Understanding image labels

In regard to the assignment on classifying images from the PatchCamelyon dataset, it is important to understand why images are labelled as tumour-containing or not. I.e. what features lead to an image being labelled as one or the other?

Although many variations of deep learning architectures can perform well in making the classification, we cannot exactly say how the model made the decision (hence they are black-boxes).

Understanding domain-specific features helps in developing more interpretable models, which is essential in healthcare: we need to understand why a model makes a particular prediction.

The following issues can arise if one does not understand why images have certain labels:

- Interpreting model decisions: Even if we have an interpretability method to investigate why a model assigned a certain label to an image, where it can e.g. highlight the features in the image by which it predicted a certain label, it is almost useless if we ourselves cannot pinpoint the features in the image that give rise to that label. What if the model is making decisions based on features that we cannot see or understand? How can we know whether our model is picking up an artefact that correlates with a certain class?
- Ineffective model improvement: Without knowing the features that contribute to a decision, it is difficult to know how to improve the model.
- Problem solving difficulties: If the model does not perform as expected, not understanding the features may make it challenging to diagnose and solve the problem effectively.

Additionally, understanding the labels and how they correspond with certain features can help to:

- Improve data efficiency: By having a better understanding of the features, it can guide how data augmentation or other pre-processing strategies are performed, that can help preserve or enhance critical features. Therefore, the available data is utilised better and should help the model to generalise better.
- Analyse model errors: Knowing which features are indicative of tumour vs or non-tumour regions helps in analysing model errors, which in turn can guide improvements in making the model more robust.
- Improve resource efficiency: Deep architectures often require substantial computational resources. Understanding domain-specific features sometimes leads to more efficient model designs that take into account the specific needs of the task, and may therefore reduce computational requirements while maintaining high classification metrics.

A simple test to try, to help address some of the points mentioned above: Choose several batches of images randomly, e.g 40-50 images in total, and see what proportion of these images you can correctly identify the label, and indicate and/or describe the features by which you have chosen that label.

Try to develop the CNN model after you have correctly identified at least ~90% of the labels corresponding to the images.