Temporal Context Matters: Enhancing Single Image Prediction with Disease Progression Representations
— Supplementary Material —

Aishik Konwer¹, Xuan Xu¹, Joseph Bae², Chao Chen², Prateek Prasanna²
¹Department of Computer Science, Stony Brook University
²Department of Biomedical Informatics, Stony Brook University

{akonwer, xuaxu}@cs.stonybrook.edu {joseph.bae, Chao.Chen.1, Prateek.Prasanna}@stonybrook.edu

In this supplementary material, we provide additional information to further understand our proposed approach. In section 1, we provide an architectural overview of how to calculate a 512 dimension vector from the last layer of Temporal Convolutional Network (TCN). We provide dataset and preprocessing details in section 2. Finally in section 3, t-SNE plots and additional class activation maps provide insights into OA severity prediction from knee radiographs.

1. Architecture details - TCN output

A row-wise summation operation is applied on the self-attention weights obtained from the third and final self-attention block in our TCN architecture. This results in a ‘T’ length attention vector, the softmax of which gives the attention scores. When these attention scores are multiplied with the output of TCN, they generate an optimal 512 dimension representation. The steps are illustrated in Supplementary Fig 1.

2. Assets and preprocessing

Chest Radiograph dataset: For snapshot pretraining, we used 28433 chest radiographs (comprising multiple pulmonary diseases). CovidProg dataset, which contained 942 scans from 150 COVID-19 patients, comprised the temporal data. The duration between the CXR scans are variable (1-5 days). The number of timepoints per patient varies from 4 to 16. Out of the total 150 patients, 23 cases were obtained from Newark Beth Israel Medical Center, 77 from Stony Brook University Hospital, and 50 from University Hospitals Cleveland Medical Center.

OA Radiograph dataset: For snapshot pretraining, we used 23008 images. 2474 knee scans from 426 patients comprised the temporal data. The images in the ‘train’ folder of Kaggle [2] were a fraction of the snapshot cohort used in pretraining the transformer. All the images in ‘validation’ and ‘test’ folder [2] were jointly used in finetuning stage.

The experiments were performed in a 5-fold cross validation setting in the finetuning stage where the pretrained transformer model was finetuned on 4 folds and tested on the remaining fold. Details about the data used for each stage can be found in Supplementary Table 1.

<table>
<thead>
<tr>
<th>Stage</th>
<th>COVID</th>
<th>OA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temporal</td>
<td>942 (CovidProg)</td>
<td>2474 [6]</td>
</tr>
</tbody>
</table>

Table 1. Data utilized in different stages

COVID-19 preprocessing: Lung region segmentation was first performed using a Residual UNet model [10]. All chest scans were aligned to the same intensity range through an average histogram matching method.

OA preprocessing: We utilized BoneFinder tool [5] to localize and crop the knee joint landmarks. Following [8], histogram clipping and global contrast normalization were applied to each localized knee joint image.

Samples of COVID and OA images after pre-processing are shown in Supplementary Fig 2 and Supplementary Fig 3, respectively.
3. Insights from OA severity prediction

Supplementary Fig 4 demonstrates that utilizing temporal representations in our architecture results in better defined clusters for the three severity grades (0, 1, 2) on the t-SNE plot. It may be observed that intermediate grades, such as 1 vs 2, which are more difficult to predict (left) can benefit from the proposed temporal approach (right). Additional CAMs of OA affected knees are compared in Supplementary Fig 5. Each row corresponds to knee radiographs from different severity grades, from 0 to 4. As may be observed, the attention maps from DeepKnee [8], CNN + Ordinal loss [1] and SE block [4] are very sparse and sometimes react to unnecessary areas (bone texture, joint centre). On the contrary, our method provides more focused attention on the osteophytes and joint narrowing - the two important indicators of osteoarthritis.

4. Recalibration using matching data

In the COVID-19 cohort, we included some matched data. 100 of 150 patients in the CovidProg temporal dataset also have their ventilation status known. We use the 100 patients, take the baseline scans (the first image) of their temporal sequences as matched snapshot images. We evaluated the distance between these matched temporal/snapshot data in the representation space through training. In Supplementary Fig 6, Curve A (blue) shows the average distance \( d \) between the matched pair of snapshot and temporal representation. \( d \) is reduced to only 0.10 after 40 epochs. For reference, we also show \( d \) between any snapshot of positive ventilation status (S+) and any temporal sequence of positive ventilation status (Tm+). The result is Curve B (orange). Meanwhile, we also report \( d \) between S+ and any temporal sequence with negative ventilation status (Tm-) as in Curve C (green). After 40 epochs, \( d \) in A,B,C are 0.10, 0.67 and 4.26, respectively. C >> B > A shows that (1) the matched snapshots and temporal sequences are automatically aligned very well during training, thanks to the MMD loss; (2) generally a positive temporal sequence is aligned much closer to a positive snapshot than a negative snapshot, although not as close as the matched pairs.

5. Longitudinal comparison

We also compared our method with other temporal models, namely CNN+LSTM, CNN+biLSTM and CNN+biLSTM+Attention. It may be observed from Supplementary Table 2 that our approach outperforms all these longitudinal models.

<table>
<thead>
<tr>
<th>Name</th>
<th>Ventilation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>AUC</td>
</tr>
<tr>
<td>CNN+LSTM</td>
<td>0.82</td>
</tr>
<tr>
<td>CNN+biLSTM</td>
<td>0.83</td>
</tr>
<tr>
<td>CNN+biLSTM+Attention</td>
<td>0.85</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>0.88</strong></td>
</tr>
</tbody>
</table>

Table 2. Comparison with longitudinal methods

6. Limitations

In our temporal analysis, the images are not registered. Registration might result in learning better representations. We aim to address this in the future by using a spatial transformer network as a pre-processing stage before extracting temporal features. Also, due to lack of sufficient temporal data, we did not use transformer architectures to learn disease progression. This research direction can be pursued
with the availability of more temporal cases in future.

Figure 4. Comparison between t-SNE plots before and after using temporal modeling for severity grades (0,1,2)

Figure 5. Qualitative comparisons of knee CAMs depicting OA severity grades 0 to 4 (top to bottom)

Figure 6. Distance between feature means across training

References