Scientific Reports

https://doi.org/10.1038/s41598-025-33178-1

Article in Press

An efficient dual path deep learning framework for COVID-19 classification using lung CT scans with explainable AI

Received: 3 August 2025

Accepted: 16 December 2025

Published online: 24 January 2026

Cite this article as: Rahat M.M.A., Islam M.I., Miah M.S.U. *et al.* An efficient dual path deep learning framework for COVID-19 classification using lung CT scans with explainable AI. *Sci Rep* (2025). https://doi.org/10.1038/s41598-025-33178-1

Md. Mahid Arfan Rahat, Md. Imamul Islam, Md Saef Ullah Miah & Talal Alharbi

An Efficient Dual Path Deep Learning Framework for COVID-19 Classification Using Lung CT Scans with Explainable AI

Md. Mahid Arfan Rahat[1], Md. Imamul Islam[2], Md Saef Ullah Miah[3], and Talal Alharbi[4*]

[1]Department of Electrical and Electronic Engineering, Green University of Bangladesh, Purbachal American City, Rupganj, Narayanganj, 1461, Bangladesh
[2]Department of Electrical and Electronic Engineering, Bangladesh University of Business and Technology, Dhaka, 1216, Bangladesh
[3]American International University-Bangladesh, Dhaka, 1229, Bangladesh
[4]Department of Electrical Engineering, College of Engineering, Qassim University, Buraydah, 52571, Saudi Arabia
[*]atalal@qu.edu.sa

ABSTRACT

While the global burden of COVID-19 has eased due to widespread vaccination and public health efforts, the virus has not been eradicated. New variants continue to emerge, and localized outbreaks remain a concern, particularly in regions with limited healthcare resources. This highlights the ongoing need for rapid, accurate, and scalable diagnostic tools. In this study, a comprehensive deep learning framework for detecting COVID-19 from lung CT scans is presented, aimed at improving diagnostic reliability and computational efficiency. An extensive and diverse CT dataset was curated by combining images from nine publicly available datasets, with a total of 25,408 samples in COVID-19 and normal classes. Multiple state-of-the-art convolutional neural networks (CNNs) and vision transformer models were fine-tuned and evaluated under consistent conditions to build a strong performance benchmark. Based on these findings, a new lightweight parallel model was developed, combining a custom CNN and a pretrained backbone. Both networks process the input image independently, and their extracted features are fused at the final stage for classification. The proposed model demonstrated higher accuracy (97.46%) compared to other models tested in this study, while maintaining low computational complexity. Additionally, explainable AI techniques, including Grad-CAM and LIME, were employed to provide visual interpretations of the model's predictions.

Keywords: COVID-19 classification, Lung CT scan, Deep learning, Medical image analysis, Computer-aided diagnosis, Feature fusion, Parallel architecture, Dual-path network, Explainable AI

Introduction

COVID-19, caused by the novel coronavirus SARS-CoV-2, emerged in late 2019 and rapidly developed into a global pandemic[1]. The widespread transmission of the virus caused substantial morbidity and mortality, along with profound socioeconomic disruption worldwide. In recent years, global vaccination campaigns and enhanced public health strategies have led to fewer severe cases and deaths. By 2025, the situation has improved, but COVID-19 has not yet disappeared. New variants of the virus continue to emerge, and occasional outbreaks remain common, particularly in regions with lower vaccination rates or weaker health systems. Recent studies suggest that SARS-CoV-2 is likely to remain a long-term public health issue, with future waves possible unless robust surveillance and immunity levels are maintained[2,3].

Computed tomography (CT) imaging plays a critical role in detecting COVID-19 by revealing characteristic patterns and abnormalities in the lungs that are indicative of early infection[4]. Since these features are not typically present in healthy lungs, CT serves as a valuable tool for differentiating COVID-19 from normal cases[5]. Consequently, many researchers have turned to artificial intelligence and deep learning to automatically analyze CT scans, facilitating faster and more reliable COVID-19 diagnosis[6]. Deep learning-based imaging is best viewed as a complement to RT-PCR, especially in resource-limited settings[7]. It also aids in monitoring disease progression, quantifying lung damage, and providing decision support in emergencies, thereby enabling a more efficient allocation of critical resources.

With the recent advances in deep learning, convolutional neural networks (CNNs) have shown impressive performance on medical image classification. CNNs extract relevant features from images in a hierarchical manner, automatically identifying low-level textures and high-level semantic patterns. In parallel with CNNs, vision transformer (ViT) models have become viable contenders to convolutional networks in image recognition tasks. Originally designed for natural language processing

(NLP) tasks, these models leverage a self-attention mechanism to model long-range dependencies in image data, allowing for a powerful approach to capture contextual information on a global scale[8].

Even with these advances, a key challenge in applying deep learning to medical image analysis is finding the right balance between model size and performance[9]. Standard CNNs are strong at learning local features but often struggle to capture broader, long-range relationships that are important for complete image understanding[10]. In contrast, Transformer based models are capable of modeling such global dependencies, but they tend to be computationally demanding and depend on very large datasets for effective pretraining[11]. Some recent works have attempted to combine local and global feature extraction paths to benefit from both approaches, yet many of these hybrid designs remain heavy in computation[12–14]. In addition, many studies on COVID-19 diagnosis have used relatively small datasets, which can limit the model from diverse feature learning. To address these gaps, this study presents a lightweight, parallel dual-path feature extraction architecture for CT image classification. The dual-path structure enriches image representation by processing fine-grained local details alongside broader contextual cues in parallel. The proposed model was evaluated on a large combined dataset of 25,408 CT images collected from nine public sources. The results indicate that the lightweight model achieved higher diagnostic accuracy compared to several baseline pretrained models. The contributions in this study are summarized as follows:

1. A large-scale lung CT dataset was constructed by integrating nine publicly available sources, resulting in 16,341 COVID-19 and 9,067 normal images.

2. Several state-of-the-art CNN and transformer models (ResNet50, Xception, MobileNetV2, DeiT, BEiT, Swin-ViT, and CaiT) were fine-tuned using unified training procedures for a fair comparative analysis on the COVID-19 classification task.

3. A custom lightweight parallel CNN architecture was designed that combines a convolutional network with a pretrained backbone for joint feature extraction. The features from both branches are fused and passed through a series of fully connected layers for classification. The proposed model contains only 3.5 million parameters and outperformed all pretrained models evaluated in this study.

4. The model's robustness was validated by evaluating it on chest X-ray data, and explainable AI methods (Grad-CAM and LIME) were applied to ensure the interpretability and clinical relevance of the predictions.

Related work

Here, relevant work devoted to the detection of COVID-19 from CT scan images using deep learning methods is summarized. The emergence of COVID-19 as a global health crisis has accelerated the demand for rapid and accurate diagnostic tools. Consequently, the integration of artificial intelligence (AI) with medical imaging techniques, especially lung CT scans, has gained immense traction to enable the early diagnosis of COVID-19. A review by Mahanty et al.[15] provides a survey of the use of deep learning for COVID-19 detection from various types of data, such as CT images. To begin with, Elmuogy et al.[16] proposed a transfer learning-based approach called Worried Deep Neural Network (WDNN), where they fine-tuned multiple convolutional neural net architectures such as InceptionV3, ResNet50, and VGG19. Among the evaluated models, VGG19 achieved the highest classification accuracy of 99.05% with impressive precision and recall. This established the feasibility of using pretrained architectures with transfer learning to conduct rapid COVID-19 detection using CT scans. Despite the promising results, the dataset's small size of 2623 images could constrain the model's applicability to broader populations. Similarly, Rahimzadeh et al.[17] developed a deep learning pipeline combining ResNet50V2 and FPN such that it mitigates overfitting. This architecture allowed multi-scale detection of the manifestations of COVID-19 across CT slices. In a dataset comprising more than 60,000 images, the model achieved 98.49% accuracy with a specificity of 98.7%. However, despite the substantial number of images, the dataset was derived from a comparatively small number of patients, which may limit the generalizability of the model.

Salama et al.[18] proposed an end-to-end segmentation-classification framework by integrating U-Net for lung region segmentation and VGG16, ResNet50 for the ultimate classification task. Their preprocessing pipeline incorporated methods such as Wiener filtering and ISDNT normalization, which improved image clarity, allowing this work to attain a classification accuracy of 98.98%. The study emphasized the significance of preprocessing and segmentation for the downstream objective of classification. However, this study is limited by the small size of the original dataset, which contained only 746 CT images. To enhance the dataset's usability for training, data augmentation was applied, increasing the total number of images to approximately 2,312. The strongest results were obtained when using a ResNet50-based model, although such a complex design inevitably increases the number of parameters and raises concerns about computational cost and efficiency. Soni et al.[19] tackled segmentation and classification in a unified model called BUF-Net. The performance on small datasets was enhanced by using CGAN for data augmentation and residual blocks within the architecture. Overall, their method achieved an accuracy

of 93.1% and demonstrated the efficacy of generative models for augmenting small medical datasets. It is essential to note that the initial work was developed on a small dataset, necessitating augmentation through Conditional Generative Adversarial Networks (CGANs); this reliance on synthetic data may temper the model's performance stability when deployed in a broader clinical setting.

Islam et al.[20] proposed a deep CNN architecture called IDConv-Net, using five convolutional layers and batch normalization. IDConv-Net has been adapted for X-ray and CT images. For CT scans, they achieved a training accuracy of 99.53%, test accuracy of 98.41% and an F1-score of 98.48% with their model. This approach functioned well in terms of diagnostic accuracy while the model itself was computationally efficient. A closer examination reveals that the discussion on computational efficiency mostly focuses on training time, while details such as inference speed and the number of floating-point operations (FLOPs) are not reported. In addition, the study does not address model interpretability using Explainable AI (XAI) techniques, nor does it clearly mention whether a patient-wise data split was used during evaluation. Foysal et al.[21] employed deformable convolutions in a ResNet-50 backbone to capture non-uniform infection patterns in CT images. Their deformable CNN outperformed the standard versions with an accuracy of 97.6% and a specificity of 98.5% in this test. Additionally, Grad-CAM was applied to highlight areas of interest, which enhanced interpretability. The enhanced performance of the deformable architecture, while notable, typically introduces greater computational overhead compared to standard convolutional models.

Afterward, Haennah et al.[22] employed a Pix2Pix GAN for data augmentation alongside a ResNet101 classifier optimized using Tuna Swarm Optimization (TSO). They achieved 97.2% internal validation accuracy and 95.4% external validation accuracy for their pipeline. It emphasized an innovative way of enhancing not just dataset variability but also hyperparameterization of the models through GANs and swarm-driven tuning. It is worth noting that the incorporation of a GAN for augmentation and a swarm optimization algorithm for tuning substantially increases the computational requirements and training duration beyond that of a standard model. Rajinikanth et al.[23] employed MobileNet variants (V1, V2, V3_Small) for feature extraction in a lightweight pipeline. These features were also processed using Shannon's entropy-driven preprocessing and further optimized with the Brownian Butterfly Algorithm (BBA). The proposed method employed three modes of feature use—individual, dual, and ensemble; the ensemble features achieved an accuracy of 99.10%. This approach underscored the capacity to implement high-performing but lightweight approaches in clinical settings. While the reported accuracy is impressive, this methodology requires a multi-stage training process that is not end-to-end, and the final computational complexity, including parameter count and inference metrics, is not explicitly quantified for the full model.

Kordnoori et al.[24] implemented a ten-layer CNN architecture with three dense layers, trained on several datasets. They utilized datasets such as MedSeg, MosMedData, and COVIDx-CT, among others, and conducted extensive data augmentation to maintain the distributions of different classes. Their model achieved an accuracy of 89.00% and an AUC of 0.92, indicating its ability to generalize well across various CT data distributions. While the model surpassed several pre-trained counterparts, its overall accuracy remains relatively modest, leaving scope for further refinement to reach state-of-the-art performance. Rajpoot et al.[25] employed an ensemble method on VGG16, ResNet50, and DenseNet169 for classification, analyzing results using LIME, SHAP, and Grad-CAM. They achieved an accuracy of 96.18% on CT scans using their model, emphasizing the importance of explainable AI in all healthcare AI applications. A significant drawback is that the ensemble of these heavy architectures incurs substantial computational cost. Nikam et al.[26] developed an architecture of stacked CNN using the backbone of VGG19 to enhance classification performance. Their model appears to function consistently between training and validation phases, achieving 98.44% training accuracy and 96.88% on validation data with the integrated stacked configuration. The study also compared standard CNNs and VGG19 alone, suggesting that deeper and more customized CNN structures can significantly enhance the diagnostic performance. Despite the high performance, the design inherently faces a challenge: the heavy VGG19 backbone, combined with the stacked configuration, gives much computation, and this is further constrained by the relatively small size of the dataset employed. Chowa et al.[27] introduced a novel self-supervised federated learning framework that employs BYOL and Paillier Homomorphic Encryption for data privacy. Their optimization was driven by both labeled and unlabeled data among distributed nodes, and the model comprises a VGG19 with an enhanced attention encoder. Training through few labeled data resulted in 97.19% accuracy with 98.18% recall, suitable for immersive medical contexts that present privacy challenges. Nevertheless, the computational demands of the architecture and the absence of explainable AI components were not explored. The study also does not specify if a patient-wise data split was used, which is important for validating model generalizability.

