# Supplementary Materials

## Deep learning for predicting COVID-19 malignant progression

Cong Fang[a,1], Song Bai[b,1], Qianlan Chen[c,1], Yu Zhou[a], Liming Xia[c], Lixin Qin[d], Shi Gong[a], Xudong Xie[a], Chunhua Zhou[d], Dandan Tu[e], Changzheng Zhang[e], Xiaowu Liu[e], Weiwei Chen[c,*], Xiang Bai[a,*], Philip H.S. Torr[b]

[a]School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China.

[b]Department of Engineering Science, University of Oxford, Parks Road, Oxford OX1 3PJ, United Kingdom.

[c]Department of Radiology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430030, China.

[d]Department of Radiology, Wuhan Pulmonary Hospital, Wuhan 430030, China.

[e]HUST-HW Joint Innovation Lab, Wuhan 430074, China.

∗Corresponding author.

Email addresses: chenweiwei_tjh@163.com (Weiwei Chen), xbai@hust.edu.cn (Xiang Bai)

[1]These authors contributed equally to this work and should be considered as co-first authors

## Supplementary methods

**Attention weights updating.** In each mini-batch of the training process, the exponential moving average is used to update the attention weights in the training set. In the testing process, the average weights obtained in the training process are used to update the clinical data. The attention weights are updated as:

$$A_i = \begin{cases} A_c, & i = 1 \\ \alpha \cdot A_c + (1-\alpha) \cdot A_{i-1}, & i > 1 \end{cases} \quad (1)$$

where $\boldsymbol{A_i}$ is the average attention weights; $\boldsymbol{A_c}$ is the current attention weights; the coefficient $\alpha$ represents the degree of weighting decrease, a constant smoothing factor between 0 and 1; $i$ is the number of iterations in the training process. In our experiment $\alpha$ is 0.1.

**Statistic metric.** The following 4 metrics are used to evaluate the performance.

1. **AUC:**

$$AUC = \sum_{i=1}^{n} TPR_i \times (FPR_i - FPR_{i-1}) \quad (2)$$

where $T = [th_1, th_2, \ldots, th_n]$, $0 \leq th_i \leq 1$, $th_i < th_{i+1}$, $TPR_i = \frac{TP}{TP+FN}$ $and\ FPR_i = 1 - \frac{TN}{FP+TN}$ $both\ with\ a\ threshold = T[i]$.

2. **accuracy:**

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN}, with\ a\ threshold = 0.5 \quad (3)$$

3. **sensitivity:**

$$sensitivity = \frac{TP}{TP+FN}, with\ a\ threshold = 0.5 \quad (4)$$

4. **specificity:**

$$specificity = \frac{TN}{FP+TN}, with\ a\ threshold = 0.5 \quad (5)$$

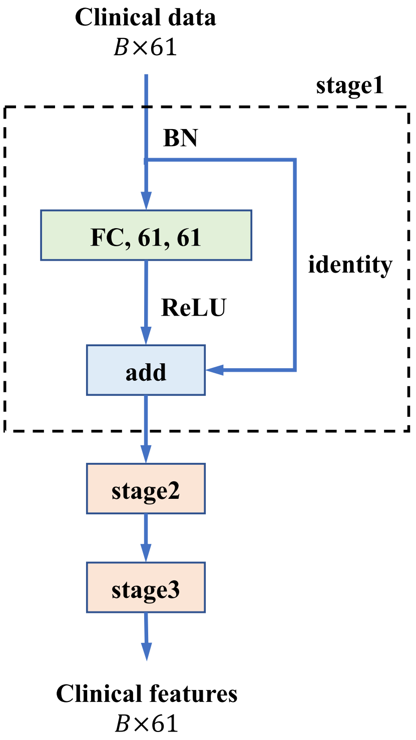
where TP is true positive, TN is true negative, FP is false positive and FN is false negative.

Supplementary Table 1. The detailed parameters of the 3D ResNet.

