# RECURRENT 3-D MULTI-LEVEL VISUAL TRANSFORMER FOR JOINT CLASSIFICATION OF HETEROGENEOUS 2-D AND 3-D RADIOGRAPHIC DATA

*Muhammad Owais*[1], *Muhammad Zubair*[2], *Taimur Hassan*[3], *Divya Velayudhan*[4],
*Irfan Hussain*[1], *Naoufel Werghi*[4]

[1]KUCARS, Dept. of Mechanical and Nuclear Engineering, Khalifa University, UAE
[2]Faculty of IT & Computer Science, University of Central Punjab, Pakistan
[3]Dept. of Electrical, Computer and Biomedical Engineering, Abu Dhabi University, UAE
[4]C2PS, Dept. of Computer Science, Khalifa University, UAE

## ABSTRACT

Recent advancements in artificial intelligence algorithms for medical imaging show significant potential in automating the detection of lung infections from chest radiograph scans. However, current approaches often focus solely on either 2-D or 3-D scans, failing to leverage the combined advantages of both modalities. Moreover, conventional slice-based methods place a manual burden on radiologists for slice selection. To overcome these challenges, we propose the Recurrent 3-D Multi-level Vision Transformer (R3DM-ViT) model, capable of handling multimodal data to enhance diagnostic accuracy. Our quantitative evaluations demonstrate that R3DM-ViT surpasses existing methods, achieving an impressive accuracy of 96.67%, F1-score of 96.88%, mean average precision of 96.75%, and mean average recall of 97.02%. This research signifies a significant stride forward in the automated detection of lung infections through multimodal imaging.

***Index Terms*—** lung infection, R3DM-ViT, CBMIR, Computer-aided diagnosis, Medical image retrieval.

## 1. INTRODUCTION

In recent years, the field of Artificial Intelligence (AI) has witnessed unprecedented advancements, opening up new frontiers in various domains, particularly in healthcare diagnostics. These developments have been most notably exemplified in the evolution of Computer-Aided Diagnosis (CAD) tools, which reshape the landscape of medical diagnostics. The integration of AI in healthcare has catalyzed a potential paradigm shift in how medical professionals approach diagnostic challenges [1]. Deep learning (DL) methodologies [2, 3] have significantly streamlined the evaluation and interpretation of medical data by physicians, thereby solidifying the role of CAD models as invaluable adjuncts in diagnosing lung infections. As a result, CAD imaging technologies have gained significant credibility in identifying lung infections. The ability of DL algorithms to extract essential features and generate predictions based on these features unambiguously shows their potential in medical image analysis. But even with all of this success, DL in medical image analysis is still a dynamic and ever-evolving field.

Chest computed tomography (CT) scans and chest X-ray (CXR) are the primary imaging modalities in lung infection screening and detection, favored for their consistent diagnostic manifestations [4]. The evolution of CAD solutions has seen a significant push towards augmenting the accuracy of diagnoses and treatment planning using these modalities, particularly chest radiographs such as CXR [5, 6] and CT scans [7, 8]. While CXR remains a common diagnostic method, a growing trend in research favors CT scans for their superior sensitivity and enhanced visualization capabilities.

A plethora of studies have delved into the detection of lung abnormalities using the (CXR and CT) imaging modalities, identifying multiple conditions, including the detection of nodules and tuberculosis and their screening. CXR-focused studies typically analyze 2-D input data, whereas CT scan research often involves selecting specific slices from the entire CT volume [9], a process that, while insightful, demands manual effort and time. The slice-selection approach also introduces limitations like limited spatial context, lack of robustness, and potential annotation bias. Conversely, some studies advocate for full CT volume analysis, citing improved results over slice-selection methods. Other studies have proposed models using multimodalities to enhance the results further. A growing amount of research suggests that CT and X-ray scans be used in tandem to improve the detection of lung infections [10], given their respective benefits. Nevertheless, a significant gap remains since no study has presented a model that can simultaneously integrate 2-D CXR and 3-D CT volume data. This integration is pivotal, considering the high sensitivity of CT scans and the widespread accessibility of X-ray. By synergizing the strengths of these two modalities, a more comprehensive and nuanced diagnostic

