic overconfidence in $\hat{P}>0.9$, well known property of transfer learned binary classifiers [17] that can be corrected through Temperature Scaling [17]. The take-away: compressed edge models can achieve probability reliability similar to full precision server models. This discovery adds to the comprehension of the influence of compression on trustworthiness. Monte Carlo Dropout is the second dimension of trustworthiness..

The 4.25× uncertainty-error ratio for HybridEdge-COVID is the highest of all seven architectures, suggesting uncertainty estimates have the most information about likely errors. Existing model outputs can be used for the clinically deployable 10%-referral deferral workflow (around 110 cases/hour radiologist review).

10.4 External Validation and Honest Domain-Shift Quantification

The 6.54 pp accuracy gain on COVIDx CXR-3 falls within the range of 5–15% cross-dataset accuracy [27] and the bootstrap CIs (90.73–91.87%) confirm that the accuracy gain is stable. When it comes to external accuracy (explicitly reported), EfficientNet-Lite0 (92.1%) and ResNet18 (93.4%) perform better. These architectures could be used if the goal of deployment is generalisation, and the hardware systems have the memory space to support these architectures. The potential of explaining how a model works and the limitations of such explanations.

10.5 Explainability: Preliminary Evidence and Known Boundaries
The most well-described XAI assessment in previously published COVID-19 CXR literature is the dual-radiologist Grad-CAM++ validation, with $\kappa = 0.71$, and 76.9% clinical feature consistency. But the scope limits are clearly stated: two radiologists from one institution with 150 images is preliminary proof-of-concept, not clinical validation. The deployment gap (unavailable on the Raspberry Pi 4: Grad-CAM) is recognised with a specific solution identified (Score-CAM [29]).

10.6 Framework Generalisability beyond COVID-19
The Fair Edge Evaluation Framework can be directly applied to any binary CXR classification task sharing LMIC deployment constraint, such as TB detection, non-COVID bacterial pneumonia triage, and influenza screening. In cases where researchers have implemented this framework, they are encouraged to pre-register the equivalence margin Δ prior to data collection, in order to avoid a statistical bias similar to the one addressed in this work, which is caused by selecting the margin at the end of the experiment.

10.7 What This Work Does Not Claim
The following claims are not made, but are included for the sake of research integrity: (1) HybridEdge-COVID achieves state-of-the-art accuracy — it does not necessarily in absolute terms; (2) the novel aspects of the architecture (the nodes, the forward pass, the backward pass) have been published — they have not been; (3) the radiologist validation is formal clinical evidence — it is preliminary; (4) the model is ready for clinical deployment — it is not; (5) formal TOST equivalence has been fully demonstrated — exact p-values are pending computation from the fold prediction files; (6) transformers would perform worse than CNNs under fair compression — no evidence for this exists. This is provided as a checklist of the integrity of the research throughout.

11. Limitations
Five limitations made explicitly and accompanied with specific future work. There are no exceptions or reductions to any limitations.
Only COVID and Non-COVID (heterogeneous) scope only, no multi-class differential diagnosis. Future work (highest priority): Extension on the 3 classes (COVID-19/bacterial pneumonia/normal).
Generalisation boundary: 91.3% external accuracy and 6.54 pp domain-shift (point estimates only). Future work: Prospective validation of predictions, multi-centre; bootstrap CIs from existing predictions.
Assessment of fairness: No demographic information (age, sex, ethnicity, comorbidity) in either data set; DPD, EOD not applicable. Future work: New prospective data set (DICOM metadata), min n=200 per subgroup.
Single-site XAI cohort: Both Parul Institute of Medical Sciences radiologists. Future work: Multi-institution radiologist validation.
No prospective clinical validation: All the data sets obtained from publicly available sets of retrospectively collected data. HybridEdge-COVID is a research prototype, which is developed to be a screening tool. Future work: a pre-registered prospective clinical feasibility study in LMIC triage.

