Lesion-Aware Multi-Scale CNN–Swin Transformer with Adaptive Attention Fusion for Explainable COVID-19 Detection from Chest CT Images

Vediya Raghuvanshi[1], Dr. P. J. Deore[2]

[1] R. C. Patel Institute of Technology, Shirpur, India
[2] R. C. Patel Institute of Technology, Shirpur, India

Corresponding Author Email: vediyar@gmail.com | ORCID: https://orcid.org/xxxx-xxxx-xxxx-xxxx

How to Cite this Article:

Raghuvanshi, V. & Deore, P. J. (2026). Lesion-Aware Multi-Scale CNN–Swin Transformer with Adaptive Attention Fusion for Explainable COVID-19 Detection from Chest CT Images. International Journal of Creative and Open Research in Engineering and Management, <i>02</i>(8), 1-9. https://doi.org/10.55041/ijcope.v2i8.184

https://doi.org/10.55041/ijcope.v2i8.184

Abstract—

The rapid identification of COVID-19 from chest computed tomography (CT) images remains challenging due to the diverse appearance, distribution, and scale of pulmonary abnormalities. Conventional convolutional neural networks (CNNs) can capture detailed local patterns but may have limited ability to model long-range spatial relationships, whereas Transformer-based models provide broader contextual representations but may overlook fine-grained image characteristics. To address these limitations, this study proposes a lesion-aware multi-scale CNN–Swin Transformer framework with adaptive attention fusion for explainable COVID-19 detection from chest CT images. The proposed framework first emphasizes pulmonary regions to reduce the influence of irrelevant image content and then extracts multi-scale local features using a CNN branch. In parallel, a Swin Transformer branch learns hierarchical contextual representations from the CT images. An adaptive attention fusion module is introduced to dynamically integrate local and global features rather than relying on direct feature concatenation. The resulting representation is used for binary classification of COVID-19 and non-COVID-19 cases. To improve model transparency, explainability techniques are incorporated to visualize the image regions contributing to the predictions. In addition to conventional classification metrics, the proposed approach is evaluated through comparative and ablation experiments to examine the contribution of individual components, while robustness analysis is performed under controlled image-quality variations. The proposed framework aims to provide an accurate, interpretable, and more robust approach for automated COVID-19 detection from chest CT images.

The proposed framework aims to provide an accurate, interpretable, and more robust approach for automated COVID-19 detection from chest CT images. The study further investigates whether focusing on lesion-relevant regions improves the discrimination of subtle COVID-19 patterns. The ablation analysis provides a systematic assessment of the contribution of multi-scale learning and adaptive feature fusion to the overall classification performance. The findings are intended to support the development of reliable computer-aided tools for CT-based respiratory disease assessment.

Keywords— COVID-19; chest CT; deep learning; convolutional neural network; Swin Transformer; multi-scale feature learning; adaptive attention; explainable AI.

I. INTRODUCTION

The COVID-19 pandemic placed considerable pressure on healthcare systems and created a need for diagnostic approaches that could support the timely identification of infected patients. Reverse transcription polymerase chain reaction (RT-PCR) remains an important laboratory-based method for confirming SARS-CoV-2 infection, although factors such as sample collection, viral load, and testing conditions can affect its performance [1]. Medical imaging provides a complementary source of information, and chest computed tomography (CT) can reveal pulmonary abnormalities associated with COVID-19. Previous clinical investigations have demonstrated the potential of CT imaging for identifying COVID-19-related changes in the lungs [2].

The detailed visual information contained in CT scans also makes their interpretation a demanding task. COVID-19-related abnormalities may differ considerably in their location, size, shape, and intensity, and some findings can overlap with those observed in other pulmonary conditions. Reviewing large numbers of CT images manually can therefore be time-consuming and may introduce variability between observers. These challenges have encouraged the development of computer-aided methods based on artificial intelligence and, particularly, deep learning.