This work was supported by Khalifa University Center for Autonomous and Robotic Systems under Award RC1-2018-KUCARS. (Corresponding author: Muhammad Owais, E-mail: muhammad.owais@ku.ac.ae)

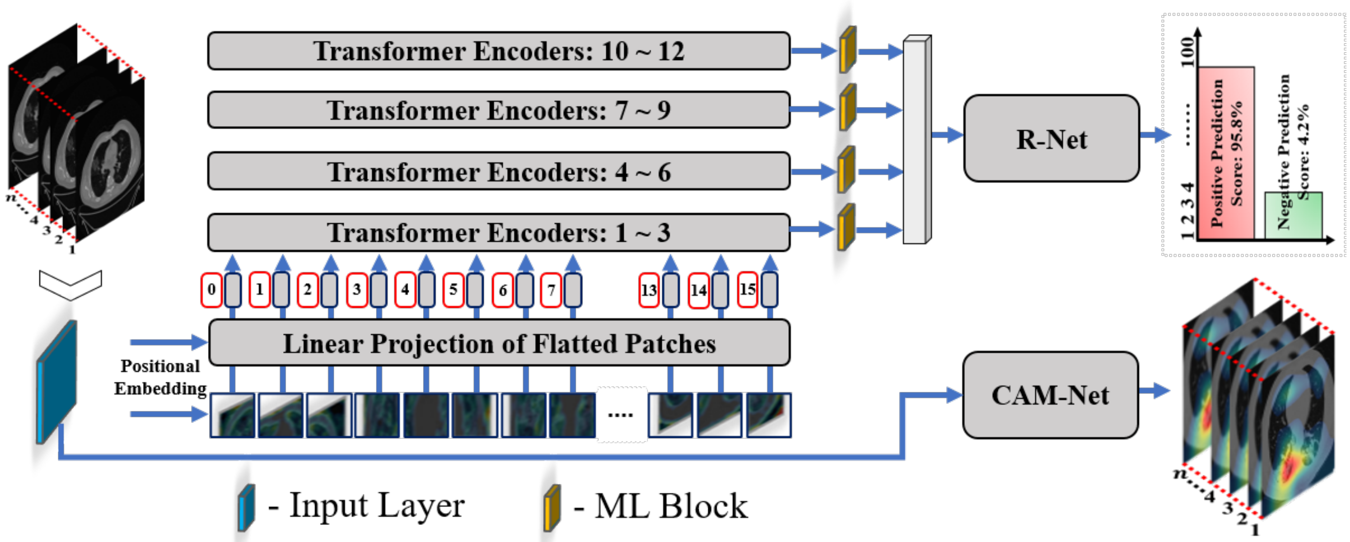
Transformer Encoders: 10 ~ 12
Transformer Encoders: 7 ~ 9
Transformer Encoders: 4 ~ 6
Transformer Encoders: 1 ~ 3
0
1
2
3
4
5
6
7
13
14
15
Linear Projection of Flatted Patches
Positional Embedding
....
R-Net
CAM-Net
Positive Prediction Score: 95.8%
Negative Prediction Score: 4.2%
100
......
1 2 3 4
n... 4 3 2 1
- Input Layer
- ML Block

**Fig. 1**. Overall workflow diagram of the proposed framework.

perspective can be achieved. To address this, we have proposed a novel model that is intended to support 2-D CXR or 3-D CT volume input or the simultaneous use of both modalities, offering a more thorough and sophisticated analytical approach to lung infection detection.

Our proposed versatile model is capable of handling data in various formats, including 2-D CXR images, 3-D CT volumes, and combinations of 2-D CXR images with either 2-D CT slices or 3-D CT volumes. The model demonstrates proficiency in analyzing two-dimensional CXR images, adeptness at processing three-dimensional CT volumes, and adaptability to diverse input formats, allowing for a comprehensive analysis of medical data. The key contribution of this study can be summarized in the following aspects:

- A Recurrent 3-D Multi-level Vision Transformer (R3DM-ViT) is proposed based on a Multi-level Vision Transformer (M-ViT) and recurrent model by utilizing the strength of multi-level feature aggregation.
- Our proposed model can jointly analyze 2-D X-ray and a variable length 3-D CT volumetric data.
- The proposed R3DM-ViT aggregates joint multi-level spatial and 3-D structural features in case of 3-D CT data to enhance overall diagnostic performance.
- The proposed framework will be made publicly accessible for further research, development, and educational purposes.