Ferraz et al.[28] conducted a comprehensive survey of various architectures, such as CNNs, Vision Transformers (ViT), and Swin Transformers. Remarkably, their analysis demonstrated that the Swin Transformer was evaluated using datasets such as SARS-CoV-2 CT, HCV-UFPR-COVID-19, and COVID-QU-Ex, consistently surpassing others with an AUC of 0.94, whereas balanced accuracy, precision, and F1-scores persisted at 94%. In contrast, the ViT model generalized poorly, with substantially lower metrics. Their work validated that attention-based architectures, particularly hierarchical architectures like Swin Transformers, generalized effectively and delivered superior performance on cross-dataset COVID-19 CT scan classification. While effective, the computational requirements of such transformer-based models and their inherent lack

of explainability remain aspects for further consideration. Taye et al.[29] proposed a two-stage ViT model for COVID-19 identification and severity grading from CT scans, achieving 99.7% detection accuracy and high segmentation performance. Although the ViT model outperformed CNNs, it heavily relied on computational resources, and their study lacks explainable AI. The study also indicated that the model performance was limited by the small dataset size. Rashed et al.[30] proposed a conditional cascaded network (CCN) for automatic diagnosis of COVID-19 on chest X-ray and CT images with the help of transfer learning. They experimented on seven pre-trained CNN models like GoogleNet, SqueezeNet, ResNet-18, ResNet-50, AlexNet, DarkNet-19, and ShuffleNet individually over multiple data sets. The best performance on the Mendeley CT dataset was obtained with RMSprop optimization for DarkNet-19. The method achieved high diagnostic performance, but it heavily depended on pretrained models with high computational cost, and limited CT data (2628 COVID-19 images) were used in training the model. Also, the paper does not present key computational components such as inference time and floating-point operations. Ghaffar et al.[31] conducted a comparative evaluation of ten state-of-the-art CNN architectures, including EfficientNet, MobileNet, and InceptionV3, for automatic detection of COVID-19 from chest X-ray images. Their study used a multi-class dataset (COVID-19, pneumonia, and normal) and found that MobileNet and EfficientNet achieved the best performance (95% accuracy). However, the work relied on limited COVID-19 samples (176 training and 24 test images) and did not analyze computational efficiency metrics such as FLOPs or inference time.

Despite the widespread use of 2D CT slice–based methods for COVID-19 detection, a few studies have attempted to use the 3D volumetric CT analysis method, which can capture richer spatial context. Compared with two-dimensional CT, 3D CT further enables the model to capture inter-slice dependencies and provides a better way to visualize lung abnormalities. Still, these methods involve heavy computations and are restricted by the lack of annotated volumetric datasets. Fouad et al.[32] designed an explainable 3D deep learning model under the British Society of Thoracic Imaging (BSTI) guidelines to categorize CT volumes into Classic, Probable, Indeterminate and Non-COVID. Their ResNet-50 model achieved accuracy of 75%, which improved to 90% after removing the Indeterminate class. Despite its methodological soundness, the study was constrained by a small dataset and showed relatively poor performance in identifying the indeterminate class. The majority of early approaches employed chest X-rays due to their wide availability; the higher sensitivity of CT for visualizing key pulmonary features, such as ground-glass opacities (GGOs), makes it an important modality for detailed analysis. To situate this study within the broader methodological context, Liu et al.[33] introduced a Controllable Ensemble CNN and Transformer (CECT) spatial context network to extract local as well as global information of COVID-19 from X-ray images. Their proposed method achieved 98.1% and 90.9% intra-dataset and cross-dataset accuracy which is competitive with other hybrid networks. Despite its high performance, the method is computationally expensive due to multiple network branches and the study does not report model size, FLOPs, inference time, or any explainable AI analysis, which hinder its transparency and deployment feasibility.

While these studies have significantly advanced the field, multiple limitations are also revealed. A common pattern is that the training datasets are small-scale, usually ranging from several hundred to a few thousand images, which may restrict a model's ability to learn representative features. Moreover, there is a large body of literature focusing on performance metrics, with insufficient emphasis on model interpretability and patient-wise validation. This is frequently coupled with the adoption of computationally intensive architectures, such as ensembles of heavy pre-trained networks (e.g., VGG16, ResNet50, DenseNet169) or complex vision transformers. On the other hand, lightweight architectures with a reduced number of network parameters might achieve more balance between accuracy and efficiency. To address these gaps, a large-scale CT dataset was constructed by combining nine publicly available sources in this study. State-of-the-art pre-trained CNN and transformer models were thoroughly assessed. Based on these insights, a lightweight CNN architecture with 3.5 million learnable parameters was designed. The performance metrics are comparable to those reported in previous studies. The robustness was then verified by testing on chest X-ray datasets, and transparency was ensured by using explainable AI techniques. By proposing a novel framework, the method addresses limitations in dataset size, computational cost, and model explainability—addressing these three components that contribute to delivering an efficient and robust solution for COVID-19 diagnosis from lung CT scans. A summary of the reviewed literature is presented in Table 1.

Methodology

This section outlines the working procedures of the proposed Methodology, as shown in Fig. 1, covering the dataset, preprocessing steps, models utilized, and evaluation metrics.

Data Collection

In this study, a combined dataset was constructed using nine publicly available sources from various countries, including Iran, Italy, Russia, China, Brazil, and Japan. Seven of these datasets were previously combined by Maftouni et al.[34], and two new datasets from Kaggle were added. These two datasets consist of images collected from actual patients in medical centers in Tehran (Iran) and São Paulo (Brazil)[35,36]. Altogether, this dataset contains 16,341 COVID-19 images and 9,067 normal images. The number of images from the COVID-19 and normal classes in each dataset is summarized in Table 2. By including samples

Table 1. Summary of Related Works in the Literature.

Author	Method	Database	Total Patients (Class-wise)	Performance Parameters (%)	Splitting Protocol
Elmuogy et al. (2021) [16]	WDNN with InceptionV3 ResNet50, VGG19	COVID-19 CT dataset (2623 images)	Not specified	Accuracy: 99.046 Precision: 98.684 Recall: 99.12 F1-score: 98.90	60:20:20 No mention about patient-wise split
Rahimzadeh et al. (2021) [17]	ResNet50V2 + Feature Pyramid Network (FPN)	COVID-CTset: 15,589 COVID, 48,260 normal	COVID: 95, Normal: 282	Accuracy: 98.49 Sensitivity: 94.96 Specificity: 98.7	K-fold No mention about patient-wise split
Salama et al. (2022) [18]	U-Net + VGG16/ResNet50 with preprocessing and TL	2312 CT images (augmented)	Not specified	Accuracy: 98.98 Precision: 97.99 F1-score: 97.88	70:30 No mention about patient-wise split
Soni et al. (2022) [19]	BUF-Net: U-Net + MLP with CGAN augmentation	746 CT images (COVID/Normal)	Not specified	Accuracy: 93.1 Precision: 97.1 Sensitivity: 87.6 AUC: 0.932	70:30 No mention about patient-wise split
Ghaffar et al. (2022) [31]	Comparative Analysis of Deep Learning Models for Detecting COVID-19 Lung Infection from Chest X-ray Images	Combined Source Normal: 8,851, Pneumonia: 6,096, COVID-19: 200	Not reported	Best Accuracy: 95% (MobileNet and EfficientNet)	90:10 No mention about patient-wise split
Islam et al. (2023) [20]	IDConv-Net (Deep CNN)	X-ray: 15,967 CT: 16,968 images	COVID: 466, Normal: 604	CT Accuracy: 98.41 Precision: 98.64 Recall: 96.31	80:10:10 No mention about patient-wise split
Haennah et al. (2023) [22]	DETS-optimized ResNet101 Pix2Pix GAN	Kaggle COVID-19 CT (2482 images)	Not specified	Accuracy: 97.2 Precision: 96.7 Recall: 95.9	70:30 No mention about patient-wise split
Foysal et al. (2023) [21]	Deformable ResNet-50	Kaggle CT dataset (1252 COVID, 1229 non-COVID)	Not specified	Accuracy: 97.6 Precision: 98.2 Sensitivity: 96.5	80:10:10 No mention about patient-wise split
Rajpoot et al. (2024) [25]	Ensemble: VGG16 + ResNet50 + DenseNet169 (with XAI tools)	COVIDx CXR-3 (X-ray), SARS-CoV-2 CT	Not specified	X-ray Accuracy: 99.00 CT Accuracy: 96.18 Sensitivity: 99.00	70:10:20 No mention about patient-wise split
Rajinikanth et al. (2024) [23]	Lightweight DL + BBA + SET	10,000 CT images (COVID/Normal)	Not specified	Accuracy: 99.10 (ensemble features)	80:10:10 No mention about patient-wise split
Kordnoori et al. (2024) [24]	CNN with 10 conv layers + 3 dense layers	Augmented CT sets (various datasets)	Not specified	Accuracy: 89.00 Sensitivity: 0.95 AUC: 0.92	70:30 No mention about patient-wise split
Taye et al. (2024) [29]	Two-stage ViT: ViT_B/32 for detection, ViTBIS for segmentation & severity classification	Combined CT datasets (5,494 images: 3,801 COVID, 1,693 non-COVID)	Not specified	Detection Acc: 99.7 Lung IoU: 95.8 Lesion IoU: 94	5-fold cross-validation No mention about patient-wise split
Rashed et al. (2024) [30]	DarkNet19 with RMSprop	Mendeley CT dataset (8,055 CT images: 2,628 COVID-19, 5,427 Non-COVID)	Not reported	Accuracy: 97.7 Precision: 97.6 F1-score: 97.7 AUC: 0.98	5-fold cross-validation No mention about patient-wise split
Liu et al. (2024) [33]	Controllable Ensemble CNN and Transformer for COVID-19 classification Using X-Ray Images	COVID-19 Radiography dataset: 3,616 COVID, 10,192 Normal COVIDx CXR-3: 16,194 COVID, 14,192 Normal	COVID-19 Radiography: 3,616 Positive, 10,192 Negative COVIDx CXR-3: 12,795 Positive, 11,194 Negative	Accuracy: 98.1 (Intra), 90.9 (Inter) F1-score: 96.4 (Intra), 97.2 (Inter)	Intra- and inter-dataset validation 8:1:1 (COVID-19 dataset) 8:2 (COVIDx CXR-3)
Chowa et al. (2025) [27]	FL-SSL: VGG19 + Attention CNN + BYOL	MosMed (1110 CT) + Curated CT (14,486)	COVID: 466, Normal: 604	Accuracy: 97.19 Precision: 97.43 Recall: 98.18	80:10:10 No mention about patient-wise split
Ferraz et al. (2025) [28]	Swin Transformer, ViT, CNN	COVID-QU-Ex, HCV-UFPR, SARS-COV-2 CT	Not specified	Accuracy: 94.00 AUC: 0.94	70:15:15 No mention about patient-wise split
Nikam et al. (2025) [26]	VGG-19 with stacked CNN	461 CT images (232 COVID, 229 Normal)	COVID: 232, Normal: 229	Training Acc: 98.44 Validation Acc: 96.88	80:10:10 No mention about patient-wise split
Fouad et al. (2025) [32]	3D ResNet-50 (pre-trained on Kinetics-700) + Grad-CAM	56 CT scans Classic: 10 Probable: 21 Indeterminate: 10 Non-COVID: 15	Classic: 10 Probable: 21 Indeterminate: 10 Non-COVID: 15	4-class Acc: 75, F1: 75 3-class Acc: 90, F1: 92	60:20:20 Patient-based split

from various populations, a more complete view of diversity is provided in the training set, supporting the model in learning features that reflect population diversity. For chest X-ray samples, the COVID-19 Radiography Database from Kaggle was used, which contains 3,616 COVID-19 samples and 12,000 normal chest X-ray images[45,46].

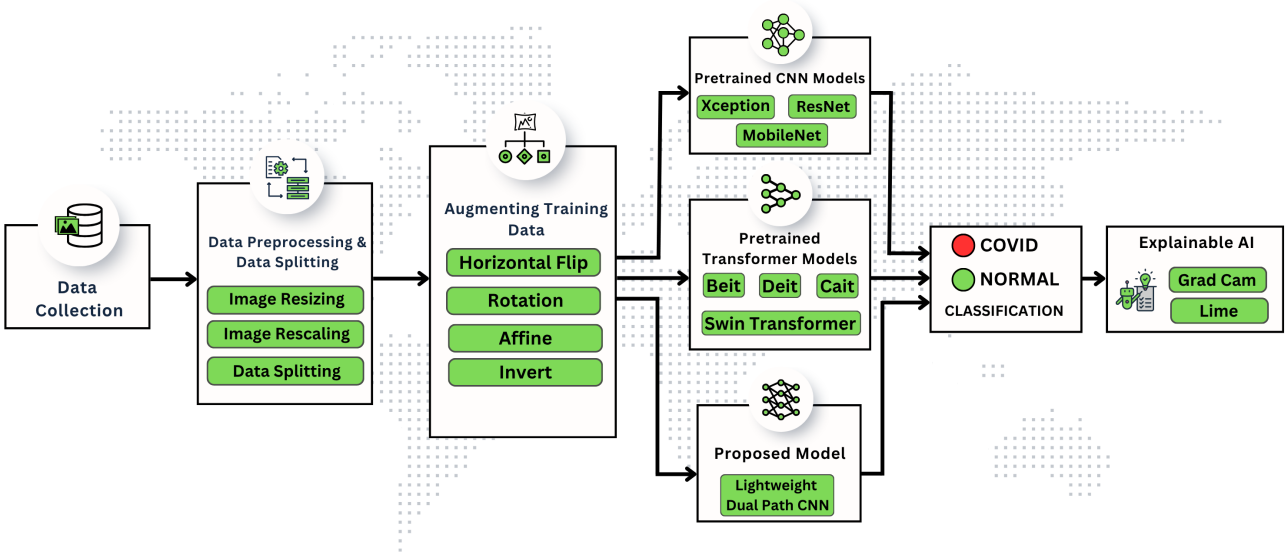

Figure 1. Overview of the proposed methodology pipeline, highlighting key stages including dataset collection, preprocessing, model training, and Explainable AI.

Table 2. Summary of Datasets Used for COVID-19 Image Classification.

Dataset	Country	COVID-19 Images	Normal Images	COVID-19 Cases	Normal Cases
[37]	Iran	666	1053	68	274
[38]	Italy	100	NA	43	NA
[39]	Multiple	1844	NA	20	NA
[40]	Multiple	34	NA	17	NA
[41]	Russia	785	5080	50	254
[42]	China, Japan	349	NA	213	NA
[43]	Iran	3815	760	55	76
[35]	Brazil	1252	1230	NA	NA
[44]	Iran	7496	944	190	59
Ours	Multiple	16341	9067	>656	>663

Preprocessing and Augmentation

Preprocessing is an essential step in deep learning approaches, which is a prerequisite to training a model. In this study, the integrated lung CT scan dataset images had different dimensions, thus all images were resized to the fixed dimensions of 224×224 pixels. After resizing, pixel values were normalized to maintain uniform input intensity across the dataset. The dataset was divided based on patient information rather than image count. A total of 1066 patient records were available, with 161 patients assigned for validation and another 161 for testing. The detailed patient-based split is presented in Table 3. To minimize the risk of data leakage caused by duplicate or near-identical images across the nine source datasets, a systematic duplicate detection procedure was conducted. Feature embeddings for all images were extracted using a pre-trained ResNet-50 model, and pairwise cosine similarity scores were computed. All image pairs across the training, validation, and test sets showed similarity scores below 0.95, confirming the absence of near-duplicate images between different subsets of the final dataset. Data augmentation was performed dynamically during the training process, including random horizontal flipping with a 50% probability, random rotation of up to 10° in either direction, random affine transformations with up to 15° rotation, translations of up to 10% in each direction, scaling between 80% and 120% of the original size, and random color inversion to simulate multiple imaging conditions.

