**Supplemental Digital Content**

**Supplemental Methods**

Generating Individual Tumor Regions from Predicted Heatmap

Application of the deep learning algorithm across the slide generates a two-dimensional matrix of tumor likelihood predictions (heatmap). Each value *p* in the heatmap corresponds to the likelihood of a 32 x 32 micrometer tissue area containing tumor. However, we discovered that creating a sensible (ranked) list of predicted tumor regions using this heatmap was non-trivial.

The most straightforward solution involves applying a single threshold *t* to produce a binary mask such that the entire heatmap is divided into tumor (p>=t) and benign (p<t) regions. Next, we can label each contiguous region (also termed connected component) a tumor. However, we found that any given threshold tended to produce both false positives and false negatives, resulting in highlighting of tumor regions that was difficult to interpret.

Therefore, we devised an adaptive threshold approach that uses a different threshold ti for each predicted region and slide. Our approach involves repeating the following steps:

1. For the *i-*th region, select the maximum value in the heatmap, pi.
2. Use the threshold k\*pi to produce a tumor-benign binary mask.
3. Define region *i* to be the contiguous region (connected component) that contains pi, and the predicted value for that region to be pi (which will also be the maximum value in that region).
4. Exclude all predicted values in the heatmap belonging to region *i* for subsequent steps.

We found empirically that k=0.5 worked well in selecting an appropriate threshold for each predicted region regardless of the predicted value for that slide.

Finally, we classified the list of regions into two categories, high confidence and moderate confidence using thresholds obtained from an independent set of images, the Camelyon16 test dataset. The high confidence (high specificity) threshold was chosen to achieve 66% sensitivity across individual tumors with no false positive predictions. The moderate confidence (high sensitivity) threshold was chosen to achieve 90% sensitivity across individual tumors; false negative tumors were typically less than 20 cells in size.

To avoid biasing the study by artificially inflating the algorithm’s accuracy, we performed a thorough visual comparison of the algorithm outlines and immunohistochemistry(IHC)-stained slides only after the completion of the study. The algorithm correctly highlighted the vast majority of the individual tumor foci larger than 200 microns detected by IHC. Exceptions (i.e., foci-level false negatives) contained necrosis or apoptosis, or presented as small infiltrative tumor foci admixed with individual tumor cells in a background of reactive stroma, bland morphology with poor contrast due to H&E staining and white balance setting on scanning, severe out-of-focus, crushed artifact combined with out-of-focus, or non-classical morphology on H&E. For foci smaller than 200 microns, foci that displayed classical morphology on H&E, such as pleomorphic cells mimicking ducts in the capsule were correctly outlined, while rare cytokeratin-positive tumor cells were inconsistently detected. These observations are consistent with our designed use case involving fairly aggressive filtering on the algorithmic output (see “Digital Assistant Design” section in Methods) to best assist image review without excessive distractions. In real use cases, both the number of displayed regions and the number of confidence categories can be adjusted for more thorough algorithmic assistance and further reduction of foci-level false negatives.

