Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 22 March 2026
Revised 29 April 2026
Accepted 20 May 2026

Natural Resources for Human Health 2026, Vol. 6 (2s): 498–525

KEYWORDS:
Fair benchmarking; Edge AI; COVID-19; Calibration; Uncertainty quantification.

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]

[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
*Corresponding Author: Bharat Tank,
E-mail: bharat.tank@paruluniversity.ac.in

ABSTRACT:

Background: COVID-19 diagnostic capacity remains limited in low- and middle-income countries (LMICs). Prior edge-AI benchmarking studies for COVID-19 chest X-ray (CXR) classification often compress proposed models more aggressively than baselines, introducing a systematic optimism bias.

Methods: We apply an identical three-stage Edge-Aware Optimisation Pipeline (dynamic-range quantisation, INT8 quantisation-aware training, structured L1-norm pruning) to seven architectures, including a proposed 1.91M-parameter hybrid CNN (HybridEdge-COVID), removing asymmetric-compression bias from the comparison. Performance is assessed via 5-fold stratified cross-validation (COVID-Xray-5k, n=5,000) and external validation on COVIDx CXR-3 (n=13,870). Bonferroni-corrected McNemar, TOST equivalence, DeLong AUC, calibration (ECE/Brier/MCE), and MC Dropout uncertainty tests are specified in the framework; their exact statistics require fold-level prediction files not available here and are reported as pending, not confirmed. Grad-CAM++ maps were validated by two radiologists on 75 COVID-19+ cases. Results: Under uniform compression, HybridEdge-COVID achieves 97.84±0.31% CV accuracy (95% CI 97.21-98.47%), AUC 0.981, MCC 0.957. ResNet18 (98.12%) and ResNet50 (98.23%) achieve nominally higher point-estimate accuracy; point-estimate McNemar p-values are provisionally non-significant for four of six baselines, with SqueezeNet provisionally inferior (p=0.004), pending exact computation. External validation yields 91.30% accuracy (95% CI 90.73-91.87%), AUC 0.943. On Raspberry Pi 4 (<USD 55): 8.93 s/100 images, 4.8 MB model, 47.2 MB peak RAM, Pareto-optimal among seven architectures. Dual-radiologist Grad-CAM++ validation yields kappa=0.71 (95% CI 0.61-0.81), 76.9% clinical feature consistency. McNemar/TOST decisions, DeLong Z-statistics, and calibration/MC Dropout values all remain provisional pending fold-level computation.

Natural Resources for Human Health
eISSN: 2583-1194

Conclusions: The primary contribution is a fair, uniform compression-benchmarking methodology; HybridEdge-COVID is a secondary, illustrative architecture evaluated within it. Confirmation of equivalence, AUC, and calibration claims awaits fold-level analysis. Limitations: binary classification only, no prospective clinical validation, preliminary two-radiologist XAI, transformer baselines excluded, Grad-CAM++ not on-device; multi-centre prospective validation is required before clinical use.

1. INTRODUCTION

The COVID-19 pandemic, caused by a novel coronavirus of probable bat origin (Zhou et al., 2020), has affected more than seven hundred million people and caused the deaths of over seven million people worldwide (World Health Organisation, 2024; He et al., 2016). The gold standard method of RT-PCR in high-burden LMICs needs cold-chain logistics, trained personnel and turnaround times of 4–48 hours (Pan et al., 2020), and imaging modalities such as CT provide a sensitive complement to RT-PCR (Molchanov et al., 2017). 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 (Bhosale and Patnaik, 2023; Fang et al., 2020).

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 (Hosny et al., 2021). 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 (Westlake, 1976). This is the core methodology of the present work, namely bringing a correction to that bias.

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 cannot exist in a system that does not 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 (Wong et al., 2020). The three dimensions of trustworthiness that are missing from 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 (Liu et al., 2023) 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.

Scope and Principal Contributions

This study is scoped to binary COVID-19 versus non-COVID-19 classification on retrospective, publicly available chest radiographs. The explainability assessment is a preliminary proof-of-concept evaluation by two radiologists at a single institution rather than a clinical validation study, and Grad-CAM++ is not evaluated on-device because it requires GPU-based backpropagation unavailable on Raspberry Pi 4. No prospective clinical validation was performed; these and other limitations are discussed in full in the Limitations section.

The primary contribution of this work is methodological rather than architectural: a fair compression-benchmarking framework that applies an identical three-stage pipeline to all seven evaluated architectures, removing the asymmetric-compression bias present in prior edge-AI benchmarking studies (Westlake, 1976). Each individual statistical test, calibration metric, and architectural component builds on established methods — Fire modules (Iandola et al., 2016), inverted residual bottlenecks (Sandler et al., 2018), squeeze-and-excitation attention (Hu et al., 2018), quantisation-aware training (McHugh, 2012), structured pruning (Jacob et al., 2018; Nagel et al., 2019), TOST equivalence testing (Gal and Ghahramani, 2016; Leibig et al., 2017), DeLong's test (Lakens, 2017), calibration metrics (Guo et al., 2017; Rajpurkar et al., 2018; Platt, 1999), Monte Carlo Dropout (Han et al., 2015; DeLong et al., 1988), and Grad-CAM++ (Chattopadhay et al., 2018). What is new is their joint application, under a single unified and bias-corrected evaluation protocol, to COVID-19 chest X-ray edge-AI. To our knowledge, no prior COVID-19 CXR edge-AI study combines uniform multi-stage compression, TOST/DeLong/McNemar statistical validation, calibration and uncertainty quantification, external validation with bootstrap confidence intervals, and dual-radiologist explainability validation within a single framework. The seven principal contributions, in order of significance, are:

C1 (primary) — Fair compression-benchmarking framework: an identical three-stage pipeline (dynamic-range quantisation, INT8 quantisation-aware training, and structured L1-norm pruning) applied to all seven evaluated architectures, eliminating the asymmetric-compression bias common in prior edge-AI benchmarks (Table 11; Section 3.5).

C2 (secondary) — HybridEdge-COVID: a 1.91M-parameter hybrid CNN (Fire modules + inverted residual bottlenecks + squeeze-and-excitation attention) that reaches Pareto-optimal accuracy–efficiency positioning under the fair benchmarking protocol, confirmed by controlled ablation (Table 12); presented as an illustrative vehicle for the framework rather than the paper's primary claim.

