Deep Learning for Identifying Metastatic Breast Cancer

Alex Martinez, Ronjon Nag, Alex Tamkin, Pol Rosello, Pranav Sriram

December 5, 2016

1 Introduction

Diagnosing diseases based on images taken with a microscope is a difficult, time-consuming, and error-prone process currently performed almost exclusively by human specialists. In recent years, there has been an increasing interest in developing computational techniques that can classify microscopic slides automatically, assisting in diagnosing diseases such as cancer faster, cheaper, and more accurately.

In their paper, Wang et al. use a deep convolutional neural network to automate detection of metastatic breast cancer in whole slide images of sentinel lymph node biopsies [1]. They focus on two tasks: (1) slide-based classification – whether or not a slide contains metastatic cells – and (2) tumor-localization – determining where in the slide the tumor is located. Their approach uses an initial pre-processing stage to identify the tissue in the slide. They then train a deep convolutional neural network to make patch-level predictions to discriminate tumor-patches from normal-patches. Finally, they aggregate the patch-level predictions to create tumor probability heatmaps and perform post-processing over these heatmaps to make predictions for the slide-based classification task and the tumor-localization task.

Their model won the International Symposium on Biomedical Imaging's Camelyon Grand Challenge in 2016 with a performance approaching human-level accuracy, obtaining a whole slide image classification AUC of 0.966 and a tumor localization score of 0.733. The authors also introduce a method for combining their model's predictions with a professional pathologist's and show that the resulting predictions are better than either agent's predictions alone.

2 Dataset and Evaluation Metrics

The Camelyon16 dataset consists of a total of 400 whole slide images (WSIs) split into 270 for training and 130 for testing. Both splits contain samples from two institutions: Radbound UMC and UMC Utrecht. Ground truths consist of delineated regions of metastatic cancer on the WSIs.

Submissions to the competition were evaluated on the following two metrics:

- Slide-based Evaluation: Judges measured AUC score derived from competitor submissions of predicted likelihood of containing cancer for each WSI.
- Lesion-based Evaluation: Judges measured average sensitivity for detecting all true cancer lesions in a WSI across 6 given false positive rates. Submissions consisted of probability and (x, y) location for each predicted cancer lesion within a WSI.

3 Context and Related Work

The evaluation of breast sentinel lymph nodes for the presence of metastatic breast cancer is an important component of the American Joint Committee on Cancer's TNM breast cancer staging system, since the results of the evaluation affect treatment courses. Several centers have taken a biochemical approach, and implemented testing of sentinel lymph nodes with immunohistochemistry for proteins known as pancytokeratins. Limitations of this approach include a high cost, time-intensive slide preparation process, an increased number of slides required for pathological review, and low accuracy. While computer vision based systems have also been developed, they are so far not yet in clinical use. Thus, developing practical, efficient, and cost-effective automated systems for detection of metastatic nodes is still an open research area.

4 Model

4.1 Image Pre-processing

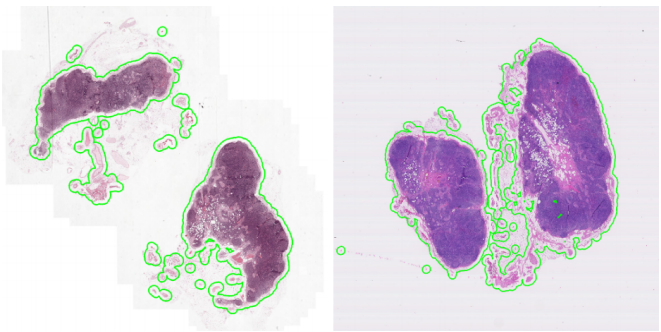


Figure 1: Tissue region (highlighted green) detection during image pre-processing.

To reduce computation time, Otsu's thresholding method is used to segment tissue data from the remaining