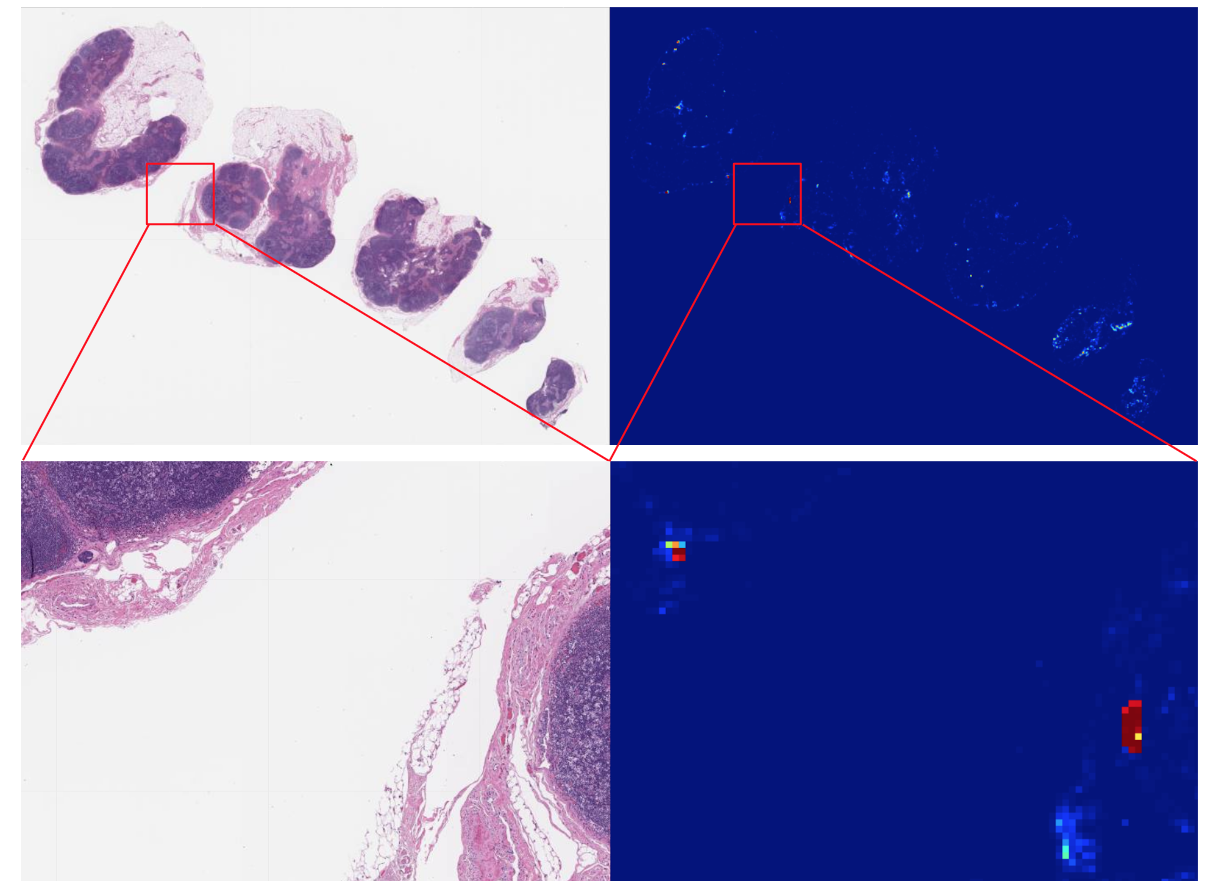
Pilot Study

An initial pilot study to establish basic study guidelines and best practices was conducted with 40 randomly selected cases. Notably these randomly selected images contained only 3 cases with micrometastases, precluding any conclusions for this case type. Although this pilot study did not provide sufficient data for power calculations regarding micrometastases, it did suggest that additional diversity of images would be necessary for case type specific evaluation. Thus, in order to allow adequate evaluation of micrometastases while maintaining a feasible number of total images, the final image set was selected with moderate enrichment for micrometastases.

