Additive MIL: Intrinsically Interpretable Multiple Instance Learning for Pathology Supplementary Material

In the supplementary material, we provide additional heatmaps from TCGA and Camelyon16 datasets which compare attention and Additive MIL heatmaps on pathology whole-slide images. We also show results for an expert evaluation of the heatmaps' utility towards a clinical-grade tool.

1 Heatmaps from TCGA-RCC Dataset

Figure 1 shows three cases of Renal Cell Carcinoma (RCC). The MIL models are trained to predict the sub-type present in the slide, namely clear cell carcinoma (KIRC), papillary cell carcinoma (KIRP), & chromophobe (KICH). In case (a), attention heatmap is shown to attend to regions predictive of both KIRC and KIRP whereas Additive MIL heatmap correctly identifies the presence of individual sub-types within the same slide. Similarly in (b), the slide shown is labeled as KIRC, however it contains areas with papillary structure as highlighted by the Additive MIL heatmap. Note that attention heatmap does not highlight these regions. In case (c), the slide is labeled as KICH and the model correctly predicts it. The attention heatmap highlights relevant KICH regions in pink. However, it misses showing patches contributing to the other two classes spuriously which are visible in the Additive MIL heatmap.

2 Heatmaps from Camelyon16 Dataset

Figure 2 shows three cases of metastatic breast cancer from the Camelyon dataset. Case (a) shows a malignant slide where the model gives the correct prediction. Attention heatmap highlights certain regions for this prediction, but it's not clear whether the patch provides excitatory or inhibitory contribution for the malignant class. In contrast, Additive MIL heatmap shows that the patches in blue are inhibitory towards the predicted class. This highlights the a key limitation of attention heatmaps which show patch importance but not their predictive value towards or against a class. Case (b) is a Benign slide which is mis-predicted as Malignant. The attention heatmap does not highlight any regions, however the Additive MIL heatmap correctly identifies and localizes the false positive failure mode of the model. This makes Additive MIL models suitable for granular model debugging. Case (c) is a malignant slide correctly predicted by the model. The attention heatmap only localizes a single cancer focus on the left side of the slide even though the whole piece of tissue is malignant. Additive MIL heatmap correctly identifies other cancer foci as well.

3 Heatmaps from TCGA-NSCLC Dataset

Figure 3 shows three cases of Non-Small Cell Lung Carcinoma (NSCLC). The MIL models are trained to predict the sub-type present in the slide, namely Adenocarinoma & Squamous Cell Carcinoma. Additive MIL heatmap in (a) shows the model picking up regions predictive of both sub-types even though the model correctly predicts the slide to be Squamous Cell Carcinoma. This information is absent from attention heatmap. Note that in this case, the attention heatmap shows high importance for regions from both sub-types. In case (b) however, the most attended patches only correspond to Adenocarcinoma shown in yellow even though regions predictive of both sub-types

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Dataset	Attention Heatmap	Additive MIL Heatmap
TCGA-RCC	6/39	33/39
Camelyon16	1/50	49/50

Table 1: Expert evaluation of Additive and attention heatmaps for highlighting regions of interest in TCGA-RCC cancer sub-typing and Camelyon16 cancer identification task. The scores indicate the proportion of slides where a board-certified pathologist prefers a particular heatmap.

exist. This ambiguous behavior of attention heatmap complicates interpretation. In (b), which is an Adenocarcinoma slide, Additive MIL heatmap shows the model being uncertain about the two classes and localizes that uncertainty to a specific region even though the final prediction is correct. In (c), we again see attention heatmap highlighting patches corresponding to both sub-types without distinguishing between them, while the Additive MIL heatmap clearly delineates the regions predictive of the two classes.

4 Expert Evaluation of Additive Heatmaps & Applicability in Decision Support Tool

We conducted an expert evaluation of the heatmaps to assess their usefulness for highlighting regions of interest in Camelyon16 and TCGA-RCC slides using both Additive MIL and attention heatmaps. For Camelyon16, we selected a random sample of 50 slides from the test set with a 1:4 distribution of benign-to-malignant class. For TCGA-RCC, we randomly selected 39 slides with equal representation from all 3 classes. A board-certified pathologist was asked to evaluate the heatmaps based on the following question:

"Which heatmap out of the two would you prefer to use in an AI+human decision-support setup for highlighting regions of interest before you give your diagnosis?"