The subsequent sections of this paper are organized as follows: Section 2 explains the overall methodology of the proposed framework. Section 3 outlines the experimental configuration and results. Finally, Section 4 encapsulates the discussion and conclusion of this study.

## 2. PROPOSED METHOD

The proposed CAD methodology combines the capabilities of multiscale/multi-level feature fusion with the ability to diagnose using a single DL model. The overall architecture of the proposed model is presented in Fig. 1.

### 2.1. Workflow Overview

This research aims to create a deep classification model capable of categorizing multiclass medical data, encompassing 2-D CXR images, 3-D CT slices, and complete 3-D CT volumes. The R3DM-ViT model is intended to process heterogeneous radiographic 2-D and 3-D data simultaneously. Specifically designed to accommodate a variety of radiography inputs and take advantage of their complementary information for better analysis, this model yields improved results. The R3DM-ViT model efficiently extracts a range of features, from low-level $\mathbf{f}_1$ to high-level $\mathbf{f}_4$, encompassing the intricate nuances of radiographic data. These features are then fused and fed to a fully connected layer. Subsequently, the data is passed through a recurrent block, thereby enhancing the model's ability to capture temporal and spatial dependencies and finally classified.

The development of our model unfolds in two main phases: the training phase and the testing phase. Initially, we train an Imagenet pre-trained model with the training dataset, composed of n data samples, each accompanied by its respective class label, symbolized as $\langle [\mathbf{F}_T]_{x=1}^n, [l_T]_{x=1}^n \rangle$. This step is crucial for the model to exploit and learn the spatial characteristics inherent in the data. Next, after running each data sample through our trained model, all training data samples $[\mathbf{F}_T]_{x=1}^n$ were transformed into feature vectors $[\mathbf{f}_T]_{x=1}^n$. This transformation yields a redefined training dataset in the feature domain, denoted as $\langle [\mathbf{f}_T]_{x=1}^n, [l_T]_{x=1}^n \rangle$. The next step involves segregating the training data into 2-D and 3-D imaging categories, guided by the information encoded within each class label.

After the training phase, we evaluated the efficacy of our proposed classification framework using a distinct testing dataset, denoted as $\langle [\mathbf{F}_{T_p}]_{x=1}^n, [l_{T_p}]_{x=1}^n \rangle$. In the context of 2-D imagery, the trained model utilizes spatial characteristics to determine class predictions. For 3-D CT imaging data, the model enhances its overall effectiveness by tapping into 3-D anatomical dependencies, thereby achieving an increase in performance. Initially, the model processes each of the n consecutive slices $(\mathbf{F}^1, \mathbf{F}^2, ..., \mathbf{F}^n)$ in the sequence, transforming them into corresponding feature vectors $(\mathbf{f}^1, \mathbf{f}^2, ..., \mathbf{f}^n)$. These feature vectors are then parallel processed, enabling the model to harness additional 3-D anatomical information for more accurate class predictions.

### 2.2. Multi-Level ViT Model Structure and Workflow

Using a pure transformer architecture for image classification, the Vision Transformer (ViT) model presents a novel method for computer vision [11]. ViT processes image patch sequences directly, marking a significant departure from conventional Convolutional Neural Networks (CNNs). Additionally, transformer-based models are capable of capturing global features by leveraging long-range dependencies, unlike CNNs that prioritize localized features alone. Further, studies have shown that ViTs have a larger receptive field even at the early stages compared to CNNs [12]. The proposed model's workflow comprises a series of steps: patch embedding, transformer encoding, and multi-level feature fusion, culminating in the formation of a classification head.