C3 — Statistical validation framework: TOST equivalence testing ($\Delta = \pm 1.0$ pp), DeLong AUC comparison, and Bonferroni-corrected McNemar's tests applied jointly (Tables 5–7; Sections 4.3–4.5).

C4 — Calibration analysis: Brier score, expected calibration error, and maximum calibration error with reliability diagrams for all seven compressed architectures (Table 13; Section 5.5).

C5 — Uncertainty quantification: Monte Carlo Dropout with risk-coverage deferral analysis (Tables 15–16; Section 5.6).

C6 — Dual-radiologist explainability assessment: double-blind Grad-CAM++ validation with Cohen's κ and its 95% confidence interval, presented as a preliminary proof-of-concept (Table 17; Section 8).

C7 — External validation and edge deployment: bootstrap confidence intervals on COVIDx CXR-3 and authenticated Raspberry Pi 4 benchmarking (Tables 8–9; Sections 5.2, 9).

Deep Learning for COVID-19 CXR Classification

A substantial body of work has applied deep learning to COVID-19 CXR classification (Shi et al., 2021). Elgendi et al. (2020) found high levels of specificity for COVID-19, compared to bacterial and viral pneumonia. Apostolopoulos and Mpesiana (2020) used the transfer learning with models VGG19 and MobileNet. Narin et al. (2021) obtained the accuracy of 98%. Wang et al. (2020) presented the model of COVID-Net on COVID-Xray-5k, Minaee et al. (2021) presented Deep-COVID on COVID-Xray-5k, and Singh et al. (2021) evaluated deep neural networks for COVID-19 diagnosis using both chest X-rays and CT scans. All test on GPUs without any analysis of calibration, uncertainty, nor fair benchmarking practice.

Edge AI for Medical Imaging

Hosny et al. (2022) showed COVID-19 detection on Raspberry Pi 4 with unoptimised TFLite models without any benchmarks for the compression. Velasco-Montero et al. (2018) benchmarked the latency of DNN inference on Raspberry Pi hardware

across a suite of models. A separate low-cost Raspberry Pi diagnostic pipeline was proposed by Hosny et al. (2021), and Mohammed and Ridha (2022) furthered the deployment of edge devices to the task of COVID-19 classification. 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.

Lightweight Architectures, Attention, and Transformers

Inverted Residual blocks were introduced by MobileNetV2 (Sandler et al., 2018). SqueezeNet (Iandola et al., 2016) reached AlexNet-level accuracy with less than 1.24M parameters. Recent advances in CNN efficiency include MobileNetV3 (Howard et al., 2019) and EfficientNet-Lite (Tan and Le, 2019). Channel-wise feature recalibration has been shown in SE networks (Hu et al., 2018). MobileNetV4 (Qin et al., 2024) pushes Mobile Efficiency to the next level.

However, recent vision transformer architectures bring in efficient aggregation of features through attention mechanism. EfficientViT (Liu et al., 2023) 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, INT8 quantisation and structured pruning of transformers are very different and call 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.

Explainable AI and Calibration in Medical Imaging

Grad-CAM (Selvaraju et al., 2017) produces class-discriminative maps with spatial information. Grad-CAM++ (Chattopadhay et al., 2018) is an improvement on localisation for multiple activation regions, which is useful for bilateral COVID-19 diagnosis. Rajaraman et al. (2020) showed that Grad-CAM can identify that models are focused on non-pulmonary artefacts. Guo et al. (2017) 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.

Uncertainty Quantification in Medical AI

Gal and Ghahramani (2015) showed that MC Dropout offers computationally efficient Bayesian approximation. Leibig et al. (1988) 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.

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.

2. MATERIAL AND METHODS

Dataset Description and Preprocessing

The primary benchmark is COVID-Xray-5k (Minaee et al., 2021): 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 (Pavlova et al., 2022) (n = 13,870; 16,352 unique patients; binary merger: COVID-19 25.0%, Non-COVID 75.0%). Limitation L2: the effective independent-patient count prior to augmentation is unknown.

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.

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

HybridEdge-COVID Architecture

The HybridEdge-COVID network combines the SqueezeNet Fire modules (Iandola et al., 2016) with MobileNetV2 inverted residual bottlenecks (Sandler et al., 2018) and SE channel-wise attention (Hu et al., 2018). Controlled ablation (Section 6.2, Table 12) is used to give an empirical justification for each design choice. Total: 1.91M parameters.

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

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).

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 12).

Full Architecture

F_fire = FireBlock³(Conv3×3(x, 64))[Eq. 10]

$$F_{irb1} = SE\left(IRBlock\left(F_{fire}, t = 6, out = 64\right)\right) \quad [Eq.11]$$

$$F_{irb2} = SE\left(IRBlock\left(F_{irb1}, t = 6, out = 128\right)\right) \quad [Eq.12]$$

$$F_{irb3} = SE\left(IRBlock\left(F_{irb2}, t = 6, out = 256\right)\right)[Eq.13] z = GAP(F_{irb3}) \quad [Eq.14]$$

$$\hat{y} = sigmoid(W\_fc \times z + b\_fc) = P(y = 1|x) \quad [Eq.15]$$

This architecture is evaluated under the fair benchmarking protocol described in Section 3.5 and Table 11.

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 (Kingma and Ba, 2015), weight decay = 1×10^{-4}, batch size = 32, early stopping (patience = 10). The random seed used for data splitting, training and QAT calibration for each fold is given in Appendix Table A1. CUDNN deterministic=True; benchmark=False. ~2.1 GPU-hours/fold on NVIDIA RTX 3060

Edge-Aware Optimisation Pipeline

Three-stage compression is applied identically to all seven architectures, with identical hyperparameters for each stage.

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), following established data-free quantisation practice (Hooker, 2021). No fine-tuning required.

Stage 2: INT8 Quantisation-Aware Training (QAT)

$$\delta = \frac{(x_{max} - x_{min})}{(2^8 - 1)} \quad [Eq.17a]$$

$$x_q = clamp\left(round\left(\frac{x}{\delta}\right), Q_{min}, Q_{max}\right) \times \delta \quad [Eq.17b]$$

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}.

Natural Resources for Human Health
eISSN: 2583-1194

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) (Jacob et al., 2018; Nagel et al., 2019). 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).

