Paper: https://www.nature.com/articles/s41467-020-18147-8

Paper summary

The authors propose the use of a deep convolutional neural network in order to detect gastric cancer. The use of artificial intelligence to help pathologists with diagnoses increases the chances of early detection, which drastically increases the chance of successful treatment.

Strengths and weaknesses
Strengths:

+ Targets one of the leading cancers. Gastric cancer is the 5th most common type of cancer and the 3rd most common cause of death by cancer worldwide.
+ The sensitivity of 0.996 extremely high. This is very good for preventing pathologists from overlooking malignancy.

Weaknesses:

- The specificity of 0.843 could be improved. Even though a false positive is of lesser concern than a false negative, it should still be improved to prevent unnecessary use of time and resources on a follow up.
- The model is trained to ignore low-quality images, but there is no discussion about how many such images were ignored.

Comments for authors

Soundness:

The authors state that other recent studies have shown the effectiveness of pathology AI for tumor detection in other systems such as lung, breast, and prostate regions. They identify multiple challenges that other studies have not met, including: sustaining a thorough test with a substantial number of slides over a continuous time period, with images produced by different machines; improving diagnostic accuracy without significantly delaying the process; and conducting a multicentre test before deployment to ensure results are consistent across multiple hospitals.

The authors state that their training samples included 958 surgical specimens (908 malignancies) and 542 biopsies (102 malignancies) with diverse tumor subtypes, with random rotations and reflections applied to the images. The proportion of surgical samples is imbalanced compared to the testing data (154 surgical specimens (118 malignancies) and 1660 biopsies (61 malignancies)), which could introduce a bias to the model. Specifically, including more samples of healthy tissue could help to improve the currently lacking specificity of the model.

The authors also applied distortions, including Gaussian and motion blurs and color jittering in brightness, saturation, contrast, and hue, in order to simulate distortions introduced by taking images from different hospital scanners. They also state that the model was trained to identify low-quality images and ignore them. It would be useful to state how many of the samples were considered low-quality to better show that the model is fit for use with real world patient data.

Significance:

The authors state that the model detected two samples of cancer cells that pathologists initially missed. While this number isn't extraordinary at first glance, it's quite significant with the sample size of less than 1000 malignancies. Thus, this model has the potential to improve the diagnostic abilities of professional pathologists.

Novelty:

The idea of using a convolutional neural network to detect tumors in tissue samples is not new. However, the challenges associated with real datasets, including low-quality samples, is discussed to some degree, so I believe the idea is novel enough.

Verifiability:

The authors provide an in-depth method, including preprocessing and analysis of the results. They also provide instructions to access the data (not public due to regulations) and a full code repository. Thus, the experiment is easily verifiable.

Presentation:

The introduction is easy to follow but some of the contents, such as the discussion of the framework used, would fit better in the method section.

The results are clearly analysed, with specific attention given to cases that pathologists missed. However, there should be more attention given to the false positives, to address some of the weaknesses of the model.

The method is very well detailed and easy to follow.

Questions for authors to respond to

1. Discuss in more detail the "ignored" class. What percentage of the data was classified in this class?
2. How were the thresholds for the distortions chosen? Is there a real-life analog to the Gaussian blurs or the color jittering, and how significant is this?
3. Discuss the false positive cases. Do these cases have anything in common? Do they show a weakness in the model?