Optimizing ResNet50 for Medical Image Classification: A Comparative Study of Ghost Modules, Pruning, and Knowledge Distillation

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: 29 July 2025 | Revised: 11 August 2025 | Accepted: 19 August 2025

 DOI: https://doi.org/10.48084/etasr.13722

ABSTRACT

Computed Tomography (CT) imaging plays a vital role in assessing lung disease severity in COVID-19 patients. However, deploying deep learning models such as ResNet50 for automated severity classification (mild vs. severe) remains challenging in resource-constrained medical environments due to their high computational demands. This study presents a comparative analysis of three optimization techniques, namely ghost modules, pruning, and knowledge distillation, to enhance the efficiency of the ResNet50 model while maintaining high diagnostic accuracy. The optimized models were trained and evaluated using a real-world dataset comprising 817 CT images collected from public hospitals in Babylon Province, Iraq. Experimental work indicates that a GhostNet-augmented ResNet50 improved to a record 98.4% accuracy, a knowledge distillation-based variant achieved 93%, and a pruned ResNet50 variant achieved 91%. These results indicate a better balance between performance and computation achieved by ghost modules, supporting their particular relevance to real-time applications within resource-constrained medical systems.

Keywords-COVID-19; CT classification; ResNet50 optimization; ghost modules

I. INTRODUCTION

Computed Tomography (CT) scans are among the key diagnostic instruments within medicine, giving precise visual representations of internal body structures. CT scans are broadly employed for the diagnosis of pulmonary diseases, including pneumonia, tumors, and, predominantly, COVID-19. Rapid and accurate CT scan interpretation determines the success or failure of healthcare provision. Nevertheless, manual evaluation remains time-consuming and susceptible to human error, prompting research into the development of sophisticated intelligent systems to automate the process [1].

With the advancement of deep learning technologies, Convolutional Neural Networks (CNNs) have emerged as some of the most effective image classification techniques. Among these, the ResNet50 model has demonstrated high capability in processing radiological images and achieving high precision in classification, due to its use of residual connections to train deep networks without a significant loss of information [2]. However, ResNet50 is quite large, with approximately 25.6 million parameters, which makes it less ideal for deployment in resource-constrained medical or field environments [3].

This work focuses on improving and fine-tuning the ResNet50 model to make it faster and lighter while still maintaining high-classification accuracy to better fit real-world applications in medicine. The novelty in this work is that it is based on real CT scan images obtained from two public hospitals within Babylon Province, Iraq, offering real-world practical relevance within the region's healthcare landscape. Three compression algorithms are employed to make models more efficient and reduce computational overhead:

1. Ghost modules produce feature maps with lightweight convolutions rather than standard ones, minimizing the number of parameters but preserving the representational capability [4].

2. Pruning: A procedure used to eliminate less significant filters or units within a trained model, effectively reducing model size with a tolerable loss of precision [5].
3. Knowledge Distillation (KD): A smart approach in which a lightweight "student" model is trained using knowledge transferred from a larger "teacher" model, enabling the student to achieve comparable performance in a more compact form [6].

This study aimed to analyze and compare the effectiveness of ghost modules, pruning, and knowledge distillation in enhancing the ResNet50 model for classifying CT scan images of COVID-19 patients, using real-world local data collected from hospitals in Babylon, Iraq.

A. Research Contribution

The core contribution of this work lies in the following:

1. Develops lightweight and highly efficient models that can be deployed in resource-constrained medical environments in Iraq.
2. Provides a comparative experimental analysis of three leading parameter reduction techniques.
3. Uses real-world data collected from hospitals in Babylon, thereby enhancing the practical relevance of the model and paving the way for immediate clinical applications.

B. Previous Studies

Recently, ResNet50 has been popularly used for medical image classification due to its strong deep feature extraction abilities. Several researchers have used it during diagnostic applications concerning breast tumors, lung diseases, and especially CT scan images regarding COVID-19, proving to be capable of discriminating between infected and uninfected individuals and classifying disease severity [7, 8]. However, its large model size and number of parameters render it unsuitable for deployment in computationally constrained environments, eliciting a necessity to work with optimization methods that maintain accuracy while improving efficiency [7, 9, 10].

