

Transformer-Based Lung Infection Severity Prediction with Cross Attention and Conditional TransMix Augmentation

Bouthaina Slika

University of the Basque Country UPV/EHU
San Sebastian, Spain

Ho Chi Minh City Open University
Ho Chi Minh City, Vietnam

bslika001@ikasle.ehu.eus

Fadi Dornaika

University of the Basque Country UPV/EHU
San Sebastian, Spain

IKERBASQUE, Basque Foundation for Science
Bilbao, Spain

fadi.dornaika@ehu.eus

Fares Bougourzi

Junia, UMR 8520, CNRS,
Centrale Lille,

Univesité Polytechnique Hauts-de-France,
59000 Lille, France

fares.bougourzi@junia.com

Karim Hammoudi

Université de Haute-Alsace, IRIMAS
Mulhouse, France

karim.hammoudi@uha.fr

Abstract

Lung infections, particularly pneumonia, pose significant health risks and can rapidly worsen, especially during pandemics. Developing advanced AI-driven tools for severity prediction based on medical imaging is essential for timely decision-making and treatment, ultimately saving lives. In this study, we introduce a novel approach applicable to multiple medical imaging modalities, including CT scans and chest X-rays, for predicting lung infection severity. Our method consists of two key components: a Transformer-based severity prediction model and an augmentation strategy called Conditional Online TransMix, designed to address data imbalance. The proposed model employs parallel encoders, integrating Pyramid Vision Transformers (PVTs) with a cross-gated attention mechanism and a feature aggregation module to generate a scalar severity score. To enhance model generalization across datasets, we introduce a tailored augmentation technique that synthesizes new mixed severity scores linked to image patches. We validate our approach using the RALO CXR and Per-COVID-19 CT datasets, demonstrating superior performance on multi-image modalities compared to several state-of-the-art deep learning models. By incorporating a customized weighted loss function, our method enhances the precision of automated lung disease severity assessment, providing a reliable

and adaptable AI tool for clinical diagnosis and treatment planning.

1. Introduction

Lung diseases, including chronic obstructive pulmonary disease and interstitial lung disease, pose major global health challenges due to their high morbidity and mortality rates [43]. Accurate diagnosis and assessment are essential for effective treatment, yet traditional methods like clinical evaluation and pulmonary function tests often lack early and detailed insights [15]. Medical imaging, particularly chest X-rays (CXRs) and computed tomography (CT) scans, plays a crucial role in diagnosing and monitoring lung diseases [1, 5]. While these imaging techniques provide noninvasive visualization of lung pathology, their interpretation requires expertise and is subject to interobserver variability, leading to potential inconsistencies [5, 26].

Deep learning has emerged as a powerful tool for medical image analysis, enabling automated and efficient diagnostic solutions [7, 30]. Convolutional Neural Networks (CNNs) have shown strong performance in detecting and classifying lung diseases from CXRs and CT scans [14, 28, 46]. However, severity prediction remains challenging due to the limited availability of pixel-wise annotations, which are costly and time-consuming to obtain.

Recent approaches have explored direct severity estimation, but robust methods that generalize well across diverse datasets are still needed.

Previous studies have proposed deep learning models for severity prediction, such as regression models for pneumonia severity [37] and multimodal approaches for COPD assessment [23]. However, challenges persist, including variability in disease presentation, dataset imbalance, and limited labeled data, which hinder model generalization and clinical adoption [20, 44, 45].

To address these issues, we propose a transformer-based model integrating cross-attention and parallel encoders to quantify lung disease severity. We validate our approach using CXRs and CT scans, introducing Conditional Online TransMix to mitigate data imbalance. Additionally, we employ an ensemble strategy to enhance performance by averaging predictions from top-performing models. The contributions of this work are listed below:

- Development of a transformer-based model using cross-attention for severity quantification.
- Performance evaluation on CXRs and CT scans to ensure robustness across multiple imaging modalities.
- Implementation of Conditional Online TransMix augmentation to address training data imbalance.
- Introduction of an ensemble method to improve prediction accuracy.

The rest of the paper is organized as follows: Section 2 provides an overview of related studies. Section 3 details the proposed lung severity quantification model. Section 4 presents the performance evaluation, including the datasets used, experimental results, and an ablation analysis. In Section 5, we discuss the obtained results. Finally, Section 6 recaps the findings and provides concluding remarks.

2. Related Work

In recent years, deep learning has advanced significantly in medical image analysis, particularly for diagnosing and assessing lung diseases using CXRs and CT scans [27]. The development of models for severity prediction has become a key research focus, providing quantitative assessments to aid clinical decision-making. Traditional deep learning methods have been widely explored for lung disease severity prediction. Tang et al. [37] introduced a regression model for pneumonia severity estimation from CXRs, achieving strong correlations with clinical severity scores. Similarly, Santosh et al. [23] proposed a multimodal deep learning approach that integrated CXRs and CT scans to improve severity assessment accuracy for COPD. Xu et al. [46] developed a framework for pneumonia severity detection using CXRs, emphasizing clinically relevant features. Numerous studies have further highlighted the critical role of deep learning in addressing healthcare challenges, particularly in lung disease evaluation [2, 4, 8, 16, 22, 29].

More recently, Vision Transformers (ViTs) have significantly improved disease detection and severity quantification using CXRs and CT scans. Shome et al. [31] introduced the COVID-Transformer, an interpretable ViT model for detecting COVID-19 and its accurate diagnosis. Kim et al. [21] applied ViTs for severity quantification and lesion localization in CXRs for monitoring disease progression. Le Dinh et al. [24] combined CNNs with transformers to enhance classification and severity assessment of COVID-19 cases. Deepa et al. [13] employed Swin Transformers to improve COVID-19 severity assessment, showcasing the model’s ability to distinguish infection levels.