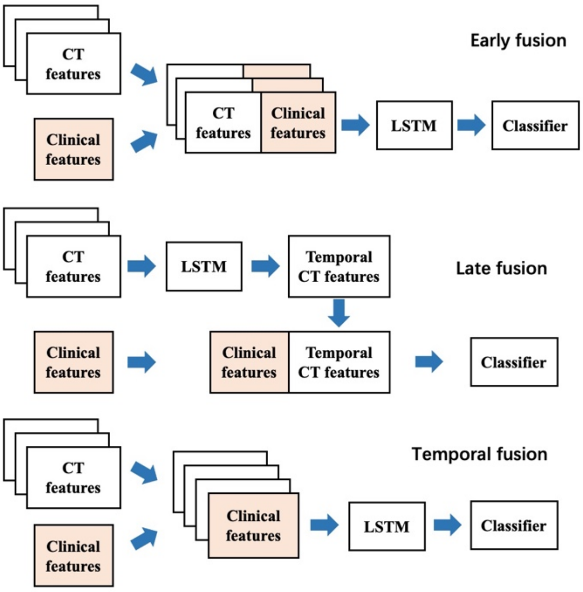
| **Layer name** | **Operation** | **Input size** | **Output size** |
|---|---|---|---|
| Conv1 | 3 × 3 × 3, 32, stride 2 | 64 × 64 × 64 × 1 | 32 × 32 × 32 × 32 |
| Block1 | 3 × 3 × 3, 32, stride 2<br>3 × 3 × 3, 32, stride 1 | 32 × 32 × 32 × 32 | 16 × 16 × 16 × 32 |
| Block2 | 3 × 3 × 3, 64, stride 2<br>3 × 3 × 3, 64, stride 1 | 16 × 16 × 16 × 32 | 8 × 8 × 8 × 64 |
| Block3 | 3 × 3 × 3, 128, stride 2<br>3 × 3 × 3, 128, stride 1 | 8 × 8 × 8 × 64 | 4 × 4 × 4 × 128 |
| Pooling | global average pooling | 4 × 4 × 4 × 128 | 1 × 1 × 1 × 128 |

Supplementary Table 2. The performance comparison of different data fusion strategies and different ratio of clinical and CT feature dimensions. 95% confidence intervals are included in brackets. The best average results are shown in **bold**. The p<0.05 indicates our method significantly improves the compared method (McNemar's test). Abbreviations: area under the receiver operating characteristic curve (AUC); accuracy (ACC); sensitivity (SENS); specificity (SPEC).