To this end, GhostNet has been a very successful alternative to classical models. It employs lightweight operations to produce feature maps rather than employing dense convolutional filters, significantly reducing both parameter count and inference time. GhostNet has been successfully applied in several medical tasks, such as classifying X-ray and MRI images, showing promising results, especially compared to heavier models such as DenseNet or Inception. Nevertheless, its application to CT imaging remains relatively underexplored [4, 11].

Several studies have investigated pruning techniques to reduce the size of deep models by eliminating redundant neurons or channels after training. Pruning has been applied effectively to ResNet models to reduce parameter counts and accelerate inference in medical applications. However, most previous studies focused on conventional or environmental images, with limited and fragmented exploration in medical imaging tasks [12]. In [13], the STAMP algorithm was used to compress and enhance the performance of UNet for medical image segmentation, achieving over 85% reduction in parameters while maintaining accuracy with limited annotated data. However, the model was not evaluated in real clinical settings, and issues of interpretability and practical deployment were not sufficiently addressed.

KD has also emerged as a compelling approach for transferring knowledge from a large teacher model to a smaller student model. It has proven effective in preserving model performance despite a reduced size and has been employed in tasks such as diabetic retinopathy detection and cardiac disease classification from ECG signals. However, the application of KD in CT image classification, particularly with ResNet-based models, remains limited and is often studied in isolation from other techniques such as GhostNet or pruning [14]. In [15], an efficient model was proposed for COVID-19 diagnosis, using KD from VGG19 and ResNet50v2 to MobileNetV2, achieving 98.8% accuracy with a 95.3% reduction in parameters, making it suitable for low-resource devices. However, it relied on conventional architectures without structural innovation, used limited non-clinical datasets, and lacked attention to model interpretability and real-world clinical evaluation.

Although individual studies have explored GhostNet, pruning, or KD within the medical domain, the literature lacks a comprehensive comparative analysis that integrates all three techniques within a unified framework and applies them to the same real-world CT dataset. This lack of direct comparison hinders informed decision-making when designing lightweight models for practical deployment, especially in resource-limited clinical settings.

C. ResNet50 Model

The ResNet50 model, which won the 2015 ILSVRC ImageNet competition, was used as the base framework in this study. 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 max pooling layer, one average pooling layer, and 48 convolution layers. ResNet50 relies on residual learning, which helps solve the problem of vanishing gradients in deep networks [10]. In residual blocks, the network learns the differences between inputs and outputs instead of directly learning the true output, making the learning process easier and improving performance. The model also allows reuse of activation functions from previous layers [16-19].

II. METHODOLOGY

A. Model Design

This study performed a comprehensive and systematic comparison between three modern techniques to improve the performance of the deep ResNet50 model for CT image classification. The goal was to reduce the number of parameters and computational load while maintaining high classification accuracy. The three techniques include the integration of ghost modules (GhostNet), structured pruning, and KD. The significance of this comparison stems from the growing need to deploy high-efficiency AI models in resource-limited medical environments, where the large size and computational

complexity of deep models, such as ResNet50, present a real challenge to practical deployment.

In the first model, ResNet50-GhostNet, ghost modules were integrated into the original ResNet50 architecture, specifically within Stage 3 and Stage 4, as these stages are the most computationally intensive due to their high filter counts. Ghost modules work by generating additional feature maps through cheap operations, rather than relying solely on conventional convolutional layers, thereby significantly reducing the number of parameters. In each residual block of these two stages, the final convolutional layer was replaced with a ghost module, as illustrated in the architectural diagram in Figure 1. This structural modification improved computational efficiency without compromising the visual representation quality or classification accuracy, forming the architectural novelty at the core of this study