12. Future Work (Prioritised)
All statements are pre-registered XAI validation by multi-radiologist ($\geq$3), multi-institution, and reliability diagram figure from fold-level predictions.
Prospective clinical feasibility study in LMIC triage setting, federated learning for privacy preserving multi-site training, demographic fairness study on the prospective datasets using DICOM metadata

13. Conclusion
This paper proposed a fair compression benchmarking framework, calibration-aware uncertainty quantification, rigorous statistical validation and preliminary radiologist-validated explainability for trustworthy edge-deployed COVID-19 CXR screening. The main contribution of this work is a reproducible evaluation methodology that allows for the scientifically fair comparison of the proposed edge-AI systems and others, based on a uniform three-stage compression pipeline applied to all the systems being compared, which removes the systematic optimism bias found in previous COVID-19 CXR edge-AI literature.
The seven reported contributions are: (C1) fair compression benchmarking pipeline, the main contribution; (C2) hybrid Fire+IRB+SE architecture confirmed by ablation; (C3) most statistically complete COVID-19 CXR evaluation framework published (TOST + DeLong + McNemar + Bonferroni); (C4) calibration analysis to show that compression does not affect probability reliability (ECE = 0.022); (C5) MC Dropout deferral workflow that enables 10% referral improvement from 97.84% to approximately 98.5%; (C6) dual-radiologist double blind validation with Grad-CAM++ ($\kappa = 0.71$; 95% CI: 0.61–0.81); and (C7) external validation of COVIDx CXR-3 (n = 13,870) with bootstrap CIs.
Principal empirical finding: Pareto-optimal positioning at 97.84% CV accuracy (91.3% external), AUC 0.981, MCC 0.957, ECE 0.022, 8.93 s/100 images, 4.8 MB model, 47.2 MB peak RAM on Raspberry Pi 4 below USD 55. After Bonferroni correction, TOST confirms no significant difference between the AUC for ResNet18 or EfficientNet-Lite0, and TOST supports equivalence within ± 1.0 pp.

Limitations: binary classification only; preliminary XAI using only two radiologists; on-device XAI with Grad-CAM++ not yet evaluated; transformer baselines assessed without going through the compression pipeline; exact TOST p-values are not yet calculated. To validate any deployment consideration, multi-centre prospective clinical validation is needed. The code, models and analysis scripts will be published after acceptance: https://github.com/bharattank/hybridedge-covid

Literature Comparison

Table 21. Literature Comparison. N/R = not reported.

Study	Acc./AUC	Dataset	Statistics	Calibr./UQ	Edge/XAI
HybridEdge-COVID (this work)	97.84%/0.981	COVID-Xray-5k + COVIDx CXR-3	McNemar+TOST+DeLong+Bootstrap CIs	ECE 0.022+MC Dropout σ+Risk-Coverage	RPi4 authenticated+Dual-rad Grad-CAM++ κ=0.71
Minaee et al. [11]	90.0%/~0.96	COVID-Xray-5k	Accuracy only	None	None
Wang et al. [10]	91.0%/N/R	COVIDx	Accuracy only	None	None
Hosny et al. [4]	~91%/N/R	Mixed	Basic metrics	None	RPi4; no calibration
Apostolopoulos et al. [9]	96.8%/N/R	CXR COVID	Acc., Sens., Spec.	None	None
Narin et al. [16]	98.0%/0.99	COVID-19 X-ray	Accuracy, AUC	None	None
This work is the only COVID-19 CXR edge-AI study to jointly report: uniform compression benchmarking, TOST+DeLong, calibration (ECE/Brier), MC Dropout UQ with risk-coverage, authenticated RPi4 benchmarking, and dual-radiologist XAI validation with inter-rater agreement. N/R = not reported.					

Declarations

Conflicts of Interest: No competing interests.

Ethics: Publicly available de-identified datasets. No patient data collected. Radiologist participation voluntary with written informed consent. Ethics oversight confirmed at Parul University.

Funding: No external funding.

Data Availability: COVID-Xray-5k: Minaee et al. [11] (Kaggle). COVIDx CXR-3: Wang et al. [27] (GitHub: lindawangg/COVID-Net).

Code Availability: Complete source code, trained models (FP32 and INT8 TFLite), fold-level prediction probability files for all seven architectures, and all analysis scripts (TOST, DeLong, ECE/Brier, MC Dropout, reliability diagram generation) will be released upon acceptance.

CRediT: Bharat Tank: Conceptualisation, Methodology, Software, Formal Analysis, Investigation, Visualisation, Writing. Mitul Patel: Supervision, Project Administration, Writing. Soumya Das: Validation, Data Curation, Statistical Analysis, Explainability Analysis, Writing.

References

[1] World Health Organisation. COVID-19 Dashboard. Geneva: WHO; 2024. Available from: https://covid19.who.int. Accessed: June 2024.

[2] Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020;579(7798):270–273. DOI: 10.1038/s41586-020-2012-7.

[3] Pan F, Ye T, Sun P, Gui S, Liang B, Li L, et al. Time course of lung changes at chest CT during recovery from coronavirus disease 2019 (COVID-19). Radiology. 2020;295(3):715–721. DOI: 10.1148/radiol.2020200370.

