Antonin Vidon

Classification for Breast Histopathology

GitHub • Original data

Early-stage breast cancer detection - the most diagnosed in women – increases the 5-year relative survival rate up to 99%. This crucial process is improved by introducing a computer-aided diagnosis system to automate cancerous tissue identification from whole slide images.

This work proposes a replicable approach for the detection of Invasive Ductal Carcinoma (IDC) - accounting for 80% of all breast cancers - and aims at outperforming diagnostic accuracy of an experienced human eye (~79%).

Data***:*** 277,524 patches scanned at x40 in resolution 50 x 50. Metadata contains patient ID (279 patients in total), coordinates of the patch in tissue slice and target (cancerous or not).

Initial investigations were performed on the distribution over patients of cancer ratio and number of patches, emphasizing the need to handle class imbalance before training. Several whole breast tissue slices were reconstructed to look at the way the invasive cancer spread across patches and the tissue color was projected into HSV space for examination. This phase emphasized the difficulty to predict the target from a single look at the patch.

Afterwards, a balanced sample of patches was extracted and a PCA was performed to keep 80% of explained variance with a 97% dimensionality reduction. Scikit-learn classifiers were then tuned using grid and randomized search.

Gradient Boosting Classifier was selected as overall best predictor according to three criteria: accuracy, training time and predicting time. This method provides a 2.5% increase in IDC detection accuracy compared to human eye. Further improvement could be achieved by implementing and tuning Convolutional Neural Network.