GoogleNet with altered stem:

-The whole slide images were tiled into 512x512 fixed sized images without any overlaps.

-Each tile was then filtered based on its pixel values to remove empty/white space captured on the WSI resulting in ~2000 tiles per WSI.

-- A fixed binary thresholding was used to highlight the white regions, and eliminating tiles if the total area of two largest connected components of the highlighted white regions is greater than half of the tile. We fine-tuned the fixed thresholding value as well as the number of connected components to get the best segregation of relevant and irrelevant tiles.

-For 193 whole slide images (495k tiles), 60% of the patients' tiles were selected for training (297k tiles) and the remaining (100k) each were used for "dev" and test sets.

-Utilize GoogLeNet 2014 (https://arxiv.org/pdf/1409.4842.pdf), a 22-layer model with additional modifications to the initial set of layers (also known as the stem of the GoogLeNet) to support 515x512) sized input images.

-Stem Modifications

-512x512x3 (input) –
-[Conv2d kernel=14x14, s=2] –
-(256x256x64) –
-[Maxpool2d kernel=3x3, s=2, p=1] –
-(128x128x64) –
-[batch norm + relu] –
-(128x128x64) –
-[Conv2d kernel=1x1, s=2] –
-(64x64x192) –
-[Conv2d kernel=3x3, s=1] –
- (32x32x192) –
-[Maxpool2d kernel=5x5, s=2, p=2] –
-(32x32x192) –
-[Native version of GoogleNet]

-Our training method used Adadelta optimization techniques which scales the learning rate based on a historical gradient while taking into account only the most recent time window.

-Only the last linear layer is modified to fit the various classification tasks at hand, e.g. final linear layer for 2-class cohort ER-Status, would be designed to take in 1024 features learned from the CNN ultimately produce a 2-node output which is further processed by a softmax layer to result in final class assignments.

-Due to a sparse final layer, we introduce a dropout layer with a 50% ratio of dropped outputs, preventing general overfitting.

Inception v3 model:

-Employed the Inception (v3) version of the GoogleNet (https://arxiv.org/pdf/1512.00567.pdf)

- Network contains 299x299 receptive field.

- Much more gradual decrease in the dimensions of the feature maps across the initial few layers, followed by the signature set of inception-v3 layers (crucial in reducing the grid size while avoiding the representational bottleneck, by preventing extreme compression)
-No modifications to the stem.
-Tile the whole slide images in accordance with the 299x299 receptive.
- ~5000 tiles on average per whole slide(Total : 960k images – 60-20-20 split) of patients.
– Filtering whitespace performed as detailed in the previous section.

Inception v3 model with pre training:

- Utilize the network detailed above, with all but last layer initialized by pre-training on ImageNet dataset of about 1.28 million images.

Inception v3 model with pre training and boosted data:

-Utilize the network detailed above
- Tiling of each WSI was performed as a sliding window, which results in some overlap between two adjacent tiles. The overlap for every tile was 75 pixels, i.e. the stride for the sliding window was 224 pixels.
- 3 Additional tiles were generated from each original tile, by rotating the original tile (90,180 and 270 degrees).

Inception v3 model with pretraining with only epithelial regions

-Utilize the network detailed above (without boosted data)
- Utilize tiles only containing primarily epithelial regions
- Classifier developed in house (Details available on request), based on segmenting homogenous regions in tissue images to superpixels (https://infoscience.epfl.ch/record/149300) and classifying these regions as epithelial or stromal
- We selected tiles that overlapped with specifically epithelial superpixels and utilized them as input.