COVID-19 Pneumonia Classification with Transformer from Incomplete Modalities

Eduard Lloret Carbonell[1], Yiqing Shen[2], Xin Yang[1], and Jing Ke[1,3(✉)]

[1] School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China
{edu.lloret,yang_xin,kejing}@sjtu.edu.cn

[2] Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA
yshen92@jhu.edu

[3] School of Computer Science and Engineering, University of New South Wales, Sydney, Australia

Abstract. COVID-19 is a viral disease that causes severe acute respiratory inflammation. Although with less death rate, its increasing infectivity rate, together with its acute symptoms and high number of infections, is still attracting growing interests in the image analysis of COVID-19 pneumonia. Current accurate diagnosis by radiologists requires two modalities of X-Ray and Computed Tomography (CT) images from one patient. However, one modality might miss in clinical practice. In this study, we propose a novel multi-modality model to integrate X-Ray and CT data to further increase the versatility and robustness of the AI-assisted COVID-19 pneumonia diagnosis that can tackle incomplete modalities. We develop a Convolutional Neural Networks (CNN) and Transformers hybrid architecture, which extracts extensive features from the distinct data modalities. This classifier is designed to be able to predict COVID-19 images with X-Ray image, or CT image, or both, while at the same time preserving the robustness when missing modalities are found. Conjointly, a new method is proposed to fuse three-dimensional and two-dimensional images, which further increase the feature extraction and feature correlation of the input data. Thus, verified with a real-world public dataset of BIMCV-COVID19, the model outperform state-of-the-arts with the AUC score of 79.93%. Clinically, the model has important medical significance for COVID-19 examination when some image modalities are missing, offering relevant flexibility to medical teams. Besides, the structure may be extended to other chest abnormalities to be detected by X-ray or CT examinations. Code is available at https://github.com/edurbi/MICCAI2023.

Keywords: COVID-19 Pneumonia Classification · Hybrid CNN-Transformer · Multi-modality

E. L. Carbonell and Y. Shen—Equal contributions.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14224, pp. 379–388, 2023.
https://doi.org/10.1007/978-3-031-43904-9_37

1 Introduction

Coronavirus disease 2019 (COVID-19) has been a highly infectious viral disease, that can affect people of all ages, with a persistently high incidence after the outbreak in 2019 [1,2]. Early detection of the lung inflammatory reaction is crucial to initiate prompt treatment decisions, where the clinical assessment typically depends on two imaging techniques in conjunction, namely X-ray and CT scans [3,4]. Specifically, CT scans can provide a three-dimensional volumetric characterization of the patient's lung; while X-rays offer a two-dimensional landscape [5]. Recently, the use of multimodality data in COVID-19 diagnosis has received increasing interest as it can significantly improve prediction accuracy with complementary information [6].

During the inference stage, modalities can be incomplete amongst some test samples, which is known as incomplete multimodal learning [7,8]. To address this, it is necessary to implement a strategy that reduces the impact of missing modalities during the inference stage while also offering clinicians the flexibility to use any possible combination of data. Various strategies have been proposed to address this issue, such as generating the missing data [9,10]. However, this approach requires training a generative model for each possible missing modality and a larger training set, making it computationally expensive. Therefore, more compact models have emerged that reduces the number of generators [7]. However, these models are prone to be biased when dealing with multiple modalities. Consequently, a modality-invariant embedding model that makes use of Transformers has been introduced [11]. Despite the excellent performance, all of the models discussed above have been applied to the datasets, where the modalities used were restricted to three-dimensional, and the structures are similar. However, in the case of COVID-19 pneumonia detection, the clinicians practically employ different dimensional data to interpret the results. One of the major problems during the analysis of medical data is that, sometimes, some data modalities are missing, e.g., the case reported negative, or the examination device was unavailable. [12] Therefore being able to have a model that is able to adapt to all possible conditions will be greatly beneficial for the detection of COVID-19 pneumonia or any other disease. To this end, a novel method to diagnose COVID-19 pneumonia which can take incomplete CT and X-Ray multimodal data is proposed. The main contributions in this model are three-fold: (1) We propose a dual feature fusion across different dimensionality data. Instead of sole pre-fusion or post-fusion, both are performed in the multi-modality prediction model. (2) We design a feature matrix dropout regularization method to improve the reliability and generalization of the model. (3) A feature correlation block is proposed between two of the given modalities to extract latent dependencies. This attention layer further improves the understanding of the patients' evolution when multi-modality images are acquired at different stages of the disease.

