The Clinical Utility of Compressed ResNetGhost vs. ResNet-50: A Comparative Study for COVID-19 CT Diagnosis

Kadhim Aseel Nadhum
Razak Faculty of Artificial Intelligence, Universiti Teknologi Malaysia, Malaysia
kadhimnadhum@graduate.utm.my (corresponding author)

Suriani Binti Mohd Sam
Razak Faculty of Artificial Intelligence, Universiti Teknologi Malaysia, Malaysia
suriani.kl@utm.my

Sahnius Bt. Usman
Razak Faculty of Artificial Intelligence, Universiti Teknologi Malaysia, Malaysia
sahnius.kl@utm.my

Received: 1 March 2026 | Revised: 9 April 2026, 16 April 2026, and 21 April 2026 | Accepted: 23 April 2026

 DOI: https://doi.org/10.48084/etasr.18468

ABSTRACT

ResNetGhost constitutes a stage-selective optimization of ResNet-50, strategically integrating Ghost modules into Stages 3–4 to minimize computational redundancy while maintaining diagnostic fidelity in classifying COVID-19 CT severity. This architecture was validated on a multi-center cohort of 577 patients to address potential domain shifts through a rigorous patient-level split. In an independent test set (n = 164), ResNetGhost demonstrated superior discriminative power with an ROC-AUC of 0.991 and a verified accuracy of 97.56% (160/164 cases). Furthermore, this method effectively resolved systematic over-confidence, achieving a high-fidelity ECE of 0.0175. McNemar's test (p = 0.1573) confirmed that the 62.29% parameter reduction and 30% FLOPs decrease did not induce statistically significant degradation in diagnostic integrity compared to the baseline. Consequently, ResNetGhost offers a robust, well-calibrated solution optimized for resource-constrained clinical environments.

Keywords-COVID-19; chest CT; severity classification; ResNet-50; Ghost modules; model compression

I. INTRODUCTION

The deployment of Deep Learning (DL) models in resource-constrained clinical settings has underscored the limitations of standard architectures [1]. Although ResNet-50 remains a benchmark for diagnostic accuracy, its high parametric density creates a critical hardware bottleneck, often hindering real-time deployment in hospital workflows. Although model compression through pruning or uniform lightweight modules has been explored, these methods frequently induce a compromise between architectural efficiency and diagnostic rigor [2]. A significant challenge remains the "reliability gap," where models achieve high technical accuracy but fail to demonstrate clinical safety and net benefit. The clinical utility of DL in radiology is fundamentally constrained by the high-dimensional nature of CT imaging, which requires significant GPU throughput. Although backbones such as ResNet-50 offer superior feature extraction [3], their static architectural density is often suboptimal for infrastructure-sensitive environments [4]. Recent surveys highlight a transition toward "resource-aware" designs; however, much of the early literature remains confined to slice-level analysis on sanitized datasets, thereby ignoring the patient-level contextual complexity required for clinical workflows [5, 6].

Traditional post-hoc strategies, particularly pruning and quantization, are increasingly scrutinized for their "destructive" impact on diagnostic integrity [7]. Aggressive weight removal can inadvertently destabilize probability estimates and compromise the model's sensitivity to subtle pathogenic textures [8]. In contrast, "constructive" alternatives, such as the Ghost module, redefine the efficiency-accuracy trade-off by synthesizing intrinsic feature maps through cheap linear transformations rather than indiscriminate pruning, preserving the high-resolution richness of CT data while effectively reducing the parametric burden [9]. Although specialized medical architectures such as MSGU-Net have addressed some efficiency concerns [10], a persistent limitation remains the application of compression modules uniformly across the network [11], which ignores the hierarchical sensitivity of

CNNs where early-stage features are vital for pathological signal preservation [12]. Building on the architectural optimization and technical benchmarking established in [13], where the ResNetGhost architecture was rigorously evaluated against various state-of-the-art lightweight models (e.g., MobileNetV2 and EfficientNet-B0) and traditional L1-norm pruning, this study advances to a clinical examination of the model. Although the technical superiority of the Ghost stage-selective module in terms of accuracy and parameter reduction was verified in [13], its probabilistic reliability and clinical net benefit compared to the ResNet-50 baseline remain the main objectives of this study. By utilizing McNemar's test and Decision Curve Analysis (DCA), this research shifts the evaluation from scalar metrics to a comprehensive validation of clinical net benefit in a multi-center clinical setting.