Other works have explored innovative transformer-based methodologies. Taye et al. [38] used ViTs to classify COVID-19 severity from thoracic CT scans, leveraging high-resolution imaging analysis. Sun et al. [35] enhanced diagnostic accuracy by integrating demographic data with a Swin Transformer model. Additionally, Huang et al. [19] introduced DeepCoVDR, a transfer learning model incorporating graph transformers and cross-attention to predict COVID-19 drug response. Collectively, these studies highlight the advantages of transformer architectures in achieving high accuracy and robustness in disease detection and severity assessment.

While many existing studies have operated transformers for diagnostic tasks, most focused on disease detection or classification rather than severity quantification. Additionally, prior work typically relies on a single imaging modality for severity assessment. In contrast, our approach aims to predict severity scores from both CXRs and CT scans within a unified model, ensuring consistency and improved generalization across multiple imaging modalities.

3. Proposed Methodology

3.1. Proposed Model

Our work aims to predict the severity of lung pneumonia by using two medical imaging modalities. We employed parallel encoders combined with a gated cross-attention mechanism and a feature aggregator to estimate a scalar value representing the severity of the lung infection. The structural framework of our proposed model is depicted in Figure 1. The proposed model is trained in an end-to-end fashion.

3.1.1. Model Encoder

The model splits an image into four quadrants, each processed independently by a Transformer-based block. However, to capture inter-region dependencies, cross-attention is applied between different quadrants before feature aggregation. The encoder is designed with a parallel architecture, where it processes multiple input pathways simultaneously. The input image I is first resized to dimensions of $2H \times 2W \times C$ and then divided into four separate regions.

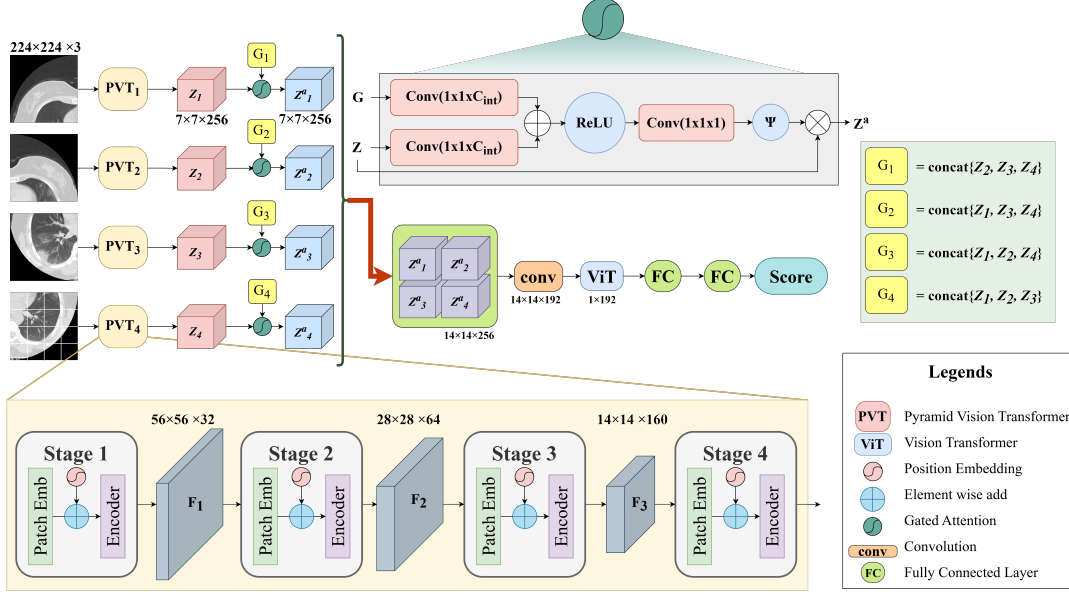


Figure 1. Illustration of the proposed model.

This resizing ensures that each quadrant matches the input size expected by the pre-trained Pyramid Vision Transformer (PVT) [41] models. Thus, each region with dimensions $H \times W \times C$ is fed into a separate PVT [41]. The PVT is an advanced model that adapts the Vision Transformer (ViT) architecture to better handle image recognition tasks by addressing some limitations in capturing multi-scale features. The PVT encoder employs Spatial-Reduction Attention (SRA) to enhance computational efficiency while maintaining strong feature representation. SRA processes queries, keys, and values, but it significantly reduces spatial complexity by downsampling keys and values before the attention operation, which allows the model to handle larger input feature maps. Each PVT model processes the image using a hierarchical structure with four stages ($k = 1, 2, 3, 4$), where each stage has different embedding dimensions, depths, and attention heads. The output size at each stage is given by:

$$\frac{H}{2^{k-1} \cdot P} \times \frac{W}{2^{k-1} \cdot P} \times C_k, \quad (1)$$

where P is the patch size and C_k is the number of channels after the k^{th} stage. At each stage, the resolution is further downsized, with each reduction capturing broader contextual information as shown in Figure 1. The final output of each PVT has a small spatial resolution with a high number of channels, providing a rich representation that combines global context with fine-grained details. To leverage pre-trained weights and mitigate the limitations of scarce training data, a pre-trained PVT model on ImageNet is utilized.

In our setup, $H = 224$, $W = 224$, $P = 4$, and

$\{C_1, C_2, C_3, C_4\} = \{32, 64, 160, 256\}$. The channel size 160 in the third stage reflects an experimentally optimized balance between feature richness and computational efficiency, particularly when used with corresponding multi-head attention settings (1, 2, 5, 8), which aligns with the scale of these channels. The input image I is divided into four sections, resulting in four 3D feature tensors Z_i . Each of these tensors is then processed using an Attention Gate (AG) defined by the following expression:

$$\begin{aligned} Z_i^a &= AG\{Z_i, G_i\} \\ &= \psi(\text{conv}[\text{ReLU}(\text{conv}(G_i) + \text{conv}(Z_i))]) \otimes Z_i \end{aligned} \quad (2)$$

The inner layers perform linear transformations using 1×1 convolutional blocks, which adjust the number of channels in the input and gating signals C_{Z_i} and C_{G_i} to intermediate feature representations C_{int} . The features derived from C_{Z_i} and C_{G_i} are combined and passed through a third 1×1 convolutional layer, producing a 2D weight map. The sigmoidal activation function ψ is then used to learn the spatial attention coefficients for each patch token. These spatial coefficients are applied to the encoder's feature maps Z_i , \otimes representing element-wise multiplication.