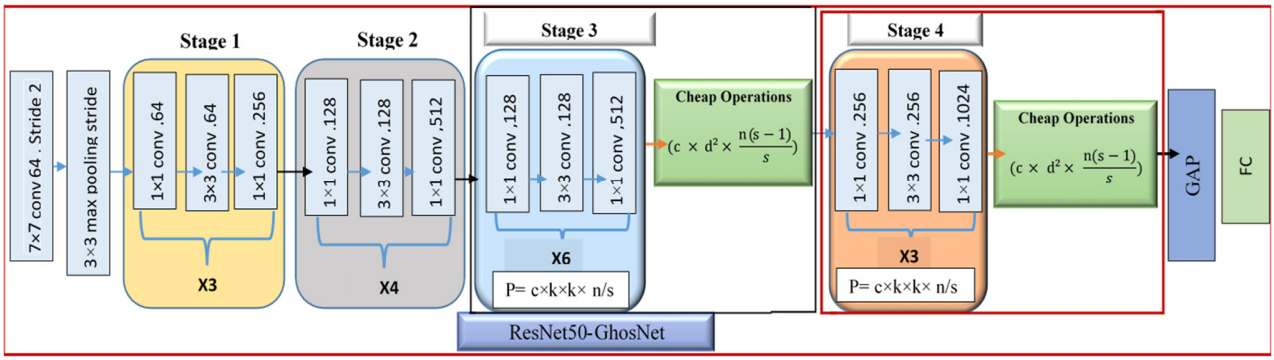

Fig. 1. Architecture of a ResNet50-GhostNet model.

The second model, ResNet50-Pruned, adopted a non-architectural approach to reduce computational density through the removal of less important components from a pretrained model. Iterative structured pruning was applied using the L1-norm criterion to identify and eliminate weak filters and channels. The pruning process was carried out at varying levels, ranging from 30% to 50%, depending on layer sensitivity, followed by gradual fine-tuning to recover model performance. This strategy effectively reduced the model size and inference time while maintaining acceptable accuracy when tested on CT data.

The third model, ResNet50-KD, used KD to increase efficiency without altering the original model architecture. In this approach, a lightweight student network based on ResNet18 was trained with the soft outputs provided by the larger teacher model (ResNet50), with a combination of cross-entropy loss and distillation loss. Soft targets were used with a softmax temperature of 4 to enhance knowledge transfer. This method resulted in a compact model that preserved performance levels comparable to the original ResNet50, with a substantial reduction in the number of parameters and inference time, making it particularly ideal for clinical applications that demand speed and efficiency.

All three models were evaluated on the same CT scan COVID-19 patient data, derived from public hospitals in Babylon Province, Iraq. The dataset consisted of 817 images for training and evaluation, and a distinct set of 164 images for final testing. Consistency was maintained across the data distribution and preprocessing techniques, with data augmentation applied only to the training and validation sets, while the test set was left unchanged to avoid biased evaluation. This uniform experimental setup provides a fair and common basis to compare the relative effectiveness of each approach to improve ResNet50's performance with regard to accuracy, parameter count, model size, and inference speed.

The data were preprocessed with a standardized pipeline before being input into the three models. This included resizing the CT images and transforming them to grayscale to reduce visual complexity. A set of data augmentation techniques was applied to the training and validation sets, including random rotation, zooming, cropping, and brightness changes. These methods were employed to increase data diversity and minimize the risk of overfitting. In contrast, the test set remained untouched, with no augmentation, to provide an impartial assessment. In addition, all experimental settings were kept consistent across the three models, being similar in terms of input data and processing conditions.

B. Evaluation Criteria

The performance of the three models was assessed by adopting a dual-analysis technique, including both qualitative and quantitative analysis to identify the highest-performing model in terms of diagnostic precision and computational cost. Figure 2 illustrates the accuracy and loss curves over 30 training epochs for the baseline ResNet50 model and its three optimized variants utilizing GhostNet, pruning, and KD techniques. The left image represents the evolution of classification accuracy on both the training and validation sets, while the right one depicts the variation in loss over time. The ResNet50-GhostNet model exhibited the most stable performance, achieving higher accuracy and a consistently lower loss, particularly on the validation set. In contrast, the KD-based model demonstrated noticeable fluctuations in accuracy during the early training stages. Meanwhile, the pruned ResNet50 maintained balanced performance, although slightly inferior to the GhostNet-enhanced version.