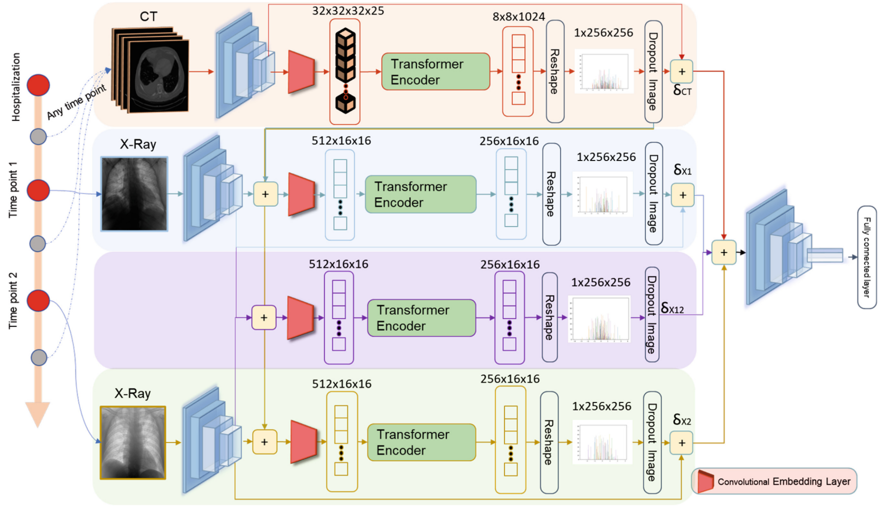


Fig. 1. Overview of the proposed model. This model is composed by three image modalities, on the first modality we use a convolutional-based feature extractor, followed by a Transformer encoder and a random features dropout. The other two modalities consist of an early fusion with the first modality features followed by a transformer encoder. Finally, all of the features have a features dropout layer followed by a Convolutional layer and a fully connected layer.

2 Methodology

2.1 Architecture Overview

We propose a model that can detect the COVID-19 pneumonia status from incomplete multi-modalities of CT scans and X-ray images. Specifically, CT scans and X-ray images are first embedded using convolutional layers and then processed by Transformer layers to obtain the global feature correlations. Then, a novel feature fusion layer can simulate incomplete modalities in the latent space while learning to fuse the features. Finally, the predictions are made using a ResNet based classification model followed by a learnable MLP layer.

2.2 Feature Fusion Layer

The objective of this layer is to combine features from different modalities, which have varying dimensions. To achieve this, we need to reduce the three-dimensional CT data to two-dimensional by convolutional layers before fusing them with two-dimensional X-ray data. We then reshape the data into smaller patches and apply a Transformer encoder to it, resulting in a two-dimensional matrix of the desired size. In this study, we investigate the data fusion at two stages, namely the early fusion and late fusion [13], where the data are fused twice in our model. Empirically, finding that dual fusion has some slight improvement compared to only early and late fusion as shown in Table 3.

2.3 Feature Matrix Dropout Layer

This layer helps regularize the model towards learning a modality agnostic representation. Specifically, during the feature fusion layer, random features are dropped from the feature matrix $\in \mathbb{R}^{X,Y}$ [14]. In our design certain percentage of data in the form of patches are dropped, where a patch is defined as a subpart of the feature matrix where all values have been set to 0. To generate the patches, we start by creating a random matrix $\mathbb{M}\delta \in \mathbb{R}^{\frac{X}{N},\frac{Y}{N}}$ with random values between 0 and 1, where N is the size of the dropout patches on the output images. We then round the values of $\mathbb{M}\delta$ using a threshold T, where all values below T are set to 0, and all values equal to or above T are set to 1. This gives us a binary matrix $\mathbb{M}_\delta$ that indicates which parts of the feature matrix should be dropped, i.e.

