Supplementary materials: DiffExplainer: Unveiling Black Box Models Via Counterfactual Generation

Yingying Fang, Shuang Wu, Zijin Zhao, Shiyi Wang, Caiwen Xu, Simon Walsh, Guang Yang

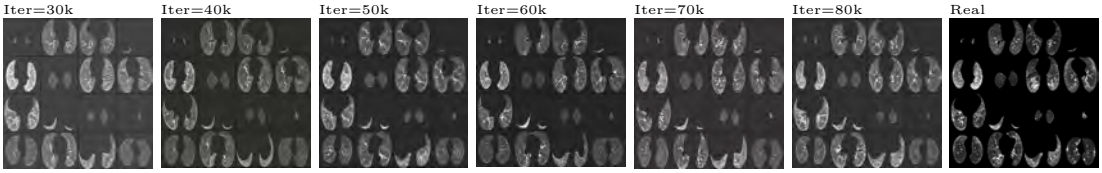


Figure 1: 16 reconstructed examples from the StyleGAN-based autoencoder. The 1st to 6th columns display reconstructions from different iterations.

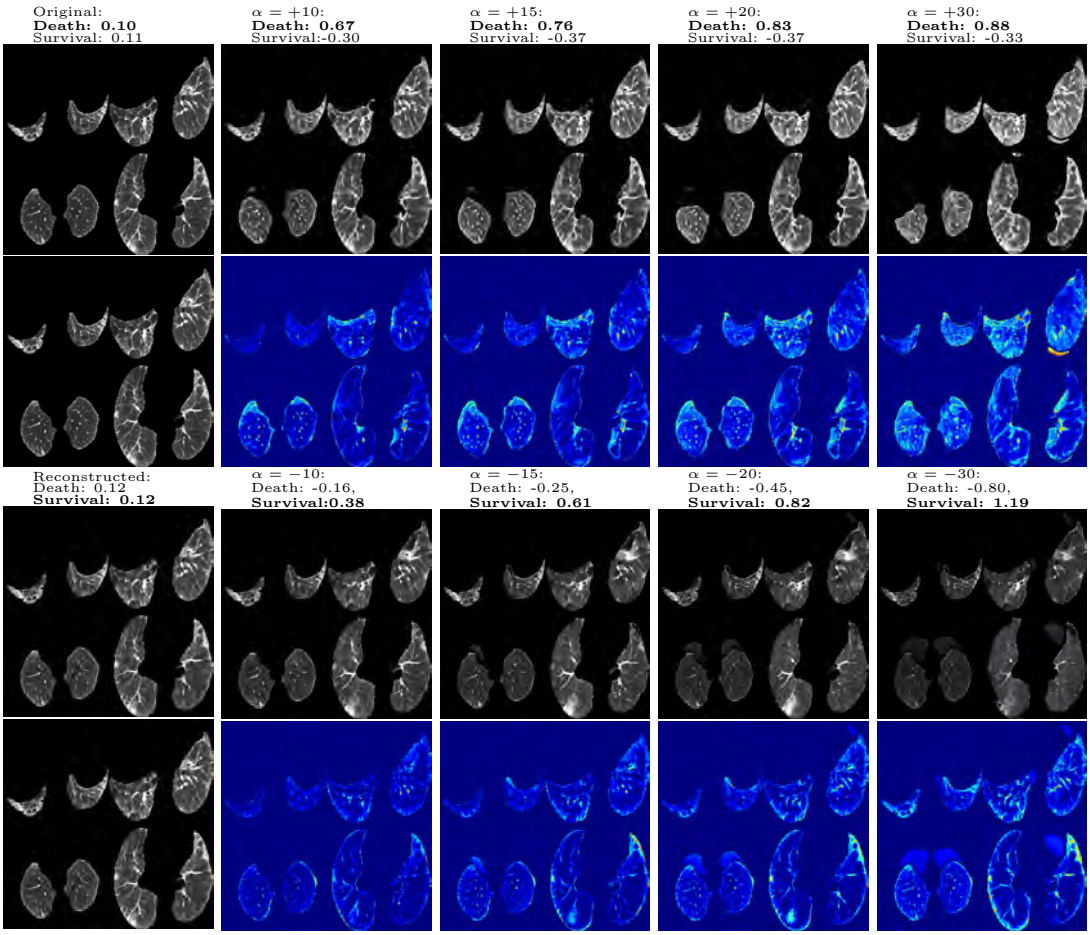
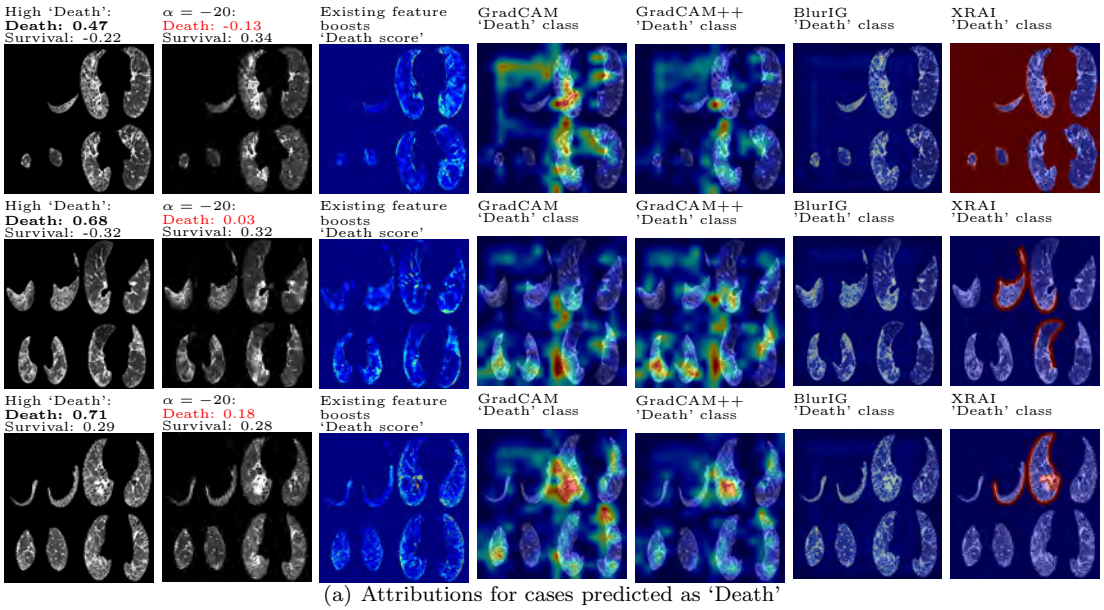
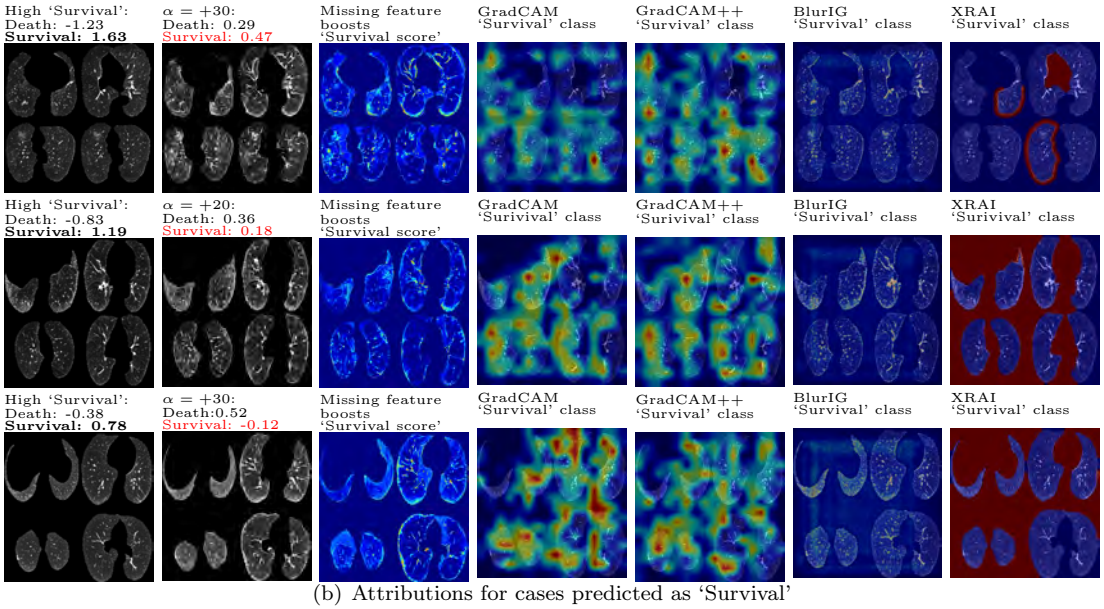