Table 3. Summary of dataset split for Normal and COVID-19 classes showing the number of patients and images used in each subset.

Class	Data Type	Train	Validation	Test	Total
COVID-19	Patients	323	70	70	463
	Images	13955	1275	1111	16341
Normal	Patients	421	91	91	603
	Images	7015	1036	1016	9067
Total	Patients	744	161	161	1066
	Images	20970	2311	2127	25408

Pretrained CNN and Transformer Models

An exhaustive analysis of a set of modern CNN architectures was performed. For comparison, three commonly used pre-trained CNN models were chosen: ResNet50[47], Xception[48], and MobileNetV2[49]. All these models were fine-tuned by transfer learning with ImageNet[50] pre-trained weights. Each model was trained with the same hyperparameter configuration to ensure a fair comparison. These settings will be discussed in detail in the section .

ResNet50 is a 50-layer deep CNN architecture known for residual learning. The main concept of ResNet50 is the idea of identity shortcut connections to alleviate the vanishing gradient phenomenon in very deep networks. This allows the network to learn residual functions relative to the layer inputs, which improves convergence speed and model accuracy. ResNet50 was selected as a strong baseline model that has been previously utilized successfully for visualization of CT scans.

Xception is a deep convolutional neural network architecture that augments Inception[51] by replacing its standard inception modules with depthwise separable convolutions. This change reduces the number of parameters but still provides high representational power. Xception is especially effective in capturing spatial patterns in medical images due to its ability to model channel-wise and spatial-wise information separately. Xception provided reasonable classification performance and computational efficiency when used on our combined dataset.

MobileNetV2 is optimized for computer vision applications on mobile and embedded systems. It uses inverted residual blocks and depthwise separable convolutions to reduce computational cost. Its architecture consists of a simple convolutional layer and 19 bottleneck layers, totaling roughly 3.4 million parameters. Considering its lightweight nature and deployment on resource-constrained systems, MobileNetV2 proved to be one of the better-performing models in this study.

In recent years, transformer architectures have become popular in computer vision tasks after proving successful in natural language processing[8]. This revolution started with the theoretical paper "Attention Is All You Need"[52], introducing the self-attention mechanism of the transformer models. Since their introduction to image classification, transformers have become the dominant approach in the field, providing an alternative to convolution-based approaches. Four popular vision transformer models, namely DeiT[53], BEiT[54], Swin-ViT[55], and CaiT[56]—were selected for evaluation. All these models were pre-trained on the ImageNet dataset and fine-tuned on the binary classification task in this study. For training those models, the PyTorch TIMM library[57] was used, and the variants used were `beit_base_patch16_224`, `cait_s24_224`, `swin_base_patch4_window7_224`, and `deit_base_patch16_224`.

DeiT (Data-efficient Image Transformer): This vision transformer architecture aims at accuracy and computational efficiency. The paper introduces a distillation token that enables the model to learn from a teacher network in training and improves performance without large-scale datasets. DeiT differs from the original Vision Transformer (ViT) by introducing a more stable training method that enables effective training with less data.

BEiT (Bidirectional Encoder Representation from Image Transformers): BEiT transfers the masked language modelling (MLM) strategy in NLP to the vision field and proposes an original self-supervised pretraining scheme. BEiT similarly first tokenizes images into discrete visual tokens through a tokenizer (e.g., discrete VAE). Then it is trained to guess these tokens in a masked setting, also the way BERT is trained on text. This enables BEiT to learn rich visual representations without the need for extensive labeled data.

Swin-ViT (Shifted Window Transformer): Swin Transformer proposes a hierarchy with non-overlapping windows shifted at each depth, achieving local and global extraction of features normally based on a less computationally intensive hierarchical framework. It operates on the images at many scales and applies attention within locally discrete windows, which are later shifted periodically to facilitate within-window connectivity. This arrangement allows Swin-ViT to efficiently capture fine-grained and contextual information.

CaiT (Class-Attention in Image Transformers): CaiT extends the vanilla ViT by introducing class-attention parameters behind the self-attention blocks. This lets the class token attend specifically to other image tokens, thus improving where the class focuses on task-relevant features. CaiT applies the LayerScale technique as well, which stabilizes the optimization of extremely deep networks and improves the training of deep transformers.

Proposed CNN Model

While transformers have a more robust representational capability, they also incur more computational complexity[58]. Currently, CNNs are still a better choice for designing lightweight models[59]. Many studies in medical image classification have relied on CNN architectures and achieved strong performance[60–62]. Despite their effectiveness in extracting local spatial features, standard CNNs often struggle to capture long-range dependencies within the image. To address the weakness of standard CNNs in modeling long-range dependencies, the proposed CNN architecture employs two parallel branches designed for local and global feature extraction. The feature maps generated by both branches are fused and passed through dense layers for classification. While all pretrained models were trained using an input size of 224×224, the custom CNN model was trained with a reduced input size of 180×180 to improve computational efficiency. The overall architecture of the proposed CNN model is presented in Figure 2.

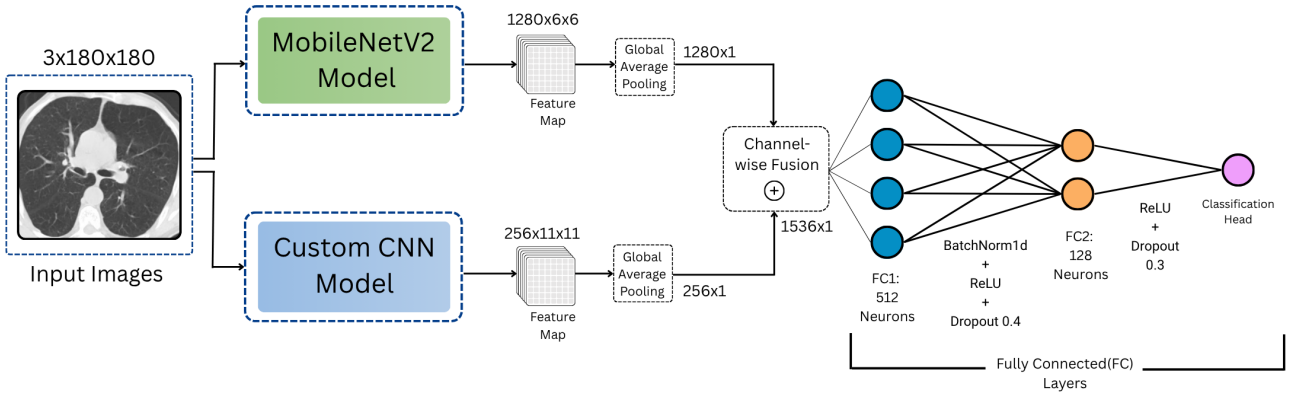

Figure 2. The architecture of the proposed CNN model for COVID-19 detection. The model employs two distinct feature extractors. One is based on the MobileNetV2 backbone and the other is a Custom CNN that operates concurrently. Features extracted from both streams are reduced via Global Average Pooling and then combined through Channel-Wise Fusion ($\oplus$) before being processed by the final Fully Connected (FC) layers for binary classification.

For the global feature extraction path, a pretrained MobileNetV2 was used. MobileNetV2 was chosen for its lightweight architecture and reliable performance in feature extraction. The model outputs a feature map of size $1280 \times 6 \times 6$, which is then processed using global average pooling to obtain a compact and discriminative feature vector suitable for fusion with the local branch.

The local feature extraction path employs a lightweight custom CNN designed to capture fine-grained and localized textural patterns in lung CT scans. It begins with a standard 2D convolutional layer that takes as input a three-channel image of dimension $180 \times 180 \times 3$. The convolution operation preserves the spatial dimensions and extracts localized edge features by applying a kernel with 32 filters of size 3×3 and padding = 1. Following the convolution, a batch normalization layer is used to stabilize and accelerate the learning process by normalizing the feature activations. A ReLU activation introduces non-linearity to the model, enabling it to learn more complex relationships within the image data. Finally, a max-pooling layer with a pool size of 2×2 and stride 2 reduces the spatial resolution to 90×90, retaining the most prominent features within each pooling region.

$$X_1 = \text{MaxPool}(\text{ReLU}(\text{BN}(\text{Conv2D}(X_0, W_1)))) \tag{1}$$

In Eq. 1, $X_0 \in \mathbb{R}^{180\times180\times3}$ is the input image, W_1 represents the learnable weights in the first convolutional layer, and X_1 is the output feature map of size $90 \times 90 \times 32$.

The second convolutional block continues this pattern, taking X_1 as input and applying a convolution with 64 filters of size 3×3, followed by batch normalization, ReLU activation, and max pooling. This stage reduces the feature map dimensions to 45×45, while increasing the channel depth to 64, thereby allowing the model to extract higher-level texture and shape patterns.

$$X_2 = \text{MaxPool}(\text{ReLU}(\text{BN}(\text{Conv2D}(X_1, W_2)))) \tag{2}$$

In Eq. 2, X_1 is the output from the first convolutional block, W_2 are the learnable weights of the second layer, and X_2 denotes the resulting feature map of size $45 \times 45 \times 64$.

The third convolutional block employs 128 filters to learn more abstract representations. With a kernel size of 3×3 and similar normalization and activation operations, it outputs a feature map of $22 \times 22 \times 128$ after max pooling. This stage effectively enhances the model's capacity to identify mid-level features relevant to the infection regions in lung CT scans.

$$X_3 = \text{MaxPool}(\text{ReLU}(\text{BN}(\text{Conv2D}(X_2, W_3)))) \tag{3}$$

In Eq. 3, X_2 is the input from the previous layer (output of the second convolution), W_3 represents the learnable weights in the third convolutional layer, and X_3 is the resulting feature map of size $22 \times 22 \times 128$.

The final convolutional block increases the filter depth to 256, further enriching the hierarchical feature representation. After convolution, batch normalization, and ReLU activation, a max-pooling operation reduces the feature map to 11×11. This final stage captures deep spatial and contextual features crucial for local-level infection discrimination.

$$X_4 = \text{MaxPool}(\text{ReLU}(\text{BN}(\text{Conv2D}(X_3, W_4)))) \tag{4}$$

In Eq. 4, X_3 is the input from the previous layer (output of the third convolution), W_4 denotes the learnable weights in the fourth convolutional layer, and X_4 is the output feature map of size $11 \times 11 \times 256$. ReLU is applied throughout the network to reduce the vanishing gradient problem and accelerate training by enabling sparse activations[63]. Batch normalization, placed after each convolutional layer, stabilizes the learning process by normalizing the intermediate feature distributions, thereby accelerating convergence and improving generalization. Max-pooling acts as a way to reduce the computational complexity while keeping only the most relevant features. The doubling of filters in each convolutional block allows the model to learn complex features, thus increasing the capacity of the model. Overall, this CNN model uses a deep but computationally effective structure, enabling the model to learn progressively from raw pixels to higher-level semantics. The architectural diagram of the custom CNN model for local feature extraction is presented in Fig. 3 to illustrate the whole layer composition and the data flow.

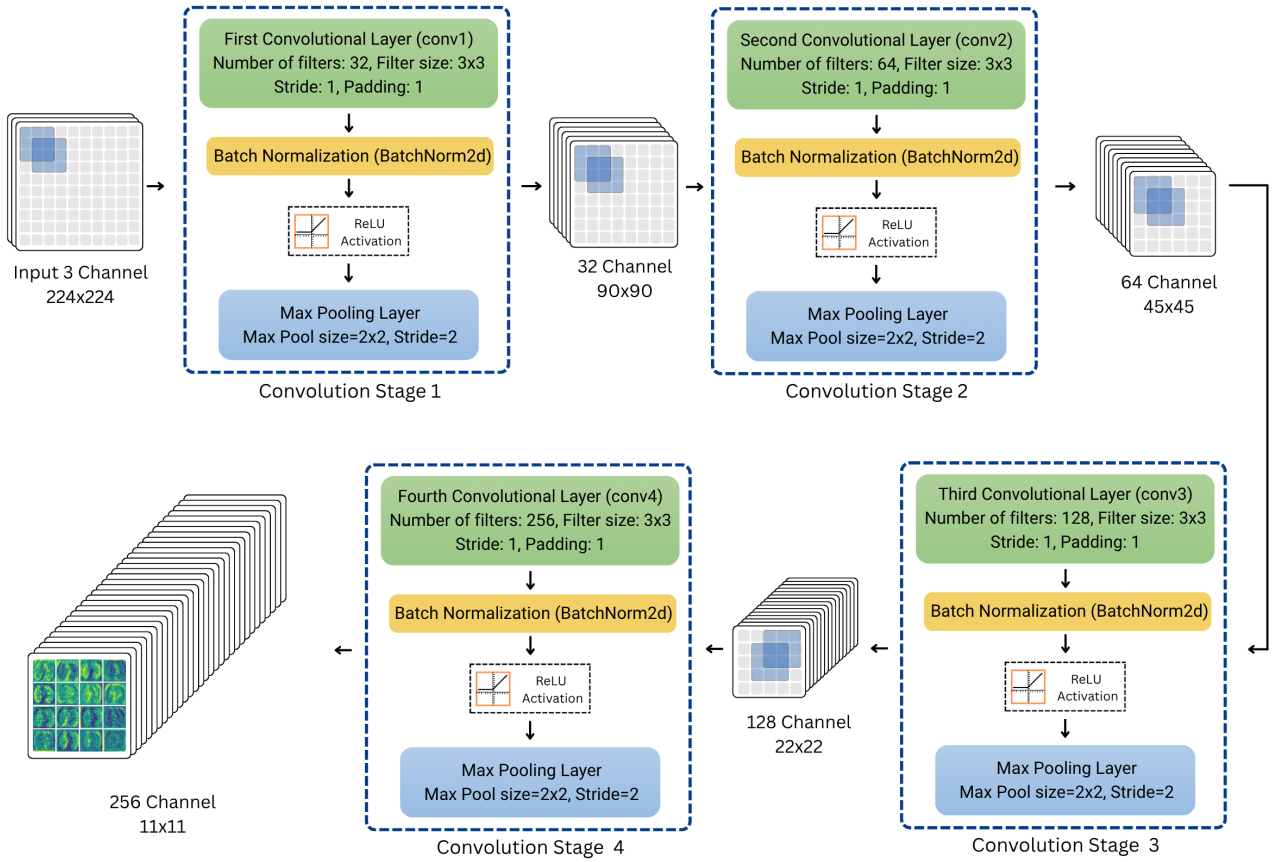

Figure 3. Structural design of the custom CNN model for local feature extraction. The network consists of 4 convolutional stages with max pooling, batch normalization, and ReLU activations for the hierarchical extraction of spatial features.