The results from the study are tabulated in Table 1. The scores for each heatmap are calculated by counting the number of times an expert pathologist would prefer one heatmap over the other. It clearly shows that Additive MIL heatmaps are almost always preferred over attention heatmaps. The main reason for this preference for TCGA-RCC was - "The Additive MIL heatmaps highlight patches for individual classes which can serve as visual reminder for pathologists to consider other differential diagnosis. "For Camelyon16, the pathologist feedback was - "Between Additive and Attention MIL, the former is preferred because the latter has more false positives and false negatives in all slides except one".

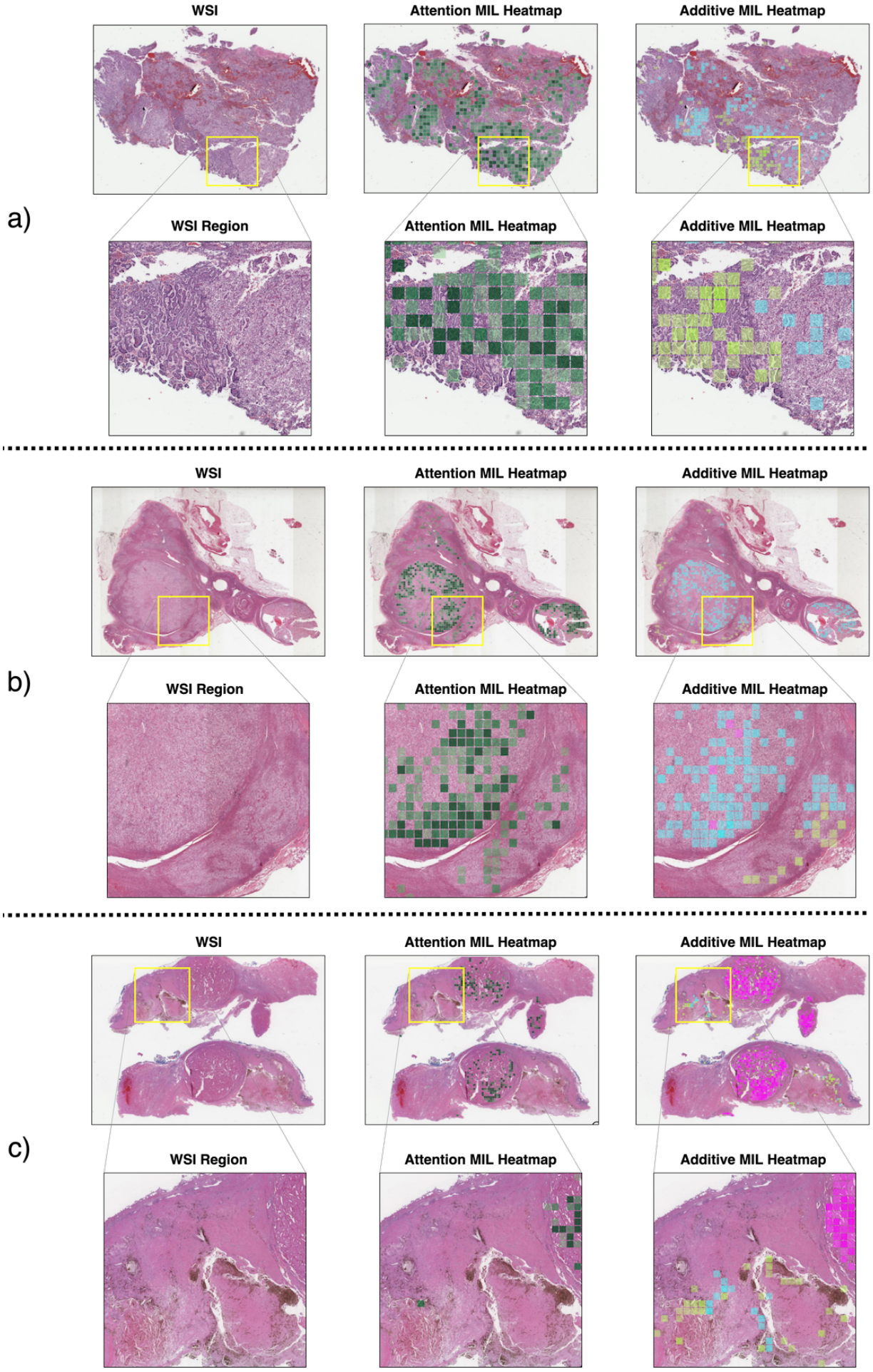

Figure 1: Comparison of Additive MIL and attention heatmaps. Cyan patches denote KIRC, lime green patches denote KIRP, and pink patches denonte KICH. Attention heatmaps are shown in green. Additive MIL heatmaps highlight regions different from attention heatmaps and offer more granularity in interpretation. See section 1 for details about the shown cases.

