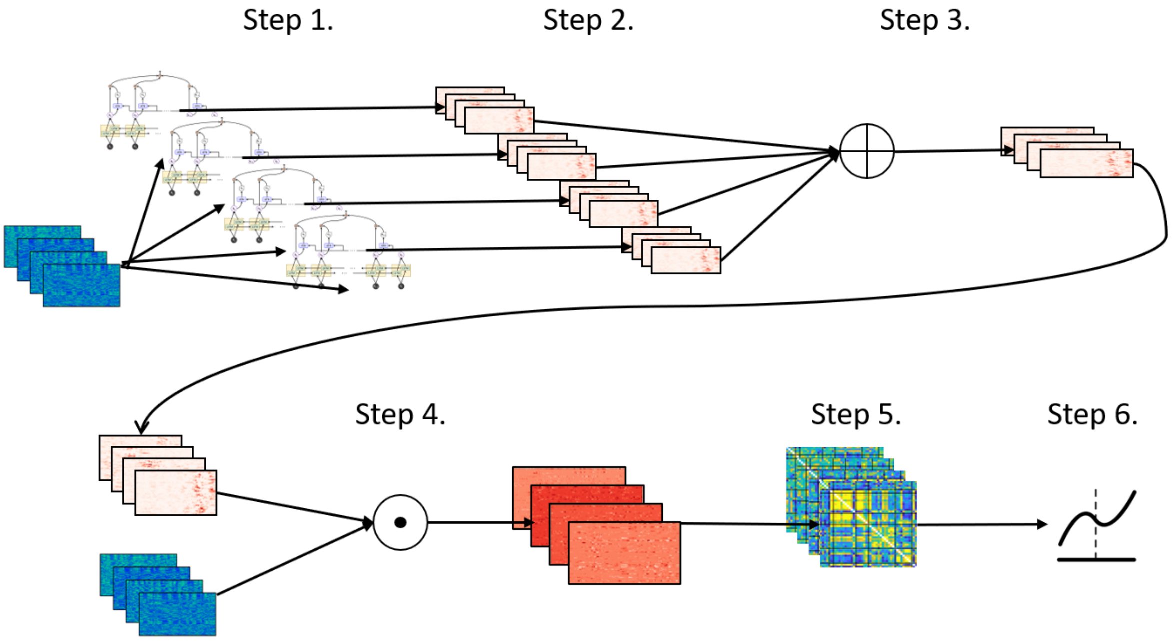
Paper ID: 2953

We thank the area chairs and the reviewers for the thorough and constructive feedback. We have fixed the typos, grammatical errors, spacing issues, and labeling errors in figures 2 and 3.

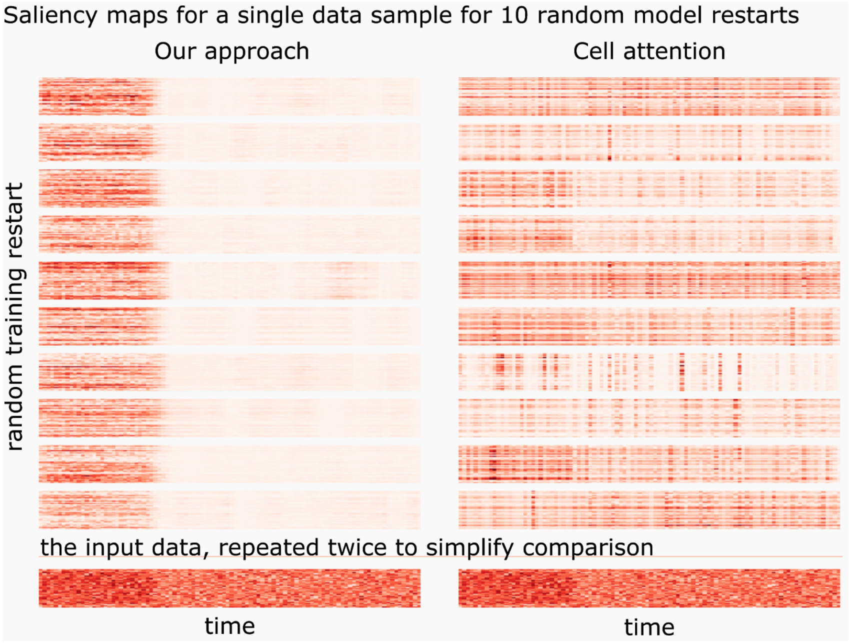
We also reduced the methods section, providing more space for the flow chart (below) and to fix the formatting issues.