II. RESNET50 MODEL

This study used the ResNet50 model, which won the 2015 ILSVRC ImageNet competition, as the baseline framework. ResNet50 is based on transfer learning techniques and efficiently processes biological images using fewer data and lower computational costs. The model consists of 50 layers, including one MaxPool layer [14], one Average Pool layer, and 48 convolution layers. ResNet50 relies on residual learning, which helps solve the vanishing gradient problem in deep networks. In the residual blocks, the network learns the differences between the inputs and outputs instead of learning the true output directly, making the learning process easier and improving performance. The model also allows the reuse of activation functions from previous layers [15], as shown in Figure 1.

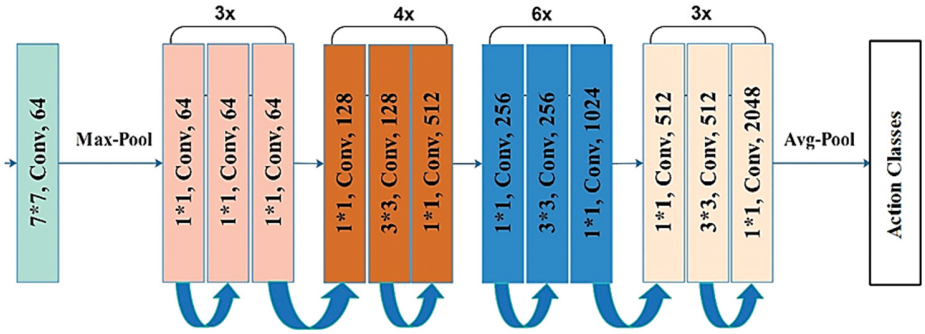

Fig. 1. ResNet50 architecture.

III. METHODOLOGY

A. Dataset, Ethics, and Labeling

1) Dataset and Clinical Labeling

This retrospective study was approved by the Iraqi Ministry of Health (Approval ID: 69/2025) and utilized 19,309 axial CT slices from 577 RT-PCR-confirmed COVID-19 patients (2020–2023). Data were acquired from three centers using Siemens, GE, and Philips scanners. Clinical ground truth was established by two board-certified radiologists using the CT Severity Score (CT-SS). Cases were categorized as Mild (CT-SS $\leq$ 19) or Severe (CT-SS > 19), with high inter-rater reliability (Cohen's $\kappa = 0.78$), as summarized in Table I.

TABLE I. PATIENT COHORT AND DATASET PARTITIONING

Split purpose	Patients (n)	Patient Balance (Severe/Mild)	Slices (n)	Slice Balance (Severe/Mild)
Training & Validation	379	188 / 191	14494	7114 / 7380
Calibration	34	18 / 16	1,299	690 / 609
Test (Held-Out)	164	89 / 75	3,516	1,658 / 1,858
Total Combined Cohort	577	295 / 282	19,309	9462 / 9847

2) Center-Based Data Partitioning

A site-specific splitting strategy was implemented to evaluate clinical transportability. The Development Cohort (379 patients; 14,494 slices) and the Calibration Set (34 patients; 1,299 slices) were sourced from the Babylon and Al-Hashimiyah Hospitals. In contrast, Al-Kifl General Hospital (164 patients; 3,516 slices) was strictly reserved as an External Test Set. This ensures that the comparative evaluation of ResNet-50 and ResNetGhost is performed on data from an unseen clinical environment and different scanner protocols.

3) Patient-Level Inference and Calibration

A deterministic Mean Probability Aggregation (MPA) protocol was adopted to derive patient-level scores from individual slice predictions. The final clinical decision was determined by averaging the probabilities of three representative slices per patient. To ensure probabilistic reliability, the decision threshold was optimized via Platt Scaling on the independent calibration set (n=34) to maximize the F1-score. This optimized threshold was then fixed for the final evaluation of both architectures on the external test set.

4) Comparative Evaluation and Reporting

The diagnostic performance of the proposed ResNetGhost was benchmarked against the standard ResNet-50 using the independent test set (n=164). Reported metrics include ROC-AUC, Sensitivity, Specificity, and Expected Calibration Error (ECE). Statistical stability was verified using 95% Confidence Intervals (CIs) derived from 1000-sample bootstrapping.