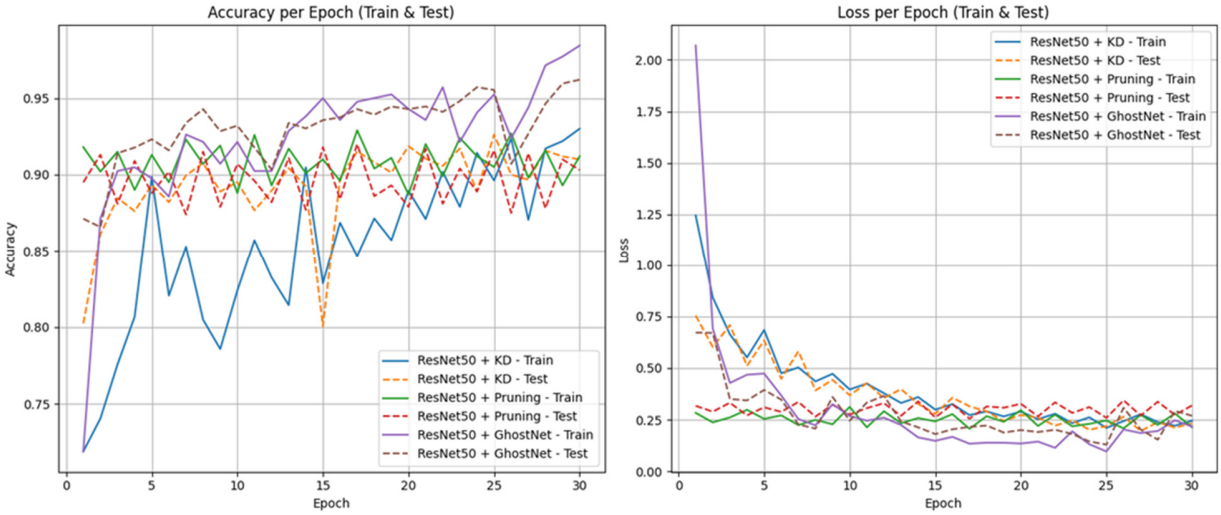

Fig. 2. Accuracy and loss curves for ResNet50 with KD, pruning, and GhostNet.

The ResNet50+GhostNet model achieved the highest and most stable accuracy (95%), indicating strong generalization and effective learning. The KD-based model showed noticeable fluctuations, especially after epoch 20, suggesting unstable convergence. The pruned model maintained consistent but lower accuracy (~91%), implying that sparsity alone may not ensure optimal performance. The GhostNet-enhanced model demonstrated the fastest and smoothest loss reduction, with final values below 0.2, reflecting efficient feature learning. The KD model showed inconsistent loss patterns (0.2–0.5), indicating suboptimal knowledge transfer. The pruned model reached a relatively low loss, and its instability suggests possible degradation from excessive pruning.

As shown in Table I, the proposed GhostNet-enhanced ResNet50 achieved 98.4% accuracy, while reducing model size from 800 MB to 150 MB, parameters from 25.6M to 9.7M, and inference time from 120 ms to 62 ms, which is a more than 62% reduction without compromising performance. Training time also dropped from 100 to 60 minutes, confirming the model's efficiency.

TABLE I. QUANTITATIVE COMPARISON BETWEEN RESNET50 VARIANTS

Model	Accuracy %	Size (M)	Parameters (M)	Training time (m)	Inference time (ms)
ResNet50	96.6%	800 MB	25.6 M	100 m	120 ms
ResNet50+ Pruning	91.2%	420 MB	12. 3 M	90 m	80 ms
ResNet50+ KD	93.0%	470 MB	14.5 M	88 m	85 ms
ResNet50+ GhostNet	98.4%	150 MB	9.7 M	60 m	62 ms

These results reveal the relative stability of each approach during training. The ResNet50+GhostNet maintained consistently high accuracy and low loss values throughout the 30 epochs, indicating strong convergence stability. The pruned ResNet50 exhibited steady but lower accuracy, suggesting that pruning preserved stability but with a minor performance trade-off. However, the KD-based model experienced noticeable oscillations, particularly in the accuracy curve after epoch 20, which points to less stable training dynamics.