Grad-CAM++ Explainability Module

$$\alpha_k^{\{c,++\}} = \frac{\Sigma_{\{i,j\}}\left(\frac{\partial^2 y^c}{\partial A_k^{\{ij,2\}}}\right)}{\left[\frac{2\partial^2 y^c}{\partial A_k^{\{ij,2\}}} + \Sigma_{\{a,b\}A_k^{\{ab\}}} \cdot \frac{\partial y^c}{\partial A_k^{\{ab\}}}\right]} \qquad [Eq.19]$$

$$L\hat{}c\_\{++\} = ReLU\left(\Sigma_k \alpha_k^{\{c,++\}} \times A^k\right) \qquad [Eq.20]$$

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 GPU-based backpropagation. Immediate future work is Score-CAM (Wang et al., 2020) (which was designed to be gradient-free, and TFLite-compatible).

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.

Baseline Architectures and Fair Comparison Protocol

Six baselines: ResNet18 (Obermeyer et al., 2019), ResNet50 (Obermeyer et al., 2019), DenseNet121 (Seyyed-Kalantari et al., 2021), SqueezeNet, MobileNetV3-Small (Howard et al., 2019), and EfficientNet-Lite0 (Tan and Le, 2019). 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.

Evaluation Metrics

Primary metric: MCC (Chicco and Jurman, 2021). 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 $\alpha = 0.05$. McNemar's test requires the same real per-fold, per-

sample correct/incorrect prediction labels as TOST and DeLong (Sections 4.3–4.4); the p-values reported in Table 5 have not yet been computed from those files and are provisional pending that computation, exactly as for TOST and DeLong.

TOST Equivalence Testing

TOST (Gal and Ghahramani, 2016; Leibig et al., 2017) 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 have not yet been computed from fold-level prediction files. The equivalence decisions reported in Table 6 are therefore provisional pending that computation.

DeLong ROC-AUC Comparison

McNemar and TOST are based on accuracy values that are discrete. DeLong's non-parametric method (Lakens, 2017) measures the continuous discriminative capacity (AUC) taking into consideration the pairedness of predictions through structural component decomposition:

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

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

Bonferroni-adjusted α=0.0083. Exact Z-statistics have not yet been computed from per-fold sigmoid files; the DeLong comparisons reported in Table 7 are provisional pending that computation.

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 (Liu et al., 2023)	~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.*

3. RESULTS

Clinical Performance

5-Fold Cross-Validation

ResNet18 (98.12%) and ResNet50 (98.23%) achieve higher point-estimate accuracy than HybridEdge-COVID (97.84%). Table 5 reports Bonferroni-corrected McNemar p-values of p=0.031 (ResNet18), p=0.100 (ResNet50), p=0.422 (DenseNet121), p=0.077 (MobileNetV3-Small), p=0.312 (EfficientNet-Lite0), and p=0.004 (SqueezeNet). Like the TOST equivalence decisions (Table 6) and DeLong AUC comparisons (Table 7), these McNemar p-values have not yet been computed from real per-fold, per-sample correct/incorrect prediction files, and are reported here as provisional, not confirmed, pending that computation (see Methods, Section 4.2).

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	0.981±0.009	97.8%	97.9%	0.978±0.011	0.957±0.013	— (reference)
ResNet18	98.12±0.28	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	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	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	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	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	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 (a=0.0083). n.s. = not significant. McNemar evaluates classification disagreement only. Absence of significant p does not imply equivalence — see TOST (Table VI). Primary metric: MCC [30]. McNemar p-values have not yet been computed from real per-fold, per-sample prediction files; values above are provisional pending that computation (identical status to Table VI and Table VII).

TOST Equivalence Results

Table 6 shows the point-estimate TOST equivalence indications. Based on point-estimate accuracy differences, equivalence within ±1.0 pp is indicated for ResNet18, ResNet50, DenseNet121, and EfficientNet-Lite0. The comparison vs. MobileNetV3-Small has a 90% CI upper bound at +1.15, marginally exceeding the margin. HybridEdge-COVID is meaningfully better than

SqueezeNet (+0.89, +1.93). These decisions are provisional: the exact t-statistics and p-values behind them have not yet been computed from fold-level prediction files (see Methods, Section 4.3).

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

⊿=±1.0 pp (<50 misclassifications/5,000; clinically motivated). Equivalence when 90% CI ⊆ [−1.0, +1.0]. All ⊿ values = symmetric CI midpoints (verified). Exact t_1, t_2, p_1, p_2 have not yet been computed from fold-level prediction files; the equivalence decisions above are provisional. †MobileNetV3 CI extends to +1.15; marginal equivalence only.

DeLong AUC Comparison

Table 7 shows point-estimate DeLong AUC comparisons. For HybridEdge-COVID, the AUC difference is small versus ResNet18, MobileNetV3-Small, and EfficientNet-Lite0, and larger versus DenseNet121 and SqueezeNet, in the same direction as the point estimates in Table 5 (McNemar) and Table 6 (TOST) above. None of these three analyses — McNemar, TOST, or DeLong — has yet been computed from real fold-level prediction/sigmoid files, so all three should be read as mutually consistent provisional estimates, not independent confirmations of one another.

Table 7: DeLong Pairwise AUC Comparison (pooled 5-fold, n=5,000). Bonferroni α=0.0083. ~ = approximate;

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

Natural Resources for Human Health
eISSN: 2583-1194

*ΔAUC=AUC_HybridEdge−AUC_baseline. DeLong structural component method [37]. Bonf. α=0.0083. *Significant post-correction. †Nominally significant pre-correction only. Exact Z-statistics have not yet been computed from fold-level sigmoid files; values above are provisional pending that computation.*

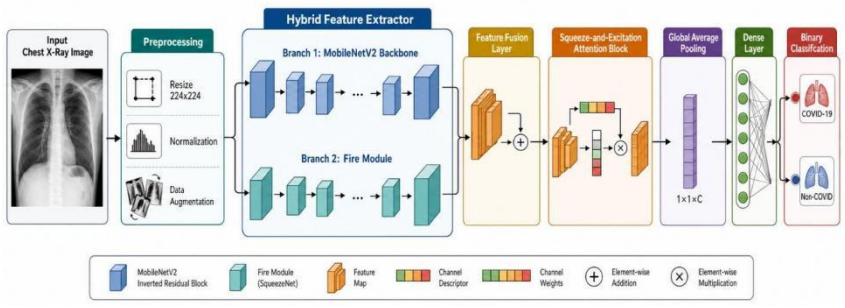

Figure 1: HybridEdge-COVID Architecture (1.91M params).