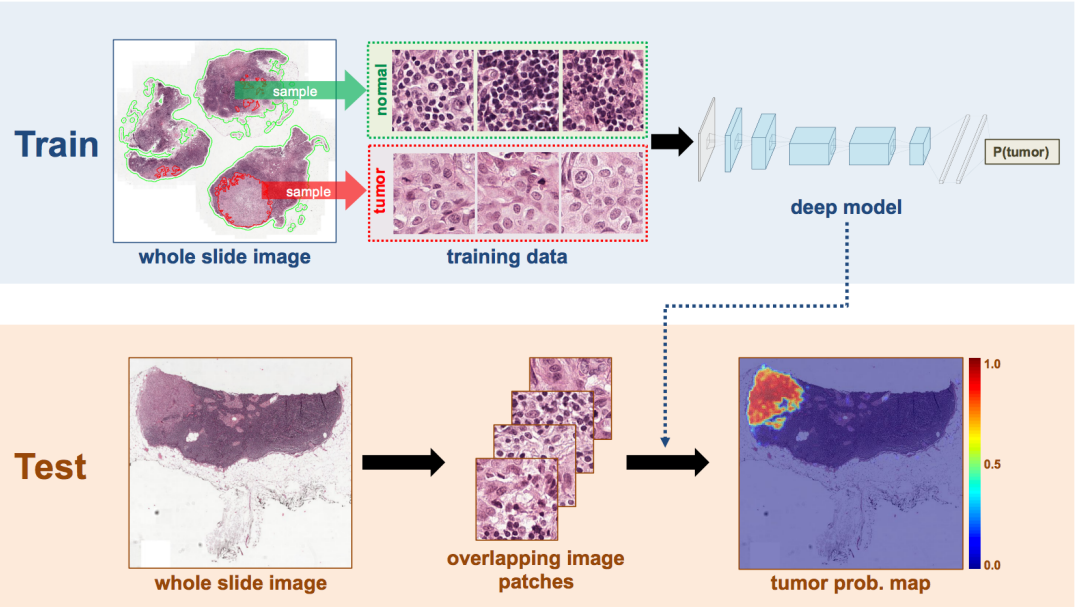


Figure 2: The framework of cancer metastases detection.

background region of the WSI, effectively removing 82% of the WSI. The detection results are visualized as green curves in Fig. 1.

4.2 Cancer Metastasis Detection Framework

The cancer metastasis detection framework consists of a **patch-based classification stage** and a **heatmap-based postprocessing stage**, as depicted in Fig. 2.

The **patch-based classification** stage takes as input WSIs and ground truth WSIs annotated with cancer regions. From these, millions of positive and negative 256x256 patches are cropped from the WSIs are used to train a GoogLeNet based discriminative model.

It's worth noting the authors evaluated three other deep learning architectures: AlexNet, VGG16, and FaceNet. A 27 layer, 6 million parameter GoogLeNet became part of the final model.

The authors found that a significant cause of false positives were histologic mimics of cancer, after which additional negative training examples featuring these mimics are added to make a second training set they call the *enriched training set*.

The **heatmap-based postprocessing stage** entails manipulating the heatmap data for the competition's two metrics:

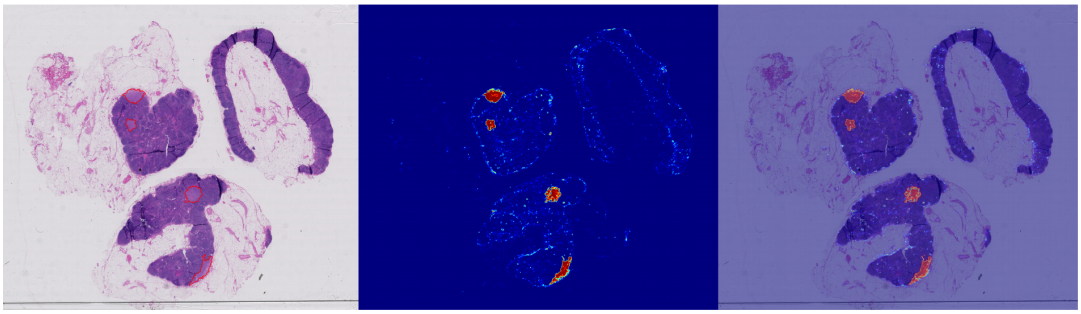
- Slide-based Evaluation: From a heatmap input, the authors extract 28 geometrical and morphological features (e.g. % of tumor region over the whole tissue region, the area ratio between tumor region and the minimum surrounding convex region, and the longest axis of the tumor region, etc.). Features extracted from the training heatmaps are used in building a random forest classifier to discriminate positive from negative WSIs.
- Lesion-based Evaluation: For this metric, the authors train two deep GoogLeNet classification models, one (D-I) using the unmodified training set and another (D-II) the *enriched training set* mentioned earlier. Noting that D-I was more prone to false positives (but more sensitive), the authors next threshold the heatmap output of D-I and identified connected components within the resultant binary mask. Defining a centroid of the connect component as (x,y), they return the average of the tumor probability predictions generated by D-I and D-II across each connected component.

5 Results

The models presented in this paper achieve state of the art performance on both slide-based and lesion-based evaluation metrics.