Z_i is used as the input signal, and the gating signal G_i is formed by concatenating the features from the other three quarters. Specifically, $G_1 = \text{concat}\{Z_2, Z_3, Z_4\}$, $G_2 = \text{concat}\{Z_1, Z_3, Z_4\}$, $G_3 = \text{concat}\{Z_1, Z_2, Z_4\}$, and $G_4 = \text{concat}\{Z_1, Z_2, Z_3\}$. This operation performs cross-attention between the tokens of the current image region and those from the other three regions. Attention scores are computed across the tokens of different regions,

enabling the model to assess the relative importance of spatial features across the image.

3.1.2. Features Processing and Regression Head

The resulting tensors are then concatenated along the spatial dimension. Specifically, the four tensors resulting from the four partial images are grouped horizontally and vertically, as shown in Figure 1. A convolution step is applied, followed by an additional Vision Transformer (ViT). This additional ViT is used to process the combined features and serves as a feature aggregator. It refines the information extracted from the different regions before predicting the severity score. The MLP head of the ViT is replaced by a regression head consisting of two fully connected layers that map the output vector of the ViT into a single scalar representing the predicted score.

3.2. Data Augmentation

In our work, we have addressed the problem of score imbalance in the dataset by using an online augmentation technique based on the TransMix method [9]. Although this technique was originally proposed for classification tasks, we have adapted it for regression by generating new mixed scores for the newly mixed images. This data augmentation technique uses attention maps to drive the score-mixing process. The mixed image is obtained using the traditional CutMix method [47], where a random patch of an image is inserted into another one to form a new mixed image, as shown in Figure 2. Its relative ground truth score is calculated following the TransMix approach. TransMix uses this attention map Att to control the process of mixing scores. In this context, λ (representing the proportion of the source image) is set to the sum of the weights in the Att that overlaps with the clipping mask, each weight corresponding to the importance assigned to a particular region of the image. Given images A and B , we update the mixing coefficients λ using the Att of the mixed image by nearest-neighbor interpolation that downsamples the mask from $H \times W$ into P pixels. In this way, we can dynamically reassign the weights of the scores depending on how much the Att is directed to each patch. The ground truth score of the mixed image \bar{y} is:

$$\bar{y} = \lambda * y_B + (1 - \lambda) * y_A, \quad (3)$$

where y_A and y_B are the ground truth scores of images A and B , respectively.

In our implementation, TransMix was applied conditionally to address dataset imbalance by focusing on underrepresented scores. The chosen threshold is derived from the training dataset’s score distribution. Figure 3 illustrates the distribution of GE scores across the dataset, revealing that images with a score greater than 4 are more prevalent than those with a score of 4 or less. This imbalance suggests a natural division within the data, prompting us to use a

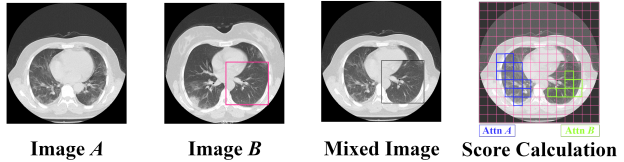


Figure 2. TransMix applied to CT images.

score of 4 as a threshold. By setting this threshold, we effectively target the less frequent data with the proposed augmentation method. Similarly, for the LO scores, images with scores below 2 and above 6 are less frequent compared to those within the mid-range scores. This highlights the scarcity of cases with either very low or very high severity, which could pose challenges for the model’s training and evaluation. We applied the augmentation to the scarce data to ensure balanced representation and accurate predictions across all severity levels.

On the other hand, Figure 4 shows the distribution of scores for CT scans, revealing that images with a score greater than 10 are less frequent. This indicates that the majority of the dataset consists of images with a score of 0, while those with non-zero scores are comparatively rare. This imbalance highlights the need for careful consideration during model training to ensure that cases with non-zero scores are adequately represented for reliable performance across the full range of scores (0-100), which is achieved by data augmentation.

3.3. Loss Function

For unbalanced data sets, a user-defined weighted loss function effectively balances the data distribution and ensures that the model focuses on critical features. With this technique, the loss contributions of the different scores are weighted differently. This is particularly useful when images with a certain severity are underrepresented in the training dataset. By weighting the loss associated with these minority categories higher, the model is instructed to pay more attention to these critical examples during training. The weights are calculated using equation (4).

$$w_l = \frac{N}{c_l \cdot k} \quad (4)$$

Here, w_l is the weight for the l^{th} score level, N is the total number of images in the training dataset, c_l is the number of samples in the l^{th} score level, and k is the total number of score levels. Since we are dealing with two modalities and their respective scores, we have different parameter values in each case. For the RALO dataset ($k = 17$), the scores range from 0 – 8 with an increment of 0.5. The score labels of the Per-COVID-19 CT ranges from 0 – 100 with an increment of 1 forming 101 score levels ($k = 101$). We use

these weights to calculate the weighted loss function $\mathcal{L}_{\mathcal{W}}$ as follows:

$$\mathcal{L}_{\mathcal{W}} = \sum_{i=1}^N w_i |y_i - \hat{y}_i|, \quad (5)$$

where w_i is the weight for the i^{th} image (computed from Eq. (4)), y_i is the ground truth value for the i^{th} image, \hat{y}_i is the predicted value for the i^{th} image, and N is the total number of images. For lung severity quantification, a custom weighted loss function ensures that the model adequately learns to distinguish between different levels of disease severity, thereby providing more reliable and fine assessments that are crucial for effective patient management.