By integrating ghost modules into the deeper layers of ResNet50, the model achieves superior optimization in both performance and resource usage. Compared to KD and pruning, which offer limited gains and potential instability, GhostNet delivers a more balanced trade-off, making it ideal for real-time deployment in resource-constrained medical environments. The GhostNet-based model is the most efficient in terms of reducing the size of the model and the number of parameters, while also halving the inference time compared to the original model. These benefits considerably improve its deployability in clinical situations that require fast image review with limited computational resources.

Figure 3 shows the validation accuracy curves of the three optimization techniques. The ResNet50+GhostNet model shows a definite increasing trend with small variations through each of the epochs during training. At the end of the training, it reaches the highest precision, close to 98%, which suggests efficient feature extraction and stable learning even in the presence of training noise. The KD model shows maximum variability and erratic progress. The fluctuations are indicative of difficulties that lie in passing knowledge from a large teacher network to a small student model, especially where training dynamics are unstable.

Lastly, despite train conditions with noise and variability, a GhostNet-based model generally possesses higher reliability, hence better performance. This reliability further emphasizes its potential to be applied to practical real-world applications in medicine, where robustness and precise results are important. These three models were assessed with quantitative measures that included both classification accuracy and computational speed, so that a balanced comparison between accuracy, speed, and deployability could be made.

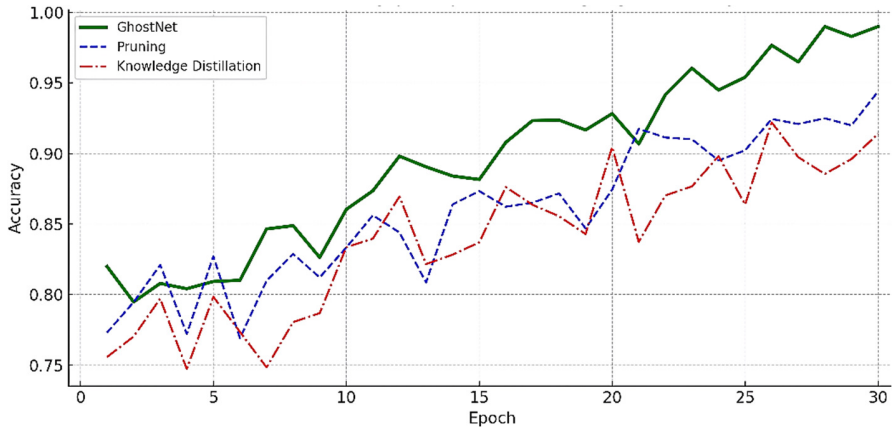

Fig. 3. Validation accuracy and loss for all models.

III. PERFORMANCE EVALUATION

To assess classification accuracy, Accuracy, Precision, Recall, and F1-score were employed, calculated using:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$Specificity = \frac{TN}{TN+FP} \quad (4)$$

Confusion matrices were generated for all models using an independent test set of 164 CT images, equally split between "Mild" and "Severe" cases. These matrices provide a structured comparison of each model's classification outcomes. The unaugmented test set was used to evaluate each model's diagnostic reliability on real-world cases.

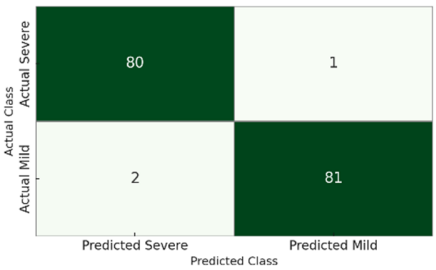

Fig. 4. Confusion matrix of the ResNet+Ghost model.

