ConsensusXAI: A framework to examine class-wise agreement in medical imaging

A. Supplementary material

Figures 2, 1, and 3 show consensus maps across varying cluster counts for ResNet50 and ResNet18 backbones. For correct predictions (Figure 1), the clearest maps are obtained with six clusters for ResNet50 and eight clusters for ResNet18 with LayerCAM (See Figure 2 in main text). In contrast, for incorrect predictions Figures 2, and 3, using two clusters consistently highlights relevant regions more effectively across both models.

Additionally, we performed further experiments using the "maximum" strategy across various combinations of ResNet backbones, considering both correct and incorrect cases (see Figures 4 and 5). These experiments demonstrate that the maximum strategy, which is a summary-based approach (Section 3.2.1), fails to produce meaningful consensus maps.

Figure 6 illustrates how the clustering patterns and corresponding consensus maps for the AMD class evolve as the number of clusters increases from 2 to 10. The visualization reveals that the ResNet50 backbone learns a dominant feature representation sufficient for correctly predicting AMD, as indicated by the consistent blue-colored cluster when using two clusters. This pattern remains largely unchanged with four and six clusters, suggesting that even two clusters are adequate for accurate AMD prediction. However, the optimal number of clusters is a tunable parameter and may vary across different classes in the Retinal-C8 dataset.

Figure 7 illustrates a toy example involving two dogs from the same class. In the leftmost heatmap, the highlighted region corresponds to the dog located in the bottom-left corner, while the adjacent heatmap to its right emphasizes the body of the second dog. These heatmaps represent distinct semantic concepts learned by a deep learning (DL) encoder. Our proposed method, "Latent Consensus", clusters such semantic concepts and generates a semantic agreement map—referred to as the consensus map shown in the rightmost heatmap, where both dogs are prominently highlighted.

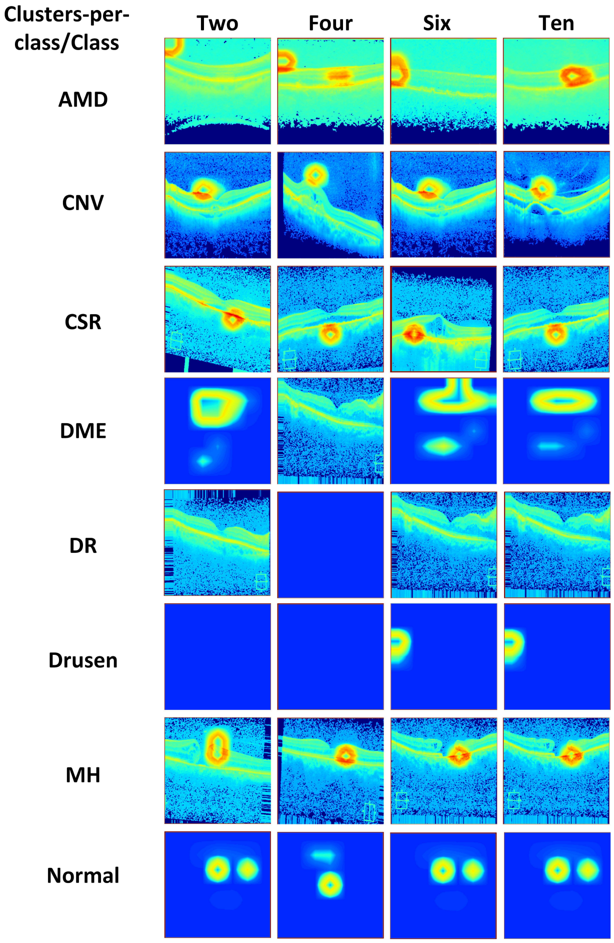


Figure 1. Consensus maps versus the number of clusters per class on the Retinal-C8 dataset using the ResNet18 backbone and LayerCAM for correctly predicted samples.