In our work, we used AdamW optimizer with a learning rate set to 10^{-5} . This choice of optimizer was made to ensure better convergence and weight decay handling. Additionally, we used a cosine annealing warm restarts scheduler with an initial restart period equal to the length of the training loader and a multiplication factor of 2.

4. Performance Evaluation

4.1. Datasets

4.1.1. RALO Dataset

The primary goal of this research is to evaluate the effectiveness of deep learning models in determining the severity of lung diseases. To accomplish this, we utilized the Radiographic Assessment of Lung Opacity Score (RALO) dataset, which consists of 2,373 CXR images [12]. These images were scored by two expert radiologists from Stony Brook Medicine. The dataset is divided into a training set of 1,878 images and a test set of 495 images. The radiological grading focuses on two key criteria: Geographic Extent (GE) and Lung Opacity (LO). GE refers to the spread of lung involvement by ground-glass opacity or consolidation, with separate scores for the right and left lungs. The GE score ranges from 0 (no consolidation) to 4 (maximum consolidation), and the overall GE score is the sum of the left and right lung scores. LO is assessed independently for each lung, with scores ranging from 0 (no opacity) to 4 (total whiteout), reflecting varying degrees of lung opacity. The total LO score, which ranges from 0 to 8 points, is calculated by summing the scores of both lungs. The final ground-truth scores are averages of the two radiologists' evaluations and range from 0 to 8 [42]. An offline combined lung and score replacement is applied to the training set as done in a previous work [32]. The resultant dataset is distributed as shown in Figure 3.

4.1.2. Per-COVID-19 Dataset

The Per-COVID-19 dataset's training and validation splits were derived from 189 CT scans, each confirming a COVID-19 infection [3, 39]. COVID-19 diagnosis in this

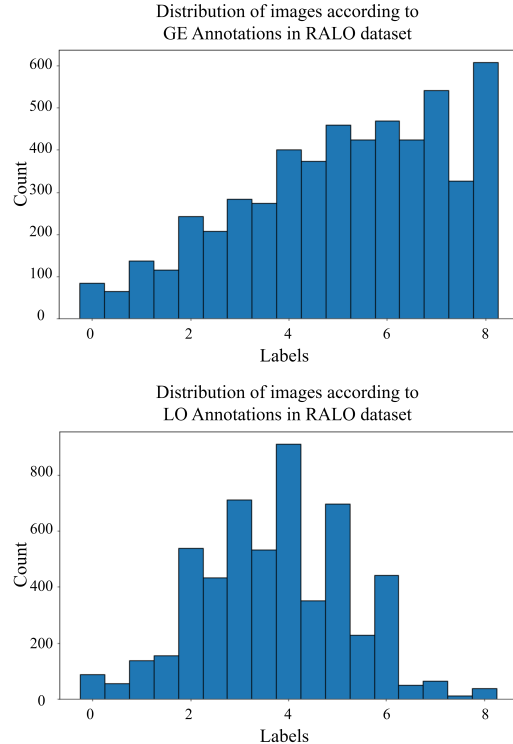


Figure 3. RALO training set scores distribution of GE and LO scores.

dataset is based on a positive reverse transcription polymerase chain reaction (RT-PCR) test and findings from CT scans, which two experienced thoracic radiologists carefully interpreted. Each CT scan comprises 40–70 slices, and the radiologists determined the percentage of lung area affected by COVID-19 infection relative to the total lung area. These COVID-19 Infection Percentage (CIP) annotations are expressed as percentages between 0% and 100%. The dataset is divided into 3,054 training slices and 1,301 validation slices [3, 39]. The testing split of Per-COVID-19 combines three different COVID-19 segmentation datasets [3]. For the test data, the ground truth for CIP is calculated by determining the proportion of infected pixels relative to the total number of lung pixels, using both the infection and lung segmentation masks, producing accurate CIP compared with the ones of training and validation data. The Per-COVID-19 dataset presents additional challenges beyond typical CIP estimation from CT scans. Specifically, the challenge arises from training models with noisy labeled data, as the CIP ground truth for both the training and validation sets was estimated by radiologists on a scale of 100. Additionally, the research addresses domain adaptation issues, as the testing data originates from three sources different from those of the training data [6]. Figure 4 shows the distribution of the training data used in our paper.

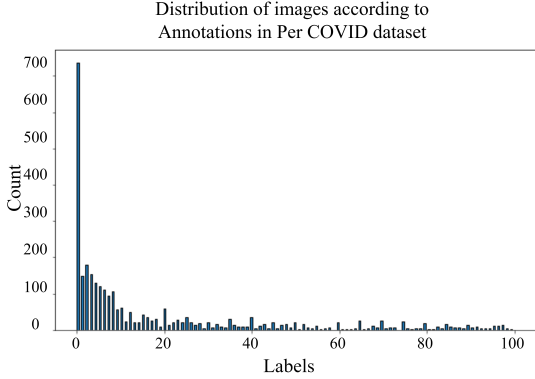


Figure 4. Per-COVID-19 training set with CIP score distribution.

4.2. Experimental Results and Comparison

To assess the severity of lung disease, we applied our model to both CXR and CT modalities. The refined RALO dataset, comprising 5,634 images, provides GE and LO scores ranging from 0 to 8, indicating the extent of disease from mild to critical. In our study, we used both the original images and the images with combined lung and score replacement augmentation from a previous work [32], but with the addition of a conditional online score-correlated TransMix augmentation strategy. The details of the applied method are described in Section 3.2.

Moreover, to estimate the generalizability of our model in assessing lung disease severity, we tested its performance on CT images. Using the Per-COVID-19 dataset, which comprises 3,054 images, we trained the proposed model to predict the infection percentage, which is a scalar value ranging from 0 to 100. The infection percentage is highly correlated with infection progression and severity. During training, conditional online score-correlated TransMix augmentation was applied to balance the dataset, following the strategy described in Section 3.2.