$$\mathbb{M}_D = R_T(\mathbb{M}_\delta) \tag{1}$$

where $R_T(\cdot)$ is the round function that converts all values of M_δ into 0 if they are under the threshold T and 1 otherwise. Finally the obtained matrix is interpolated or upsampled by a given scale N to match the size of the feature matrices $\mathbb{F}_M \in \mathbb{R}^{X,Y}$.

$$\mathbb{F}_D = I_M(\mathbb{M}_D) \odot \mathbb{F}_M \tag{2}$$

where $\mathbb{F}_D$ represents the final feature map after the patch dropout, F_M is the initial feature map, $I_M(\cdot)$ represents a nearest interpolation and $(\cdot) \odot (\cdot)$ represents an element-wise matrix multiplication. This gives us the final feature map $\mathbb{F}_D$, which has a similar structure to $\mathbb{F}_M$ but with some parts of size $N \times N$ converted to 0.

2.4 Transformer Layer

Transformers helps finding the different dependencies between the different modalities making use of their attention-based nature. Therefore, in this paper we will make use of the benefits that the transformers offers by implementing a ViT based transformer layer [15]. In this case we will use the transformer layer as a feature extractor and find the dependencies between the different embedded features. To extract the features, we will first split the image into patches of a fixed size, to afterwards be processed by an embedding layer, in this case the embedding layer will be a convolutional layer. Then each one of the embedded feature will pass through a different Self Attention Layer (SA) formulated as

$$\begin{aligned} SA_L &= \sigma\left(\frac{Q_L K^T}{\sqrt{d_L}}\right)V_L, \\ MSA_L &= \text{Concat}(SA_L^1, \cdots, SA_L^n) \cdot \mathbb{W}_0, \end{aligned} \tag{3}$$

where $\sigma(\cdot)$ represents the softmax function, d represents the size of each one of the heads, and $\mathbb{W}_0$ denotes a value embedding. Meanwhile, Q, K, V $\in \mathbb{R}^{X,d}$ represent the Query, key and embedding respectively. To constitute the ViT the values of each of the SA are concatenated forming the Multi-Headed Self Attention Layer (MSA) to afterward be normalized by a layer normalization. Finally, a Multi-Layer Perceptron (MLP) is applied to give non-linearity to the data, obtaining

$$\begin{aligned} A'_k &= MSA(LN(A_{k-1})) + A_{k-1} \\ A_k &= LN(MLP(A'_k)) + A'_k \end{aligned} \tag{4}$$

where A^{k-1} is the previous transformer layer, LN$(\cdot)$ is a Layer Normalization. In the vanilla ViT a final MLP is introduced to obtain the probabilities of each class however, in this paper the transformer layer is only used for feature extraction, therefore, the final MLP is deleted.

2.5 Dual X-Ray Attention

The Dual X-Ray attention block main idea is to find the different dependencies between the two input X-Ray images. Transformers have showed great results when looking for dependencies [11,15]. Therefore, this paper introduced a new attention layer that extracts the dependencies between two X-Ray images. The dependencies are extracted using the Transformer mentioned above Layer having as input both X-Ray images in the multimodality. This layer can also be seen as a fusion layer between two of the input modalities.

3 Experiments

3.1 Dataset

We leverage images from the BIMCV-COVID19 dataset [16], a public dataset with thousands of different positive and negative cases, for performance evaluation. This dataset is composed of 1200 unique cases from a combination of 1 CT Scan, 1 CT Scan and 1 X-Ray, 1 CT Scan and 2 X-Ray, 1 X-Ray or 2 X-Ray. One patient may have more than one image modality of CT or X-Ray. Regarding the image size of the dataset, the dimension of the CT scans is non-fixed per image, with the average image size around $500 \times 500 \times 400$. To facilitate the training on GPUs, the images perform dimension reduction with factor 2 and resulted in a final image size of $250 \times 250 \times 200$. The dataset is composed of approximately 2900 cases. From which around 1700 are negative, and 1200 are positive. The distribution is unbalanced because we wanted to extract as much data as possible with a balanced number of miss-modalities.