After feature extraction, the global and local representations obtained from MobileNetV2 and the custom CNN are passed through a global average pooling operation to achieve dimensional uniformity. This operation reduces each feature map to a single spatial representation, converting the $1280 \times 6 \times 6$ and $256 \times 11 \times 11$ outputs into compact $1280 \times 1 \times 1$ and $256 \times 1 \times 1$ tensors, respectively. The use of `GlobalAveragePooling2D` effectively minimizes computational cost while preserving the most salient channel-wise information from each branch. The resulting feature vectors are aligned in dimension and fused along the channel axis, producing a unified 1536-dimensional vector that integrates the global semantic information from MobileNetV2 with the fine-grained spatial features captured by the custom CNN. This fusion enriches the feature space and allows the model to jointly learn both texture-level and contextual patterns. Although advanced fusion strategies such as multi-scale and intermediate feature fusion were explored, simple additive fusion was adopted to maintain model simplicity and low computational overhead, which aligns with the goal of designing a lightweight framework. The fused feature vector is then fed into two fully connected layers (512 and 128 neurons), after which batch normalization, ReLU activation, and dropout (0.4 and 0.3) are applied to improve generalization and prevent overfitting. The final classification head consists of a single neuron with a sigmoid activation function for generating a binary probability corresponding to each input CT image. The mathematical formulation of the fusion and classification is given in (5)–(14).

$$x_g = \text{MobileNetV2}_{\text{feat}}(x_{\text{in}}) \in \mathbb{R}^{1280\times6\times6}, \tag{5}$$

In Eq. 5, x_g indicates the deep feature tensor extracted by the pretrained MobileNetV2 branch from the input image $x_{\text{in}} \in \mathbb{R}^{3\times180\times180}$. The output consists of 1280 channels with a spatial resolution of 6×6, representing high-level semantic information.

$$x_l = \text{CustomCNN}_{\text{feat}}(x_{\text{in}}) \in \mathbb{R}^{256\times11\times11}, \tag{6}$$

Eq. 6 defines x_l as the final feature map from the custom CNN pathway, composed of 256 channels and a spatial size of 11×11, primarily encoding low-level and structural lung patterns.

$$f_g = \text{GAP}(x_g) \in \mathbb{R}^{1280\times1\times1} \cong \mathbb{R}^{1280}, \tag{7}$$

In Eq. 7, f_g represents the compact feature descriptor obtained through global average pooling (GAP), which converts the 6×6 spatial feature map into a one-dimensional vector by averaging across each channel. This process preserves global contextual information while reducing computational cost.

$$f_l = \text{GAP}(x_l) \in \mathbb{R}^{256\times1\times1} \cong \mathbb{R}^{256}, \tag{8}$$

In Eq. 8, f_l denotes the local feature vector derived from the custom CNN after applying GAP to its 11×11 feature maps, resulting in a 256-dimensional representation that captures fine-grained texture cues.

$$f_{\text{fusion}} = (f_l + f_g) \in \mathbb{R}^{1536}, \tag{9}$$

Eq. 9 defines f_{fusion} as the fused feature embedding obtained by adding the global and local descriptors along the channel dimension, forming a 1536-dimensional representation. While alternative strategies such as multi-scale or intermediate fusion were explored, simple feature addition was selected to preserve efficiency and maintain low model complexity, as further analyzed in the ablation experiments.

$$h_1 = \rho(W_1 f_{\text{fusion}} + b_1), \quad W_1 \in \mathbb{R}^{512\times1536}; b_1 \in \mathbb{R}^{512}, \tag{10}$$

Eq. 10 describes the first dense transformation, where ρ represents the ReLU activation function. This layer projects the fused vector into a latent space of 512 neurons to enhance feature discrimination.

$$\tilde{h}1 = \text{Dropout}(h1; p = 0.4), \tag{11}$$

In Eq. 11, dropout regularization with a probability of 0.4 is applied to prevent overfitting and stabilize training.

$$h_2 = \rho(W_2\tilde{h}1 + b2), \quad W_2 \in \mathbb{R}^{128\times512}; b_2 \in \mathbb{R}^{128}, \tag{12}$$

Eq. 12 denotes the second fully connected layer containing 128 neurons, which further refines the representation prior to the final classification stage.

$$\tilde{h}2 = \text{Dropout}(h2; p = 0.3), \tag{13}$$

Eq. 13 introduces another dropout layer with a reduced rate of 0.3 to maintain regularization while preserving learned features.

$$\hat{y} = \sigma(W_3\tilde{h}2 + b3), \quad W_3 \in \mathbb{R}^{1\times128}; b_3 \in \mathbb{R}^{1}, \tag{14}$$

In Eq. 14, $\hat{y}$ is the final output, representing the predicted probability for the classification. σ is the sigmoid activation function:

$$\sigma(z) = \frac{1}{1+e^{-z}} \tag{15}$$

Therefore, the final decision rule for classification is:

$$\hat{Y} \geq 0.5 \Rightarrow \text{COVID, else Non-COVID}$$

While the local feature extraction path employs a conventional hierarchical structure (Conv $\rightarrow$ BN $\rightarrow$ ReLU $\rightarrow$ MaxPool), the fundamental novelty lies in the synergistic parallel combination of pretrained MobileNetV2 and the Custom CNN via feature fusion. The Custom CNN itself was designed under a strict computational budget, resulting in a model with only 0.6 M parameters (which complements the MobileNetV2 branch to total 3.5M for the entire fusion model). This parsimonious design maximizes feature extraction for both local texture patterns and global context, ensuring low computational overhead. This leads to superior efficiency and an accuracy that outperforms deeper, more complex networks, as detailed in the results section.

Optimization and Hyperparameter

The hyperparameters were tuned using Optuna[64]. Optuna is an automatic hyperparameter optimization framework that efficiently searches for the best parameter combinations. A batch size of 16 was used to train all models on an NVIDIA Tesla P100 GPU (16GB VRAM). The AdamW optimizer with weight decay 1×10^{-4} was used to regularize and prevent overfitting. A small learning rate of 1×10^{-4} was selected for smooth convergence. `BCEWithLogitsLoss()` was used for binary classification, which was numerically stable and incorporated both sigmoid activation and binary cross-entropy in one function[65]. All models were trained for 50 epochs with a patience of 10 epochs. The same hyperparameters were used to train all the models.

Evaluation metrics

In this study, the models' performance was evaluated using six fundamental evaluation metrics: accuracy, precision, recall, F1 score, ROC-AUC, and Brier score, defined in Eqs. (16)–(20).

Accuracy is the ratio of the correctly predicted instances (true positives and true negatives) to the total number of instances, and is calculated using the equation:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{16}$$

where TP, TN, FP, and FN represent the number of true positives, true negatives, false positives, and false negatives, respectively.

Precision measures how accurate the positive predictions are, and is defined as:

$$\text{Precision} = \frac{TP}{TP+FP} \tag{17}$$

Recall reflects the model's ability to correctly identify positive instances, and is defined as:

$$\text{Recall} = \frac{TP}{TP+FN} \tag{18}$$

F1 Score is the harmonic mean of precision and recall, providing a balanced measure between them:

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{19}$$

The ROC curve illustrates the diagnostic ability of the model by plotting the true positive rate against the false positive rate at various threshold settings. The Area Under the Curve (AUC) quantifies the model's ability to discriminate between classes, with higher values indicating better overall performance.

The Brier score evaluates the accuracy of probabilistic predictions by computing the mean squared difference between the predicted probability and the actual binary outcome. It is calculated using the formula:

$$\text{Brier Score} = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2 \tag{20}$$

where N is the total number of predictions, f_t is the predicted probability for the t-th instance, and o_t is the actual outcome (0 or 1). A lower Brier score indicates better calibrated and more accurate probability estimates.

All together, these metrics allow for comprehensive assessment of the model's performance in distinguishing between COVID-19 and normal lung CT scan images.

Results

This section compares the performance of the pretrained CNN and Transformer models along with the Proposed CNN architecture for classifying COVID-19 using lung CT scans. The detailed classification report for all tested models is presented in Table 4, containing the mean and standard deviation of the relevant metrics across different runs with distinct random seeds. Analysis of the results shows that pretrained CNN models generally performed better than the pretrained Transformer models.

Table 4. Comparison of classification performance across pretrained CNNs, Transformers, and the proposed CNN model, showing mean and standard deviation for major evaluation metrics over multiple runs.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	AUC	Brier Score
Xception	92.85 ± 0.18	**96.75 ± 0.15**	87.99 ± 0.28	92.16 ± 0.21	0.984 ± 0.003	0.056 ± 0.004
ResNet50	96.90 ± 0.11	95.24 ± 0.18	98.43 ± 0.14	96.81 ± 0.12	0.992 ± 0.002	0.025 ± 0.003
MobileNet	96.10 ± 0.14	95.16 ± 0.19	96.75 ± 0.15	95.95 ± 0.13	0.989 ± 0.003	0.030 ± 0.003
Swin	96.57 ± 0.12	94.11 ± 0.21	99.02 ± 0.09	96.50 ± 0.11	0.993 ± 0.002	0.026 ± 0.003
BEiT	93.75 ± 0.17	89.95 ± 0.26	98.43 ± 0.16	93.73 ± 0.19	0.984 ± 0.004	0.047 ± 0.004
DeiT	94.97 ± 0.16	91.28 ± 0.23	98.92 ± 0.11	94.95 ± 0.17	0.982 ± 0.003	0.037 ± 0.003
CaiT	95.77 ± 0.15	93.93 ± 0.20	97.44 ± 0.15	95.65 ± 0.14	0.986 ± 0.003	0.032 ± 0.003
Proposed CNN	**97.46 ± 0.07**	96.00 ± 0.06	**99.13 ± 0.06**	**97.53 ± 0.06**	**0.994 ± 0.002**	**0.022 ± 0.002**

Across all tested models, the proposed CNN outperformed others in most evaluation metrics, achieving $97.46 \pm 0.07\%$ accuracy and the lowest Brier score of 0.022 ± 0.002.

Trade-offs in performance between the pretrained models were also observed. For example, the Xception model had the highest precision but the lowest recall among CNNs. The training dynamics of the pretrained CNN models (ResNet50, Xception, and MobileNetV2) are illustrated in Figures 4 and 5, which show the accuracy and loss curves, respectively. Both ResNet50 and Xception converged to their optimal validation performance relatively quickly, whereas MobileNetV2 exhibited a slower convergence rate. Late overfitting was evident for all CNN models, consisting of a continuous drop in training loss while

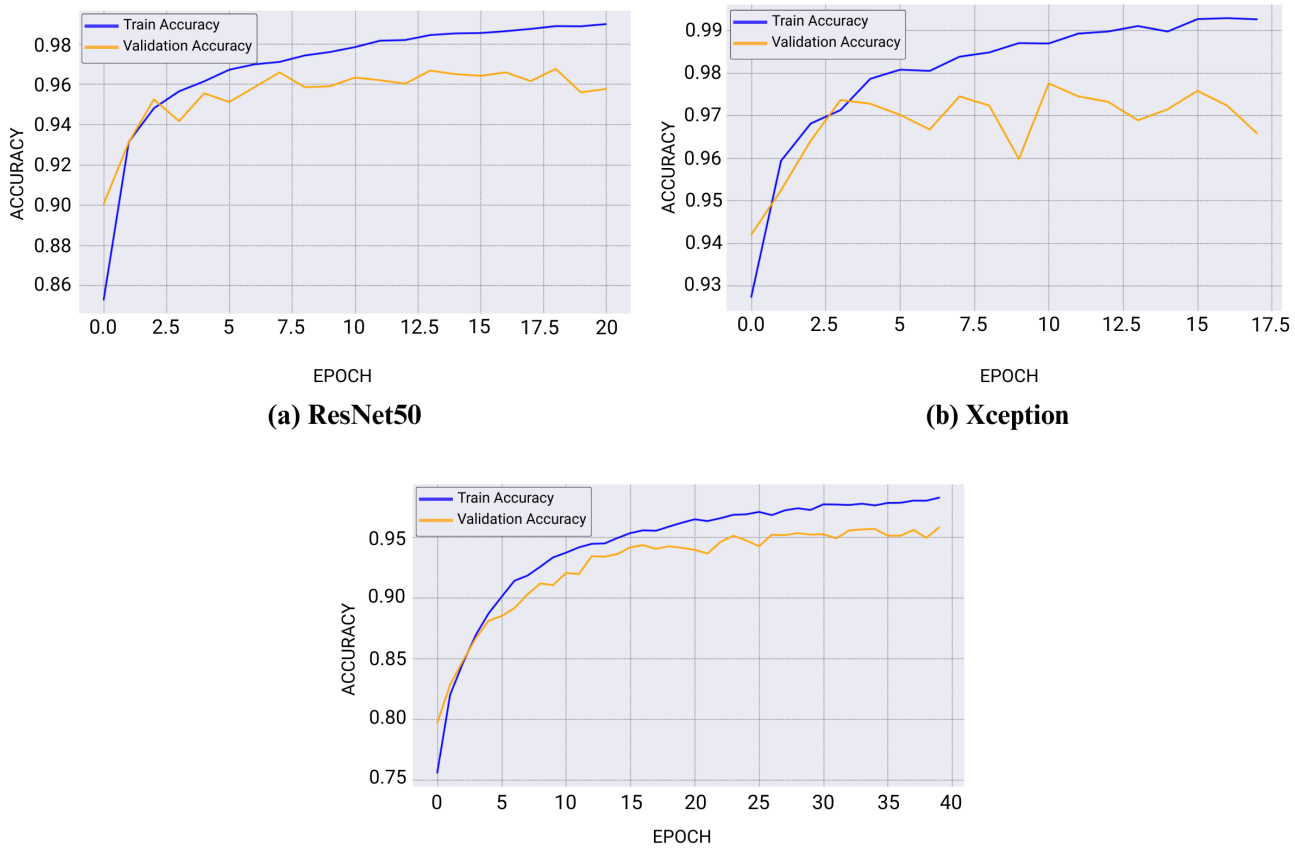

Figure 4. Training and validation accuracy curves for pretrained CNN models, illustrating convergence behavior and performance stability over epochs.

validation loss plateaued or even increased. For testing, the weights with the lowest validation loss were selected, representing each model's best performance point. The confusion matrices in Figure 6 further illustrate class-specific prediction patterns. ResNet50 produced significantly fewer false positives (16) than Xception (122), which is consistent with its superior Recall. As shown in Table 4, ResNet50 achieved the highest pretrained CNN accuracy ($96.9 \pm 0.11\%$) and an AUC of 0.992 ± 0.002.