We performed comparisons with SOTA architectures for both modalities to emphasize the superior performance of the proposed model. For the CXR modality, several deep-learning models are included in this comparison, such as COVID-NET [42], COVID-NET S [40], ResNet50 [17], Swin Transformer [25], XceptionNet [10], Feature Extraction [11], MobileNetV3 [18], InceptionNet [36], ViTReg-IP [34], MVITReg-IP [33], in addition to our proposed model. Table 1 shows the key performance metrics.

On the other hand, a comparison is conducted with leading existing methods while training on CT images. The results are collected from [6]. Tables 2 and 3 show these results for the CT dataset for both the validation CT set and the test CT set, respectively.

The performance of the models is assessed by computing the Mean Absolute Error (MAE) and the Pearson corre-

Table 1. Performance Evaluation of the Proposed Method vs. SOTA Work on the RALO Dataset.

Model	GE		LO		Number of parameters
	MAE ↓	PC ↑	MAE ↓	PC ↑	
COVID-NET[42]	4.458	0.549	2.242	0.535	12M
COVID-NET-S[40]	4.698	0.591	2.254	0.529	12M
ResNet50 [17]	1.094	0.688	1.061	0.431	23M
Swin Transformer[25]	0.916	0.817	0.803	0.697	29M
XceptionNet[10]	0.854	0.821	0.768	0.701	23M
Feature Extraction[11]	0.967	0.753	0.865	0.711	2.2M
MobileNetV3[18]	0.847	0.827	0.732	0.738	4.2M
InceptionNet[36]	0.702	0.886	0.609	0.829	24M
ViTReg-IP[34]	0.565	0.925	0.510	0.857	5.5M
MViTReg-IP [33]	0.531	0.938	0.462	0.881	11.2M
Ours	0.362	0.971	0.337	0.945	28M

Table 2. Performance Evaluation of the Proposed Method vs. SOTA Work on the Per-COVID-19 validation set [6].

Team	MAE ↓	PC ↑
ACVLab	4.99	0.9364
EIDOSlab_Unito	4.91	0.9429
Ours	5.42	0.9432
ausilianapoli94	4.95	0.9435
TAC	4.48	0.9460
SenticLab.UAIC	4.17	0.9487
Taiyuan_university_lab713	4.50	0.9490

Table 3. Performance Evaluation of the Proposed Method vs. SOTA Work on the Per-COVID-19 test set [6].

Team	MAE ↓	PC ↑
IPLab	6.53	0.7091
ACVLab	4.86	0.7287
SenticLab.UAIC	4.61	0.7634
EIDOSlab_Unito	5.02	0.7977
TAC	3.64	0.8022
Ours	4.45	0.8094
Taiyuan_university_lab713	3.55	0.8547

lation coefficient (PC) between the predicted scores and the ground truth values provided by expert radiologists. A perfect MAE is 0, which would indicate completely accurate predictions. The PC measures the strength of the correlation, ranging from -1 to +1, where 0 indicates no correlation, and +1 represents a perfect linear relationship.

4.3. Ablation Studies

This section analyzes the importance of each component of our model through various ablation studies. These studies are designed to validate the model’s robustness across different scenarios and provide deeper insights into its operational dynamics.

To evaluate the significance of the applied Conditional TransMix as an online augmentation method, we test our proposed model, both with and without the application of TransMix. The goal of these experiments is to assess how

Table 4. Evaluation of the TransMix Augmentation’s Impact on the Proposed Method Using the RALO and Per-COVID-19 Test Sets.

TransMix	RALO CXR dataset				Per-COVID-19 CT dataset			
	GE		LO		CT val		CT test	
	MAE ↓	PC ↑	MAE ↓	PC ↑	MAE ↓	PC ↑	MAE ↓	PC ↑
×	0.370	0.965	0.342	0.942	5.462	0.9412	4.467	0.8081
✓	0.362	0.971	0.337	0.945	5.421	0.9432	4.453	0.8094

the augmentation strategy affects the model’s performance across different tasks and datasets, specifically in predicting GE and LO from CXRs, as well as the CIP scores from CT images. Table 4 presents the experimental results.

By eliminating the cross-attention mechanism, the model’s ability to capture relationships and dependencies between different regions of the lung images may be diminished, potentially affecting the accuracy of the severity predictions. To evaluate the impact of gated cross-attention, we compared the model’s performance with and without this mechanism. We tested a modified version where these components were excluded. In the modified model, the encoder processes each of the four image regions independently, and the resulting tensors are directly concatenated. The outcomes of these experiments are shown in Table 5.

We propose an ensemble model that combines the strengths of the best-performing architectures to enhance prediction accuracy. We leverage multiple models by averaging their outputs to produce a final prediction. Each of the highly performing models generates its predicted scores which are then averaged and compared against the ground truth scores. By aggregating the outputs from multiple architectures, the ensemble model aims to reduce the impact of errors from any single model, leading to more robust and accurate predictions across different tasks and datasets. The individual models used in the ensemble are variations of the proposed model, where either the encoders or the aggregator transformer are modified. We present three specific models: the one presented in this paper utilizes a PVT encoder with a ViT aggregator, another one uses a ViT encoder with a PVT aggregator, and the last one combines both PVT encoders and a PVT aggregator. By averaging the outputs of these three models, the ensemble aims to enhance the overall prediction accuracy by leveraging the strengths of each configuration. Table 6 shows the results of each of the three models and the ensemble method in terms of MAE and PC.

5. Discussion

Table 1 shows that our proposed model achieves the lowest MAE of 0.362 for GE and 0.337 for LO, indicating that our model provides the most accurate predictions. Moreover, it achieves the highest PC values of 0.971 for GE and 0.945 for LO, showing a strong correlation between the predicted scores and the actual ground truth values. These results highlight the superior predictive capability and reliability of

our proposed approach. Additionally, Table 1 compares the number of parameters for each model, providing valuable insights into their complexity and computational requirements. While our model has a relatively higher number of parameters compared to SOTA approaches, it achieves superior performance. Moreover, the model size of 28M parameters remains within the average range for computer vision models. Considering the complexity of medical imaging tasks, this model size is relatively lightweight and well-suited for healthcare applications.