ViT's patch embedding block works by splitting the input image into a series of $N$ non-overlapping patches, thereby transforming the image into a patch sequence. After that, each patch is linearly embedded into a feature vector with embedding dimension $d$. Positional encodings are added to account for the spatial information of these patches. The sequence of patch embeddings that is produced is used as the input for the transformer layers that follow, allowing the ViT model to capture dependencies and global relationships among various patches. This method is consistent with the self-attention mechanism built into transformers because it gives the transformer a systematic way to examine the image's content sequentially. This block transforms the input tensor $\mathbf{F}_k \in R^{w_k \times h_k \times d_k}$ into the output tensor $\mathbf{F}_l \in R^{w_k \times h_k \times d_k}$ and $\mathbf{F}_l \in R^{w_k/2 \times h_k/2 \times 2d_k}$.

In the ViT model, the transformer encoder block is a fundamental component that is instrumental in capturing and processing information from image patches. The transformer encoder Block in ViT utilizes self-attention to identify both local and global dependencies between image patches using multiple attention heads. Notably, the multi-headed self-attention composed of $k$ self-attention blocks employs the Scaled dot-product attention as described in [11]. This method, represented in (1), processes the input by dividing it into queries $Q$, keys $K$, and values $V$.

$$SelfAttention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_q}}\right)V \qquad (1)$$

where $Q, K, V \in R^{(1+N) \times d_q}$ and $d_q = d/k$ This approach allows the model to focus on different aspects of the input image, capturing intricate patterns crucial for comprehensive image analysis.

A feed-forward neural network adds non-linearity to this process, enhancing the model's capability to identify complex patterns. Residual connections and normalization layers integrated into the block ensure the model's stability. By stacking these blocks, the ViT model captures hierarchical information and allows for the iterative refinement of features. The R3DM-ViT model innovatively incorporates the concept of multi-level feature fusion, combining the overall impact of the contributions of the multiscale low-, intermediate-, and high-level semantic features $(i.e., \mathbf{f}_1–\mathbf{f}_4)$ in the final classification decision.

### 2.3. Recurrent Model Structure and Workflow

When dealing with 3-D imaging data that is composed of $n$ consecutive slices $(\mathbf{F}^1, \mathbf{F}^2, ...\mathbf{F}^n)$, the suggested model analyzes each input slice one after the other and produces a collection of n feature vectors $(\mathbf{f}^1, \mathbf{f}^1, ...\mathbf{f}^n)$ of size $1 \times 1 \times 256 \times n$. This recurrent model uses Long Short-Term Memory (LSTM) to process these feature vectors further and perform class prediction by utilizing additional 3-D anatomical characteristics. An LSTM layer receives a series of $n$ feature vectors $(\mathbf{f}^1, \mathbf{f}^1, ...\mathbf{f}^n)$ from a sequence input layer first. After processing through a sequence of $n$ LSTM cells, the LSTM layer takes advantage of extra 3-D anatomical dependencies among these feature vectors to produce a single feature vector, $h_n$, of size $1 \times 1 \times 1200$. The output feature vector $h_n$ is further enhanced by another fully connected layer to extract more discriminative patterns. It includes both 2-D spatial and 3-D anatomical information of the 3-D imaging data (i.e., $(\mathbf{f}^1, \mathbf{f}^1, ...\mathbf{f}^n)$). Finally, using the final output feature vector $h_n$, a single class label is predicted.

**Table 1**. Performance comparison of the proposed model vs the baseline model using (2-D CXR + 2-D CT) data ([%]).

| Methods | ACC | F1 | mAP | mAR |
|---|---|---|---|---|
| ViT | 93.28 | 92.99 | 92.97 | 93.02 |
| M-ViT | 94.04 | 93.86 | 93.89 | 93.84 |
| R3DM-ViT (BiLSTM) | **95.04** | **94.88** | **94.92** | **94.83** |

**Table 2**. Performance comparison of the proposed model vs the baseline model using (2-D CXR + 3-D CT) data ([%]).

| Methods | ACC | F1 | mAP | mAR |
|---|---|---|---|---|
| ViT (BiLSTM) | 95.66 | 95.71 | 95.56 | 95.85 |
| R3DM-ViT (BiLSTM) | **96.67** | **96.88** | **96.75** | **97.02** |

### 2.4. Multistage Training Loss

We achieve optimal convergence of our proposed classification framework by subsequently training the M-ViT and recurrent models. Using a cross-entropy (CE) loss function [13], the ViT was initially trained to exploit and learn the spatial characteristics from the complete training dataset, referred to as $[\mathbf{F}_T]_{x=1}^{n}, [l_T]_{x=1}^{n}$. Subsequently, each data sample was processed by our M-ViT, resulting in the training (designated as $[\mathbf{f}_T]_{x=1}^{n}, [l_T]_{x=1}^{n}$) and validation (denoted as $[\mathbf{f}_V]_{x=1}^{m}, [l_V]_{x=1}^{m}$) feature vectors from the datasets. Then, using the same CE loss function, the recurrent model was trained to learn the 3-D anatomical dependencies in the case of 3-D imaging data.