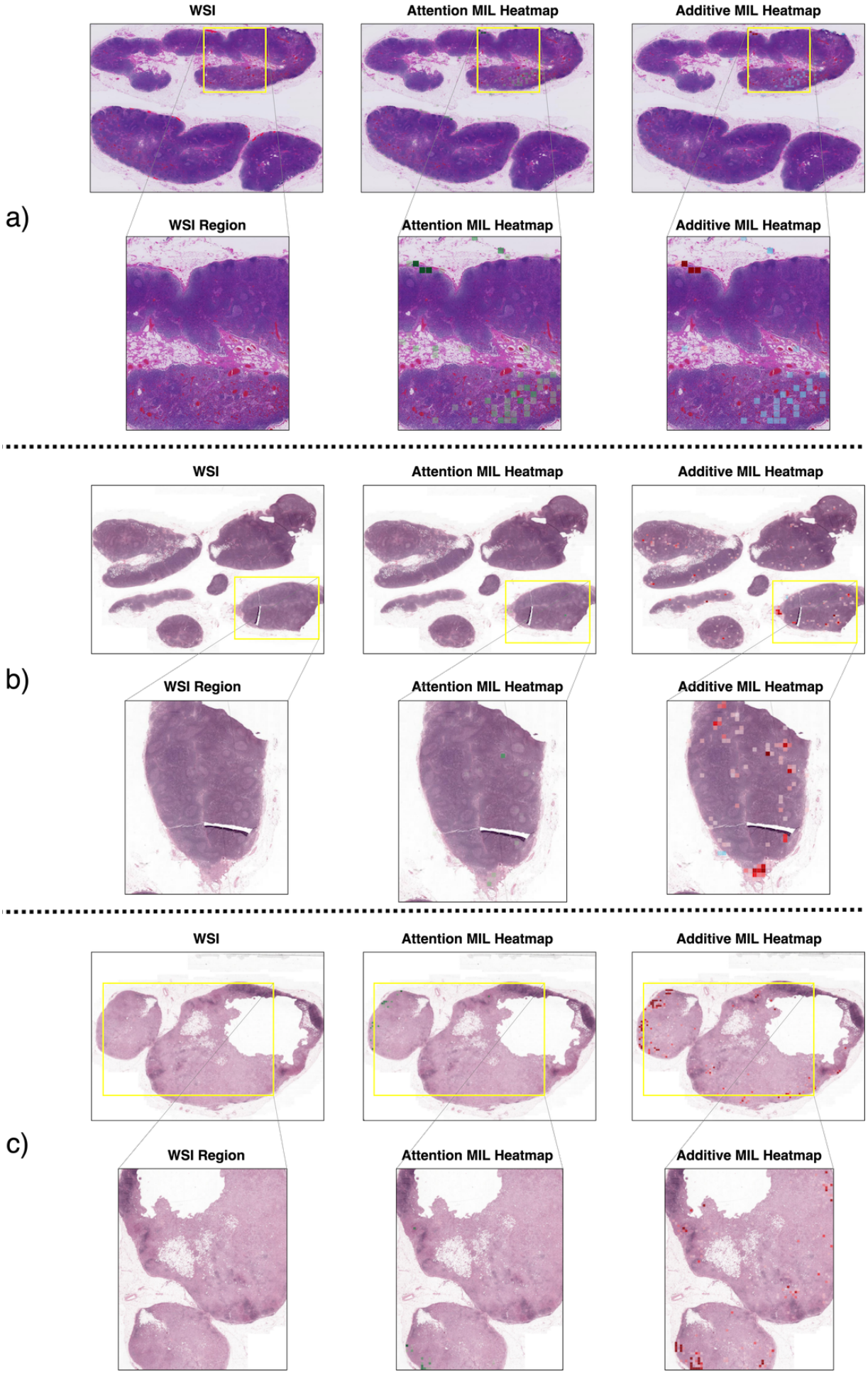

Figure 2: Comparison of Additive MIL and attention heatmaps. Red patches denote the class MALIGNANT and blue patches denote the class BENIGN. Attention heatmaps are shown in green. Attention heatmaps do not distinguish between excitatory & inhibitory patch contributions and often do not highlight false positive patches which are critical for model debugging. See section 2 for details about the shown cases.