On the other hand, Tables 2 and 3 reveal that our proposed model has a comparative performance against several SOTA models on the Per-COVID-19 dataset, both for the validation and test sets. As shown in Table 2, our model achieves an MAE of 5.421 and a PC of 0.9432 for the validation set. Although this MAE is slightly higher than that of the top-performing models such as “Taiyuan_university_lab713” (MAE of 4.50), our model’s performance in terms of PC metric is still competitive, reflecting a high degree of correlation between the predicted and ground truth values. Notably, our PC value of 0.9432 is comparable to that of the leading teams, such as “ausilianapoli94” (PC of 0.9435) which indicates that our model is accurate and its prediction consistency is strong. Similarly, this is shown in Table 3 where the results over the test set are presented. Our model exhibits a more competitive performance with an MAE of 4.45, and a PC of 0.8094. This MAE value is lower than those achieved by several other models and our PC value of 0.8094 is higher than those of many competing models, like “SenticLab.UAIC” (PC of 0.7634), reflecting a better correlation between the predicted scores and the actual values. The differences in MAE and PC values between our model and the top performers are marginal, especially considering the score range (0 to 100). It is important to highlight that the validation labels are naturally noisy. This is because they are based on radiologists’ direct estimation of the infection percentage by visually comparing the infected areas to the total lung volume. However, despite the difficulty of learning from such noisy data, our approach shows a strong capacity to generalize in domain adaptation scenarios.

Overall, these tables illustrate that our model demonstrates robust performance, achieving relatively low MAE and high PC values compared to the SOTA models. Its ability to maintain strong correlation scores indicates its relia-

Table 5. Evaluation of the Impact of Gated Cross-Attention on the Proposed Method Using the RALO and Per-COVID-19 Test Sets.

	RALO CXR dataset				Per-COVID-19 CT dataset			
	GE		LO		CT val		CT test	
Gated cross-attention	MAE ↓	PC ↑	MAE ↓	PC ↑	MAE ↓	PC ↑	MAE ↓	PC ↑
×	0.371	0.961	0.342	0.938	5.512	0.9325	4.511	0.7684
✓	0.362	0.971	0.337	0.945	5.421	0.9432	4.453	0.8094

Table 6. Evaluation of Modified Versions of the Proposed Model in Addition to the Ensemble Model.

Model		RALO CXR dataset				Per-COVID-19 CT dataset			
		GE		LO		CT val		CT test	
Encoder	Aggregator	MAE ↓	PC ↑	MAE ↓	PC ↑	MAE ↓	PC ↑	MAE ↓	PC ↑
ViT	PVT	0.399	0.955	0.328	0.930	5.623	0.9291	4.521	0.7772
PVT	PVT	0.385	0.959	0.342	0.934	5.642	0.9328	4.501	0.7799
PVT	ViT	0.362	0.971	0.337	0.945	5.421	0.9432	4.453	0.8094
Ensemble Model		0.358	0.973	0.336	0.948	5.409	0.9444	4.439	0.8107

bility and effectiveness in severity prediction tasks for both CXR and CT image analysis.

The ablation studies presented in Table 4 demonstrate the clear benefits of applying Conditional TransMix as an online augmentation method to our model. The results consistently show that the model performs better with TransMix across various datasets and prediction tasks. This improvement is reflected in the model’s ability to generate predictions that are more closely aligned with the ground truth, highlighting its enhanced generalization capabilities. These findings, demonstrated by a lower MAE and a higher PC, confirm that the augmentation method significantly boosts the model’s learning efficiency and overall performance. The results in Table 5 show that the inclusion of gated cross-attention significantly improves the model’s performance for the RALO and the Per-COVID-19 datasets, in which it leads to a lower MAE and a higher PC, indicating improved prediction accuracy and a stronger correlation with ground truth. These improvements highlight the effectiveness of gated cross-attention in capturing dependencies between different image regions, leading to more accurate and reliable predictions. By facilitating the exchange of information between different regions of the input image, gated cross-attention enables the model to better understand complex patterns and relationships.

For the ensemble model, Table 6 highlights that while each individual model configuration has unique strengths, the third model (PVT encoder and ViT aggregator) generally outperforms others in terms of MAE and PC across most datasets. However, the ensemble model, which combines the outputs of all three variations, achieves the best overall performance. It consistently delivers the lowest MAE and highest PC values for both GE and LO predictions on the RALO dataset, as well as for CIP scores on the Per-COVID-19 validation and test sets. These findings emphasize the effectiveness of an ensemble approach, as it uses the strengths of multiple models, mitigates individual

model weaknesses, and reduces the impact of errors from any single model. The ensemble model’s slight improvements in MAE and PC values demonstrate its superior generalization ability and resilience to data variability, making it a powerful strategy for enhancing performance in complex medical imaging tasks.

The ablation studies confirm that the selected parameters consistently lead to improved performance in both modalities. This validates that our model, with its specific configuration, achieves high performance in predicting severity scores for both CXRs and CT scans. These results emphasize the effectiveness of the detailed structure of our model in achieving excellent results in medical image analysis.

6. Conclusion

Accurate identification of pneumonia is critical to optimize patient care, prevent disease transmission, and initiate public health interventions to reduce the impact of this widespread and potentially serious respiratory disease. Our study presents a novel approach that combines a parallel design of PVTs with cross-gated attention as an encoder and feature processing with a ViT to improve the computational assessment of lung disease severity. The results show that our model performs superiorly when a weighted loss function is applied. The results across multiple modalities outperform various deep learning models and several SOTA techniques. In particular, our method accurately quantifies the severity of lung disease when applied to CXRs and CT scans, providing reliable predictions that allow physicians to make more direct and objective assessments. In addition, the application of a conditional online augmentation technique helps to balance the training data and further improve the prediction accuracy. Overall, our proposed approach not only improves the accuracy of automatic lung severity assessment but also provides physicians with a versatile tool for diagnosis and treatment planning.