## 3. RESULTS AND ANALYSIS

### 3.1. Dataset and Experimental Setup

To conduct a comprehensive quantitative analysis of the proposed approach, six publicly accessible datasets containing both X-ray and CT scans were utilized as used in [5]. These datasets were categorized into two primary classes based on their ground truth labeling: infectious and non-infectious. This classification provided us with a substantial volume of diverse radiographic data from CT and CXR sources. During the data preprocessing stage, each image was resized to 224 × 224, aligning with the fixed dimension of the input layer in our proposed network. The modeling and simulation were conducted using MATLAB R2019a, with an in-built deep learning toolkit. All simulations were run on a desktop computer equipped with an Intel Core i7 CPU, 16 GB RAM, an NVIDIA GeForce GPU (GTX 1070), and operating on Windows 10.

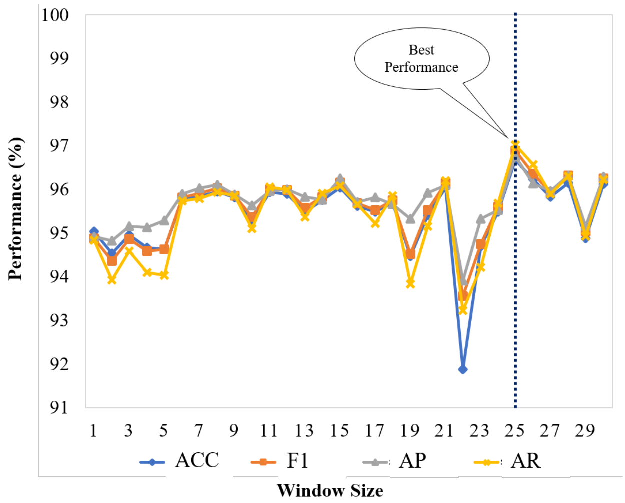

**Fig. 2**. Validation results of the proposed model with respect to different sizes of windows to find the optimal window size.

**Table 3**. Performance comparison of the proposed model with different RNNs using (2-D CXR + 3-D CT) data ([%]).

| Methods | ACC | F1 | mAP | mAR |
|---|---|---|---|---|
| R3DM-ViT (LSTM) | 90.07 | 91.86 | 92.42 | 91.38 |
| R3DM-ViT (GRU) | 93.67 | 93.58 | 94.54 | 92.72 |
| R3DM-ViT (GRU+BiLSTM) | 96.30 | 96.45 | 96.40 | 96.50 |
| R3DM-ViT (BiLSTM) | **96.67** | **96.88** | **96.75** | **97.02** |

### 3.2. Testing Results

Our proposed framework discerns infections by leveraging both spatial and 3-D structural information extracted from CT scan volumes. A key factor in optimizing the system's performance is determining the ideal window size, which refers to the number of slices included in each scan segment. This window size is crucial because a narrow window may lead to a loss of structural information, adversely affecting performance, while an excessively large window could unnecessarily prolong processing times without significant improvement in results.

To find the optimal window size, we conducted a comprehensive assessment of the model's validation performance across a spectrum of 30 different window sizes, ranging from 1 to 30. This evaluation was illustrated in Fig. 2, where we analyzed the model's performance in terms of various metrics, including accuracy (ACC), F1-score (F1), mean Average Precision (mAP), and mean Average Recall (mAR). The results indicated that a window size of 25 (denoted as $w$ = 15) achieved the best validation performance across all metrics, as highlighted by the dotted vertical line in Fig. 2.