The learning behavior of the Transformer models (Swin, BEiT, DeiT, and CaiT) is visualized in Figure 7 (accuracy) and Figure 8 (loss). The BEiT model required more epochs to converge compared to the other Transformers. Also, both the CNNs and the Transformer models showed overfitting in later epochs, demonstrating a persistent gap between the training and validation curves. The classification errors are further visualized based on the confusion matrices in Figure 9. The BEiT model showed the highest number of false negatives (111), suggesting that it was biased toward predicting the "NON-COVID" class, while the Swin Transformer showed the least (63). Based on Table 4, the CaiT outperformed the DeiT and BEiT, reaching an accuracy of $95.77 \pm 0.15\%$.

The training behavior of the proposed CNN is shown in Figure 10. The model rapidly reached its lowest validation loss at epoch 9, after which the validation loss began to increase, indicating the onset of overfitting. Therefore, the model state at epoch 9 was saved for the final evaluation. The confusion matrix and ROC curve for the proposed model are presented in Figure 11, showing a strong diagonal distribution with only 8 false positives and 45 false negatives. The model achieved a near-perfect AUC of 0.994, underscoring its superior quantitative performance as also highlighted in Table 4.

A comparative ROC curve analysis of all pretrained models is shown in Figure 12, illustrating that the CNN models generally enclose a larger area than the Transformer models. Finally, Figures 13 and 14 present the calibration curves for the pretrained and proposed models, respectively, evaluating the reliability of their confidence scores. The pretrained models exhibit varying degrees of slight miscalibration, while the calibration curve of the proposed model aligns most closely with the

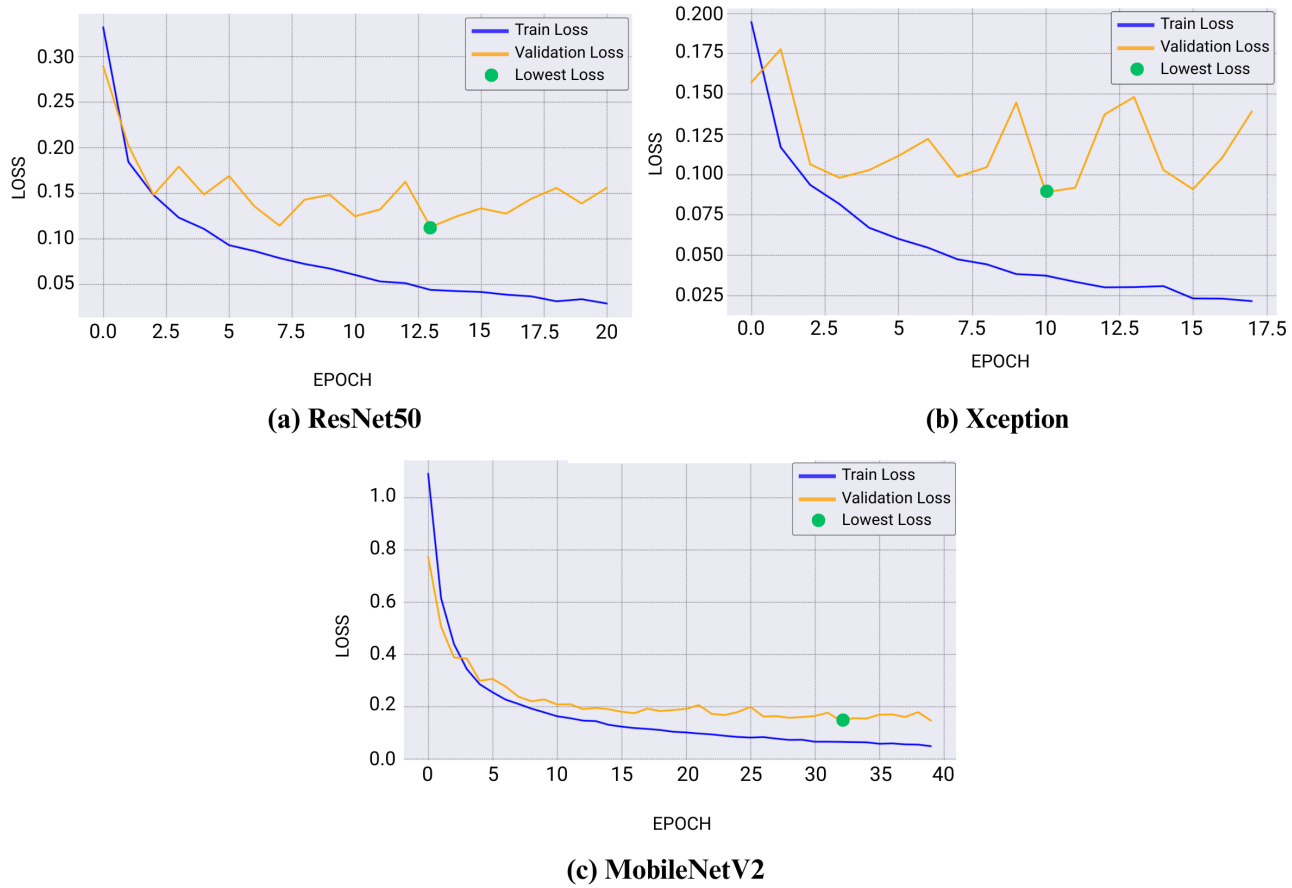

Figure 5. Training and validation loss curves of pretrained CNN models, highlighting convergence behavior and overfitting patterns observed in later epochs. The lowest validation loss point is marked with a green circle.

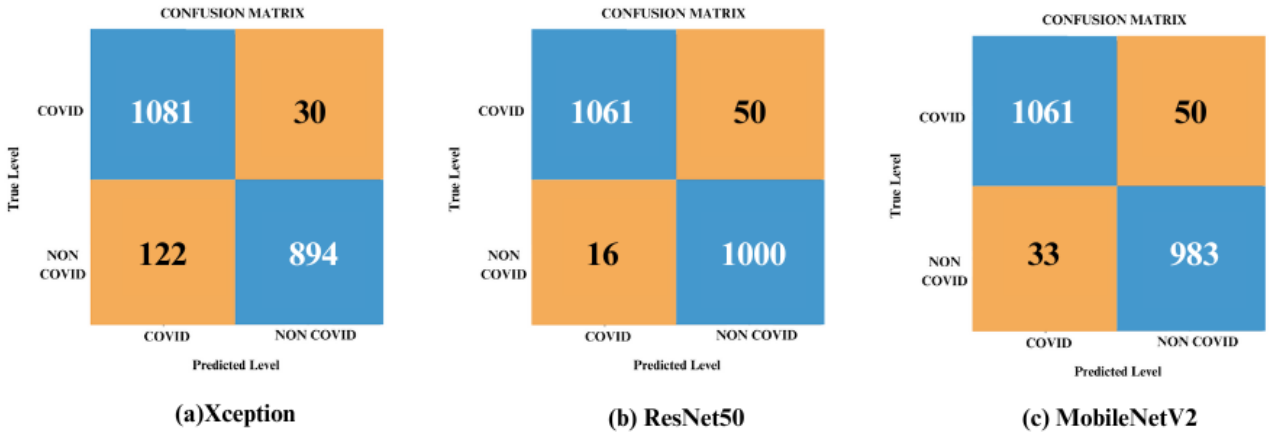

Figure 6. Confusion matrices depicting per-class performance of CNN models, where ResNet50 demonstrates notably fewer false positives compared to Xception.

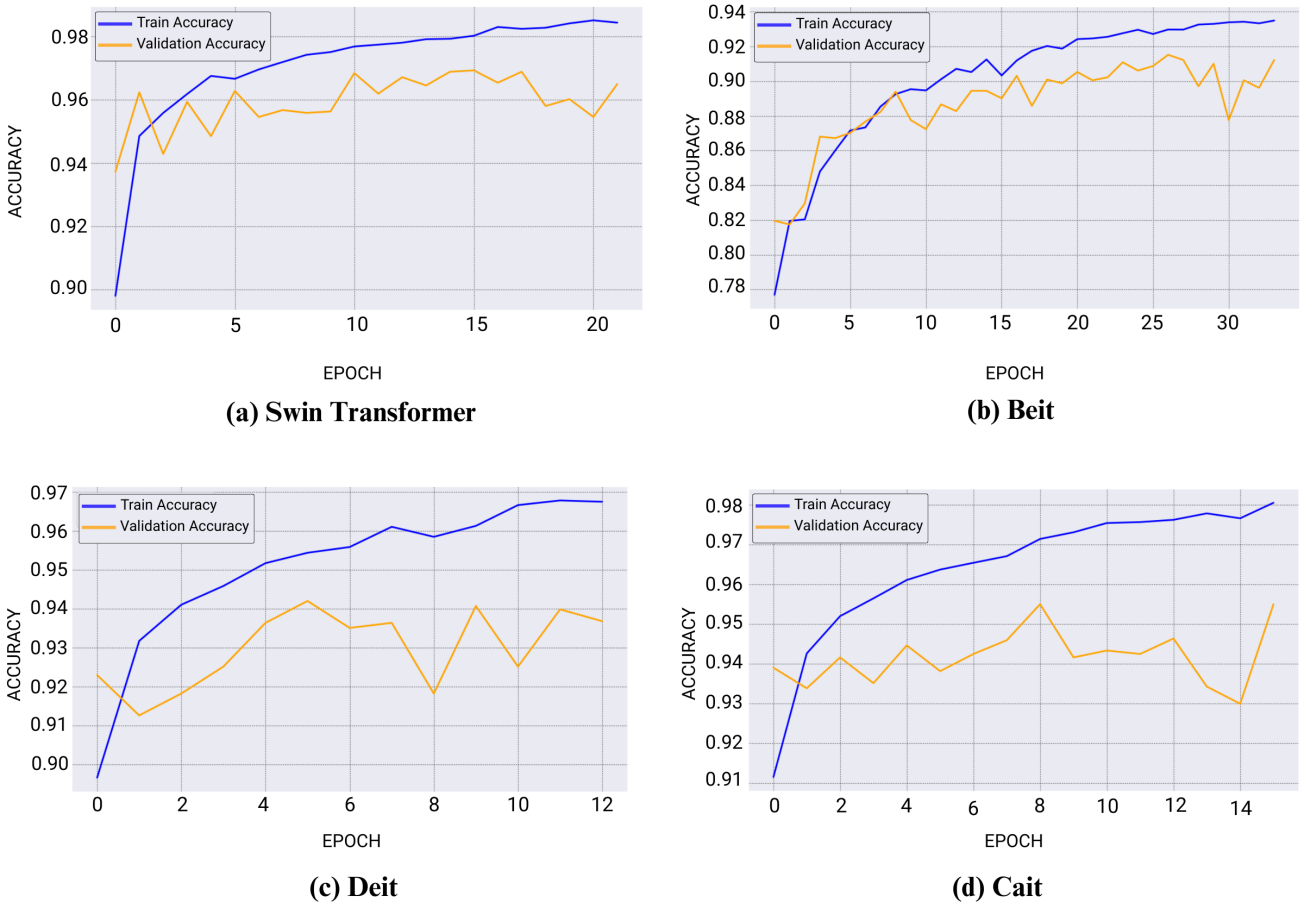

Figure 7. Training and validation accuracy trends of pretrained Transformer models, illustrating convergence rates and generalization performance.

diagonal "Perfect" line. This observation, supported by its lowest Brier score (0.021) in Figure 14, confirms that the proposed model produces the most reliable probability estimates.

Additionally, the proposed model preserves a lightweight structure and computational efficiency comparable to other efficient architectures such as MobileNet. It achieves similar complexity while ensuring faster inference and shorter training time, as shown in Table 5. All pretrained models were trained with ImageNet weights using a standardized input size of 224×224 pixels, which is conventional practice for leveraging pretrained weights effectively. The transformer architectures were specifically designed for this input dimension and do not support variable image sizes. By contrast, the custom CNN was designed to accommodate 180×180 pixel inputs for efficiency purposes. This difference in input dimensions may influence the inference speed and training time comparisons in Table 5. However, for metrics such as GFLOPs, which depend on input size, identical dummy input dimensions were used across all models to ensure consistent evaluation. In conclusion, the proposed CNN model provides exceptional performance while requiring less computational cost.