References

- [1] M. Adam, J. H. Bates, and J. Hildebrandt. Diagnosis of lung diseases using chest radiographs. *Journal of Medical Imaging*, 7(1):011002, 2020. 1
- [2] Sunday Adeola Ajagbe and Matthew O Adigun. Deep learning techniques for detection and prediction of pandemic diseases: a systematic literature review. *Multimedia Tools and Applications*, 83(2):5893–5927, 2024. 2
- [3] Fares Bougourzi, Cosimo Distanto, Abdelkrim Ouafi, Fadi Dornaika, Abdenour Hadid, and Abdelmalik Taleb-Ahmed. Per-covid-19: a benchmark dataset for covid-19 percentage estimation from ct-scans. *Journal of Imaging*, 7(9):189, 2021. 5
- [4] F. Bougourzi, F. Dornaika, and A. Taleb-Ahmed. Deep learning based face beauty prediction via dynamic robust losses and ensemble regression. *Knowledge-Based Systems*, 242:108246, 2022. 2
- [5] Fares Bougourzi, Cosimo Distanto, Fadi Dornaika, and Abdelmalik Taleb-Ahmed. Pdatt-unet: Pyramid dual-decoder attention unet for covid-19 infection segmentation from ct-scans. *Medical Image Analysis*, 86:102797, 2023. 1
- [6] Fares Bougourzi, Cosimo Distanto, Fadi Dornaika, Abdelmalik Taleb-Ahmed, Abdenour Hadid, Suman Chaudhary, Wanting Yang, Yan Qiang, Talha Anwar, Mihaela Elena Breaban, et al. Covid-19 infection percentage estimation from computed tomography scans: Results and insights from the international per-covid-19 challenge. *Sensors*, 24(5):1557, 2024. 5, 6
- [7] Bingzhi Chen, Yishu Liu, Zheng Zhang, Guangming Lu, and Adams Wai Kin Kong. Transattunet: Multi-level attention-guided u-net with transformer for medical image segmentation. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 8(1):55–68, 2024. 1
- [8] Jun Chen, Lianlian Wu, Jun Zhang, Liang Zhang, Dexin Gong, Yilin Zhao, Qiuxiang Chen, Shulan Huang, Ming Yang, Xiao Yang, et al. Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography. *Scientific reports*, 10(1):19196, 2020. 2
- [9] Jie-Neng Chen, Shuyang Sun, Ju He, Philip Torr, Alan Yuille, and Song Bai. Transmix: Attend to mix for vision transformers. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 4
- [10] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 6
- [11] Joseph Paul Cohen, Lan Dao, Karsten Roth, Paul Morrison, Yoshua Bengio, Almas F Abbasi, Beiyi Shen, Hoshmand Kochi Mahsa, Marzyeh Ghassemi, Haifang Li, et al. Predicting covid-19 pneumonia severity on chest x-ray with deep learning. *Cureus*, 12(7), 2020. 6
- [12] Joseph Paul Cohen, Beiyi Shen, Almas Abbasi, Mahsa Hoshmand-Kochi, Samantha Glass, Haifang Li, Matthew P Lungren, Akshay Chaudhari, and Tim Q Duong. Radiographic Assessment of Lung Opacity Score Dataset, 2021. 5
- [13] J Deepa, D Ramya, KS Rishab, Sachi Shome, SK Manigandan, and J Velmurugan. Swin transformer based covid-19 identification and its severity quantification. In *Recent Trends in Computational Intelligence and Its Application*, pages 304–312. CRC Press, 2023. 2
- [14] B. Ghoshal and A. Tucker. Estimating the uncertainty of deep learning predictions for medical imaging. *Journal of Medical Imaging*, 7(1):011003, 2020. 1
- [15] Global Initiative for Chronic Obstructive Lung Disease. Global strategy for the diagnosis, management, and prevention of copd 2021, 2021. 1
- [16] Karim Hammoudi, Halim Benhabiles, Mahmoud Melkemi, Fadi Dornaika, Ignacio Arganda-Carreras, Dominique Collard, and Arnaud Scherpereel. Deep learning on chest x-ray images to detect and evaluate pneumonia cases at the era of covid-19. *Journal of medical systems*, 45(7):1–10, 2021. 2
- [17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6
- [18] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3, 2019. 6
- [19] Zhijian Huang, Pan Zhang, and Lei Deng. Deepcovdr: deep transfer learning with graph transformer and cross-attention for predicting covid-19 drug response. *Bioinformatics*, 39 (Supplement_1):i475–i483, 2023. 2
- [20] Asifullah Khan, Saddam Hussain Khan, Mahrukh Saif, Asiya Batool, Anabia Sohail, and Muhammad Waleed Khan. A survey of deep learning techniques for the analysis of covid-19 and their usability for detecting omicron. *Journal of Experimental & Theoretical Artificial Intelligence*, pages 1–43, 2023. 2
- [21] Gwanghyun Kim, Sangjoon Park, Yujin Oh, Joon Beom Seo, Sang Min Lee, Jin Hwan Kim, Sungjun Moon, Jae-Kwang Lim, and Jong Chul Ye. Severity quantification and lesion localization of covid-19 on cxr using vision transformer. *arXiv preprint arXiv:2103.07062*, 2021. 2
- [22] Su Yeon Kim, James Diggans, Dan Pankratz, Jing Huang, Moraima Pagan, Nicole Sindy, Ed Tom, Jessica Anderson, Yoonha Choi, David A Lynch, et al. Classification of usual interstitial pneumonia in patients with interstitial lung disease: assessment of a machine learning approach using high-dimensional transcriptional data. *The Lancet Respiratory Medicine*, 3(6):473–482, 2015. 2
- [23] Santosh Kumar, Vijesh Bhagat, Prakash Sahu, Mithliesh Kumar Chaube, Ajoy Kumar Behera, Mohsen Guizani, Raffaele Gravina, Michele Di Dio, Giancarlo Fortino, Edward Curry, and Saeed Hamood Alsamhi. A novel multimodal framework for early diagnosis and classification of copd based on ct scan images and multivariate pulmonary respiratory diseases. *Computer Methods and Programs in Biomedicine*, 243:107911, 2024. 2
- [24] Tuan Le Dinh, Suk-Hwan Lee, Seong-Geun Kwon, and Ki-Ryong Kwon. Covid-19 chest x-ray classification and severity assessment using convolutional and transformer neural networks. *Applied Sciences*, 12(10):4861, 2022. 2