B. Preprocessing

A standardized preprocessing pipeline was implemented to ensure architectural compatibility with the ResNet backbone

and stabilize the training convergence. Each axial CT slice was resized to 224×224 pixels utilizing bilinear interpolation with anti-aliasing to prevent aliasing artifacts during downsampling. Images were converted into a 3-channel RGB representation and scaled to a floating-point range of [0, 1]. Finally, global normalization was applied using the mean (0.485, 0.456, 0.406) and standard deviation (0.229, 0.224, 0.225). This step aligns the input distribution with the pre-trained weights, facilitating effective transfer learning from the initial layers. To mitigate the risk of overfitting and ensure that the model remains invariant to the acquisition variance inherent in the multi-vendor Iraqi cohort (Siemens, GE, and Philips), a stochastic augmentation strategy was applied exclusively to the training set. This pipeline introduces synthetic diversity without altering the underlying pathological labels:

- Geometric Invariance: Random horizontal flipping (p = 0.5) and affine transformations, including rotational perturbations up to ±15°, spatial translations within ±10%, and isotropic scaling between 0.9 and 1.1.
- Intensity Robustness: To simulate variations in scanner contrast and brightness, a color jittering filter was applied with shifts in brightness (±0.2), contrast (±0.2), saturation (±0.1), and hue (±0.05). These perturbations ensure that the ResNetGhost variant learns to identify Ground-Glass Opacity (GGO) patterns rather than hardware-specific artifacts [16].

Calibration analysis was performed to ensure the probabilistic honesty of the ResNetGhost outputs, reliability was evaluated through Reliability Diagrams, Expected Calibration Error (ECE), and the Brier Score. The calibration framework, comparing Platt Scaling and Isotonic Regression, was strictly parameterized using the dedicated Calibration Set (n = 34) to prevent overfitting the model's confidence to the final benchmarking cohort. Platt Scaling was selected to map raw logit scores into well-calibrated probabilities. These frozen parameters were then applied to the independent Test Set (n = 164), where the ECE was quantified using M = 10 equally-spaced bins. To establish the statistical stability of the calibration metrics, 95% Confidence Intervals (CIs) were generated from 1,000 patient-wise stratified bootstrap resamples.

IV. IMPLEMENTATION DETAILS

The models were implemented using Python 3.10 and PyTorch 2.0. Image preprocessing and augmentation pipelines were developed using the OpenCV and Albumentations libraries to ensure efficient real-time transformation. Evaluation metrics (Accuracy, Sensitivity, Specificity, ROC-AUC) were computed using Scikit-learn, while calibration curves and decision analyses were visualized using Matplotlib and Seaborn. All experiments were executed on the Google Colab Pro+ platform utilizing a single NVIDIA Tesla T4 GPU (16 GB VRAM). Training was performed with a batch size of 32, using the AdamW optimizer (learning rate = 3e-4, weight decay = 1e-4) and a OneCycleLR scheduler for 30 epochs.

V. COMPARATIVE MODEL PERFORMANCE

The diagnostic efficacy of the proposed ResNetGhost was evaluated against the ResNet-50 baseline through a granular analysis of their respective confusion matrices. This head-to-head comparison on the held-out test set (n = 164) reveals a significant shift in diagnostic reliability. This architectural choice is further supported by the previous comparative study [13], which established that Ghost modules provide a superior balance between computational overhead and diagnostic precision compared to pruning and knowledge distillation techniques.

A. Performance of ResNet-50

The baseline architecture correctly identified 158 out of 164 subjects, yielding an aggregate accuracy of 96.34%. Despite showing a strong specificity (97.33%), the model exhibited a critical diagnostic gap with 4 False Negatives (FN). This latent vulnerability to underdiagnosing severe pathology poses a risk in high-stakes clinical triage.

B. Performance of ResNetGhost Model

The ResNetGhost model correctly identified 160 out of 164 subjects, yielding an aggregate accuracy of 97.56%. It achieved strong specificity and sensitivity with only two False Negatives (FN) and two False Positives (FP).

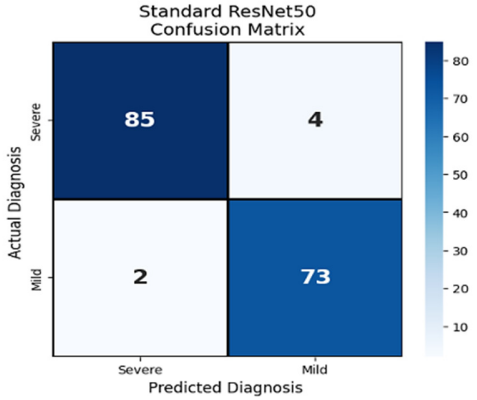

Fig. 2. Confusion matrix of the ResNet50.