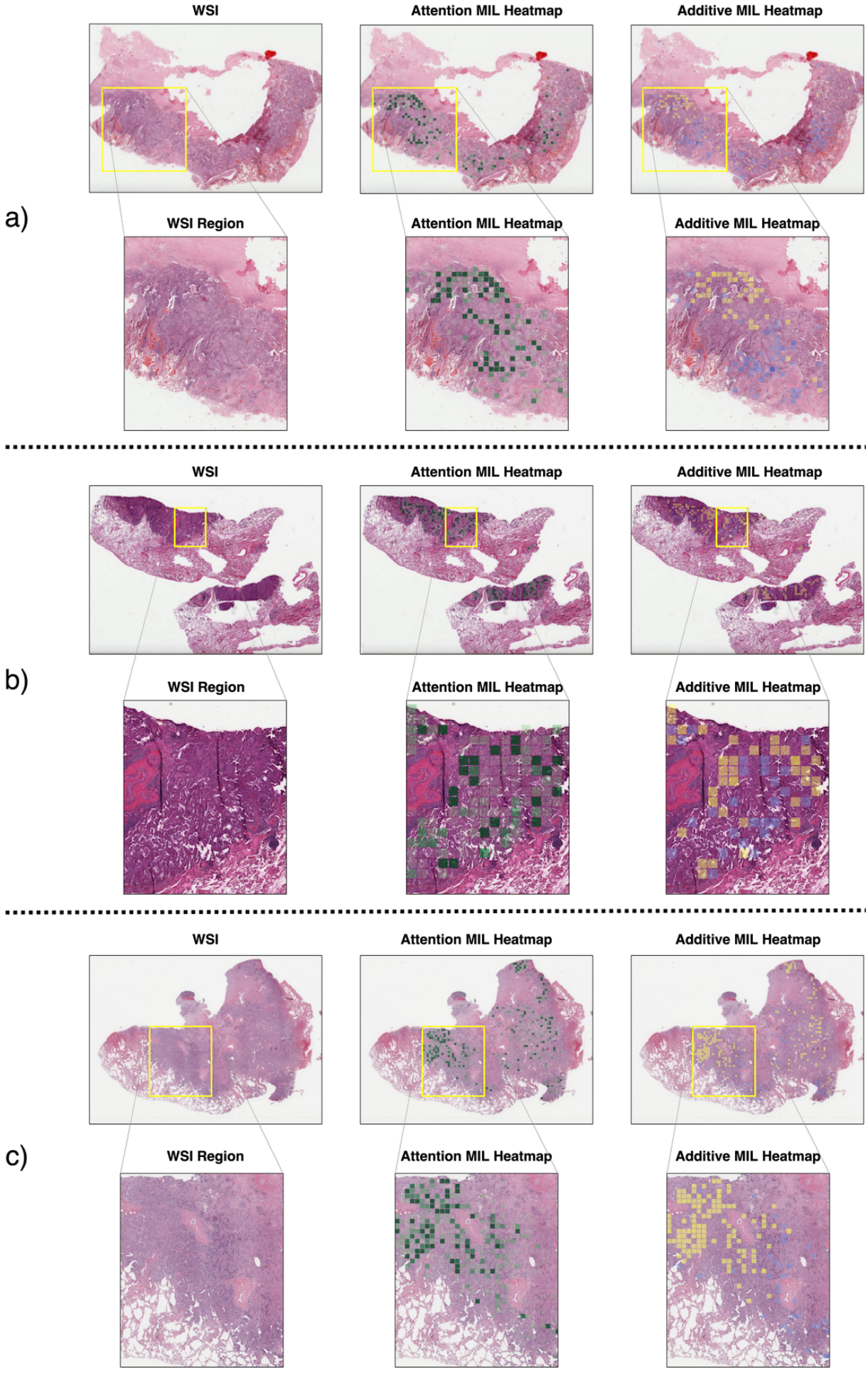

Figure 3: Comparison of Additive MIL and attention heatmaps. Yellow patches denote Adenocarcinoma and blue patches denote Squamous cell carcinoma. Attention heatmaps are shown in green. They lack class-dependent patch attribution and often differ in their patch contribution values as compared to Additive MIL heatmaps. See section 3 for details about the shown cases.