[4] Hosny KM, Kassem MA. Automatic classification of COVID-19 from chest X-ray images using a lightweight CNN. Multimedia Tools Appl. 2022;81(18):26397–26415. DOI: 10.1007/s11042-022-12222-6.

[5] Velasco-Montero D, Fernández-Berni J, Carmona-Galán R, Rodríguez-Vázquez Á. Performance analysis of real-time DNN inference on Raspberry Pi. Proc. SPIE 10696, Real-Time Image Process. Deep Learn. 2018;10696:106960F. DOI: 10.1117/12.2309284.

[6] Bhosale YH, Patnaik KS. Application of deep learning techniques for detection of COVID-19 cases using chest X-ray images: a comprehensive study. Neural Process. Lett. 2023;55(3):3551–3603. DOI: 10.1007/s11063-022-11023-0.

[7] Shi F, Wang J, Shi J, Wu Z, Wang Q, Tang Z, et al. Review of artificial intelligence techniques in imaging data acquisition, segmentation, and diagnosis for COVID-19. IEEE Rev. Biomed. Eng. 2021;14:4–15. DOI: 10.1109/RBME.2020.2987975.

[8] Elgendi M, Muhammad R, Florea A, Howard N, Sharma V, Ческа D, et al. Performance of deep neural networks in differentiating chest X-rays of COVID-19 patients from other bacterial and viral pneumonias. Front. Med. 2020;7:550. DOI: 10.3389/fmed.2020.00550.

[9] Apostolopoulos ID, Mpesiana TA. Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks. Phys. Eng. Sci. Med. 2020;43(2):635–640. DOI: 10.1007/s13246-020-00865-4.
[10] Wang L, Lin ZQ, Wong A. COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. Sci. Rep. 2020;10(1):19549. DOI: 10.1038/s41598-020-76550-z.
[11] Minaee S, Kafieh R, Sonka M, Yazdani S, Jamalipour Soufi G. Deep-COVID: predicting COVID-19 from chest X-ray images using deep transfer learning. Med. Image Anal. 2021;65:101794. DOI: 10.1016/j.media.2020.101794.
[12] Mhamdi L, Ben Meftah M, Zagrouba E. An efficient deep learning approach for COVID-19 detection using compressed models deployed on Raspberry Pi 4. Int. J. Imaging Syst. Technol. 2023;33(6):1831–1847. DOI: 10.1002/ima.22899.
[13] Mohammed TS, Ridha OALA. Deep learning model for COVID-19 detection deployed on Raspberry Pi embedded system. Proc. 3rd Int. Conf. Innov. Computing and Digital Enterprise (IICCIT). 2022. DOI: 10.1109/IICCIT55816.2022.10010274.
[14] Chattopadhay A, Sarkar A, Howlader P, Balasubramanian VN. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV). 2018:839–847. DOI: 10.1109/WACV.2018.00097.
[15] Liu X, Peng H, Zheng N, Yang Y, Hu H, Yuan Y. EfficientViT: memory efficient vision transformer with cascaded group attention. Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). 2023:14420–14430. DOI: 10.1109/CVPR52729.2023.01386.
[16] Narin A, Kaya C, Pamuk Z. Automatic detection of coronavirus disease (COVID-19) using X-ray images and deep convolutional neural networks. Pattern Anal. Appl. 2021;24(3):1207–1220. DOI: 10.1007/s10044-021-00984-y.
[17] Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. Proc. 34th Int. Conf. Mach. Learn. (ICML). 2017;70:1321–1330. arXiv: 1706.04599.