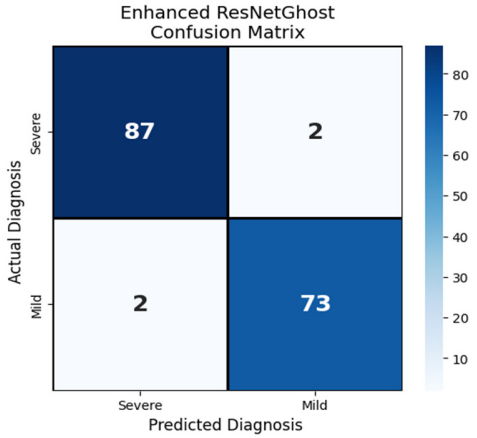

Fig. 3. Confusion matrix of the ResNetGhost model.

C. Statistical Comparison with McNemar's Test

To evaluate the statistical equivalence between ResNetGhost and the ResNet-50 baseline, a McNemar's test was conducted on the independent test set (n = 164). This paired-sample approach focused on discordant pairs to identify shifts in predictive behavior. The analysis resulted in a Chi-squared value of 2.0 (df = 1) with a p-value of 0.1573. Since the p-value exceeds the significance threshold (0.05), the null hypothesis of marginal homogeneity is sustained [17]. This indicates that the 62.29% reduction in parametric density does not lead to a statistically significant difference in diagnostic integrity. Thus, ResNetGhost achieves clinical parity with the standard architecture through performance-preserved compression [18].

VI. DECISION-CURVE ANALYSIS (DCA)

To bridge the gap between statistical probability and practical clinical deployment, DCA was implemented to evaluate the net benefit of the ResNetGhost architecture. Unlike conventional accuracy-centric metrics, DCA provides a multidimensional quantification of the model's utility across a spectrum of threshold probabilities (p_t). This framework benchmarks the proposed model against the default clinical strategies of 'Treat-All' (assuming all patients have Severe pathology) and 'Treat-None' (assuming all cases are Mild). By calculating the clinical dividend on the independent Test Set (n = 164), this analysis weighs the value of True Positive (TP) identifications against the resource-drain caused by FP interventions, utilizing the following standardized formula [19]:

$$NB(p_t) = TP/N - FP/N \cdot p_t/1 - p_t$$

where N is the total number of patients (164). Substituting the empirical results at a threshold of p_t = 0.05, the model achieved a high net benefit of 0.5299, reaching an operational efficiency of 99.7% compared to a perfect classifier. The DCA curve in Figure 4 demonstrates that the ResNetGhost model maintains a superior trajectory over default strategies across an extensive threshold range (0.05–0.95), confirming its high clinical reliability. The Receiver Operating Characteristic (ROC) curve, shown in Figure 5, validates the model's superior discriminatory power, as it converges toward the upper-left coordinate, signifying that a high True Positive Rate (TPR) is achieved alongside a near-zero False Positive Rate (FPR). The resulting Area Under the Curve (AUC) of 0.991 demonstrates the model's robust capability in distinguishing between 'Mild' and 'Severe' patient classifications.

Probabilistic reliability was significantly enhanced in the proposed architecture compared to the baseline. While the standard ResNet-50 exhibited a notable calibration gap with an ECE of 0.0313, the ResNetGhost achieved a superior ECE of 0.0175 and a Brier Score of 0.0226, as illustrated in Figure 6. This reduction in the calibration error ensures that the model's confidence levels are statistically congruent with actual diagnostic outcomes, effectively neutralizing the over-confidence bias typical of high-parameter backbones [20].

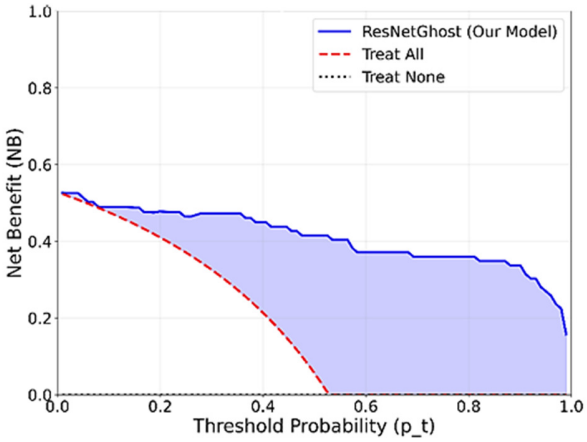

Fig. 4. Confusion Decision Curve Analysis (DCA) for the Held-Out Test Set (n=164).