Results on X-ray Images

To assess the generalization ability of the proposed CNN model with respect to a different modality from lung CT scans, an additional experiment with chest X-ray images was performed. In particular, only the proposed CNN model was fine-tuned using the weights learned from the CT-based training and subsequently tested on the COVID-19 Radiography Dataset. The performance of the model on this dataset was promising, with test accuracy, precision, recall, and F1-score being 98.48%, 98.2%, 96.08%, and 97.13%, respectively. This further strengthens the performance of the model across different modalities of medical imaging. Figure 15 shows the loss curve and confusion matrix of the proposed model on the X-ray dataset. Furthermore, a comparison (Table 6) of the model with other studies based on the same dataset was provided to demonstrate the effectiveness

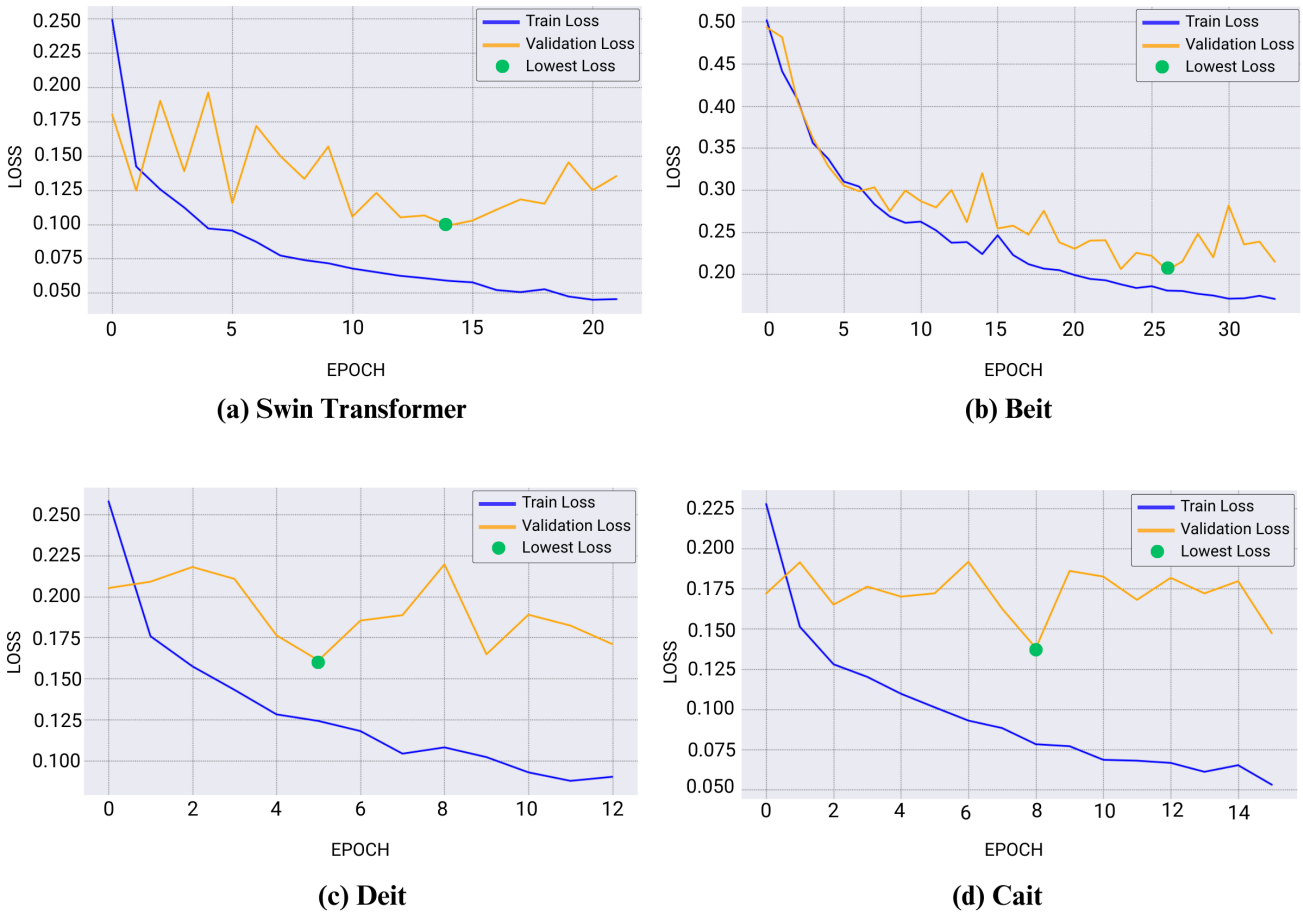

Figure 8. Loss curve comparison of Swin, BEiT, DeiT, and CaiT, showing distinct convergence patterns and overfitting tendencies in later epochs. The optimal validation loss points are marked with green circles.

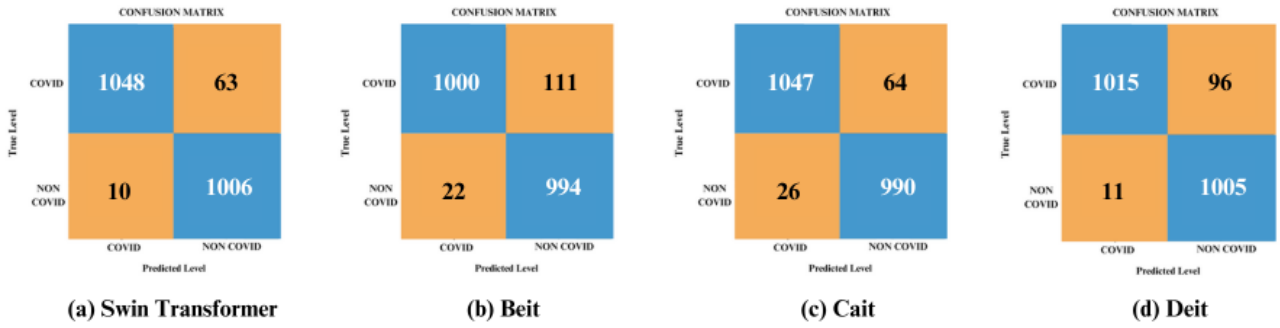

Figure 9. Confusion matrices for Transformer models showing per-class performance differences.

of the model.

Significance test

To establish the statistical significance of the proposed CNN model's performance, McNemar's test was employed for pairwise comparison against the fine-tuned pretrained architectures. This non-parametric test is appropriate for assessing differences between two classifiers on the same set of samples, using the misclassification counts of the discordant pairs. Here, n_{10}

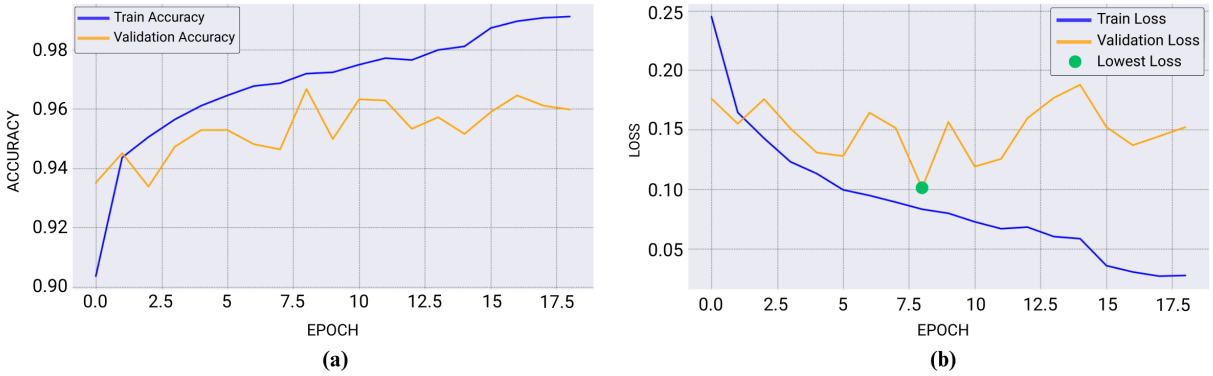

Figure 10. Training and validation accuracy and loss curves of the proposed CNN model showing rapid convergence and onset of overfitting after epoch 9. The lowest validation loss, used for model selection, is indicated with a green circle.

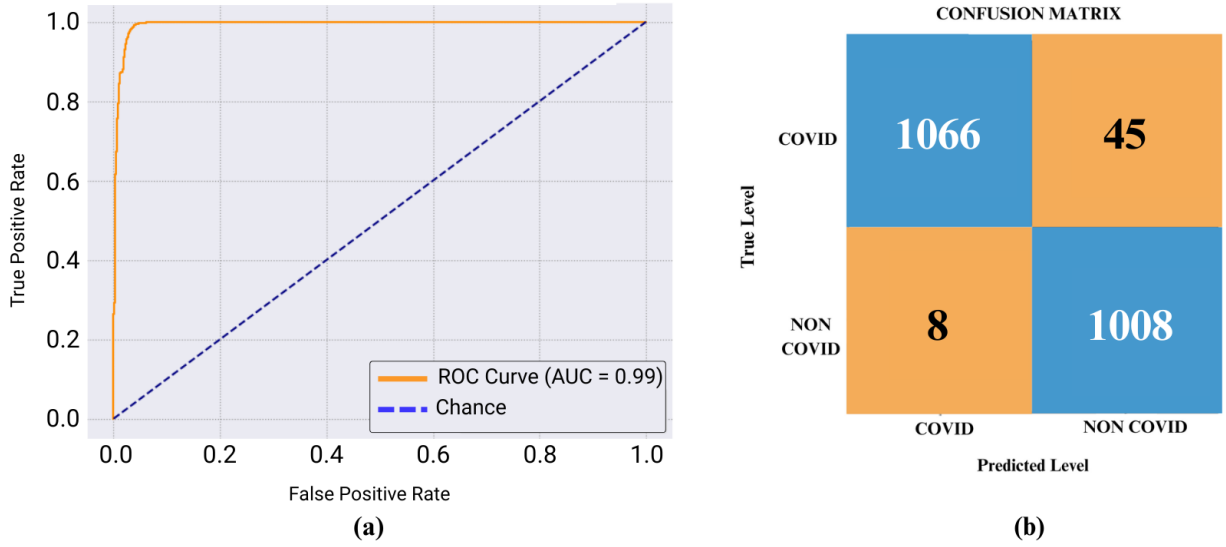

Figure 11. Confusion matrix and ROC analysis of the proposed CNN showing strong classification boundaries with minimal false predictions and near-perfect separability.

Table 5. Comparison of computational efficiency across all models, highlighting the lightweight structure and fast inference of the proposed CNN.

Model	Size (MB) ↓	Parameters (M) ↓	GFLOPs ↓	Inference Time (s) ↓	Training Time (min) ↓
Xception	79.67	20.81	2.896	0.1054	81
MobileNet	16.90	**2.23**	**1.002**	0.0496	135
ResNet50	89.99	23.51	2.900	0.0861	75
BEiT	327.22	64.39	16.866	0.3557	286
DeiT	327.35	64.39	16.867	0.3427	61.21
CaiT	177.66	46.44	9.346	0.2725	158
Swin	331.02	86.60	15.467	0.4092	156
Proposed CNN	**13.48**	3.5	1.05	**0.0059**	**37.44**

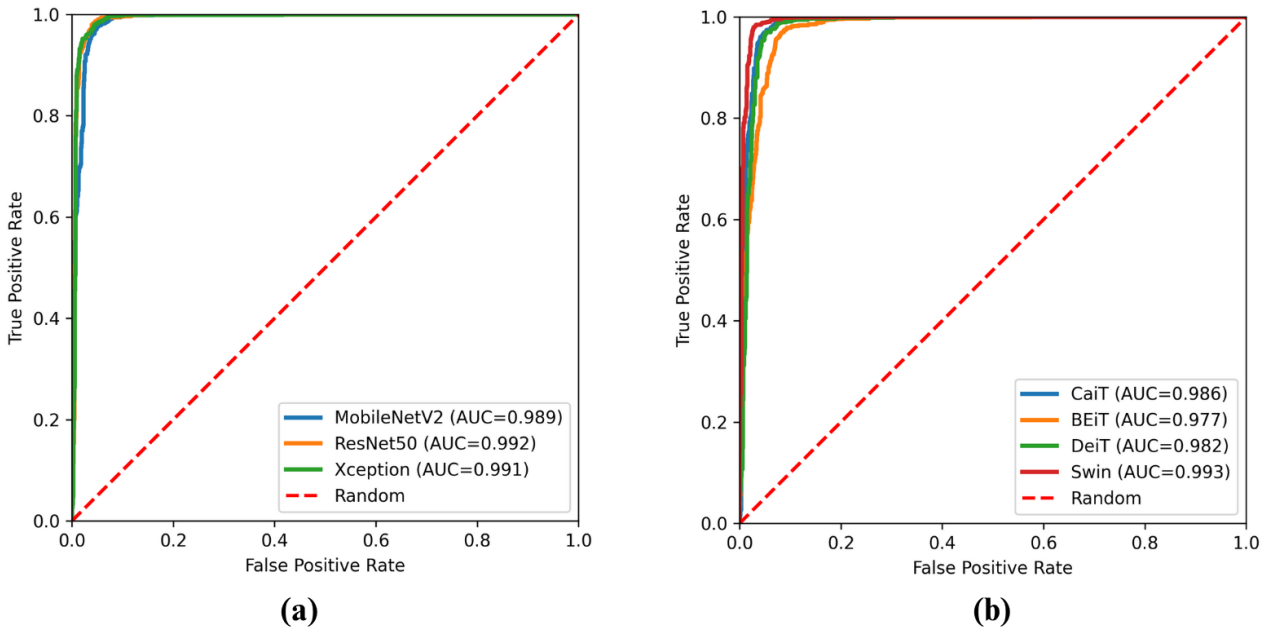

Figure 12. Combined ROC curves comparing discriminative performance of all pretrained CNN and Transformer models.