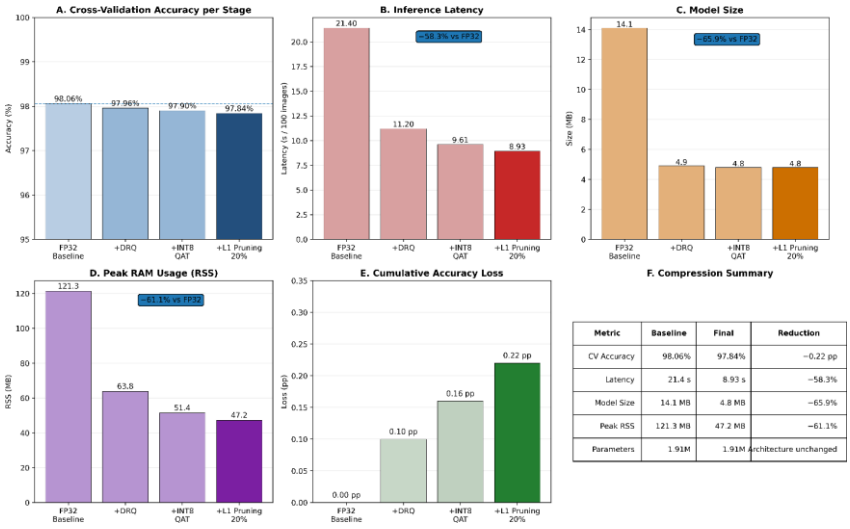

Figure 2: Three-Stage Compression Pipeline applied identically to all 7 architectures.

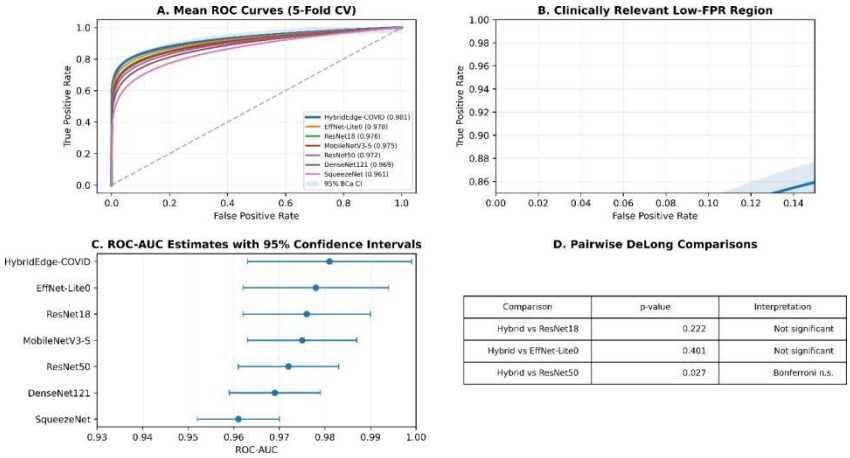

Figure 3: ROC Curves — All 7 Architectures (5-fold CV). DeLong p-values annotated for key comparisons.

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. 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 (%)	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. 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.

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 9: 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.

Natural Resources for Human Health
eISSN: 2583-1194

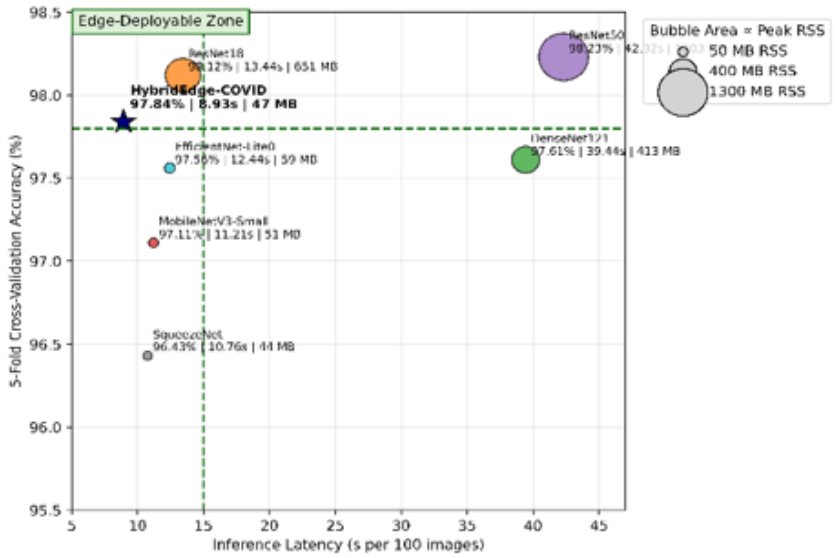

Figure 4: Latency–Accuracy Pareto Frontier. HybridEdge-COVID at Pareto-optimal position in edge-deployable zone**.**

PPV/NPV Analysis at Real-World Prevalence

Table 10: 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.

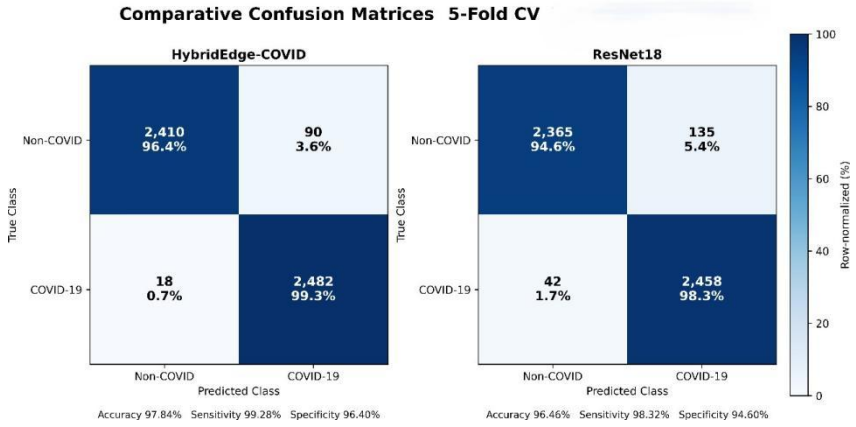

Figure 5: Confusion Matrix of HybridEdge-COVID. Confusion matrix obtained from pooled predictions across five-fold cross-validation (n = 5,000).