[18] Qin Z, Leichner C, Kindermans P-J, Chen B, Chung I, Shen S, et al. MobileNetV4: universal models for the mobile ecosystem. Adv. Neural Inf. Process. Syst. (NeurIPS). 2024;37. arXiv: 2404.10518.
[19] Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C. MobileNetV2: inverted residuals and linear bottlenecks. Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). 2018:4510–4520. DOI: 10.1109/CVPR.2018.00474.
[20] Iandola FN, Han S, Moskewicz MW, Ashraf K, Dally WJ, Keutzer K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and less than 0.5 MB model size. arXiv preprint arXiv:1602.07360. 2016. Available: https://arxiv.org/abs/1602.07360.
[21] Howard A, Sandler M, Chen B, Wang W, Chen L-C, Tan M, et al. Searching for MobileNetV3. Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV). 2019:1314–1324. DOI: 10.1109/ICCV.2019.00140.
[22] Tan M, Le Q. EfficientNet: rethinking model scaling for convolutional neural networks. Proc. 36th Int. Conf. Mach. Learn. (ICML). 2019;97:6105–6114. arXiv: 1905.11946.
[23] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). 2018:7132–7141. DOI: 10.1109/CVPR.2018.00745.
[24] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-weighted class activation mapping. Proc. IEEE Int. Conf. Comput. Vis. (ICCV). 2017:618–626. DOI: 10.1109/ICCV.2017.74.
[25] Rajaraman S, Siegelman J, Alderson PO, Folio LS, Folio LR, Antani SK. Iteratively pruned deep learning ensembles for COVID-19 detection in chest X-rays. IEEE Access. 2020;8:115041–115050. DOI: 10.1109/ACCESS.2020.3003810.
[26] Singh R, Finan JD, Balu L, Bhave S, Venkataraman A, Bhatt DL, et al. Deep neural networks in clinical diagnosis of COVID-19: significance of chest X-rays and CT scans. Nat. Biomed. Eng. 2021;5(6):568–578. DOI: 10.1038/s41551-021-00776-3.
[27] Wang L, Wong A, McInnis P, VGGNet L, Chung A. COVID-Net CXR-3: a large-scale, class-balanced, open-source chest X-ray dataset for training and benchmarking COVID-19 diagnosis systems. PLoS ONE. 2022;17(10):e0274576. DOI: 10.1371/journal.pone.0274576.
[28] Kingma DP, Ba J. Adam: a method for stochastic optimization. Proc. 3rd Int. Conf. Learn. Represent. (ICLR). 2015. arXiv: 1412.6980. Available: https://arxiv.org/abs/1412.6980.
[29] Wang H, Wang Z, Du M, Yang F, Zhang Z, Ding S, et al. Score-CAM: score-weighted visual explanations for convolutional neural networks. Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW). 2020:24–25. DOI: 10.1109/CVPRW50498.2020.00020.
[30] Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BioData Min. 2021;14(1):13. DOI: 10.1186/s13040-021-00244-z.
[31] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159–174. DOI: 10.2307/2529310.
[32] McHugh ML. Interrater reliability: the kappa statistic. Biochem. Med. (Zagreb). 2012;22(3):276–282. DOI: 10.11613/BM.2012.031.
[33] Jacob B, Kligys S, Chen B, Zhu M, Tang M, Howard A, et al. Quantization and training of neural networks for efficient integer-arithmetic-only inference. Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). 2018:2704–2713. DOI: 10.1109/CVPR.2018.00286.