B. Reproducibility

To ensure reproducibility, we provide an open-source GitHub repository at `https://github.com/a-haider1992/cas_toolbox`. The repository includes comprehensive instructions in the README file. The codebase is organized into a structured directory and features a

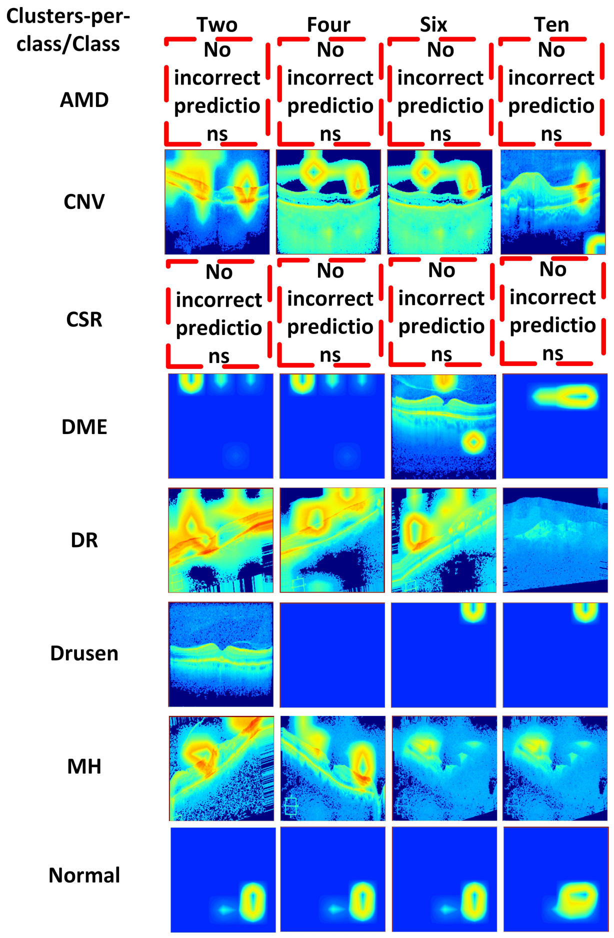


Figure 2. Consensus maps versus the number of clusters per class on the Retinal-C8 dataset using the ResNet50 backbone and Grad-CAM for incorrectly predicted samples.

primary entry point via the `main.py` script.

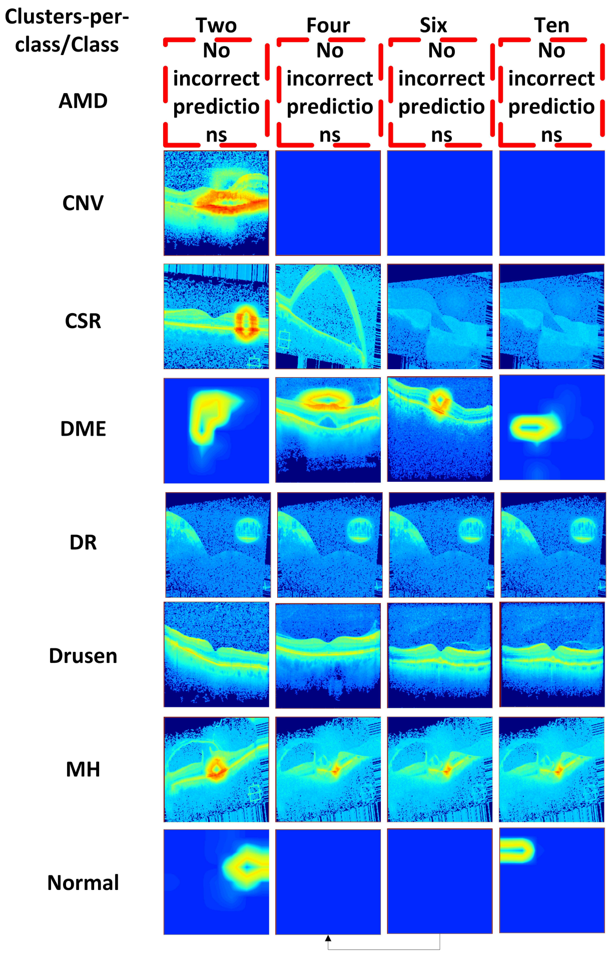


Figure 3. Consensus maps versus the number of clusters per class on the Retinal-C8 dataset using the ResNet18 backbone and LayerCAM for incorrectly predicted samples.

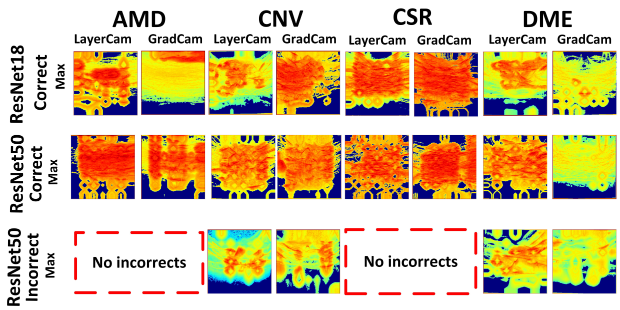


Figure 4. Consensus maps for the Retinal-C8 dataset are presented from left to right: visualization of the first four classes AMD, CNV, CSR, and DME.