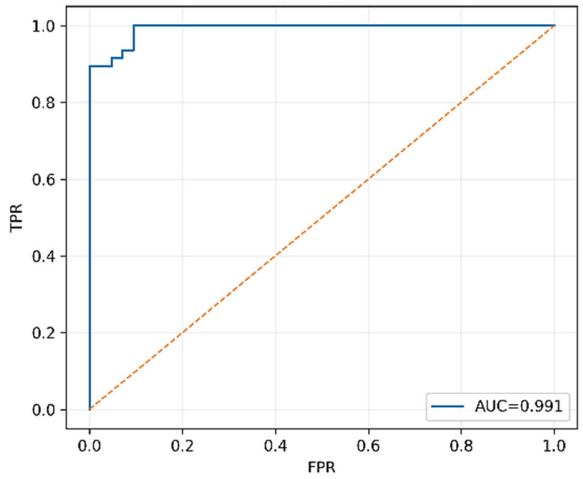

Fig. 5. Patient-level ROC curve (Held-Out Test Set, n=164).

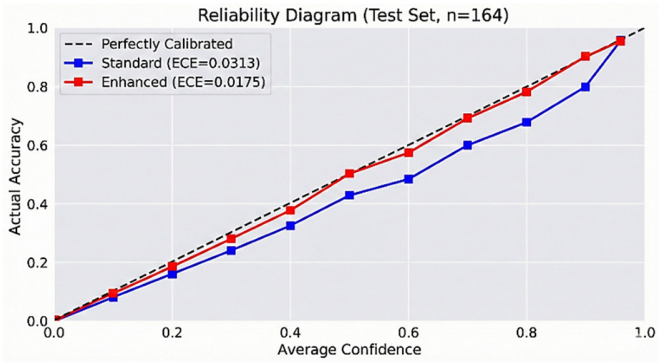

Fig. 6. Patient-Level Calibration Plot (Test Set, n=164).

VII. EFFICIENCY, LATENCY, AND STATISTICAL VALIDATION

The diagnostic efficacy of the ResNetGhost, previously established as a high-efficiency architecture, was rigorously evaluated against the ResNet-50 baseline. As shown in Table II, the optimized model achieves a diagnostic Accuracy of

97.56%, marking a distinct improvement over the 96.34% baseline performance. This increase in diagnostic accuracy is reinforced by confirmed statistical parity (p = 0.1573), establishing that the optimized model does not induce any significant diagnostic degradation.

The clinical and operational advantages of the ResNetGhost (Optimization) are evidenced through a comprehensive resource audit:

- Parametric Decimation: A 62.29% reduction in parametric density, transitioning from the baseline's 25.6 M to a streamlined 9.7 M.
- Computational Throughput: A 30% decrease in FLOPs (from 4.13 G to 2.91 G), optimizing the model for high-speed processing.
- Inference Latency: A significant reduction in execution time to 10.11 ms on the NVIDIA Tesla T4 GPU, outperforming the baseline's 16.54 ms.

The reduction in model size from 98 MB to 38 MB confirms that ResNetGhost overcomes the complexity-accuracy bottleneck. This efficiency stems directly from a 62.29% drop in parametric density, where the replacement of redundant convolutions with stage-selective Ghost modules eliminated nearly 15.9 million non-essential parameters. Such massive weight reduction removes the prohibitive computational overhead of the original ResNet-50. These findings validate the model's readiness for edge deployment in resource-constrained Iraqi clinical centers, providing a lean and medically reliable diagnostic tool.