- [25] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 6
- [26] R. J. Mason, V. C. Broaddus, T. R. Martin, T. E. King, D. E. Schraufnagel, J. F. Murray, and J. A. Nadel. Image analysis in lung disease: current challenges and future directions. *Lung Imaging Journal*, 8(2):123–135, 2019. 1
- [27] Saad I Nafisah, Ghulam Muhammad, M Shamim Hossain, and Salman A AlQahtani. A comparative evaluation between convolutional neural networks and vision transformers for covid-19 detection. *Mathematics*, 11(6):1489, 2023. 2
- [28] A Nair, JCL Rodrigues, S Hare, A Edey, A Devaraj, J Jacob, A Johnstone, R McStay, Erika Denton, and G Robinson. A british society of thoracic imaging statement: considerations in designing local imaging diagnostic algorithms for the covid-19 pandemic. *Clinical Radiology*, 75(5):329–334, 2020. 1
- [29] Sangjoon Park, Gwanghyun Kim, Yujin Oh, Joon Beom Seo, Sang Min Lee, Jin Hwan Kim, Sungjun Moon, Jae-Kwang Lim, and Jong Chul Ye. Multi-task vision transformer using low-level chest x-ray feature corpus for covid-19 diagnosis and severity quantification. *Medical Image Analysis*, 75: 102299, 2022. 2
- [30] Dinggang Shen, Guorong Wu, and Heung-II Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221–248, 2017. 1
- [31] Debaditya Shome, Tejaswini Kar, Sachi Nandan Mohanty, Prayag Tiwari, Khan Muhammad, Abdullah AlTameem, Yazhou Zhang, and Abdul Khader Jilani Saudagar. Covid-transformer: Interpretable covid-19 detection using vision transformer for healthcare. *International Journal of Environmental Research and Public Health*, 18(21):11086, 2021. 2
- [32] B Slika, F Dornaika, K Hammoudi, and VT Hoang. Automatic quantification of lung infection severity in chest x-ray images. In *2023 IEEE Statistical Signal Processing Workshop (SSP)*, pages 418–422. IEEE, 2023. 5, 6
- [33] Bouthaina Slika, Fadi Dornaika, and Karim Hammoudi. Multi-score prediction for lung infection severity in chest x-ray images. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2024. 6
- [34] Bouthaina Slika, Fadi Dornaika, Hamid Merdji, and Karim Hammoudi. Lung pneumonia severity scoring in chest x-ray images using transformers. *Medical & Biological Engineering & Computing*, pages 1–19, 2024. 6
- [35] Yunlong Sun, Jingge Lian, Ze Teng, Ziyi Wei, Yi Tang, Liu Yang, Yajuan Gao, Tianfu Wang, Hongfeng Li, Meng Xu, et al. Covid-19 diagnosis based on swin transformer model with demographic information fusion and enhanced multi-head attention mechanism. *Expert Systems with Applications*, 243:122805, 2024. 2
- [36] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 6
- [37] Y. Tang, M. D. Li, E. Capparelli, and S. Patel. Quantifying pneumonia severity using deep learning regression models. *Medical Image Analysis*, 60:101630, 2020. 2
- [38] Gizatie Desalegn Taye, Zewdie Habtie Sisay, Genet Worku Gebeysu, and Fisha Hailelassie Kidus. Thoracic computed tomography (ct) image-based identification and severity classification of covid-19 cases using vision transformer (vit). *Discover Applied Sciences*, 6(8):1–16, 2024. 2
- [39] Edoardo Vantaggiato, Emanuela Paladini, Fares Bougourzi, Cosimo Distanto, Abdenour Hadid, and Abdelmalik Taleb-Ahmed. Covid-19 recognition using ensemble-cnns in two new chest x-ray databases. *Sensors*, 21(5):1742, 2021. 5
- [40] Linda Wang, Zhong Qiu Lin, and Alexander Wong. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Scientific Reports*, 10(1):1–12, 2020. 6
- [41] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578, 2021. 3
- [42] A Wong, ZQ Lin, L Wang, AG Chung, B Shen, A Abbasi, M Hoshmand-Kochi, and TQ Duong. Towards computer-aided severity assessment via deep neural networks for geographic and opacity extent scoring of sars-cov-2 chest x-rays. *Scientific reports*, 11(1):1–8, 2021. 5, 6
- [43] World Health Organization. Global health estimates 2020, 2020. 1
- [44] Zunyou Wu and Jennifer M McGoogan. Characteristics of and important lessons from the coronavirus disease 2019 (covid-19) outbreak in china: summary of a report of 72 314 cases from the chinese center for disease control and prevention. *jama*, 323(13):1239–1242, 2020. 2
- [45] Laure Wynants, Ben Van Calster, Gary S Collins, Richard D Riley, Georg Heinze, Ewoud Schuit, Elena Albu, Banafsheh Arshi, Vanesa Bellou, Marc MJ Bonten, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *bmj*, 369, 2020. 2
- [46] Xiaowei Xu, Xiangao Jiang, Chunlian Ma, Peng Du, Xukun Li, Shuangzhi Lv, Liang Yu, Qin Ni, Yanfei Chen, Junwei Su, et al. A deep learning system to screen novel coronavirus disease 2019 pneumonia. *Engineering*, 6(10):1122–1129, 2020. 1, 2
- [47] Sangdoon Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 4