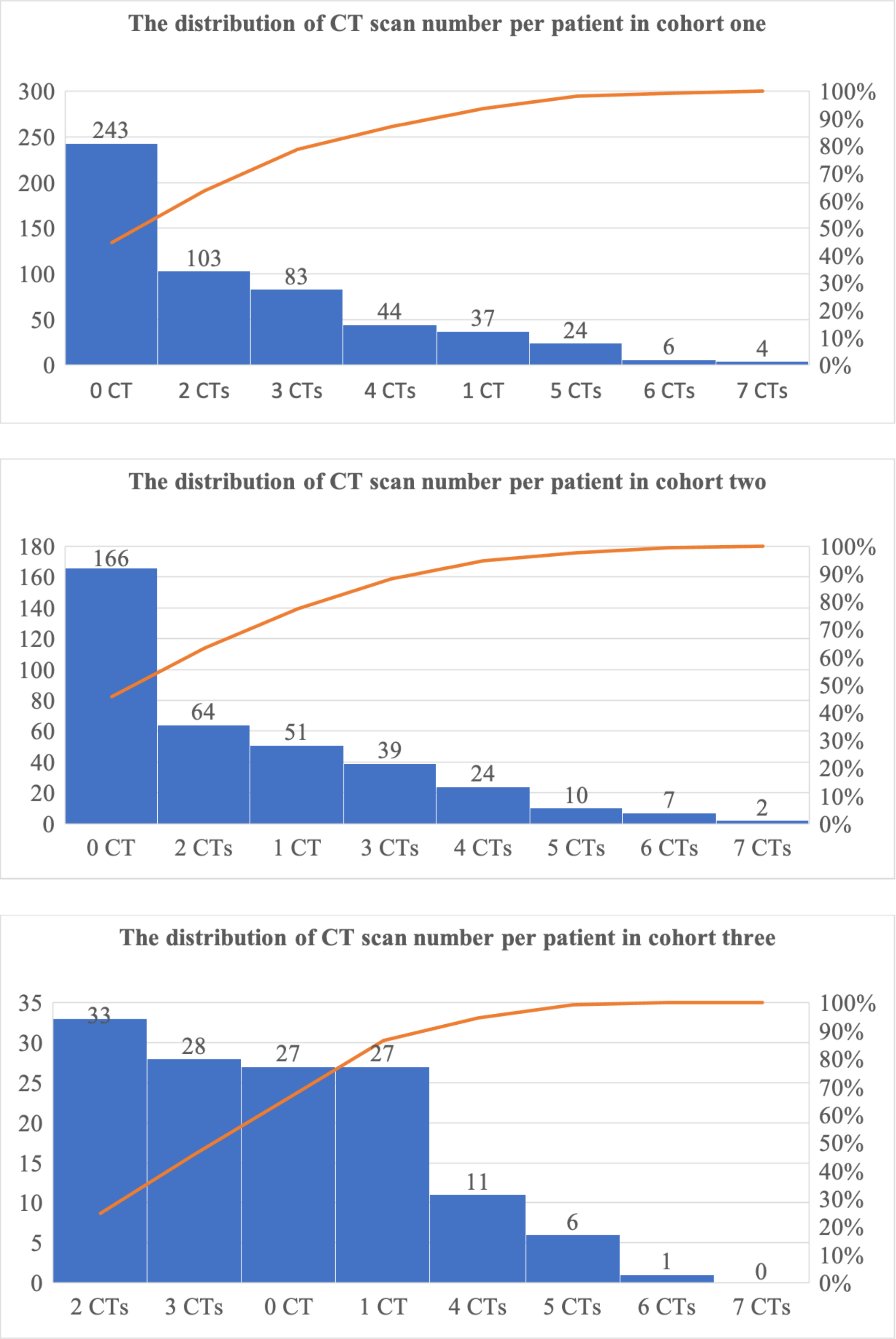
| Fusion strategy | Feature dimension | AUC | ACC (%) | SEN (%) | SPEC (%) | p-value |
|---|---|---|---|---|---|---|
| Early fusion | 61 / 64 (Clinical / CT) | 0.871[0.839,0.904] | 84.7[81.5,87.5] | 72.7[64.4,79.6] | 88.5[85.0,91.2] | 0.001 |
| Early fusion | 61 / 128 (Clinical / CT) | **0.920[0.861,0.979]** | **87.7[84.7,90.2]** | **89.1[82.5,93.4]** | 87.3[83.7,90.1] | *(base) |
| Early fusion | 61 / 256 (Clinical / CT) | 0.752[0.710,0.794] | 78.9[75.2,82.1] | 28.1[21.1,36.5] | **94.5[91.8,96.3]** | <0.001 |
| Late fusion | 61 / 64 (Clinical / CT) | 0.883[0.851,0.914] | 84.2[80.9,87.0] | 78.1[70.2,84.4] | 86.1[82.4,89.1] | 0.001 |
| Late fusion | 61 / 128 (Clinical / CT) | 0.860[0.827,0.894] | 84.6[81.3,87.4] | 74.2[66.0-81.0] | 87.7[84.2,90.6] | <0.001 |
| Late fusion | 61 / 256 (Clinical / CT) | 0.844[0.809,0.879] | 81.6[78.1,84.6] | 65.6[57.0,73.3] | 86.5[82.9,89.5] | <0.001 |
| Temporal fusion | 128 / 128 (Clinical / CT) | 0.787[0.747,0.827] | 77.9[74.3,81.2] | 58.6[49.9,66.8] | 83.9[80.1,87.1] | <0.001 |



**Supplementary Figure 1. Clinical data encoder.** This encoder has three stages, each of which consists of a fully connected layer and an identity connection. BN: batch normalization, add: pixel-wise addition, identity: identity connection. FC, 61, 61 represents a fully connected layer, the size of input features, and the size of output features. $B \times 61$ represents the batch size and the length of the vector.



**Supplementary Figure 2. Fusion strategies of clinical features and CT features.** Early fusion: CT features at each time point are concatenated with clinical features before fed into LSTM. Late fusion: The output of LSTM is concatenated with clinical features before fed into the classifier. Temporal fusion: The clinical features are considered as preliminary information before the CT scan sequence and fed into LSTM as the features at the first time point.



**Supplementary Figure 3. The distribution of CT scan numbers per patient in three cohorts.** The ordinate is the number of patients and the abscissa is the number of CT scans per patient.