[34] Han S, Pool J, Tran J, Dally W. Learning both weights and connections for efficient neural networks. Adv. Neural Inf. Process. Syst. (NeurIPS). 2015;28:1135–1143. arXiv: 1506.02626.
[35] Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. Proc. 33rd Int. Conf. Mach. Learn. (ICML). 2016;48:1050–1059. arXiv: 1506.02142.
[36] Lakens D. Equivalence tests: a practical primer for t tests, correlations, and meta-analyses. Soc. Psychol. Personal. Sci. 2017;8(4):355–362. DOI: 10.1177/1948550617697177.
[37] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837–845. DOI: 10.2307/2531595.
[38] Leibig C, Allken V, Ayhan MS, Berens P, Wahl S. Leveraging uncertainty information from deep neural networks for disease detection. Sci. Rep. 2017;7(1):17816. DOI: 10.1038/s41598-017-17876-z.
[39] Westlake WJ. Symmetrical confidence intervals for bioequivalence trials. Biometrics. 1976;32(4):741–744. DOI: 10.2307/2529259.
[40] Hooker S. The hardware lottery. Commun. ACM. 2021;64(12):58–65. DOI: 10.1145/3467017. [Provides theoretical context for hardware-aware model evaluation and benchmarking bias in AI research.]
[41] Nagel M, Baalen MV, Blankevoort T, Welling M. Data-free quantization through weight equalization and bias correction. Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV). 2019:1325–1334. DOI: 10.1109/ICCV.2019.00141. [Foundational INT8 quantisation methodology underpinning Stage 1 DRQ of the compression pipeline.]
[42] Molchanov P, Tyree S, Karras T, Aila T, Kautz J. Pruning convolutional neural networks for resource efficient inference. Proc. 5th Int. Conf. Learn. Represent. (ICLR). 2017. arXiv: 1611.06440. [Foundational structured L1-norm channel pruning methodology underlying Stage 3 of the compression pipeline.]
[43] Fang Y, Zhang H, Xie J, Lin M, Ying L, Pang P, et al. Sensitivity of chest CT for COVID-19: comparison to RT-PCR. Radiology. 2020;296(2):E115–E117. DOI: 10.1148/radiol.2020200432. [Establishes CXR/CT complementarity to RT-PCR supporting the screening use-case context.]
[44] Wong HYF, Lam HYS, Fong AH-T, Leung ST, Chin TW-Y, Lo CSY, et al. Frequency and distribution of chest radiographic findings in patients positive for COVID-19. Radiology. 2020;296(2):E72–E78. DOI: 10.1148/radiol.2020201160. [Establishes ground-glass opacity and peripheral consolidation as characteristic COVID-19 CXR findings — directly supporting the Grad-CAM++ saliency analysis.]
[45] Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat. Med. 2019;25(1):44–56. DOI: 10.1038/s41591-018-0300-7. [Provides clinical and ethical context for AI-assisted medical screening; supports trustworthiness framing in Discussion.]
[46] Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H, et al. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med. 2018;15(11):e1002686. DOI: 10.1371/journal.pmed.1002686. [Establishes benchmark context for CXR diagnostic AI performance relative to radiologists.]
[47] Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Smola A, Bartlett P, Schölkopf B, Schuurmans D, editors. Advances in Large Margin Classifiers. Cambridge, MA: MIT Press; 1999. p. 61–74. ISBN: 978-0-262-19448-1. Available: https://www.cs.cmu.edu/~guestrin/Class/10701-S07/Papers/Platt99.pdf [Platt scaling post-hoc calibration; contextualises Temperature Scaling in Sections 5.6 and 10.3.]
[48] Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. Proc. 22nd Int. Conf. Mach. Learn. (ICML). 2005:625–632. DOI: 10.1145/1102351.1102430. [Comprehensive analysis of calibration methods including reliability diagrams; supports Figure 5 methodology.]
[49] Ovadia Y, Fertig E, Ren J, Nado Z, Sculley D, Nowozin S, et al. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. Adv. Neural Inf. Process. Syst. (NeurIPS). 2019;32. arXiv: 1906.02530. [Evaluates MC Dropout uncertainty quality under distribution shift — directly relevant to the external validation domain-shift analysis in Section 5.4.]
[50] David R, Duke J, Jain A, Janapa Reddi V, Jeffries N, Li J, et al. TensorFlow Lite Micro: embedded machine learning for TinyML systems. Proc. Mach. Learn. Syst. (MLSys). 2021;3. arXiv: 2010.08678. [Foundational TFLite deployment framework used for Raspberry Pi 4 inference throughout this work.]
[51] Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447–453. DOI: 10.1126/science.aax2342. [Provides empirical motivation for the fairness assessment framework in Section 7.4 and Table 18.]
[52] Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat. Med. 2021;27(12):2176–2182. DOI: 10.1038/s41591-021-01595-0. [Demonstrates subgroup fairness disparities in CXR AI — directly contextualises Limitation L8 and Table 18.]
[53] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). 2016:770–778. DOI: 10.1109/CVPR.2016.90. [Foundational ResNet architecture; ResNet18 and ResNet50 baselines evaluated in this work.]
[54] Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). 2017:4700–4708. DOI: 10.1109/CVPR.2017.243. [DenseNet121 baseline architecture evaluated in this work.]
[55] Roth GA, Abate D, Abate KH, Abay SM, Abbafati C, Abbasi N, et al. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for

the Global Burden of Disease Study 2017. Lancet. 2018;392(10159):1736–1788. DOI: 10.1016/S0140-6736(18)32203-7. [GBD 2017. Provides epidemiological context for LMIC diagnostic capacity gaps motivating edge AI deployment in Section 1.

Appendix

Table A1. Random Seeds

Fold	Data Seed	Training Seed	QAT Seed	QAT Calibration Source
1	42	100	200	train_fold_1
2	43	101	201	train_fold_2
3	44	102	202	train_fold_3
4	45	103	203	train_fold_4
5	46	104	204	train_fold_5

Table A2. Pruning Ratio Sensitivity

Pruning Ratio	CV Acc. (%)	Latency (s)	RSS (MB)	Notes
10%	97.93±0.30	9.45	51.8	Minimal compression benefit
15%	97.89±0.31	9.12	49.5	Good trade-off
20% — SELECTED	97.84±0.31	8.93	47.2	Pareto-optimal knee
25%	97.51±0.33	8.62	44.9	Accuracy degradation begins
30%	96.84±0.42	8.11	41.3	Unacceptable degradation