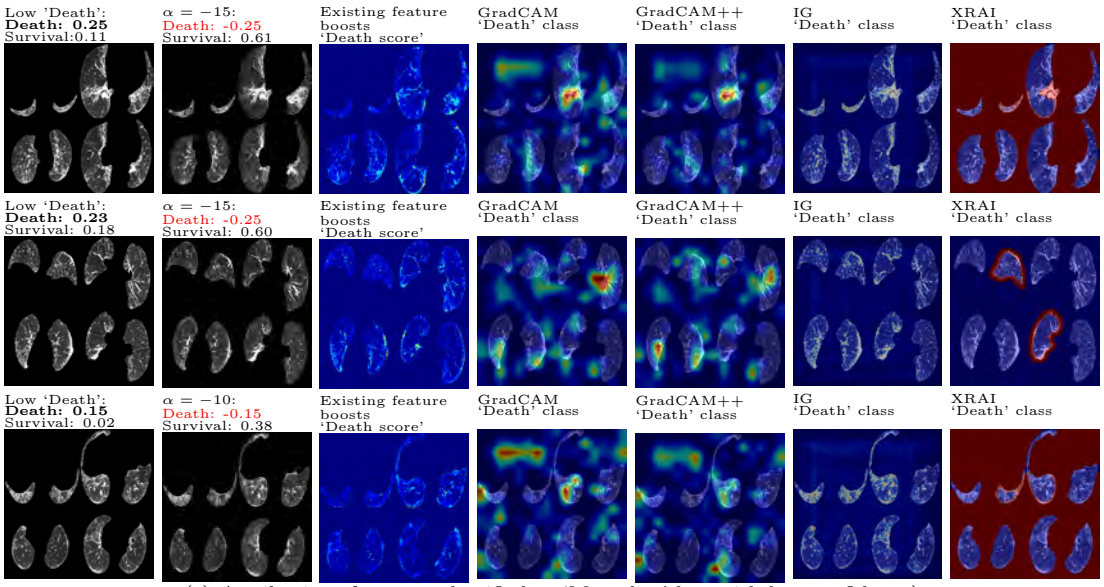
Figure 2: Counterfactual generation for gradually confident classification results.



(a) Attributions for cases predicted as 'Death'



(b) Attributions for cases predicted as 'Survival'



(c) Attributions for cases classified as 'Mortality' but with low confidence)

Figure 3: Comparison of feature identification methods for cases with confident predictions of 'Mortality in one year', confident predictions of 'Survival in one year', and indeterminate predictions from FLD datasets.