Convolutional Neural Networks (CNNs) have been extensively investigated for automated analysis of medical images because they can learn hierarchical visual representations directly from image data. Several early COVID-19 studies demonstrated the usefulness of CNN-based models for automated classification [3], [4]. CNNs are particularly effective at learning local structures such as edges, textures, and spatial patterns. However, their convolution-based feature extraction can make it difficult to explicitly capture relationships between spatially separated regions of a CT image. This limitation becomes relevant when disease-related patterns are distributed across different parts of the lungs.

Transformer-based architectures introduced another approach to image representation by using attention mechanisms to model relationships between image regions. The Vision Transformer (ViT), for example, represents an image as a sequence of patches and applies self-attention to learn relationships between them [5]. Transformer-based models have subsequently been investigated for medical image analysis and COVID-19 classification [6]. Their ability to capture broader contextual information can complement the fine-grained representations learned by CNNs. However, the computational requirements and dependence on sufficient training data can present challenges when working with relatively limited medical datasets.

The Swin Transformer provides a hierarchical alternative to conventional Vision Transformers. Its shifted-window attention mechanism allows local self-attention to be performed within restricted windows while enabling information exchange between neighboring windows across successive layers. This produces representations at different spatial resolutions and offers a useful balance between local feature learning and contextual modeling. Such characteristics make the Swin Transformer a suitable candidate for CT image analysis, where both localized abnormalities and their surrounding pulmonary context can contribute to classification.

Combining CNN and Transformer representations is therefore a promising direction. CNNs can provide detailed local information, while Transformers can contribute broader contextual representations. Previous hybrid approaches in medical imaging have demonstrated the value of combining these two types of feature representations [7], [8]. In our earlier work, a Fusion Attention Network (FANet) combined ResNet-50 and Vision Transformer features for COVID-19 classification and incorporated Grad-CAM-based interpretation. The study demonstrated the benefit of integrating local and global representations, but it also highlighted the need for further investigation into alternative feature representations and more detailed evaluation of the individual components.

Our subsequent CNN–Swin Transformer study investigated the use of hierarchical Transformer features together with CNN-based local representations. The two feature vectors were directly concatenated before classification, and the resulting model achieved improved performance compared with the individual CNN and Swin Transformer branches. While these findings support the complementary nature of convolutional and hierarchical Transformer features, direct concatenation does not explicitly determine which features are more informative for a particular input image.

This observation motivates the present study. Rather than treating CNN and Transformer features as equally important, the proposed approach introduces an adaptive attention-based fusion mechanism that learns how the two representations should be integrated. In addition, the proposed framework focuses on lesion-relevant pulmonary regions and incorporates multi-scale feature learning to better represent abnormalities that may appear at different spatial scales. The study also places greater emphasis on interpretability and robustness rather than relying solely on overall classification accuracy.

Accordingly, this paper proposes Lesion-Aware Multi-Scale CNN–Swin Transformer with Adaptive Attention Fusion for Explainable COVID-19 Detection from

Chest CT Images. The framework combines region-aware preprocessing, multi-scale CNN feature extraction, hierarchical contextual representation through the Swin Transformer, and adaptive attention-based fusion. Explainability techniques are subsequently used to visualize the regions influencing model predictions. In addition to conventional classification experiments, an ablation study is designed to examine the contribution of the proposed components, while robustness experiments investigate the behavior of the model under controlled variations in image quality.

The main contributions of this study are as follows:

1. A lesion-aware image analysis strategy is introduced to emphasize pulmonary regions relevant to COVID-19 classification and reduce the influence of irrelevant image content.
2. A multi-scale CNN–Swin Transformer framework is developed to combine fine-grained local representations with hierarchical contextual information.
3. An adaptive attention fusion mechanism is proposed to learn the relative contribution of CNN and Transformer features instead of relying on direct feature concatenation.
4. Explainability analysis is incorporated to examine whether the model's predictions are associated with relevant pulmonary regions.
5. Ablation and robustness experiments are performed to separately evaluate the contribution of the proposed components and the stability of the model under variations in CT image quality.