TABLE II. CLINICAL AND TECHNICAL VALIDATION OF THE OPTIMIZED FRAMEWORK

Metric	ResNet-50 (Baseline)	ResNetGhost (Optimization)
Accuracy (%)	96.34	97.56
ROC-AUC	0.982	0.991
Parameters (M)	25.6	9.7
Model Size (MB)	98	38
Training Time (m)	100	60
Latency (ms) (GPU)	16.54	10.11
Latency (ms) (CPU)	51.12	20.45
ECE (Calibration)	0.0313	0.0175

VIII. CONCLUSION

This research provided a rigorous clinical and statistical evaluation of ResNetGhost, a stage-selective lightweight architecture designed to address the computation-precision trade-off in medical imaging. By transitioning from architectural development to an extensive diagnostic validation framework, this study validated the model's performance on a robust, multi-center Iraqi cohort. On an independent, strictly held-out test set (n = 164), ResNetGhost demonstrated exceptional generalization, achieving a diagnostic Accuracy of 97.56%, Sensitivity of 97.75%, and a superior ROC-AUC of 0.9910. The implementation of a dedicated calibration protocol resolved probabilistic reliability issues, yielding a remarkably low ECE of 0.0175 and a Brier score of 0.0226. These metrics confirm that the optimized model is not only precise but also clinically safe for real-time decision support.

The primary contribution of this work is the empirical verification of performance-preserved compression. The results demonstrated that the 62.29% reduction in parametric density (from 25.6M to 9.7M) did not induce diagnostic degradation, as confirmed by McNemar's test (p = 0.1573). Furthermore, DCA verified a high clinical net benefit of 0.5299, reaching an operational efficiency of 99.7% compared to a perfect classifier. These results establish ResNetGhost as a robust, interpretable, and computationally efficient engine, uniquely suited for deployment on storage-constrained edge devices within the resource-sensitive Iraqi healthcare infrastructure.

Limitations include the regional Iraqi focus and hardware-specific latency. Future research will focus on multi-national validation and edge-computing deployment (e.g., NVIDIA Jetson Nano) in resource-constrained clinical settings.

ABBREVIATIONS

AdamW: Adaptive Moment Estimation with Decoupled Weight Decay; AUC: Area Under the Curve; BN: Bottleneck Unit; CI: Confidence Interval; CNN: Convolutional Neural Network; CT: Computed Tomography; DCA: Decision Curve Analysis; df: Degrees of Freedom; ECE: Expected Calibration Error; FLOPs: Floating Point Operations; FN: False Negative; FP: False Positive; GGO: Ground-Glass Opacity; NB: Net Benefit; OneCycleLR: One Cycle Learning Rate Scheduler; PR–AUC: Precision–Recall AUC; RGB: Red, Green, Blue; ROC: Receiver Operating Characteristic; RT-PCR: Reverse Transcription Polymerase Chain Reaction; TN: True Negative; TP: True Positive; TPR/FPR: True/False Positive Rate; TTA: Test-Time Augmentation; X2: Chi-squared statistic.

DECLARATION OF COMPETING INTERESTS

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

ACKNOWLEDGMENT

This study was approved by the Iraqi Ministry of Health (Approval ID: 69/2025). The authors would also like to thank Universiti Teknologi Malaysia (UTM) for providing research facilities and support. They also express their gratitude to the medical staff at the clinical facilities in Babylon, Iraq, for their invaluable assistance in the data collection process.

DATA AVAILABILITY

The dataset utilized in this study is private and belongs to clinical facilities in Babylon, Iraq. Due to patient confidentiality and institutional privacy policies, the raw data are not publicly available but may be provided by the corresponding author upon reasonable request.

REFERENCES

[1] G. Rasul *et al.*, "Socio-Economic Implications of COVID-19 Pandemic in South Asia: Emerging Risks and Growing Challenges," *Frontiers in Sociology*, vol. 6, Feb. 2021, https://doi.org/10.3389/fsoc.2021.629693.

[2] K. A. Nadhum, S. M. Sam, and S. Usman, "Prediction Model Using Deep Learning for Lung Illness Severity Among Covid-19 Patients in Iraq," in *2024 5th International Conference on Smart Sensors and Application (ICSSA)*, Sept. 2024, pp. 1–6, https://doi.org/10.1109/ICSSA62312.2024.10788660.

[3] M. Chhabra and R. Kumar, "An Efficient ResNet-50 based Intelligent Deep Learning Model to Predict Pneumonia from Medical Images," in *2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, Apr. 2022, pp. 1714–1721, https://doi.org/10.1109/ICSCDS53736.2022.9760995.

[4] D. Suganya. and R. Kalpana., "Prognosticating various acute covid lung disorders from COVID-19 patient using chest CT Images," *Engineering Applications of Artificial Intelligence*, vol. 119, Mar. 2023, Art. no. 105820, https://doi.org/10.1016/j.engappai.2023.105820.