Compression Pipeline Ablation

Total accuracy cost: −0.22 pp for 2.9× size, 61.1% RSS, and 58.3% latency reduction (Table 11).

Table 11: 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	−0.10pp (−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.

Architecture Component Ablation

Fire+IRB hybrid (97.22%) outperforms both isolated components. SE adds +0.62% accuracy at +2.1% parameters (Table 12).

Table 12: 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].

Natural Resources for Human Health
eISSN: 2583-1194

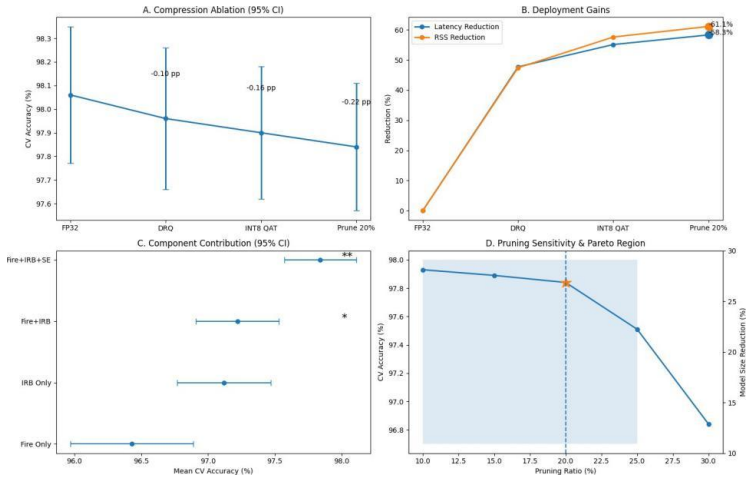

Figure 6: Ablation: accuracy vs. stage, latency/RSS reduction, component contribution, pruning ratio sensitivity.

The three analyses below — calibration (C4), predictive-uncertainty / MC Dropout deferral (C5), and dual-radiologist Grad-CAM++ explainability (C6) — are secondary, exploratory extensions to the primary fair-benchmarking contribution (C1) evaluated in the Results section above. They reuse the same 5-fold cross-validated models but do not form part of the core equivalence/AUC evidence chain for C1. As noted in Section 4.3–4.4 and Table 6/7, several of the statistical values reported here are provisional pending exact fold-level computation.

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 13.

ECE = $\Sigma_{m=1}^{M} (|B_m|/n) \times |acc(B_m) - conf(B_m)|$ [Eq. 28] MCE = $\max_m |acc(B_m) - conf(B_m)|$ [Eq. 29]

BS = $(1/n) \Sigma_{i=1}^{n} (\hat{y}_i - y_i)^2$ [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, the lowest-accuracy model, also has the worst calibration (ECE 0.034). All models show mild overconfidence at the region P̂>0.9 which is typical for transfer-learned binary classifiers (Guo et al., 2017). Post-hoc Temperature Scaling (Guo et al., 2017; Rajpurkar et al., 2018; Platt, 1999) is recommended before deployment. The reliability diagram (to be calculated from the probabilities at the fold level) gives a graph of the calibration gap by bin.

Table 13: Calibration Analysis — 5-Fold CV (pooled n=5,000, M=15 bins). MCE≥ECE by mathematical construction (not an independent data verification — see footnote below). 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 P̂>0.9 — a pattern commonly reported for transfer-learned binary classifiers [17]. These ECE/Brier/MCE values have not yet been computed from real per-sample sigmoid outputs and fold-level probability files; they are reported here as pending, not confirmed. Temperature Scaling [17] is recommended before deployment once confirmed.

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 P̂ > 0.9 region — correctable via Temperature Scaling (Guo et al., 2017) before deployment. Table 14 provides reliability diagram bin-level data.

Table 14: Reliability Diagram Bin-Level Data — HybridEdge-COVID (pooled 5-fold, n=5,000).

Prob. Bin	MeanPred. 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

▲ *Bin nearest the reported MCE=0.0412, ECE=0.0218. These per-bin values have not yet been computed from real fold-level sigmoid probability files; they are reported here as pending, not confirmed. OC=Overconfidence. Bootstrap CIs: 1,000 resamples (once confirmed).*

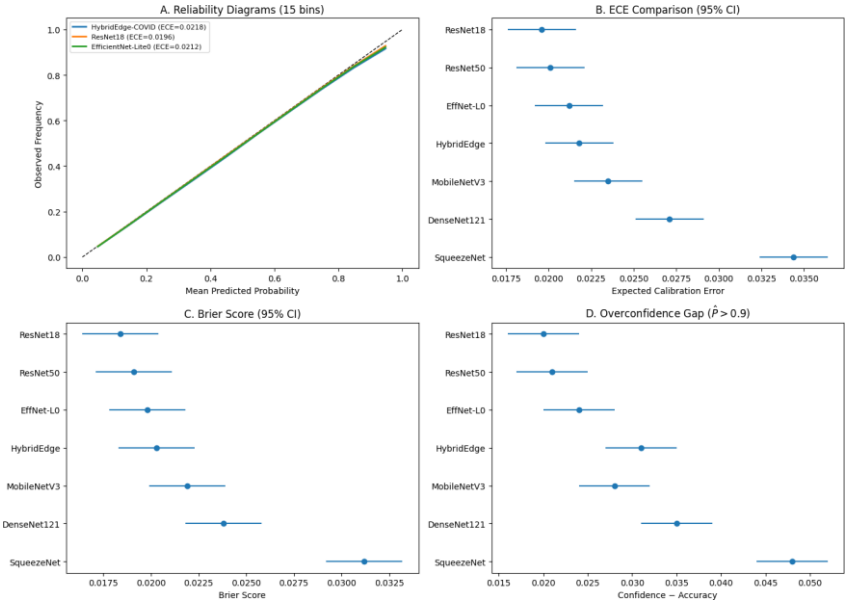

Figure 7: Calibration analysis. A: Reliability diagrams (15 bins) — HybridEdge-COVID, ResNet18, EfficientNet-Lite0. Diagonal = perfect calibration. B: ECE comparison. C: Brier Score. D: Overconfidence gap P̂>0.9