3.2 Implementation Details

This framework is implemented with Pytorch 1.13.0 using an NVIDIA A10 GPU with 24 GB of VRAM. The input size is $250 \times 250 \times 200$ for the CT scan and

2048 × 2048 with a batch size of 4 for the X-Ray. For the proposed model, we applied Adam optimizer with an cosine annealing scheduler of an initial learning rate 1e-5, and 100 epochs for training. The data partition for training, validation, and testing is 80%, 10%, and 10% on patient level.

3.3 Results

We use the following metrics to evaluate the performance of the binary classification model: the AUC score, Recall and Precision. To obtain the values on the Table 1, we made a split form the original dataset without miss-modalities. This table shows the difference between the possible combinations of models we can build to interpret the data. Taking in account the AUC, the worst model is the one using the CT as its only input with only a 65.74% in the AUC score, followed by the model with two X-Ray images as its input, obtaining a 67.98%. The result obtained in the Dual X-Ray multimodality performs around 2% worse than just having one X-Ray as an input. Finally, when having all the modalities performs 3% better in the AUC score than when only having one CT and one X-Ray which has around 71.26% in the AUC metric.

Table 1. The performance of AUC-score, Precision and Recall of the different modalities is reported. The baseline does not include any additional model proposed by us.

CT	X-Ray 1	X-Ray 2	AUC	Recall	Precision
X	✓	X	70.15	66.66	64.28
X	✓	✓	67.98	65.9	70.58
✓	X	X	65.74	61.11	75.0
✓	✓	X	71.26	74.28	69.76
✓	✓	✓	74.49	66.66	62.22

A comparison is made between our model and others with different possible variations, shown in the Table 2. The first model is used as a comparison is a fully convolutional model where all of its transformer components are converted into convolutions. This variant shows a significant reduction in all of its parameters with a sharp decrease in the performance of 70.18%, 59.9% and 68.93% in the

Table 2. The result comparisons between different variations of the proposed model.

Model	AUC	Recall	Precision
Convolutional Model	70.18	59.9	68.93
ViT Classificator	71.38	76.80	59.84
Final model	79.93	86.62	67.40

AUC, Recall, and Precision metrics, respectively. Another model is tested, where the final convolutional block is changed into a ViT base block with a final MLP layer. The AUC score remains similar to the one found in the previously tested model, obtaining 71.38%, 76.80% and 59.84% in AUC, Recall and Precision respectively. Due to the multi-modality and missing modality nature of this model and this dataset, the comparison between this model and other existing models was not feasible.

Table 3. Ablation study between the three different proposed methods by adding Dual Fusion block, Dual X-Ray attention block, and Feature matrix dropout in our method.

Dual Fusion	Dual X-Ray attention	Feature matrix dropout	AUC	Recall	Precision
X	X	X	75.28	71.14	69.87
✓	X	X	76.49	61.43	83.55
✓	✓	X	77.27	72.72	72.84
✓	✓	✓	79.93	86.62	67.40

3.4 Ablation Study

We show the classification performance on Table 3, of the different model components that introduced in the Fig. 1. We make use of the original dataset with missing modalities. From the Table 3, the addition of the Fusion Layer gives an slight increase of 0.19% in the AUC-score metric compared with the baseline. Meanwhile, we see an observable increase in the recall metric of 2.93%. The addition of the Dual X-Ray layer further increases the performance of the model increasing of 1.8% and 3.43% in the AUC-score and Precision respectively, yet a slight reduction in the recall metric by around 1.35%. Finally, the Regularization layer is topped up to further increases the AUC by a 2.66% and recall by a 13.90%, yet lower Precision score of 5.44%. Thus, compared to the baseline, the final model shows an increase of 4.65% in the AUC metric, a boost of 15.58% in the recall score and only a 2.47% decrease in the precision metric.