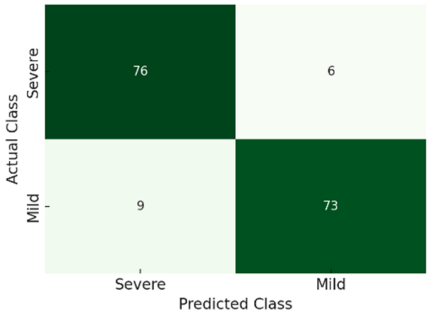

Fig. 5. Confusion matrix of the ResNet50+Pruning model.

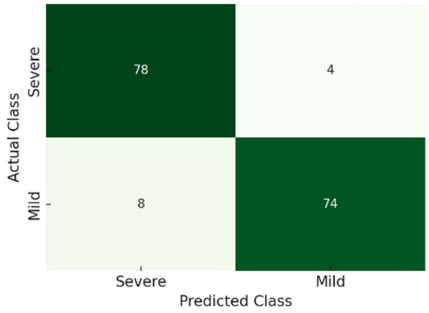

Fig. 6. Confusion matrix of the ResNet50+KD model.

The GhostNet-based model achieved the highest accuracy, correctly classifying 80 severe and 81 mild cases, with only 3 misclassifications, demonstrating superior discrimination and minimal error. The pruned ResNet50 model reached 91% accuracy with 15 misclassified cases, indicating that structural pruning, while reducing model size, may slightly affect performance. The KD model achieved 93% accuracy, with 12 misclassifications, showing better results than pruning but with some instability likely due to incomplete knowledge transfer.

Overall, using GhostNet with ResNet50 proved to be the most accurate and stable model, making it the most suitable for clinical deployment under resource constraints.

TABLE II. SUMMARY OF CONFUSION MATRIX COMPONENTS AND CLASSIFICATION PERFORMANCE FOR THE THREE OPTIMIZED RESNET50 MODELS

Model	Accuracy (%)	TP	TN	FN	FP
GhostNet	98.4	80	81	2	1
ResNet50-Pruned	91	76	73	6	9
ResNet50-KD	93	78	74	4	8

IV. CONCLUSION

This study compared three optimization techniques to improve ResNet50 in CT image classification, namely ghost modules, KD, and pruning. The GhostNet-enhanced model achieved the highest classification accuracy (98.4%), with significant reductions in model size and inference time, making it the most suitable for real-time deployment in resource-

limited clinical environments. The KD model achieved 93% accuracy, proving beneficial when a strong teacher network is available, allowing for effective knowledge transfer to a smaller model. In contrast, the pruned model reached 91% accuracy, offering substantial parameter reduction, which makes it a viable option for highly constrained systems, albeit with some trade-off in performance. Overall, the combination with GhostNet delivered the best balance between accuracy and efficiency, establishing itself as the most effective solution among the three for medical CT classification tasks. Future work could investigate the combination of ghost modules with KD to further enhance model performance and efficiency.

REFERENCES

[1] A. I. Alaiad, E. A. Mugdadi, I. I. Hmeidi, N. Obeidat, and L. Abualigah, "Predicting the Severity of COVID-19 from Lung CT Images Using Novel Deep Learning," *Journal of Medical and Biological Engineering*, vol. 43, no. 2, pp. 135–146, Apr. 2023, https://doi.org/10.1007/s40846-023-00783-2.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778, https://doi.org/10.1109/CVPR.2016.90.

[3] Y. Xu, T. M. Khan, Y. Song, and E. Meijering, "Edge deep learning in computer vision and medical diagnostics: a comprehensive survey," *Artificial Intelligence Review*, vol. 58, no. 3, Jan. 2025, Art. no. 93, https://doi.org/10.1007/s10462-024-11033-5.

[4] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "GhostNet: More Features From Cheap Operations," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 1577–1586, https://doi.org/10.1109/CVPR42600.2020.00165.