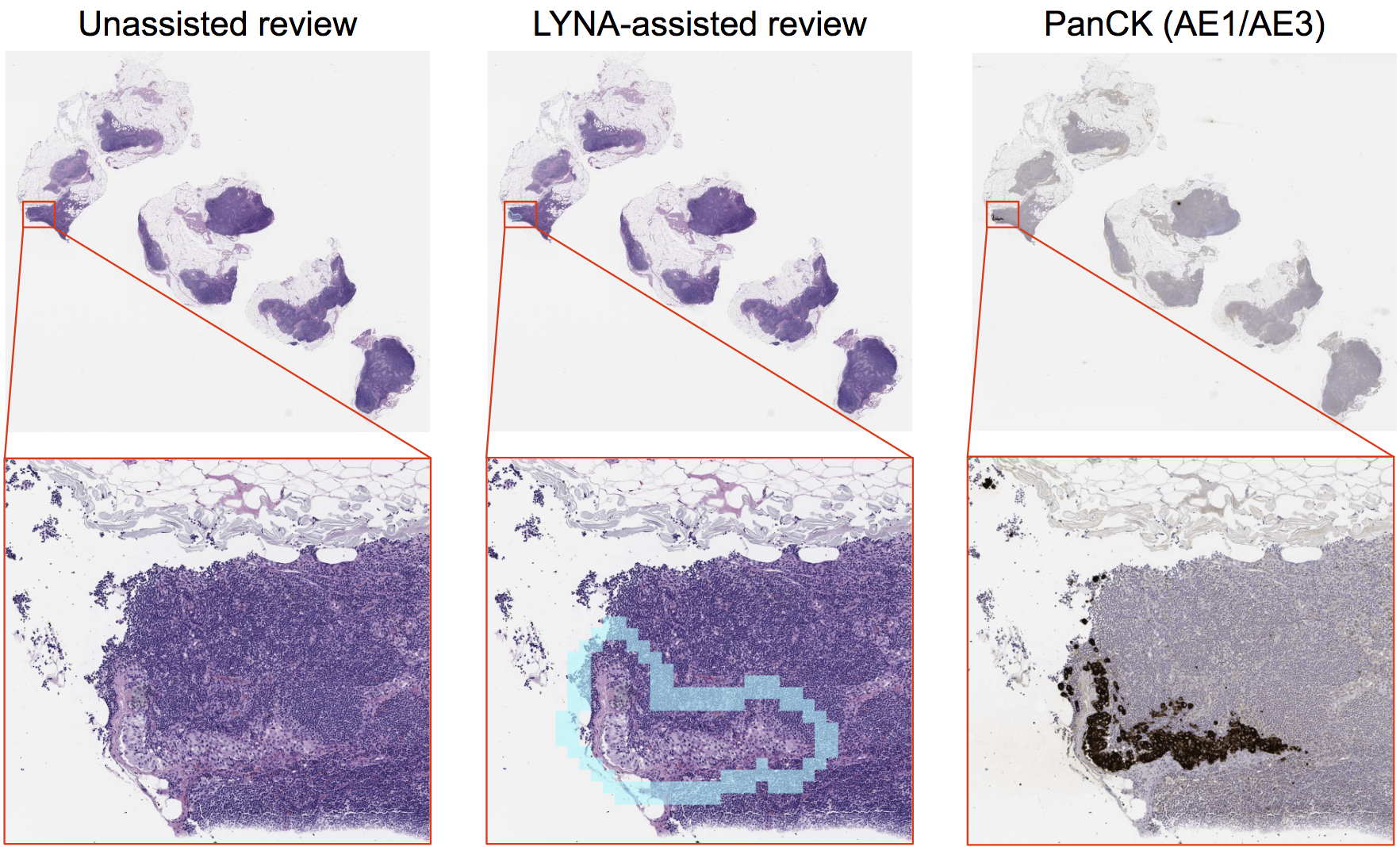
This pilot study also involved the use of a patch-level heatmap display of algorithm predictions across the entire image, rather than the limited number of moderate and high confidence outlines described in the main manuscript. In the interest of displaying the most complete information, we displayed predictions in the pilot study using a “jet” color heatmap, side-by-side with the H&E image (Supplemental Fig. 1). In this color scheme, blue indicated low likelihood of being tumor, red indicated high likelihood, and colors in-between indicated intermediate likelihood.

However, we found that distinguishing relative likelihood based on color alone was difficult and made it difficult for pathologists to calibrate understanding and trust of the algorithm predictions. This, together with the natural tendency of pathologists to review all data of potential concern, caused pathologists to review many regions (with low predicted value by the algorithm) at high power. In addition, the side-by-side heatmap either required two side-by-side monitors or occupied half of the computer monitor’s viewing area, reducing the amount of tissue that could be seen at any one time. This “screen real estate” problem was exacerbated if using smaller monitors, such as laptop screens. Finally, pathologists commented that although the tissue and heatmap were aligned and the navigation was synchronized, it was still difficult to visually gauge precisely which cells on the H&E image corresponded to a given location on the heatmap. The use of a crosshair centered at the same location on both the tissue and heatmap panels was moderately helpful in this respect, but the left-right image matching remained unintuitive.