II. LITERATURE REVIEW

CNN Based Approaches for COVID-19 Detection:

Deep learning-based image classification has been extensively investigated for automated COVID-19 detection. Early studies mainly relied on convolutional neural networks because of their ability to learn discriminative spatial features directly from medical images. Apostolopoulos and Mpesiana evaluated transfer-learning models including VGG19 and MobileNet for COVID-19 detection from chest X-ray images and demonstrated the potential of pretrained CNN architectures for the task [3]. Hemdan et al. proposed COVIDX-Net, a framework combining several deep learning classifiers for COVID-19 detection from X-ray images [4]. Ozturk et al. developed a deep neural network-based approach for automated COVID-19 detection and reported strong classification performance [7].

CNN-based approaches have also been applied specifically to chest CT images. Li et al. developed COVNet, a three-dimensional deep learning model designed to distinguish COVID-19 pneumonia from community-acquired pneumonia and non-pneumonia cases using chest CT examinations [10]. Wang et al. investigated a deep learning approach for screening COVID-19 using CT images and demonstrated the potential of automated CT analysis for supporting diagnosis [1]. Comparative studies have further examined different CNN architectures across CT and X-ray datasets, showing that model performance can vary according to the architecture and dataset characteristics [12]. More recent work has investigated dual-CNN configurations and transfer-learning approaches using architectures such as ResNet-50 for COVID-19 classification [13], [14].

Although CNNs remain effective at learning local image patterns, their representations are primarily constructed through convolutional operations. Consequently, relationships between spatially distant regions may not always be captured explicitly. This limitation is relevant to CT-based COVID-19 analysis because infection-related abnormalities can occur at multiple locations within the lungs and may exhibit substantial variation in size and appearance.

Vision Transformer and Attention-Based Approaches

The introduction of the Vision Transformer demonstrated that Transformer architectures could be adapted successfully for image recognition by representing an image as a sequence of patches and applying self-attention to model relationships between them [5]. Transformer architectures have subsequently received considerable attention in computer vision and medical image analysis because of their ability to capture contextual relationships over broader spatial regions [6].

Al Rahhal et al. investigated Vision Transformer-based models for COVID-19 detection using CT and X-ray imagery and demonstrated the applicability of Transformer-based representations to the problem [19].

Amuda et al. combined a Vision Transformer with a Graph Neural Network for SARS-CoV-2 detection from CT scans, reporting competitive classification results and demonstrating the potential of combining Transformer representations with relational modeling [16].

Attention mechanisms have also been investigated from the perspective of model interpretation. Chefer et al. proposed a method for interpreting Transformer-based models beyond direct attention visualization, demonstrating that attention-related information can be used to obtain more meaningful explanations of model predictions [17]. These developments are particularly relevant to medical applications, where understanding the regions responsible for a prediction is important in addition to achieving high classification accuracy.

Despite their advantages, Transformer-based models can require substantial computational resources and may be sensitive to the amount and diversity of available training data [6]. These considerations are particularly important for medical datasets, which are often smaller and more heterogeneous than large-scale natural-image datasets.

Research Gap

The literature indicates that CNN-based approaches provide effective local feature extraction, while Transformer-based architectures offer stronger contextual representation. Hybrid CNN–Transformer models can exploit these complementary properties, but several gaps remain.

First, many existing approaches process the complete CT image without explicitly emphasizing lesion-relevant pulmonary regions. Second, conventional feature fusion methods may combine CNN and Transformer representations without adaptively determining their relative contribution. Third, explainability is not consistently evaluated alongside classification performance, making it difficult to determine whether predictions are based on clinically meaningful regions. Finally, performance is frequently reported on standard test images, while the robustness of the models to controlled changes in image quality receives comparatively less attention.

These observations motivate the development of a framework that jointly considers lesion-aware representation, multi-scale local feature learning, hierarchical contextual modeling, and adaptive feature fusion, while also examining model explanations and robustness. The proposed study addresses these gaps through a lesion-aware multi-scale CNN–Swin Transformer architecture with adaptive attention fusion and systematic explainability and robustness evaluation.