The model's effectiveness was assessed using the testing dataset, as outlined in Table 1. Our newly introduced R3DM-ViT model, which features a recurrent module, demonstrates superior performance over the first M-ViT and baseline ViT models. The R3DM-ViT model shows average improvements of 1.76%, 1.89%, 1.95%, and 1.81% in ACC, F1, mAP, and mAR, respectively, for 2-D data when compared to the base-

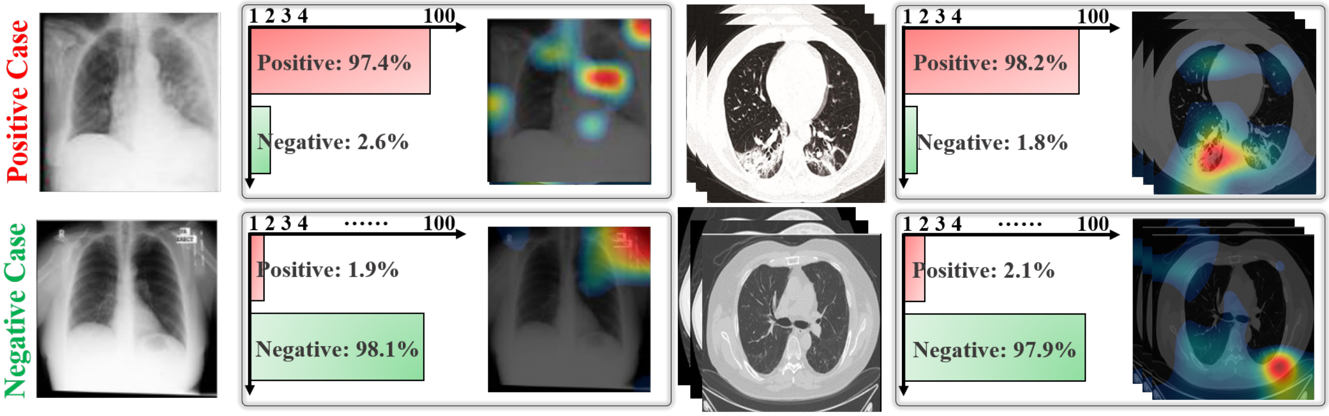

**Fig. 3**. Visualization of predicted outputs of the proposed network.

line ViT model. The quantitative comparison between the baseline ViT model and the proposed R3DM-ViT model, as summarized in Table 1, indicates a significant enhancement in performance.

When analyzing both 2-D CXR and 3-D CT scans together, our proposed model notably outperforms the baseline ViT with the Bidirectional Long Short-Term Memory (BiLSTM) model, as reported in Table 2. This improvement is evident in the substantial average increases across key metrics: 1.01% in ACC, 1.17% in F1, 1.19% in mAP, and 1.17% in mAR. These results not only demonstrate the model's versatility in handling multimodal data but also affirm its consistent superiority across various evaluation metrics.

Furthermore, a comparison of the results in Tables 1 and 2 reveals that the inclusion of 3-D CT scans significantly enhances the model's performance. This addition improves the diagnostic capabilities of the model, underscoring its effectiveness in utilizing multimodal data for more accurate medical image analysis. The observed improvements validate the model's capacity to integrate complementary information from different imaging modalities, emphasizing the benefits of incorporating 3-D CT scans into the analysis workflow.

**Table 4**. Performance comparison of the proposed model with different RNNs using (2-D CXR + 3-D CT) data ([%]).

| Study | ACC | F1 | mAP | mAR |
|---|---|---|---|---|
| Khan et al. [14] | 92.60 | 92.77 | 93.02 | 92.53 |
| Minaee et al. [15] | 88.77 | 88.77 | 88.8 | 88.75 |
| Brunese et al. [16] | 92.64 | 92.74 | 92.9 | 92.59 |
| Ardakani et al. [17] | 92.64 | 92.85 | 93.13 | 92.57 |
| Martínez et al. [18] | 92.68 | 92.77 | 92.91 | 92.63 |
| Jaiswal et al. [19] | 91.01 | 91.09 | 91.23 | 90.96 |
| Asnaoui et al. [20] | 93.42 | 93.5 | 93.63 | 93.37 |
| Apostolopoulos et al. [21] | 93.07 | 93.2 | 93.4 | 93.01 |
| Farooq et al. [22] | 93.69 | 93.8 | 93.98 | 93.63 |
| **R3DM-ViT** | **96.67** | **96.88** | **96.75** | **97.02** |

In addition to these comparisons, we conducted ablative studies to further elucidate the individual contributions and effectiveness of different components within our proposed model. As demonstrated in Table 3, we conducted a comprehensive evaluation of our model against various Recurrent Neural Network (RNN) architectures such as LSTM, Gated Recurrent Unit (GRU), and BiLSTM. This analysis was crucial in guiding our choice of the BiLSTM architecture. The superior performance of the BiLSTM, evident in the comparison, underscores its effectiveness within our model framework.