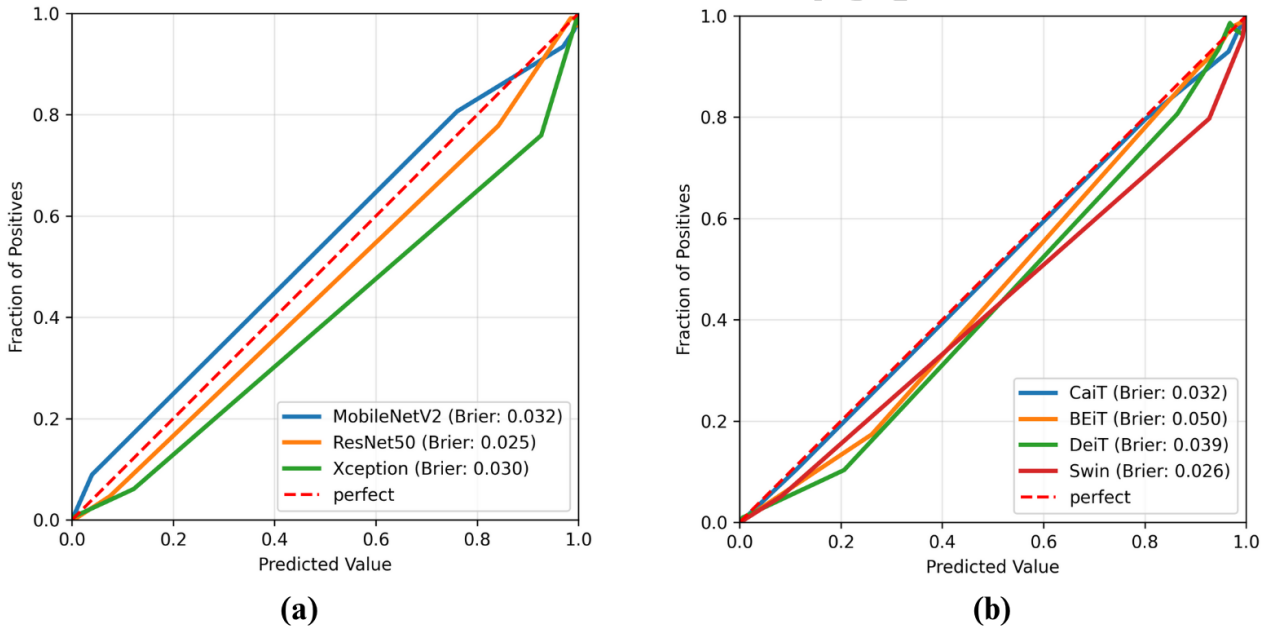

Figure 13. Reliability plots showing calibration performance of pretrained CNNs and Transformers, indicating varying degrees of probability alignment.

denotes the number of samples correctly classified by the proposed model but misclassified by the competitor, while n_{01} represents the opposite case. The null hypothesis (H_0) posits that the two models have equivalent error rates. As detailed in Table 7, the calculated χ^2 statistic and corresponding p-values show that the proposed CNN model achieves a statistically significant improvement ($p < 0.05$) when compared to Xception, MobileNet, Swin, BEiT, DeiT, and CaiT. Notably, the proposed CNN model demonstrates the most substantial improvement over BEiT ($\chi^2 = 52.889$, $n_{10} = 99$, $n_{01} = 19$) and Xception ($\chi^2 = 61.172$, $n_{10} = 128$, $n_{01} = 29$). The only comparison where the improvement is not statistically significant is with ResNet50 ($p = 0.091$), although the proposed CNN model maintains a numerical advantage in accuracy (97.41% vs. 96.90%). Furthermore, the proposed CNN model achieves this performance efficiently. It uses 3.5 million parameters and

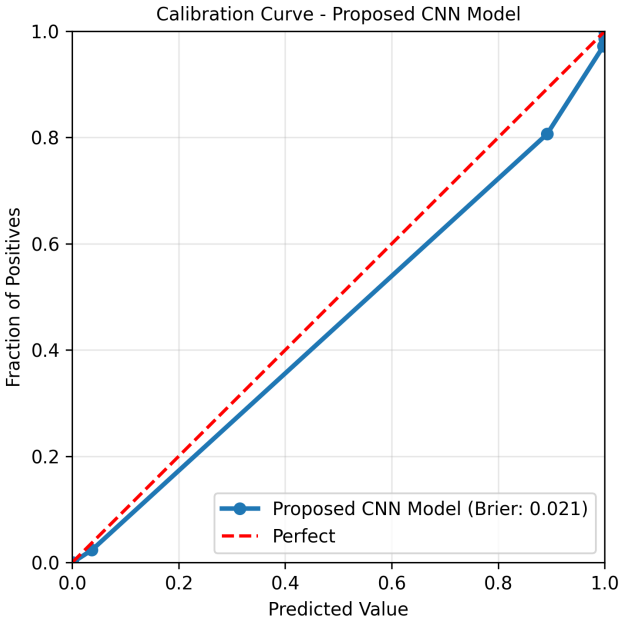

Figure 14. Calibration curve of the proposed CNN model showing near-perfect alignment with the ideal diagonal, indicating well-calibrated confidence estimates.

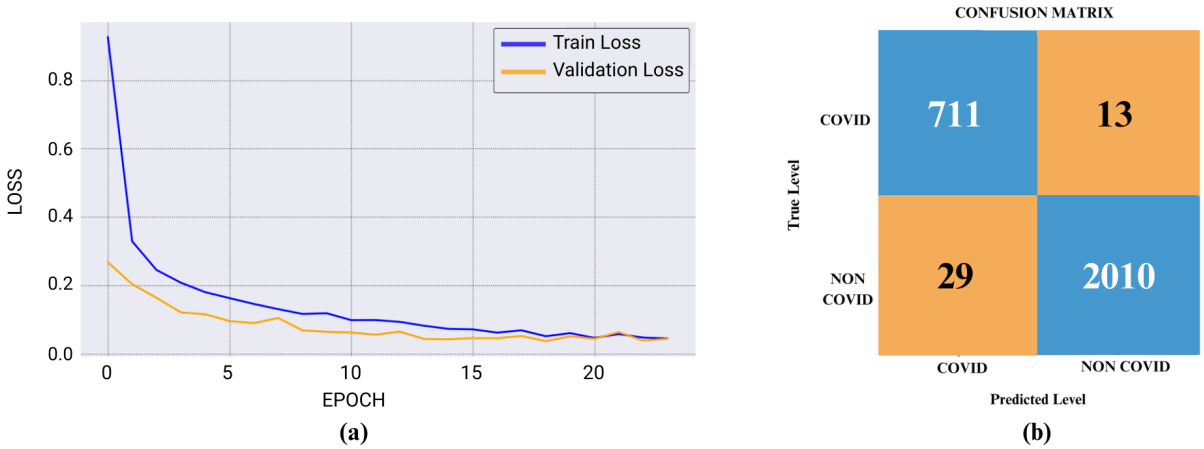

Figure 15. Training loss curve and confusion matrix of the proposed CNN on the COVID-19 Radiography X-ray dataset, illustrating strong classification consistency across classes.

requires 1.05 GFLOPs, with an inference time of 0.0059 seconds, while ResNet50 has 23.51 million parameters, requires 2.9 GFLOPs, and takes nearly 15 times longer (0.0861 seconds). These results confirm that the proposed architecture provides an efficient and competitive alternative.

Explainable AI

Explainable Artificial Intelligence (XAI) is employed to enhance complex ML models, making them more explainable to human beings. In the domain of medical imaging, XAI helps determine whether models are truly recognizing medically pertinent features or merely relying on spurious correlations. According to Samek et al.[72], explainable AI is crucial for building trust in AI, especially in sensitive application areas like healthcare. Doshi-Velez et al.[73] also emphasize explainable models for responsibility and clinical adoption. Two of the most popular XAI approaches are Gradient-weighted Class Activation Mapping (Grad-CAM) and Local Interpretable Model-Agnostic Explanations (LIME). Both were used in this study. The Grad-CAM introduced by Selvaraju et al.[74] produces heatmaps highlighting certain areas of an input image that are most critical to the predictions made by a model. LIME, by Ribeiro et al.[75], gives understandable local approximations of a model's prediction by

Table 6. Comparison of reported performances from recent studies using the COVID-19 Radiography dataset, showing that the proposed Custom Parallel CNN achieves accuracy comparable to or exceeding other state-of-the-art methods.

Author	Method	Database	Performance Parameters (%)
Hafeez et al. (2022) [66]	CODISC-CNN	COVID-19 Radiography Database	Accuracy: 97.2
Sahin et al. (2022) [67]	Custom CNN	COVID-19 Radiography Database	Accuracy: 96.71 F1 Score: 97%
Niloy et al. (2024) [68]	CovRoot (custom 42-layer CNN)	COVID-19 Radiography Database	Accuracy: 93.33
El Houby et al. (2024) [69]	VGG19 + CLAHE	COVID-19 Radiography Database	Accuracy: 95 Recall: 96 Specificity: 94
Bani Baker et al. (2024) [70]	Xception + Enhancement	COVID-19 Radiography Database	Accuracy: 98.13 Precision: 98.14 Recall: 97.65 F1 Score: 97.89
Wang et al. (2024) [71]	Dense MobileNetV3	COVID-19 Radiography Database	Accuracy: 98.71 Precision: 98.74 Recall: 97.78
Our Study	Proposed CNN (Custom CNN + MobileNetV2)	COVID-19 Radiography Database	Accuracy: 98.48 Precision: 98.2 Recall: 96.08 F1 Score: 97.13

Table 7. McNemar's Test Results Comparing the Proposed Model with Different Pretrained Architectures. Here, n_{10} denotes the number of samples correctly classified by the proposed model but misclassified by the compared model, while n_{01} represents the opposite case.

Model	n_{10}	n_{01}	Discordant Sum	χ^2 Statistic	p-value	Significant ($\alpha = 0.05$)
Xception	128	29	157	61.172	<0.001	Yes
ResNet50	40	27	67	2.149	0.091	No
MobileNet	58	28	86	9.779	<0.001	Yes
Swin	38	18	56	6.446	0.011	Yes
BEiT	99	19	118	52.889	<0.001	Yes
DeiT	72	18	90	31.211	<0.001	Yes
CaiT	62	25	87	14.896	<0.001	Yes

modifying input features and studying their influence. In the LIME visualizations, green regions show parts that reinforce the model's prediction, and red regions show parts that counter it.

To assess the interpretability of the model's predictions, two CT scan images were selected from a radiology journal[76]. In these scans, a radiologist had annotated the presence of ground-glass opacities (GGOs) using red arrows (Figure 16 a). When

visualized using Grad-CAM and LIME, the attention maps indicated that both methods focused on the regions corresponding to the GGOs (Figure 16 b and Figure 16 c, respectively). This alignment with radiologist-marked regions suggests that the

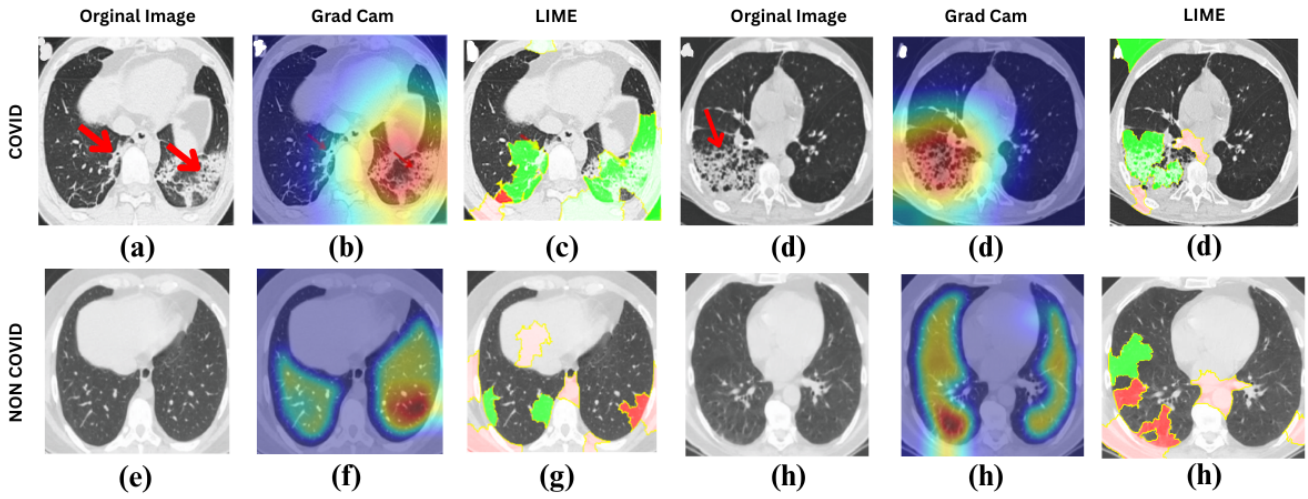

Figure 16. Visualization of model interpretability using Grad-CAM and LIME for CT scan images. Both methods focus on ground-glass opacity regions identified by radiologists, confirming that the model's attention aligns with clinically relevant features.

model's decision-making process is consistent with expert knowledge. In contrast, similar CT images without GGOs were also analyzed, and the model correctly predicted them as non-COVID. In this case, the Grad-CAM and LIME explanations primarily highlighted non-specific regions within the lungs, further supporting the model's ability to differentiate between COVID-positive and COVID-negative cases based on relevant radiological features.

Ablation Study

An ablation study was conducted to evaluate the contribution of each component and the effectiveness of different feature fusion strategies. The performance of each variant is detailed in Table 8. The standalone MobileNetV2 baseline achieved an

Table 8. Ablation analysis of model variants, including standalone baselines and feature fusion schemes between the Custom CNN and MobileNetV2 branches. The late fusion method attains the highest classification accuracy of 97.51%.

Variant	Acc (%)	Pre (%)	Recall (%)	F1 (%)
MobileNet Only	96.10	95.16	96.75	95.95
Custom CNN Only	90.74	83.70	98.30	90.42
Mid-Level Fusion	97.18	95.68	98.88	97.25
Multi-Scale Fusion	97.08	95.60	98.80	97.17
Late Fusion (Proposed)	97.51	95.95	99.26	97.57

accuracy of 96.10%, confirming the effectiveness of its pre-trained features. In contrast, the custom CNN alone achieved a lower accuracy of 90.74%, highlighting its limitations as a standalone model. However, its strength was evident in its high recall of 98.30%. To leverage the complementary traits of MobileNetV2 (high precision) and the Custom CNN (high recall), several fusion strategies were evaluated. A mid-level fusion scheme achieved an accuracy of 97.18%, while a multi-scale fusion approach yielded a comparable result of 97.08%. Ultimately, the proposed late fusion method, which aggregates the final predictions, achieved the best performance across all metrics with an accuracy of 97.51% and a recall of 99.26%. These results confirm that late fusion most effectively combines the strengths of both branches, yielding superior and more robust classification performance.

Limitation and Future Work

Compared to fine-tuned pretrained CNN and transformer models, the proposed CNN model achieved the best accuracy and the best F1 score. However, there are also a few drawbacks that should be further investigated in the future. Cross-dataset validation was not performed in this study. Since the primary aim was to benchmark multiple models under consistent training conditions and to design a lightweight, effective CNN architecture, the focus remained on achieving strong intra-dataset performance first. Future work will include cross-dataset experiments to better assess real-world generalizability across imaging centers and different populations. Furthermore, to better understand the model's clinical utility, a blinded evaluation comparing its diagnostic performance with that of expert radiologists will be prioritized. Evaluating the model's robustness and reliability in safety-critical healthcare settings will require a systematic examination of its responses to adversarial perturbations and edge cases.