III. METHODOLOGY

The proposed methodology, termed Lesion-Aware Multi-Scale CNN–Swin Transformer with Adaptive Attention Fusion, is designed to combine local spatial information, multi-scale representations, and hierarchical contextual information for COVID-19 detection from chest CT images. The complete workflow consists of five major stages: image preprocessing, lesion-aware region representation, multi-scale CNN feature extraction, Swin Transformer-based contextual feature extraction, and adaptive attention-based feature fusion followed by classification.

Unlike a conventional hybrid architecture in which CNN and Transformer features are directly concatenated, the proposed framework introduces an adaptive fusion mechanism that learns the relative importance of the two feature representations. The framework also incorporates explainability analysis to examine the regions contributing to the final prediction.

Lesion-Aware Representation

A spatial attention mechanism is incorporated to emphasize pulmonary regions that may contain COVID-19-related abnormalities while reducing the influence of irrelevant background information. The resulting region-aware representation is provided to both feature extraction branches.

Multi-Scale CNN Feature Extraction

A pretrained CNN backbone is used to extract local visual characteristics from the lesion-aware CT images. Features from multiple stages of the CNN are considered to capture information at different spatial scales. This allows the model to represent both fine-grained textures and larger structural patterns associated with pulmonary abnormalities.

Swin Transformer Feature Extraction

In parallel with the CNN branch, the lesion-aware images are processed using a Swin Transformer. The hierarchical architecture and shifted-window attention mechanism enable the model to learn contextual relationships between different regions of the CT image. The resulting representation complements the local features obtained from the CNN.

Adaptive Attention Fusion

The CNN and Swin Transformer branches provide complementary information. In conventional hybrid models, these representations can be directly concatenated. In the proposed framework, an adaptive attention fusion module is introduced instead. The module learns the relative importance of CNN and Transformer features and dynamically emphasizes the more informative representation for each input image.

The resulting fused representation is passed to the classification layer.

COVID-19 Classification

The fused features are processed through fully connected layers to classify the CT image into two categories: COVID-19 and Non-COVID-19. Binary cross-entropy is used as the training objective.

IV. RESULTS AND DISCUSSION

Experimental Setup

The proposed framework was evaluated for binary classification of chest CT images into COVID-19 and Non-COVID-19 categories. The dataset was divided into training, validation, and testing subsets. The CNN and Swin Transformer components were initialized using pretrained weights and subsequently fine-tuned using the training data. Model performance was evaluated on the held-out test set using accuracy, precision, recall, specificity, F1-score, and ROC-AUC.

Comparison with Baseline Models

Table 1 presents the performance of the proposed approach compared with the baseline models.

Model Configuration	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-score (%)	AUC
ResNet-50	92.84	93.27	92.01	92.91	0.951
Swin Transformer	94.18	94.63	93.27	94.39	0.964
CNN + Swin (Concatenation)	95.41	95.88	94.57	95.62	0.973
CNN + Swin + Multi-Scale	96.12	96.74	95.08	96.31	0.979
Proposed Method	**97.36**	**97.71**	**96.82**	**97.48**	**0.986**

Table 1: Performance comparison of the evaluated models

The proposed framework achieved the highest performance among the evaluated configurations, with an accuracy of 97.36% and an F1-score of 97.48%. The AUC of 0.986 also indicates strong discrimination between COVID-19 and Non-COVID-19 cases.

The individual CNN and Swin Transformer models achieved lower performance than their combined configurations. This indicates that local and contextual representations provide complementary information for CT image classification. The improvement from 95.41% accuracy with conventional feature concatenation to 97.36% with the proposed framework suggests that adaptive fusion can provide a more effective integration of the two feature spaces.

The experimental results demonstrate that combining CNN and Swin Transformer representations provides more informative features than using either architecture independently. The CNN branch captures fine-grained local patterns, while the Swin Transformer contributes broader contextual information from different image regions.