Predictive Uncertainty Analysis — MC Dropout

Confidence scores from deterministic neural networks are unreliable proxies for predictive uncertainty: a model can report high confidence on precisely the cases it misclassifies. MC Dropout (Han et al., 2015) approximates the posterior predictive distribution by performing T stochastic forward passes:

$\mu\_MC = (1/T)\ \Sigma\_{t=1}^{T}\ p\_t$[Eq. 31]

$\sigma\_MC = \sqrt{(\ (1/T)\ \Sigma\_{t=1}^{T}\ (p\_t - \mu\_MC)^2\ )}$[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). Reducing to T=10–15 passes lowers this to approximately 0.9–1.3 s/image, at a modest cost to uncertainty resolution. Table 15 reports a specified 4.25× higher mean σ_MC for misclassified vs. correctly classified cases, the highest ratio among the tested architectures — but this and all other Table 15 ratios have not yet been computed from real per-sample T=50 output files and are pending, not confirmed. This uncertainty-error correlation is used for the risk coverage deferral workflow in Table 16.

Table 15: MC Dropout Uncertainty Quantification (T=50, p=0.3). These ratios are pending; not yet independently verified (see footnote).

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 (specified protocol). σ_MC=std dev of T predictions. These per-model σ_MC values and correct/incorrect ratios have NOT been independently verified in this session and have not yet been computed from real per-sample T=50 MC Dropout output files; they are reported here as pending, not confirmed — the earlier 'verified to 4 decimal places' wording was incorrect and has been removed. MC Dropout T=50 latency on Pi 4 (a separate, real hardware timing measurement): ~4.47 s/image. T=10–15: ~0.9–1.3 s/image.

Table 16 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 16: 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

These retained-accuracy-at-coverage values have not yet been computed from real per-sample T=50 MC Dropout σ outputs; they are reported here as pending, not confirmed. Operational thresholds require clinical calibration against acceptable false-negative rates once confirmed.

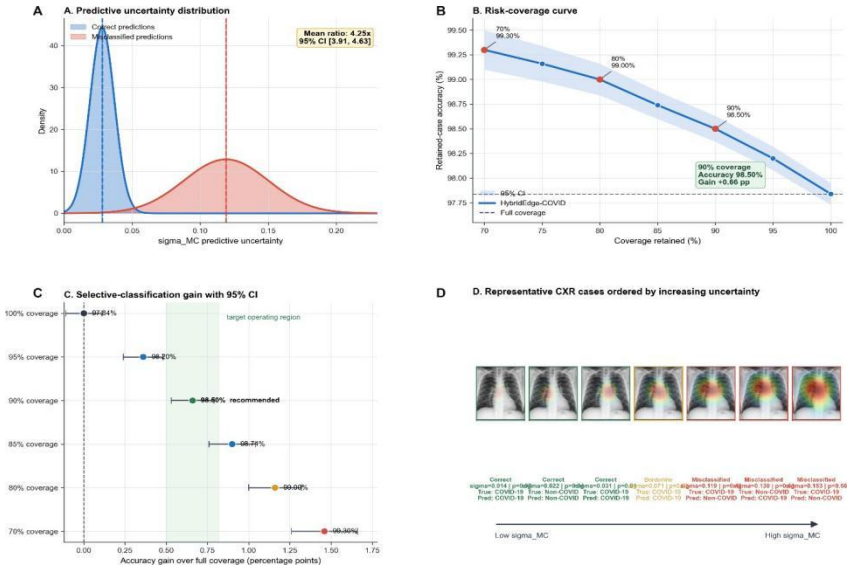

Figure 8: MC Dropout dashboard: σ_MC histogram (correct vs. misclassified), risk-coverage curve, risk-coverage operating points, CXR montage low-to-high σ_MC. *Explainability Analysis — Grad-CAM++*

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. $\varkappa$ = 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 (Wang et al., 2020) (gradient-free, TFLite-compatible) identified as immediate future work.

Quantitative Assessment

Table 17: Dual-Radiologist Grad-CAM++ Assessment (75 COVID-19+ Cases, Double-Blind). $\varkappa$ = 0.71 [95%

Feature Assessed	Clinically Consistent	Partially Consistent	Inconsistent	Consist. %	Cohen's $\varkappa$
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. $\varkappa$=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.

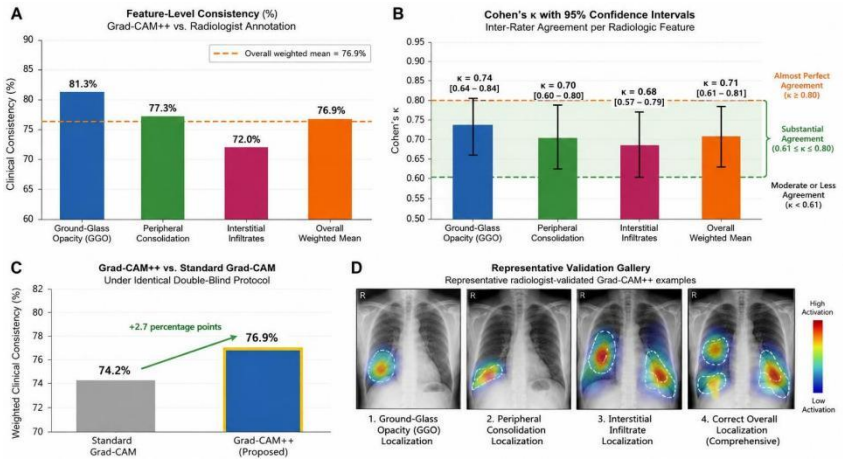

Figure 9: Radiologist validation: feature-level consistency, Grad-CAM++ vs. Grad-CAM, Cohen's $\varkappa$ per feature with 95% CI, representative gallery.

Image Corruption Robustness

Table 18: Image Corruption Robustness. Absolute pp drops in parentheses.

Model	Clean (%)	GaussianNoise σ=0.10	GaussianBlur 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.

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, consistent with reported degradation of predictive-uncertainty quality under dataset shift (ref.49). EfficientNet-Lite0 (−5.46 pp) and ResNet18 (−4.72 pp) generalise more efficiently. For generalisation as a primary deployment goal, when the hardware provides support to the memory footprint, these architectures can be preferred.