Conclusion

An efficient deep learning architecture for COVID-19 classification using lung CT scans is presented, utilizing a large dataset of about 25,000 images collated from nine different sources. Extensive benchmarking on pre-trained state-of-the-art CNN and Vision Transformer models showed that the designed CNN model attained the highest classification accuracy among the tested models. The CNN model has only 3.5 million parameters and uses low computational resources. The model also achieved reliable performance when fine-tuned on a chest X-ray dataset. Interpretability methods (Grad-CAM and LIME) were used to visualize model decisions and increase transparency. Most notably, the generalizability of the model, especially its robustness across diverse CT datasets, requires further validation through external, cross-dataset testing. Rigorous external validation across independent, multi-center CT datasets is still required to fully establish robustness and generalizability before clinical adoption. Future work will prioritize such cross-dataset evaluations and prospective testing to confirm the model's clinical utility.

Acknowledgement

The authors would like to thank Green University of Bangladesh, Universiti Malaysia Pahang Al-Sultan Abdullah, American International University-Bangladesh, and Qassim University for providing a collaborative platform and their technical facilities, computational resources, and research infrastructure.

Author Contribution

Md. Mahid Arfan Rahat (M.M.A.R.) conceived the study, designed the methodology, conducted the experiments, analyzed the data, and wrote the original draft. Md. Imamul Islam (M.I.I.) contributed to the conceptualization, methodology development, data validation, and manuscript review and editing. Md Saef Ullah Miah (M.S.U.M.) assisted with data analysis, visualization, and manuscript preparation. Talal Alharbi (T.A.) supervised the project, provided critical review and editing of the manuscript, secured funding, and administered the research project. All authors reviewed and approved the final manuscript.

Competing interests

The author(s) declare no competing interests.

Funding

The Researchers would like to thank the Deanship of Graduate Studies and Scientific Research at Qassim University for financial support (QU-APC-2025).

Data Availability

The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

References

1. Zhu, N. *et al.* A novel coronavirus from patients with pneumonia in china, 2019. *New Engl. journal medicine* **382**, 727–733 (2020).
2. Bartha, I. *et al.* Morbidity of sars-cov-2 in the evolution to endemicity and in comparison with influenza. *Commun. Medicine* **4**, 244 (2024).

3. Cavalli, M. *et al.* Next generation sequencing of multiple sars-cov-2 infections in the omicron era. *Sci. Reports* **15**, 3372 (2025).
4. Kwee, T. C. & Kwee, R. M. Chest ct in covid-19: what the radiologist needs to know. *Radiographics* **40**, 1848–1865 (2020).
5. Liu, X.-P. *et al.* Development and validation of chest ct-based imaging biomarkers for early stage covid-19 screening. *Front. Public Heal.* **10**, 1004117 (2022).
6. Santosh, K., GhoshRoy, D. & Nakarmi, S. A systematic review on deep structured learning for covid-19 screening using chest ct from 2020 to 2022. In *Healthcare*, vol. 11, 2388 (MDPI, 2023).
7. Sahin, M. E. Deep learning-based approach for detecting covid-19 in chest x-rays. *Biomed. Signal Process. Control.* **78**, 103977 (2022).
8. Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
9. Li, M., Jiang, Y., Zhang, Y. & Zhu, H. Medical image analysis using deep learning algorithms. *Front. public health* **11**, 1273253 (2023).
10. Wu, L. & Wang, H. Global and pyramid convolutional neural network with hybrid attention mechanism for hyperspectral image classification. *Geocarto Int.* **38**, 2226112 (2023).
11. Khan, S. H. *et al.* Transformers in vision: A survey. *ACM Comput. Surv. (CSUR)* **54**, 1 – 41, DOI: 10.1145/3505244 (2021).
12. Chen, X. *et al.* Cmfuse: a hierarchical feature fusion model combining convolution and mamba for medical image classification. *Clust. Comput.* **28**, 662 (2025).
13. Huo, X. *et al.* Hifuse: Hierarchical multi-scale feature fusion network for medical image classification. *Biomed. Signal Process. Control.* **87**, 105534 (2024).
14. Hussain, S. S. *et al.* A swin transformer and cnn fusion framework for accurate parkinson disease classification in mri. *Sci. Reports* **15**, 15117 (2025).
15. Mahanty, C. *et al.* A comprehensive review on covid-19 detection based on cough sounds, symptoms, cxr and ct images. *IEEE Access* (2024).
16. Elmuogy, S., Hikal, N. A. & Hassan, E. An efficient technique for ct scan images classification of covid-19. *J. Intell. & Fuzzy Syst.* **40**, 5225–5238 (2021).
17. Rahimzadeh, M., Attar, A. & Sakhaei, S. M. A fully automated deep learning-based network for detecting covid-19 from a new and large lung ct scan dataset. *Biomed. Signal Process. Control.* **68**, 102588 (2021).
18. Salama, W. M. & Aly, M. H. Framework for covid-19 segmentation and classification based on deep learning of computed tomography lung images. *J. Electron. Sci. Technol.* **20**, 100161 (2022).
19. Soni, M., Singh, A. K., Babu, K. S., Kumar, S. *et al.* Convolutional neural network based ct scan classification method for covid-19 test validation. *Smart Heal.* **25**, 100296 (2022).
20. Islam, M. K., Rahman, M. M., Ali, M. S., Miah, M. S. & Rahman, M. H. An enhanced technique of covid-19 detection and classification using deep convolutional neural network from chest x-ray and ct images. *BioMed Res. Int.* **2023**, 6341259 (2023).
21. Foysal, M., Hossain, A. A., Yassine, A. & Hossain, M. S. Detection of covid-19 case from chest ct images using deformable deep convolutional neural network. *J. healthcare engineering* **2023**, 4301745 (2023).
22. Haennah, J. J., Christopher, C. S. & King, G. G. Prediction of the covid disease using lung ct images by deep learning algorithm: Dets-optimized resnet 101 classifier. *Front. Medicine* **10**, 1157000 (2023).
23. Rajinikanth, V. *et al.* Covid-19 detection in lung ct slices using brownian-butterfly-algorithm optimized lightweight deep features. *Heliyon* **10** (2024).
24. Kordnoori, S., Sabeti, M., Mostafaei, H. & Banihashemi, S. S. A. A deep learning framework for accurate covid-19 classification in ct-scan images. *Mach. Learn. with Appl.* 100628 (2025).
25. Rajpoot, R., Gour, M., Jain, S. & Semwal, V. B. Integrated ensemble cnn and explainable ai for covid-19 diagnosis from ct scan and x-ray images. *Sci. Reports* **14**, 24985 (2024).

26. Nikam, N., Ganorkar, S. & Raut, V. Very deep convolutional networks based transfer learning approach for sars-cov-2 recognition from chest ct images. *J. Integr. Sci. Technol.* **13**, 1009–1009 (2025).

27. Chowa, S. S. *et al.* An automated privacy-preserving self-supervised classification of covid-19 from lung ct scan images minimizing the requirements of large data annotation. *Sci. Reports* **15**, 226 (2025).

28. Ferraz, A. & Betini, R. C. Comparative evaluation of deep learning models for diagnosis of covid-19 using x-ray images and computed tomography. *J. Braz. Comput. Soc.* **31**, 99–131 (2025).

29. Taye, G. D., Sisay, Z. H., Gebeyhu, G. W. & Kidus, F. H. Thoracic computed tomography (ct) image-based identification and severity classification of covid-19 cases using vision transformer (vit). *Discov. Appl. Sci.* **6**, 384, DOI: 10.1007/s42452-024-06048-0 (2024).

30. et al., A. E. E. R. Conditional cascaded network (ccn) approach for rapid automated diagnosis of covid-19 in chest x-ray and ct images using transfer learning. *Comput. Biol. Medicine* DOI: 10.1016/S1746-8094(23)00996-5 (2024).

31. Ghaffar, Z. *et al.* Comparative analysis of state-of-the-art deep learning models for detecting covid-19 lung infection from chest x-ray images. *arXiv preprint arXiv:2208.01637* (2022).

32. Fouad, S. *et al.* Explained deep learning framework for covid-19 detection in volumetric ct images aligned with the british society of thoracic imaging reporting guidance: A pilot study. *J. Imaging Informatics Medicine* DOI: 10.1007/s10278-025-01444-3 (2025).

33. Liu, Z. & Shen, L. Cect: Controllable ensemble cnn and transformer for covid-19 image classification. *Comput. Biol. Medicine* **173**, 108388, DOI: 10.1016/S0010-4825(24)00472-4 (2024).

34. Maftouni, M. *et al.* A robust ensemble-deep learning model for covid-19 diagnosis based on an integrated ct scan images database. In *IIE annual conference. Proceedings*, 632–637 (Institute of Industrial and Systems Engineers (IISE), 2021).

35. Soares, E., Angelov, P., Biaso, S., Froes, M. H. & Abe, D. K. Sars-cov-2 ct-scan dataset: A large dataset of real patients ct scans for sars-cov-2 identification. *MedRxiv* 2020–04 (2020).

36. Ghaderzadeh, M. *et al.* Deep convolutional neural network–based computer-aided detection system for covid-19 using multiple lung scans: design and implementation study. *J. Med. Internet Res.* **23**, e27468 (2021).

37. Rahimzadeh, M., Attar, A. & Sakhaei, S. M. A fully automated deep learning-based network for detecting covid-19 from a new and large lung ct scan dataset. *Biomed. Signal Process. Control.* **68**, 102588 (2021).

38. Segmentation, M. Covid-19 ct segmentation dataset (2025). Accessed: April 8, 2025.

39. Ma, J. & et al. Covid-19 ct lung and infection segmentation dataset, DOI: 10.5281/zenodo.3757476 (2020). Accessed: April 8, 2025.

40. Cohen, J. P. *et al.* Covid-19 image data collection: Prospective predictions are the future. *arXiv preprint arXiv:2006.11988* (2020).

41. Morozov, S. P. *et al.* Mosmeddata: Chest ct scans with covid-19 related findings dataset. *arXiv preprint arXiv:2005.06465* (2020).

42. Yang, X. *et al.* Covid-ct-dataset: a ct scan dataset about covid-19. *arXiv preprint arXiv:2003.13865* (2020).

43. Afshar, P. *et al.* Covid-ct-md, covid-19 computed tomography scan dataset applicable in machine learning and deep learning. *Sci. Data* **8**, 121 (2021).

44. Aria, M., Ghaderzadeh, M., Asadi, F. & Jafari, R. Covid-19 lung ct scans: A large dataset of lung ct scans for covid-19 (sars-cov-2) detection. *Kaggle. URL: https://www. kaggle. com/mehradaria/covid19-lung-ct-scans [accessed 2021-04-20]* (2021).

45. Chowdhury, M. E. *et al.* Can ai help in screening viral and covid-19 pneumonia? *Ieee Access* **8**, 132665–132676 (2020).

46. Rahman, T. *et al.* Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images. *Comput. biology medicine* **132**, 104319 (2021).

47. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016).

48. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1251–1258 (2017).

49. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4510–4520 (2018).

50. Deng, J. *et al.* Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255 (Ieee, 2009).

51. Szegedy, C. *et al.* Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–9 (2015).

52. Vaswani, A. *et al.* Attention is all you need. *Adv. neural information processing systems* **30** (2017).

53. Touvron, H. *et al.* Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, 10347–10357 (PMLR, 2021).

54. Bao, H., Dong, L., Piao, S. & Wei, F. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254* (2021).

55. Liu, Z. *et al.* Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022 (2021).

56. Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G. & Jégou, H. Going deeper with image transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 32–42 (2021).

57. Wightman, R. Pytorch image models. *GitHub repository* DOI: 10.5281/zenodo.4414861 (2019).

58. Maurício, J., Domingues, I. & Bernardino, J. Comparing vision transformers and convolutional neural networks for image classification: A literature review. *Appl. Sci.* **13**, 5521 (2023).

59. Zhao, Y. *et al.* A battle of network structures: An empirical study of cnn, transformer, and mlp. *arXiv preprint arXiv:2108.13002* (2021).

60. Khan, M. *et al.* Iomt-enabled computer-aided diagnosis of pulmonary embolism from computed tomography scans using deep learning. *Sensors* **23**, 1471 (2023).

61. Shah, P. M., Zeb, A., Shafi, U., Zaidi, S. F. A. & Shah, M. A. Detection of parkinson disease in brain mri using convolutional neural network. In *2018 24th international conference on automation and computing (ICAC)*, 1–6 (IEEE, 2018).

62. Hussain, S. S. *et al.* Classification of parkinson's disease in patch-based mri of substantia nigra. *Diagnostics* **13**, 2827 (2023).

63. Glorot, X., Bordes, A. & Bengio, Y. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 315–323 (JMLR Workshop and Conference Proceedings, 2011).

64. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2623–2631 (2019).

65. PyTorch. *BCEWithLogitsLoss* (2025). Documentation for the BCEWithLogitsLoss class in PyTorch.

66. Hafeez, U. *et al.* A cnn based coronavirus disease prediction system for chest x-rays. *J. Ambient Intell. Humaniz. Comput.* **14**, 13179–13193 (2023).

67. Sahin, M. E. Deep learning-based approach for detecting covid-19 in chest x-rays. *Biomed. Signal Process. Control.* **78**, 103977 (2022).

68. Niloy, A. H. *et al.* Covroot: Covid-19 detection based on chest radiology imaging techniques using deep learning. *Front. Signal Process.* **4**, 1384744 (2024).

69. El Houby, E. M. Covid-19 detection from chest x-ray images using transfer learning. *Sci. Reports* **14**, 11639 (2024).

70. Bani Baker, Q. *et al.* Enhanced covid-19 detection from x-ray images with convolutional neural network and transfer learning. *J. Imaging* **10**, 250 (2024).

71. Wang, S., Ren, J. & Guo, X. A high-accuracy lightweight network model for x-ray image diagnosis: A case study of covid detection. *Plos one* **19**, e0303049 (2024).

72. Samek, W., Wiegand, T. & Müller, K.-R. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296* (2017).

73. Doshi-Velez, F. & Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

74. Selvaraju, R. R. *et al.* Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626 (2017).

75. Ribeiro, M. T., Singh, S. & Guestrin, C. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144 (2016).

76. Machnicki, S. *et al.* The usefulness of chest ct imaging in patients with suspected or diagnosed covid-19: a review of literature. *Chest* **160**, 652–670 (2021).