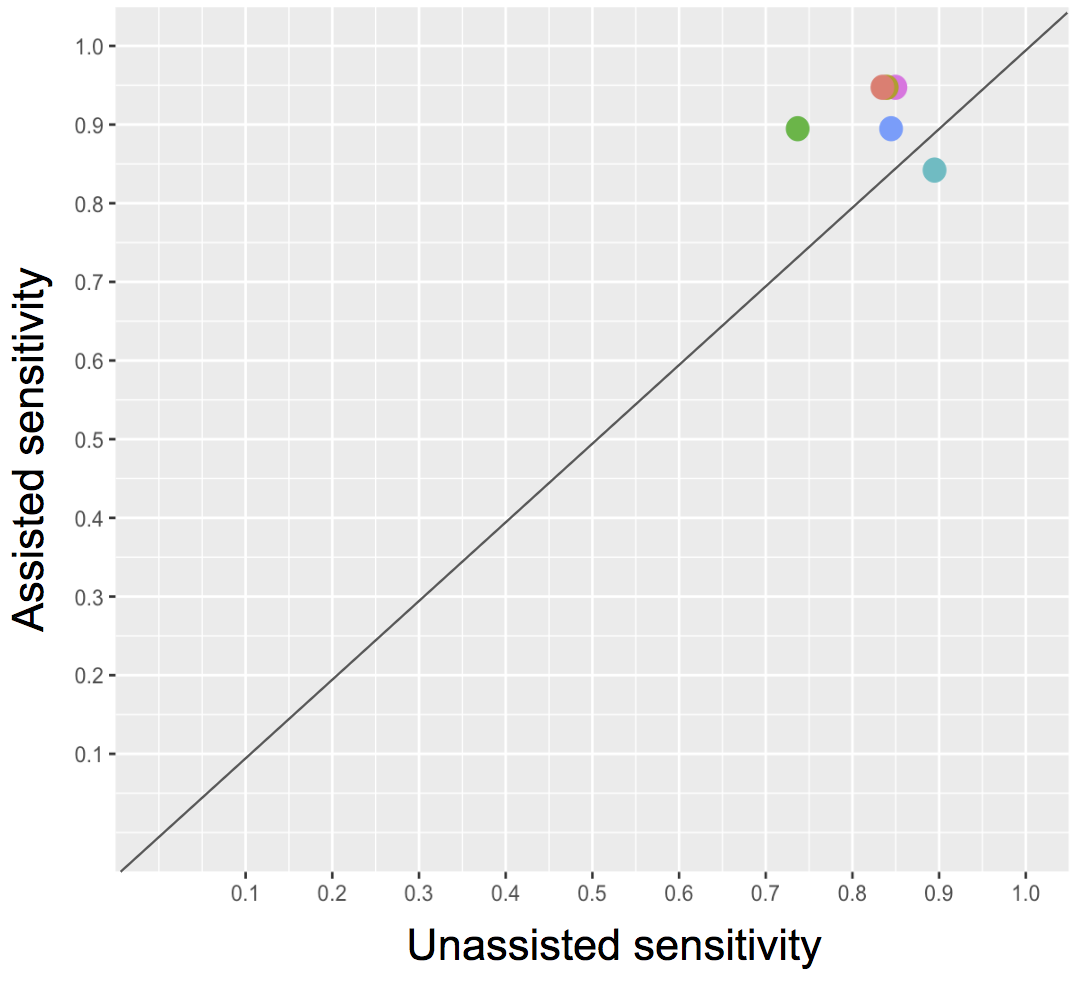
**Supplemental Figures**

****

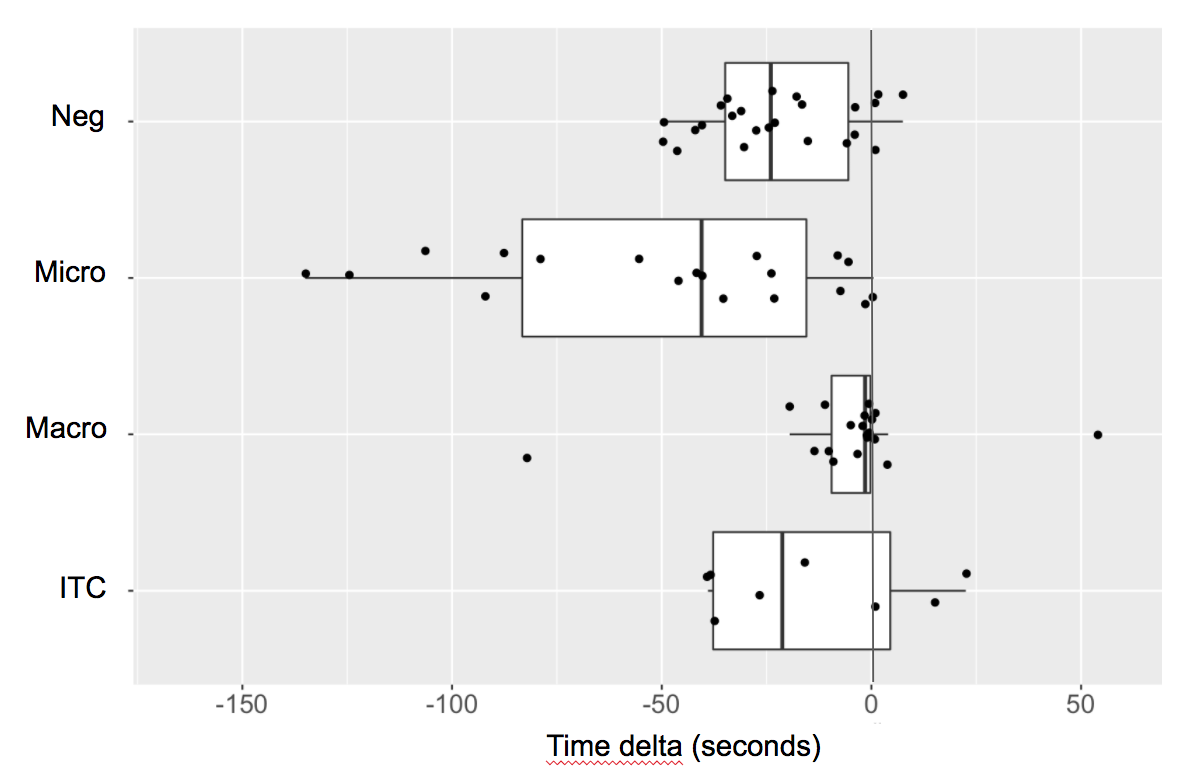
**Supplemental Figure 1**. Side-by-side heatmap interface used in the pilot study for the same image shown in Fig. 2 of the main text. The left shows the original H&E image, while the right displays a “heatmap” rendition of the predictions of the algorithm for the entire slide. High likelihood of tumor is indicated by red, low likelihood by blue, and intermediate likelihood by colors in between. These registered, side-by-side images were initially presented on adjacent, full size monitors. In a pilot study, this interface was more distracting than helpful as it often resulted in time-consuming review of many intermediate or low likelihood regions at high power, but that were not reviewed as thoroughly in the absence of the heatmap.