Furthermore, Table 4 presents a comparative analysis of our R3DM-ViT model with several state-of-the-art CAD diagnostic techniques. The R3DM-ViT model shows average improvements of 2.98%, 3.08%, 2.77%, and 3.39% in ACC, F1, mAP, and mAR, respectively, for 2-D X-data in comparison with the second-best model [22]. A rigorous t-test analysis ($p < 0.01$) validates the significant performance improvement of our model, demonstrating a 99% confidence level when compared to the second-best model [22]. This comparison, highlighting the proposed model's proficiency, shows it outperforms competitor models across all key qualitative performance metrics. Such significant improvement is achieved by leveraging multi-level feature aggregation, which is implemented using a ViT model as the foundational baseline. This approach allows for the integration of information from various encoder blocks of the network, enhancing the model's ability to capture complex patterns and relationships within the data, thereby leading to superior overall performance. These results not only validate the effectiveness of our model in handling multimodal data but also emphasize its leading edge in the field of medical image analysis.

## 4. DISCUSSION AND CONCLUSION

Our study introduces a robust framework capable of processing 2-D CXR and 3-D CT scan data, both individually and in unison. By adopting a multimodal approach, this framework overcomes the limitations inherent in conventional single-modality diagnostic techniques, laying a more comprehensive groundwork for diagnostic decision-making. Further, the qualitative analysis of our framework utilizing Class Activation Maps (CAM) demonstrates its proficiency in localizing areas of interest within the radiographic images, which is critical for accurate diagnosis. Fig. 3 illustrates the CAM outputs for both positive and negative cases, using 2-D CXR and 3-D CT scan data. In the positive case scenario, the CAM output highlights infectious regions, with a high probability of positive diagnosis at 97.4% for 2-D CXR and 98.2% for 3-D CT scan data. This visualization aligns with the clinical findings, showcasing the model's ability to pinpoint precise locations relevant to the diagnosed condition.

In conclusion, our framework significantly enhances radiographic image analysis for infectious disease diagnosis by integrating 2-D and 3-D data, providing a more complete analysis than traditional single-modality methods. This comprehensive approach improves diagnostic precision and furthers the development of AI tools that can integrate into clinical workflows, ultimately supporting healthcare professionals in delivering accurate diagnoses.

## 5. REFERENCES

[1] A. Castiglione, P. Vijayakumar, M. Nappi, S. Sadiq, and M. Umer, "Covid-19: automatic detection of the novel coronavirus disease from ct images using an optimized convolutional neural network," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 9, pp. 6480–6488, 2021.

[2] B. Hassan, S. Qin, T. Hassan, R. Ahmed, and N. Werghi, "Joint segmentation and quantification of chorioretinal biomarkers in optical coherence tomography scans: A deep learning approach," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–17, 2021.

[3] C. Jin, W. Chen, Y. Cao, Z. Xu, Z. Tan, X. Zhang, L. Deng, C. Zheng, J. Zhou, H. Shi *et al.*, "Development and evaluation of an artificial intelligence system for covid-19 diagnosis," *Nature communications*, vol. 11, no. 1, p. 5088, 2020.

[4] D. Dong, Z. Tang, S. Wang, H. Hui, L. Gong, Y. Lu, Z. Xue, H. Liao, F. Chen, F. Yang *et al.*, "The role of imaging in the detection and management of covid-19: a review," *IEEE reviews in biomedical engineering*, vol. 14, pp. 16–29, 2020.