The AUROC of their slide-based classification model was 0.925, beating out the next best model by over a percentage point. This high score can be credited in large part to the model's high performance when the false-positive rate is required to be low; in these cases, the model does far better than its competitors. This is likely due to the authors' creation of the "enriched" dataset to improve model performance in these conditions.

The performance of the lesion-based classification model was even more striking. The authors use the free-response receiver operating characteristic (FROC) curve to evaluate the models, which plots sensitivity against average number of false positives per image. The AUC



(a) Tumor Slide (b) Heatmap (c) Heatmap overlaid on slide

Figure 3: Visualization of tumor region detection.

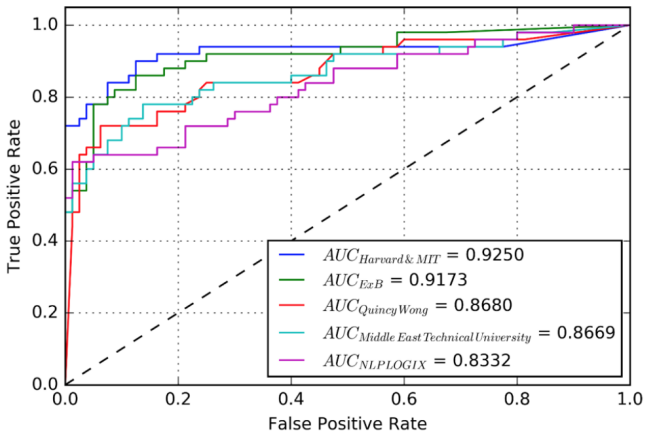


Figure 4: ROC Curve of Slide-based Classification

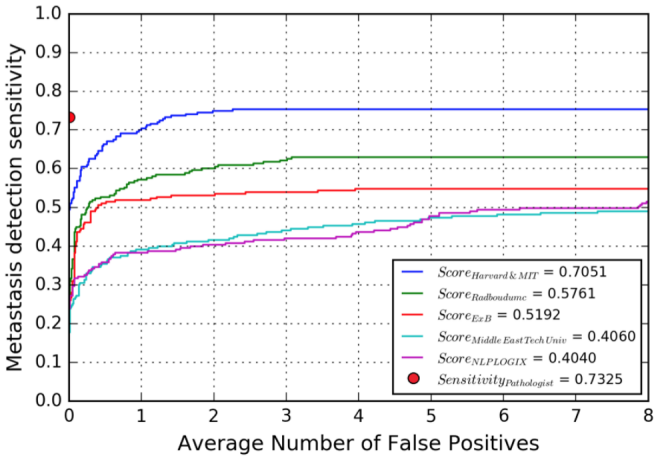


Figure 5: Average number of false positives

of the authors' model was 0.705, compared to the next-best model which achieved 0.576. A trained pathologist achieved a sensitivity of 0.733 with zero false positives (the red dot in Figure 5); when allowing two or more false positives, the authors' model achieves greater precision than the pathologist.

Finally, when the authors augmented the predictions of the pathologist with those of their model, the error rate dropped from 3.4 percent to 0.52 percent.

6 Discussion and Criticisms

This paper makes a significant contribution to the field of computational medical image analysis, and provides the best currently known automated system for detecting metastatic breast cancer in whole slide images of sentinel lymph node biopsies. Although their model does not outperform expert pathologists, combining it with pathologists reduced the pathologist error rate from over three percent to almost half a percent.

Their framework differs from end-to-end neural network systems for computer vision in that their framework incorporates considerable pre-processing, post-processing, and domain knowledge. In particular, the authors augment their training procedure to focus more heavily on negative training patches that are hard to classify correctly. This drives down false positives, which is particularly important in the domain of medicine, since negative test cases are much more common than positive ones. The post-processing stage incorporates various domain-specific metrics, thereby effectively combining the power of general deep feature learning with biomedical insights. The authors also test various hyperparameters, including magnification level. However, we would like to see a little bit more analysis of the false postives and false negative characteristics, discussing precision, recall, and F1 scores.

One shortcoming of this paper's approach is that the model used is built using the GoogLeNet deep neural network, which was trained on ImageNet. The image data in ImageNet, which comprises multiple classes of animals, outdoor scenes, and man-made objects, is significantly different structurally from tumor images. Although transfer learning has proven successful in certain settings, the qualitative difference in datasets makes it unlikely that GoogLeNet would be the best possible feature extractor on tumor slide images, and the authors could have experimented with training their own model.

References

[1] Wang, D., Khosla, A., Gargeya, R., Irshad, H., & Beck, A. H. (2016). *Deep Learning for Identifying Metastatic Breast Cancer*. arXiv preprint arXiv:1606.05718