****

**Supplemental Figure 2**. Representative image of a micrometastasis predicted by LYNA with high confidence. Comparison of digital slide review without assistance (left) versus with assistance (middle). Suspicious regions of interest (ROI) are highlighted in cyan for high confidence and green for moderate confidence. Cytokeratin IHC is provided as reference on the right. Assistance improved the sensitivity of metastasis detection on H&E alone across pathologists in this case.

****

**Supplemental Figure 3.** Assisted vs unassisted sensitivity for detecting micrometastases. Each colored dot represents sensitivity across 19 micrometastases for an individual pathologist, unassisted (x-axis) vs. algorithm-assisted (y-axis). Dots above the diagonal line represent improved sensitivity with assistance.

****

**Supplemental Figure 4.** Plot of the time difference per image with algorithm assistance vs. without assistance, grouped by category. A negative time delta indicates a shorter review time with assistance. Dots are median values for each image.

****

**Supplemental Figure 5**. Left: LYNA highlights the larger clusters of tumor cells on the H&E image but does not completely circumscribe the contiguous tumor in this specimen, which demonstrates fibrotic change. This image was labeled as macrometastasis and correctly classified as such in both assisted and unassisted modes by all readers. This image was the only macrometastasis to take longer on average with assistance than without. Right: Corresponding IHC stain.

**Supplemental Tables**

**Supplemental Table 1**. LYNA’s case-level performance on the images in this study, stratified by the two types of outlined regions: moderate confidence and high confidence.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Confidence** | **Outline color** | **Maximum number of regions** | **Sensitivity (% of metastasis-containing cases detected)** | | **Specificity (% of negative cases labeled correctly)** | **Positive predictive value (% of predicted positive cases that contain metastasis)** | |
| **Macro + micro** | **Micro** | **Independent of macro/micro** | **Macro + micro** | **Micro** |
| High (almost certainly tumor) | Cyan | No maximum | 95% (36/38) | 95% (18/19) | 96% (23/24) | 97% (36/37) | 95% (18/19) |
| Moderate (might contain false positives) | Green | Up to three (including any high confidence regions) | 100% (38/38) | 100% (19/19) | 67% (16/24) | 83% (38/46) | 70% (19/27) |

**Supplemental Table 2**. Images with initial disagreement and reference standard classification after adjudication with additional review and measurement of tumor foci on the IHC image. The initial discrepancy, IHC measurement, and final category classification are described. Regions of interest for Image 68 are also shown in Supplemental Fig. 5.

|  |  |  |  |
| --- | --- | --- | --- |
| **Image** | **Discrepancy** | **Size measurement on IHC** | **Adjudicated classification** |
| 12 | Micro vs. Macro | 2.3 mm | Macro |
| 49 | ITC vs. Micro | 0.2mm (>200 cells) | Micro |
| 68 | ITC vs. Micro | 200-250 isolated cells | Micro |