The saliency maps are visualized to pinpoint the diagnostic areas in a CT or X-ray image. Clinically, the saliency maps are helpful to assist radiologists. In the Fig. 2, (a) shows an example of a COVID-19 positive slice extracted from a complete CT scan. Its saliency map is situated next to the image, where a big opacity is found in the left patient's lung. The top-right figure gives an example of COVID-19 positive X-Ray which its correspondent saliency map, which mainly focus on the bottom contour of the lung. Moving to the bottom-left figure we can see a COVID-19 negative CT scan using a different angle compared to the first introduced figure. In this case, similar to what has been seen in the positive case, some opacities have been found however, in this case, the image is negative. Finally, the bottom-right figure is a negative X-Ray case, in this case, similar

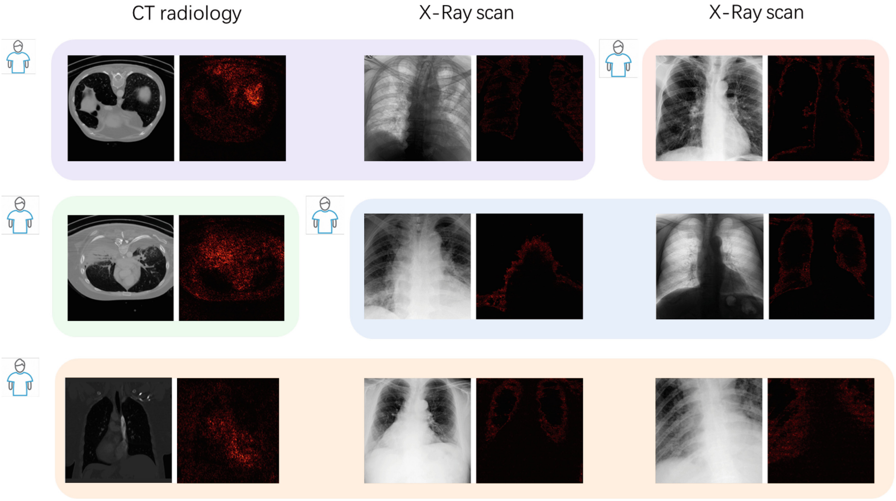
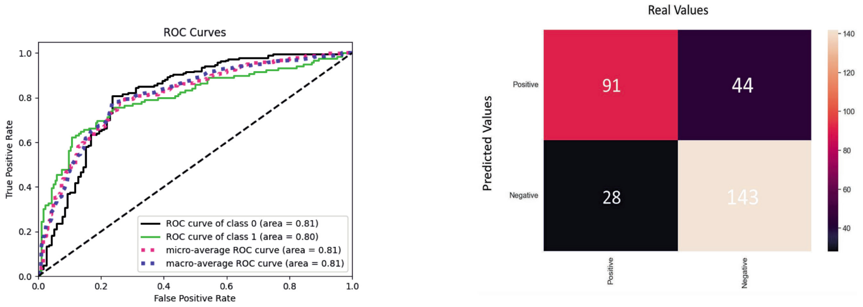


Fig. 2. X-ray and CT images with their saliency visualization.



(a) AUC Curve for the proposed model

(b) Confusion matrix for the proposed model

Fig. 3. Performance of proposed model. (a) AUC Curve, (b) Confusion matrix

to what has been found in the positive X-Ray figure the model mainly looks for the contours of the lungs. Therefore, through this images, a difference can be seen between the extracted features using the CT scans and the ones extracted using the X-Ray images. Finally, in the Fig. 3 the AUC Curve obtained with the proposed model together with the confusion matrix, obtained as a result of classifying the test dataset with the model, are shown as Sub-figure a and b respectively.

4 Conclusion

In this paper, we propose a novel multi-modality framework for COVID-19 pneumonia image diagnosis. A Dual fusion layer is introduced to help establish the dependencies between the different input modalities, as well as to take in different dimensionality inputs. The proposed Dual X-Ray attention layer makes it possible to effectively extract dependencies from the two X-Ray images focusing on the salient features between multi-modality images, whereas irrelevant features are ignored. The feature deletion layer helps to regularize the model dropping random features and improving the generalization of the model. Consequently, we provide the possibility to use one modality of CT or X-Ray for COVID-19 pneumonia diagnosis. Moreover, this model has the potential to be applied to other chest abnormalities in clinical practice.

Acknowledgments. This work was supported by National Natural Science Foundation of China (Grant No. 62102247) and Natural Science Foundation of Shanghai (No. 23ZR1430700).