[5] M. A. Mezher, S. B. Alrifai, and W. M. Raoof, "Analysis of Proinflammatory Cytokines in COVID-19 Patients in Baghdad, Iraq," *Archives of Razi Institute*, vol. 78, no. 1, pp. 305–313, Feb. 2023, https://doi.org/10.22092/ARI.2022.359356.2411.

[6] J. F. Abdulkareem and H. K. Aljobouri, "Chest CT images analysis with deep learning algorithms for COVID-19 diagnostic for Iraqi center," *AIP Conference Proceedings*, vol. 2414, no. 1, Feb. 2023, Art. no. 060004, https://doi.org/10.1063/5.0117655.

[7] T. Liang, J. Glossner, L. Wang, S. Shi, and X. Zhang, "Pruning and quantization for deep neural network acceleration: A survey," *Neurocomputing*, vol. 461, pp. 370–403, Oct. 2021, https://doi.org/10.1016/j.neucom.2021.07.045.

[8] Y. He and L. Xiao, "Structured Pruning for Deep Convolutional Neural Networks: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 5, pp. 2900–2919, Feb. 2024, https://doi.org/10.1109/TPAMI.2023.3334614.

[9] X. Jiang, X. Bian, and C. Guo, "Ghost-Stereo: GhostNet-based Cost Volume Enhancement and Aggregation for Stereo Matching Networks." arXiv, May 23, 2024, https://doi.org/10.48550/arXiv.2405.14520.

[10] H. Hussain, P. S. Tamizharasan, and P. K. Yadav, "LCRM: Layer-Wise Complexity Reduction Method for CNN Model Optimization on End Devices," *IEEE Access*, vol. 11, pp. 66838–66857, 2023, https://doi.org/10.1109/ACCESS.2023.3290620.

[11] Z. Wang and T. Li, "A Lightweight CNN Model Based on GhostNet," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–12, July 2022, https://doi.org/10.1155/2022/8396550.

[12] S. H. Khan, "COVID-19 Detection and Analysis From Lung CT Images using Novel Channel Boosted CNNs." arXiv, Sept. 01, 2022, https://doi.org/10.48550/arXiv.2209.10963.

[13] K. A. Nadhum, S. B. M. Sam, and S. B. Usman, "Optimizing ResNet50 for Medical Image Classification: A Comparative Study of Ghost Modules, Pruning, and Knowledge Distillation," *Engineering, Technology & Applied Science Research*, vol. 15, no. 6, pp. 28544–28549, Dec. 2025, https://doi.org/10.48084/etasr.13722.

[14] Y. Fu, Y. Lei, T. Wang, W. J. Curran, T. Liu, and X. Yang, "Deep learning in medical image registration: a review," *Physics in Medicine & Biology*, vol. 65, no. 20, Oct. 2020, Art. no. 20TR01, https://doi.org/10.1088/1361-6560/ab843e.

[15] Y. Tang, K. Han, J. Guo, C. Xu, C. Xu, and Y. Wang, "GhostNetV2: Enhance Cheap Operation with Long-Range Attention," *Advances in Neural Information Processing Systems*, vol. 35, pp. 9969–9982, Dec. 2022.

[16] C. Shorten, T. M. Khoshgoftaar, and B. Furht, "Text Data Augmentation for Deep Learning," *Journal of Big Data*, vol. 8, no. 1, July 2021, Art. no. 101, https://doi.org/10.1186/s40537-021-00492-0.

[17] O. A. Adedokun and W. D. Burgess, "Analysis of Paired Dichotomous Data: A Gentle Introduction to the McNemar Test in SPSS," *Journal of MultiDisciplinary Evaluation*, vol. 8, no. 17, pp. 125–131, Jan. 2012, https://doi.org/10.56645/jmde.v8i17.336.

[18] H. K. Hamarashid, "Utilizing Statistical Tests for Comparing Machine Learning Algorithms," *Kurdistan Journal of Applied Research*, vol. 6, no. 1, pp. 69–74, July 2021, https://doi.org/10.24017/science.2021.1.8.

[19] Z. Zhang *et al.*, "Decision curve analysis: a technical note," *Annals of Translational Medicine*, vol. 6, no. 15, Aug. 2018, Art. no. 308, https://doi.org/10.21037/atm.2018.07.02.

[20] R. Vasilev and A. D'yakonov, "Calibration of Neural Networks." arXiv, 2023, https://doi.org/10.48550/ARXIV.2303.10761.