*Flowchart describing our pipeline. Step 1: Train 30 separate models using different initializations for the same dataset. 2: Get saliency maps for all samples/models. 3: Calculate average per-sample saliency over all models and select the maps closest to the average. 5: Compute Hadamard between input and normalized maps. 6: Compute static FNC for each saliency-masked subject. 7: FNCs used for new classification.*

The reviewers pointed out that our work lacked comparisons with other methods. We are not aware of any other publications in the saliency of recurrent models but [20], as cited in our paper. It came out just 4 months before the MICCAI submission deadline and the code was not yet available. The major structural difference is [20] modifies the model heavily by adding parameters at each time step, requiring all data time steps to be fed into the model simultaneously. In contrast, our approach acts as a model introspection add-on, preserving all dynamic features of the model - a much less intrusive method. Our method allows for all possible encoding and decoding architectures, which would be difficult with [20]. Importantly, our method requires many fewer parameters.

Now that the code of [20] is available, we have compared our models using their simulations: Gaussian noise samples with one of the classes endowed with a raised rectangle 1) in the beginning and 2) at the end of the time series (see [20]). We train both models 10 times with a random restart. A sample of how saliency changes for the same input relative those re-trainings:



Our model is much better than cell attention at capturing salient features (above), but both exhibit some instability relative to random restart. A novelty of our work is the pipeline (see above diagram) that cleans and stabilizes the saliency maps so that they can be reliably used to analyze dynamic patterns of the input data, especially brain imaging data. For the above experiments, we calculated the average Euclidean distance and weighted Jaccard similarity and their standard deviation (see the table below). Our model is significantly more stable than cell attention in both.

| Class feature: | in the beginning | | at the end | |
|---|---|---|---|---|
| | Weighted Jaccard (larger is better) | Euclidean Distance (smaller is better) | Weighted Jaccard (larger is better) | Euclidean Distance (smaller is better) |
| Our model: | **0.229, $\sigma$: 0.014** | **0.780, $\sigma$: 0.012** | **0.23, $\sigma$: 0.009** | **0.786, $\sigma$: 0.012** |
| Cell attention: | 0.1, $\sigma$: 0.02 | 0.93, $\sigma$: 0.02 | 0.088, $\sigma$: 0.014 | 0.94, $\sigma$: 0.01 |

We’ve extended the manuscript with a brief description of the FBIRN data, a large multisite schizophrenia imaging study [16]. The parameters are precise yet the pipeline is standard and described elsewhere [9]. The length limitations prevent us from including it here.

In the cluster centroids, the y-axis is the component list, or regions of the brain with known functional roles, the x-axis is time. Using an elbow criteria, we found that 3 clusters was most appropriate. All typical controls fell into 1 cluster, the patients were split between the 2 other clusters. In the typical control centroid, component 25, the inferior parietal lobe was the most relevant to the disorder classification. In a patient centroid, components 14 (superior parietal lobe), 25, and 27 (superior medial frontal gyrus) were all highly salient. There is previous research to support this information [1][2][3].

[1] Palaniyappan L et al. *Eur Arch Psychiatry Clin Neurosci*. 2012

[2] Zhou SY et al. *Schizophr Res*. 2007

[2] Harms MP et al. *Br J Psychiatry*. 2010