References

1. Percentage of Visits for COVID-19-Like Illness: Covid data page. https://www.cdc.gov/coronavirus/2019-ncov/covid-data/covidview/index.html
2. Griffin, D.O., et al.: The importance of understanding the stages of covid-19 in treatment and trials. AIDS Rev. **23**(1), 40–47 (2021). https://doi.org/10.24875/aidsrev.200001261
3. Luo, N., et al.: Utility of chest CT in diagnosis of covid-19 pneumonia. Diagn. Interv. Radiol. **26**(5), 437–442 (2020). https://doi.org/10.5152/dir.2020.20144. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7490028/. pMID: 32490829; PMCID: PMC7490028
4. Alyasseri, Z.A.A., et al.: Review on covid-19 diagnosis models based on machine learning and deep learning approaches. Exp. Syst. **39**(3), e12759 (2022). https://doi.org/10.1111/exsy.12759. https://onlinelibrary.wiley.com/doi/abs/10.1111/exsy.12759
5. Li, B., et al.: Diagnostic value and key features of computed tomography in coronavirus disease 2019. Emerg. Microbes Infect. **9**(1), 787–793 (2020). https://doi.org/10.1080/22221751.2020.1750307. pMID: 32241244
6. Abdelaziz, M., Wang, T., Elazab, A.: Alzheimer's disease diagnosis framework from incomplete multimodal data using convolutional neural networks. J. Biomed. Informatics **121**, 103863 (2021). https://doi.org/10.1016/j.jbi.2021.103863. https://www.sciencedirect.com/science/article/pii/S1532046421001921
7. Azad, R., Khosravi, N., Dehghanmanshadi, M., Cohen-Adad, J., Merhof, D.: Medical image segmentation on MRI images with missing modalities: a review (2022). https://doi.org/10.48550/ARXIV.2203.06217. https://arxiv.org/abs/2203.06217
8. Ma, M., Ren, J., Zhao, L., Tulyakov, S., Wu, C., Peng, X.: SMIL: multimodal learning with severely missing modality. Proc. AAAI Conf. Artif. Intell. **35**(3), 2302–2310 (2021). https://doi.org/10.1609/aaai.v35i3.16330. https://ojs.aaai.org/index.php/AAAI/article/view/16330
9. Jin, L., Zhao, K., Zhao, Y., Che, T., Li, S.: A hybrid deep learning method for early and late mild cognitive impairment diagnosis with incomplete multimodal data. Frontiers Neuroinf. (2022). https://doi.org/10.3389/fninf.2022.843566. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8965366/

10. Gao, X., Shi, F., Shen, D., Liu, M.: Task-induced pyramid and attention Gan for multimodal brain image imputation and classification in Alzheimer's disease. IEEE J. Biomed. Health Inform. **26**(1), 36–43 (2022). https://doi.org/10.1109/JBHI.2021.3097721
11. Zhang, Y., et al.: mmFormer: multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation (2022). https://doi.org/10.48550/ARXIV.2206.02425. https://arxiv.org/abs/2206.02425
12. Altman, D.G., Bland, J.M.: Missing data. BMJ **334**(7590), 424 (2007). https://doi.org/10.1136/bmj.38977.682025.2C. https://www.bmj.com/content/334/7590/424
13. Gadzicki, K., Khamsehashari, R., Zetzsche, C.: Early vs late fusion in multimodal convolutional neural networks. In: 2020 IEEE 23rd International Conference on Information Fusion (FUSION), pp. 1–6 (2020). https://doi.org/10.23919/FUSION45008.2020.9190246
14. Choi, J.H., Lee, J.S.: EmbraceNet: a robust deep learning architecture for multimodal classification. Inf. Fusion **51**, 259–270 (2019). https://doi.org/10.1016/j.inffus.2019.02.010. https://www.sciencedirect.com/science/article/pii/S1566253517308242
15. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale (2020). https://doi.org/10.48550/ARXIV.2010.11929. https://arxiv.org/abs/2010.11929
16. de la Iglesia Vayá, M., et al.: BIMCV covid-19+: a large annotated dataset of RX and CT images from covid-19 patients with extension Part II (2023). https://doi.org/10.21227/mpqg-j236