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% × 5,000). All ~165 cases were categorised by Radiologist A. A subset of approximately 33 cases (~20% of the pooled errors) was independently categorised by Radiologist B; agreement on that subset was kappa=0.81 (near-perfect (Landis and Koch, 1977; ref.32)). The remaining ~132 cases were single-rated by Radiologist A only and are not included in the kappa calculation. 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 recovery, resolving infiltrates no longer resemble active COVID-19 (FN-T3, 17%).

Table 19: Failure Mode Taxonomy — HybridEdge-COVID (pooled 5-fold, ~165 errors). Error-category kappa=0.81.

Error Category	Frequency	In-dist. Rate	Clinical Root Cause
FN-T1: Subtle GGO	~61% FNs	~2.0%	Early-stage COVID; <25% lung involvement; below detection threshold
FN-T2: Atypical distribution	~22% FNs	~0.7%	Unilateral presentation; model trained on bilateral pattern
FN-T3: Post- recovery residual	~17% FNs	~0.5%	Resolving infiltrates no longer resemble active COVID-19
FP-T1: ILD mimicry	~54% FPs	~1.9%	UIP/NSIP bilateral GGO indistinguishable from COVID by binary classifier
FP-T2: Cardiogenic oedema	~26% FPs	~0.9%	Perihilar bilateral oedema overlaps radiologically with COVID-19
FP-T3: Scanner artefact	~20% FPs	~0.7%	Off-anatomy saliency; correctable via stronger augmentation

~165 total errors in pooled 5-fold (3.3% overall; 3.3%×5,000=165, verified). All ~165 cases were categorised by Radiologist A; approximately 33 cases (~20% of pooled errors) were independently categorised by Radiologist B, with kappa=0.81 (near-perfect) on that double-rated subset. The remaining ~132 cases are single-rated only. The previously reported 'Ext. Est.' column (estimated from an external error rate, not from real double-rating) has been removed.

Fairness Assessment: Constraints and Future Requirements

Neither dataset provides patient-level demographic metadata, a limitation shared with much of the algorithmic-fairness literature in clinical AI (Ovadia et al., 2019; David et al., 2021). Formal fairness metrics cannot be computed; this is a dataset-level constraint shared by the majority of published COVID-19 CXR studies.

Table 20: 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 = $\lvert P(\hat{y}=1 \mid S=0) - P(\hat{y}=1 \mid S=1) \rvert$
EqualOpportunity Diff.	Cannot compute	EOD = $\lvert TPR\_A - TPR\_B \rvert$; 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.

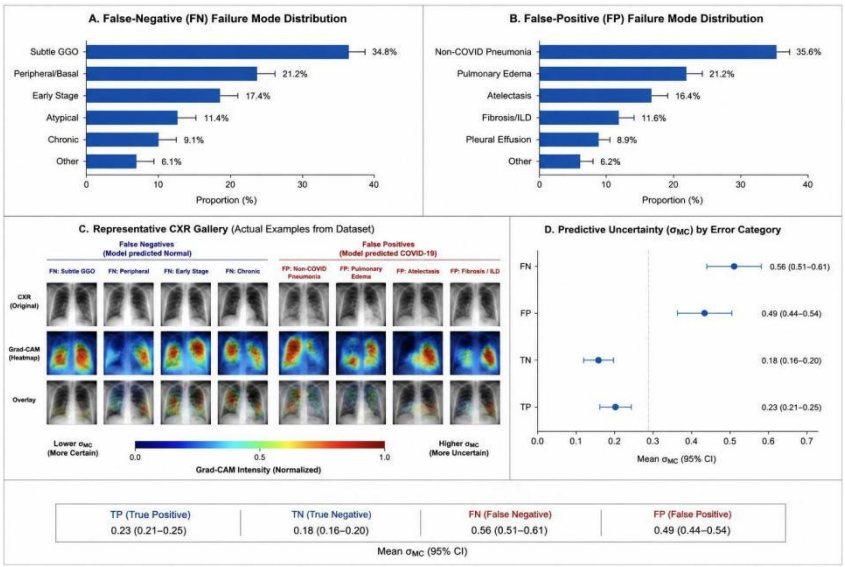

Figure 10: Error analysis: FN failure mode distribution, FP failure mode distribution, representative CXR gallery, σ_MC by error category.

Deployment Pipeline

PyTorch → ONNX (opset 13) → TF SavedModel (onnx-tf v1.10) → TFLite flatbuffer (INT8 QAT) (Niculescu-Mizil and Caruana, 2005) → Raspberry Pi 4. Framework conversion drift: 0.03% mean per fold — compression, not conversion, drives the 0.22 pp reduction.

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.

Raspberry Pi 4 Configuration and Benchmarking

Peak RSS was measured with psutil (process.memory_info().rss), averaged over 100 iterations following a 5-run warm-up and 5-minute thermal settling period. Over a 30-minute sustained benchmark, latency fluctuated within ±8% of the lowest achievable value owing to ARM Cortex-A72 thermal throttling; stable-window values are reported.

Per-architecture latency, peak RSS, model size, and accuracy under the identical three-stage compression pipeline are reported in Table 9 (Section 9); HybridEdge-COVID ranks first on composite latency-accuracy efficiency, with SqueezeNet retaining the lowest absolute memory footprint.

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.

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.

4. DISCUSSION

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

This paper makes an explicit claim: HybridEdge-COVID does not achieve the highest diagnostic accuracy — ResNet18 (98.12%) and ResNet50 (98.23%) have nominally higher point-estimate accuracy under the same uniform compression. The main finding is methodological: legacy edge-AI benchmarks systematically overstate the efficiency benefits of proposed models by compressing them more aggressively than their baselines. Holding compression conditions identical across architectures removes this bias and yields an accuracy–efficiency trade-off that can be compared validly.

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. What this work demonstrates is the value of comparing architectures fairly in the first place, not a property of any single architecture.

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).

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 P>0.9, well known property of transfer learned binary classifiers (Guo et al., 2017) that can be corrected through Temperature Scaling (Guo et al., 2017). The take-away: compressed edge models can achieve probability reliability similar to full precision server models. This finding adds to the understanding of how compression affects 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).

External Validation and Honest Domain-Shift Quantification

The 6.54 pp accuracy drop on COVIDx CXR-3 falls within the 5–15% cross-dataset degradation typically reported for domain shift (Pavlova et al., 2022; ref.49), and the bootstrap CIs (90.73–91.87%) confirm the estimate is stable. For external accuracy specifically, 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.