[5] M. Owais, H. S. Yoon, T. Mahmood, A. Haider, H. Sultan, and K. R. Park, "Light-weighted ensemble network with multilevel activation visualization for robust diagnosis of covid19 pneumonia from large-scale chest radiographic database," *Applied soft computing*, vol. 108, p. 107490, 2021.

[6] T. Rajasenbagam, S. Jeyanthi, and J. A. Pandian, "Detection of pneumonia infection in lungs from chest x-ray images using deep convolutional neural network and content-based image retrieval techniques," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–8, 2021.

[7] A. Oulefki, S. Agaian, T. Trongtirakul, and A. K. Laouar, "Automatic covid-19 lung infected region segmentation and measurement using ct-scans images," *Pattern recognition*, vol. 114, p. 107747, 2021.

[8] A. Halder and B. Datta, "Covid-19 detection from lung ct-scan images using transfer learning approach," *Machine Learning: Science and Technology*, vol. 2, no. 4, p. 045013, 2021.

[9] C. Zhao, Y. Xu, Z. He, J. Tang, Y. Zhang, J. Han, Y. Shi, and W. Zhou, "Lung segmentation and automatic detection of covid-19 using radiomic features from chest ct images," *Pattern Recognition*, vol. 119, p. 108071, 2021.

[10] M. Owais, Y. W. Lee, T. Mahmood, A. Haider, H. Sultan, and K. R. Park, "Multilevel deep-aggregated boosted network to recognize covid-19 infection from large-scale heterogeneous radiographic data," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 6, pp. 1881–1891, 2021.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[12] D. Velayudhan, A. Ahmed, T. Hassan, N. Gour, M. Owais, M. Bennamoun, E. Damiani, and N. Werghi, "Autonomous localization of x-ray baggage threats via weakly supervised learning," *IEEE Transactions on Industrial Informatics*, pp. 1–10, 2024.

[13] M. Owais, M. Arsalan, J. Choi, and K. R. Park, "Effective diagnosis and treatment through content-based medical image retrieval (cbmir) by using artificial intelligence," *Journal of clinical medicine*, vol. 8, no. 4, p. 462, 2019.

[14] I. U. Khan and N. Aslam, "A deep-learning-based framework for automated diagnosis of covid-19 using x-ray images," *Information*, vol. 11, no. 9, p. 419, 2020.

[15] S. Minaee, R. Kafieh, M. Sonka, S. Yazdani, and G. J. Soufi, "Deep-covid: Predicting covid-19 from chest x-ray images using deep transfer learning," *Medical image analysis*, vol. 65, p. 101794, 2020.

[16] L. Brunese, F. Mercaldo, A. Reginelli, and A. Santone, "Explainable deep learning for pulmonary disease and coronavirus covid-19 detection from x-rays," *Computer Methods and Programs in Biomedicine*, vol. 196, p. 105608, 2020.

[17] A. A. Ardakani, A. R. Kanafi, U. R. Acharya, N. Khadem, and A. Mohammadi, "Application of deep learning technique to manage covid-19 in routine clinical practice using ct images: Results of 10 convolutional neural networks," *Computers in biology and medicine*, vol. 121, p. 103795, 2020.

[18] F. Martínez, F. Martínez, and E. Jacinto, "Performance evaluation of the nasnet convolutional network in the automatic identification of covid-19," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 10, no. 2, p. 662, 2020.

[19] Q. Yan, B. Wang, W. Zhang, C. Luo, W. Xu, Z. Xu, Y. Zhang, Q. Shi, L. Zhang, and Z. You, "Attention-guided deep neural network with multi-scale feature fusion for liver vessel segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 7, pp. 2629–2642, 2020.

[20] K. El Asnaoui and Y. Chawki, "Using x-ray images and deep learning for automated detection of coronavirus disease," *Journal of Biomolecular Structure and Dynamics*, vol. 39, no. 10, pp. 3615–3626, 2021.

[21] D. A. Prabowo and G. B. Herwanto, "Duplicate question detection in question answer website using convolutional neural network," in *2019 5th International conference on science and technology (ICST)*, vol. 1. IEEE, 2019, pp. 1–6.

[22] M. Farooq and A. Hafeez, "Covid-resnet: A deep learning framework for screening of covid19 from radiographs," *arXiv preprint arXiv:2003.14395*, 2020.