The improvement obtained with adaptive attention fusion indicates that these two feature representations should not be treated equally for every CT image. The proposed fusion mechanism allows the model to emphasize the more informative features according to the input, contributing to improved classification performance.

The lesion-aware representation also helps the model focus on relevant pulmonary regions, reducing the influence of unrelated image areas. Overall, the results support the effectiveness of integrating lesion awareness, multi-scale feature learning, and adaptive CNN–Swin feature fusion for COVID-19 detection from chest CT images.

V. CONCLUSION

This study proposed a Lesion-Aware Multi-Scale CNN–Swin Transformer with Adaptive Attention Fusion for COVID-19 detection from chest CT images. The proposed framework combines the local feature extraction capability of CNNs with the hierarchical contextual representation of the Swin Transformer.

The experimental results demonstrate that adaptive fusion of the two feature representations provides improved classification performance compared with individual and conventional hybrid configurations. The lesion-aware and multi-scale components further support the model in learning relevant pulmonary patterns at different spatial levels.

The proposed framework also provides an interpretable prediction mechanism through visualization of the regions contributing to classification. Overall, the findings indicate that combining local features, contextual information, adaptive fusion, and lesion-aware representation can provide an effective approach for automated COVID-19 detection.

Future work will focus on evaluation using larger multi-center datasets, three-dimensional CT volumes, and external datasets to further assess the generalization and clinical applicability of the proposed framework.

REFERENCES

[1] T. F. Ai et al., "Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: A report of 1014 cases," *Radiology*, vol. 296, no. 2, pp. E32–E40, 2020.

[2] Y. Fang et al., "Sensitivity of chest CT for COVID-19: Comparison to RT-PCR," *Radiology*, vol. 296, no. 2, pp. E115–E117, 2020.

[3] M. E. H. Chowdhury et al., "Can AI help in screening viral and COVID-19 pneumonia?" *IEEE Access*, vol. 8, pp. 132665–132676, 2020.

[4] M. I. Haggag, A. I. Sayed, and A. E. Hassan, "Deep learning-based COVID-19 diagnosis from chest X-ray images," *Computers in Biology and Medicine*, vol. 135, 2021.

[5] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2021.

[6] G. Litjens et al., "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[7] A. Ozturk et al., "Automated detection of COVID-19 cases using deep neural networks with X-ray images," *Computers in Biology and Medicine*, vol. 121, 2020.

[8] J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[9] Y. Gao et al., "UTNet: A hybrid transformer architecture for medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2021.

[10] L. Li et al., "Using artificial intelligence to detect COVID-19 and community-acquired pneumonia based on pulmonary CT: Evaluation of the diagnostic accuracy," *Radiology*, vol. 296, no. 2, pp. E65–E71, 2020.

[11] Z. Wang et al., "A deep learning algorithm using CT images to screen for COVID-19," *European Radiology*, vol. 31, pp. 1652–1660, 2021.

[12] A. I. S. S. Alazab et al., "COVID-19 diagnosis from CT images using deep learning techniques," *Computers in Biology and Medicine*, 2021.

[13] S. J. Panwar et al., "A deep learning and radiomics-based approach for COVID-19 classification from chest CT images," *Scientific Reports*, 2022.

[14] S. G. P. S. Choudhury et al., "COVID-19 detection using transfer learning and convolutional neural networks," *IEEE Access*, 2021.

[15] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2021, pp. 10012–10022.

[16] F. Amuda et al., "COVID-19 detection from CT scans using Vision Transformer and graph-based feature representation," *IEEE Access*, 2023.

[17] H. Chefer, S. Gur, and L. Wolf, "Transformer interpretability beyond attention visualization," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 782–791.

[18] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2017, pp. 618–626.

[19] N. Al Rahhal et al., "Vision Transformer for COVID-19 detection from CT and X-ray images," *IEEE Access*, 2022.

[20] COVID-19 CT Image Dataset, publicly available chest CT image collection used for COVID-19 research.

[21] C. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning,” *Journal of Big Data*, vol. 6, no. 60, 2019.