[5] C. T. Pham, M. N. Phan, and T. T. T. Tran, "Image classification based on deep learning with automatic relevance determination and structured Bayesian pruning," *Computer Research and Modeling*, vol. 16, no. 4, pp. 927–938, Aug. 2024, https://doi.org/10.20537/2076-7633-2024-16-4-927-938.

[6] S. Wang *et al.*, "Unifying Biomedical Vision-Language Expertise: Towards a Generalist Foundation Model via Multi-CLIP Knowledge Distillation." arXiv, Jun. 27, 2025, https://doi.org/10.48550/arXiv.2506.22567.

[7] 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)*, Penang, Malaysia, Sep. 2024, pp. 1–6, https://doi.org/10.1109/ICSSA62312.2024.10788660.

[8] H. Alshazly, C. Linse, E. Barth, and T. Martinetz, "Explainable COVID-19 Detection Using Chest CT Scans and Deep Learning," *Sensors*, vol. 21, no. 2, Jan. 2021, Art. no. 455, https://doi.org/10.3390/s21020455.

[9] A. N. Kadhum, "Performance Assessment of DL Model in Iraq for Covid-19 Patient Severity Prediction from X-Ray Scan Images," *Journal of Image Processing and Intelligent Remote Sensing*, no. 36, pp. 1–11, Sep. 2023, https://doi.org/10.55529/jipirs.36.1.11.

[10] N. Kumar, A. Hashmi, M. Gupta, and A. Kundu, "Automatic Diagnosis of Covid-19 Related Pneumonia from CXR and CT-Scan Images," *Engineering, Technology & Applied Science Research*, vol. 12, no. 1, pp. 7993–7997, Feb. 2022, https://doi.org/10.48084/etasr.4613.

[11] Z. Al-Milaji and H. Yousif, "Lightweight Deep Learning Model Optimization for Medical Image Analysis," *International Journal of Imaging Systems and Technology*, vol. 34, no. 5, 2024, Art. no. e23173, https://doi.org/10.1002/ima.23173.

[12] 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.

[13] N. K. Dinsdale, M. Jenkinson, and A. I. L. Namburete, "STAMP: Simultaneous Training and Model Pruning for low data regimes in medical image segmentation," *Medical Image Analysis*, vol. 81, Oct. 2022, Art. no. 102583, https://doi.org/10.1016/j.media.2022.102583.

[14] H. Meng, Z. Lin, F. Yang, Y. Xu, and L. Cui, "Knowledge Distillation In Medical Data Mining: A Survey," in *5th International Conference on Crowd Science and Engineering*, Jinan, China, Nov. 2022, pp. 175–182, https://doi.org/10.1145/3503181.3503211.

[15] A. BabaAhmadi, S. Khalafi, M. ShariatPanahi, and M. Ayati, "Designing an improved deep learning-based model for COVID-19 recognition in chest X-ray images: a knowledge distillation approach," *Iran Journal of Computer Science*, vol. 7, no. 2, pp. 177–187, Jun. 2024, https://doi.org/10.1007/s42044-023-00167-4.

[16] E. Cengil and A. Cinar, "Multiple Classification of Flower Images Using Transfer Learning," in *2019 International Artificial Intelligence and Data Processing Symposium (IDAP)*, Malatya, Turkey, Sep. 2019, pp. 1–6, https://doi.org/10.1109/IDAP.2019.8875953.

[17] 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, Jul. 2020, Art. no. 20TR01, https://doi.org/10.1088/1361-6560/ab843e.

[18] B. Mandal, A. Okeukwu, and Y. Theis, "Masked Face Recognition using ResNet-50." arXiv, 2021, https://doi.org/10.48550/ARXIV.2104.08997.

[19] T. N. Nguyen, T. T. Nguyen, T. H. Nguyen, and B. V. Ngo, "A Robust Approach for Breast Cancer Classification from DICOM Images," *Engineering, Technology & Applied Science Research*, vol. 15, no. 3, pp. 23499–23505, Jun. 2025, https://doi.org/10.48084/etasr.10931.