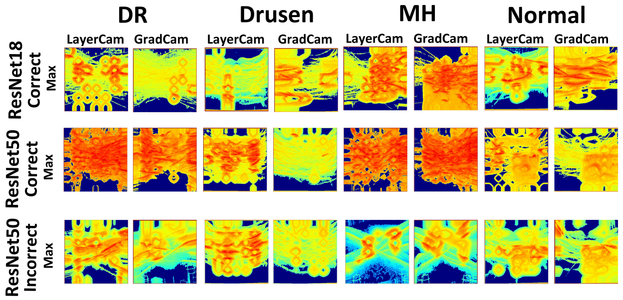


Figure 5. Consensus maps for the Retinal-C8 dataset are presented from left to right: visualization of the remaining four classes DR, Drusen, MH, and Normal.

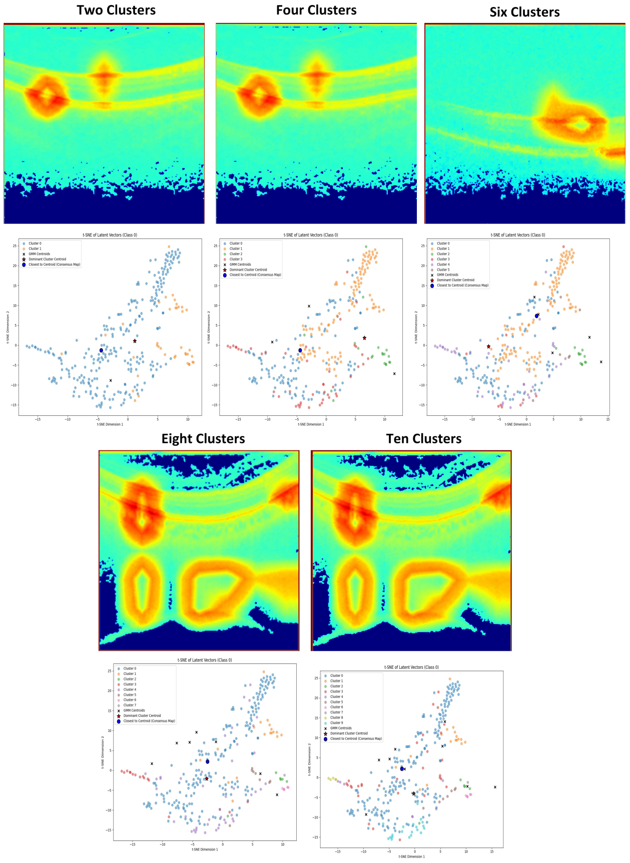
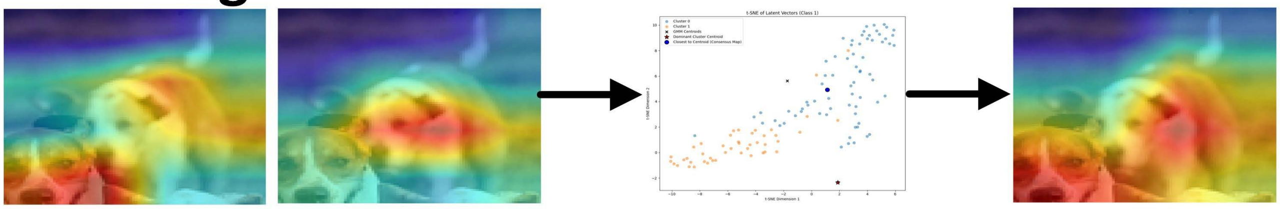


Figure 6. Consensus maps versus the number of clusters for AMD class of Retinal-C8 dataset using the ResNet50 backbone and GradCAM for correctly predicted samples.


t-SNE of Latent Vectors (Class 1)
Cluster 0
Cluster 1
GMM Centroids
Dominant Cluster Centroid
Closest to Centroid (Consensus Map)
t-SNE Dimension 1
t-SNE Dimension 2
−10
−8
−6
−4
−2
0
2
4
6
8
10

Dog class

Semantic concept1: Dog in bottom left corner is highlighted

Semantic concept 2: Center dog body is highlighted

t-SNE of Latent Consensus **Cluster1:** Face features **Cluster 2:** Body features

Consensus map: Dog's face and body both are semnatic concepts a DL model could use to correctly identify a dog

Figure 7. An illustrative example of how consensus mapping captures collective agreement across multiple semantic concepts.