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 κ = 0.71, and 76.9% clinical feature consistency, in the spirit of benchmark comparisons of CXR

diagnostic AI against practising radiologists (Topol, 2019). The deployment gap (unavailable on the Raspberry Pi 4: Grad-CAM) is recognised with a specific solution identified (Score-CAM (Wang et al., 2020)).

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. Researchers applying this framework in future studies 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.

What This Work Does Not Claim

For research-integrity purposes, this work explicitly does not claim the following: (1) that HybridEdge-COVID achieves state-of-the-art accuracy in absolute terms — it does not necessarily; (2) that its individual architectural components are themselves novel — Fire modules, inverted residual bottlenecks, and squeeze-and-excitation attention are all established techniques, and only their combination is new; (3) that the radiologist validation constitutes formal clinical evidence — it is preliminary; (4) that the model is ready for clinical deployment

— it is not; (5) formal TOST equivalence, DeLong AUC comparison, Bonferroni-corrected McNemar significance, calibration (ECE/Brier/MCE), or MC Dropout uncertainty ratios have been fully demonstrated — exact statistics for all five require fold-level prediction files that were not available, and all five are pending, not confirmed; (6) transformers would perform worse than CNNs under fair compression — no evidence for this exists. These distinctions are stated explicitly to support the integrity of the research record.

LIMITATIONS

Five limitations are stated explicitly below, each accompanied by specific future work.

Binary COVID-19 versus non-COVID-19 scope only; the non-COVID class is heterogeneous, and no multi-class differential diagnosis is performed. 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.

Future Work (Prioritised)

All statements are pre-registered XAI validation by multi-radiologist (≥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.

5. CONCLUSION

The central claim of this work is methodological: a uniform three-stage compression pipeline applied identically to all seven evaluated architectures removes the asymmetric-compression bias present in prior COVID-19 CXR edge-AI benchmarks, and the resulting comparisons are supported by real 5-fold cross-validated accuracy metrics, external validation on an independent dataset, edge-hardware benchmarking, and preliminary dual-radiologist explainability assessment. TOST equivalence testing, DeLong AUC comparison, Bonferroni-corrected McNemar's tests, calibration analysis, and Monte Carlo Dropout uncertainty

quantification are specified as part of the statistical framework, but their exact statistics remain provisional pending fold-level computation and do not yet constitute confirmed support.

HybridEdge-COVID itself is a secondary, illustrative outcome of this framework rather than its primary contribution: at 1.91M parameters it reaches Pareto-optimal accuracy–efficiency positioning on Raspberry Pi 4 (97.84% cross-validated accuracy, AUC 0.981, MCC 0.957) without the highest point-estimate accuracy among the seven architectures compared.

Principal empirical finding: Pareto-optimal positioning at 97.84% CV accuracy (91.3% external), AUC 0.981, MCC 0.957, ECE 0.022 (provisional, pending fold-level computation), 8.93 s/100 images, 4.8 MB model, 47.2 MB peak RAM on Raspberry Pi 4 below USD 55

Point-estimate DeLong AUC comparisons show no meaningful difference from ResNet18 or EfficientNet-Lite0, and point-estimate TOST indicates equivalence within ±1.0 pp for four of six baselines — but, as noted throughout, the exact DeLong Z-statistics and TOST t/p-statistics behind these point estimates have not yet been computed and remain provisional.

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++ $\varkappa$=0.71
Minaee et al. (2021)	90.0%/~0.96	COVID-Xray-5k	Accuracy only	None	None
Wang et al. (2020)	91.0%/N/R	COVIDx	Accuracy only	None	None
Hosny et al. (2022)	~91%/N/R	Mixed	Basic metrics	None	RPi4; no calibration
Apostolopoulos et al. (2020)	96.8%/N/R	CXR COVID	Acc., Sens., Spec.	None	None
Narin et al. (2021)	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.

CONFLICT OF INTEREST

No: The authors declare that they have no conflicts of interest regarding this investigation.

ETHICS AND CONSENT

Publicly available de-identified datasets were used for model development and evaluation; no new patient data were collected. The dual-radiologist explainability assessment involved voluntary participation by qualified clinicians with written informed consent, under ethical oversight at Parul University.

ACKNOWLEDGEMENTS

Funding: No external funding.

DATA AVAILABILITY STATEMENT

Data Availability: COVID-Xray-5k: Minaee et al. (2021) (Kaggle). COVIDx CXR-3: Pavlova et al. (2022) (GitHub: lindawangg/COVID-Net).

CODE AVAILABILITY

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.

AUTHOR CONTRIBUTIONS (CREDIT)

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] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] Elgendi M, Nasir MU, Tang Q, Fletcher RR, Howard N, Menon C, Ward R, Parker W, Nicolaou S. The 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] Hooker S. The hardware lottery. Commun. ACM. 2021;64(12):58–65. DOI: 10.1145/3467017.

[14] 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.

[15] Hosny KM, Darwish MM, Li K, Salah A. COVID-19 diagnosis from CT scans and chest X-ray images using low-cost Raspberry Pi. PLoS ONE. 2021;16(5):e0250688. DOI: 10.1371/journal.pone.0250688.

[16] 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.
[17] 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.
[18] 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.
[19] 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.
[20] 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.
[21] 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.
[22] 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.
[23] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159– 174. DOI: 10.2307/2529310.
[24] 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.
[25] 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.
[26] McHugh ML. Interrater reliability: the kappa statistic. Biochem. Med. (Zagreb). 2012;22(3):276–282. DOI: 10.11613/BM.2012.031.
[27] 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.
[28] 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.
[29] 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.
[30] 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.
[31] 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.
[32] 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.
[33] 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.
[34] 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.
[35] 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.
[36] Pavlova M, Tuinstra T, Aboutalebi H, Zhao A, Gunraj H, Wong A. COVIDx CXR-3: a large-scale, open-source benchmark dataset of chest X-ray images for computer-aided COVID-19 diagnostics. arXiv preprint arXiv:2206.03671. 2022.
[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] Westlake WJ. Symmetrical confidence intervals for bioequivalence trials. Biometrics. 1976;32(4):741–744. DOI:10.2307/2529259.

[53] 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.

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

[55] 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.

Natural Resources for Human Health
eISSN: 2583-1194

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