Classification for Breast Histopathology

Antonin Vidon

I. Why this topic

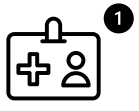
Fast and accurate detection can allow patients to have proper treatment and consequently **reduce rate of morbidity**

II. What to aim at

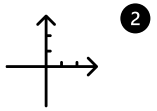
- **Automating** the detection of cancerous tissue
- Achieving **better classification accuracy** than human eye **(~79% for experienced biologists upon balanced data)**
- **Reducing the delay** before starting medical treatment

III. Exploratory Data Analysis

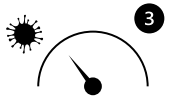
277,524 patches (50 x 50 px each scanned at x40) of **279 patients**. Metadata contains the following features :



PatientID

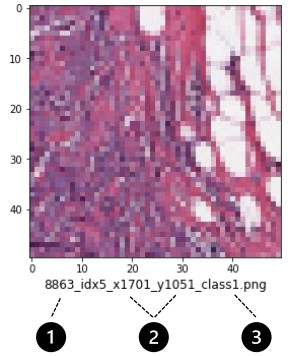


Coordinates of the patch in breast tissue slice

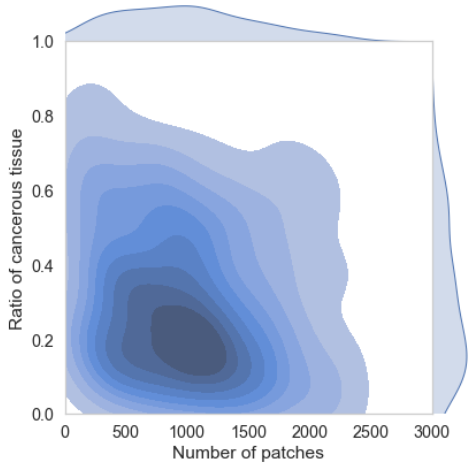


Binary target : cancerous or not

Cancerous patch for PatientID 8863

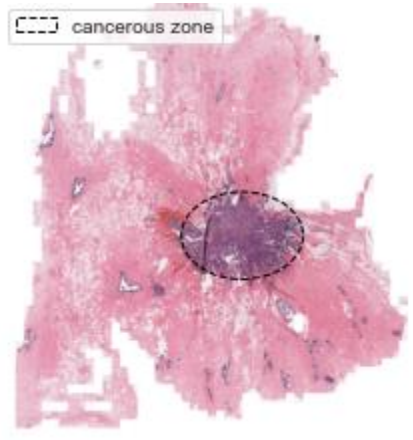


Kernel density over patients of #patches vs. cancer ratio



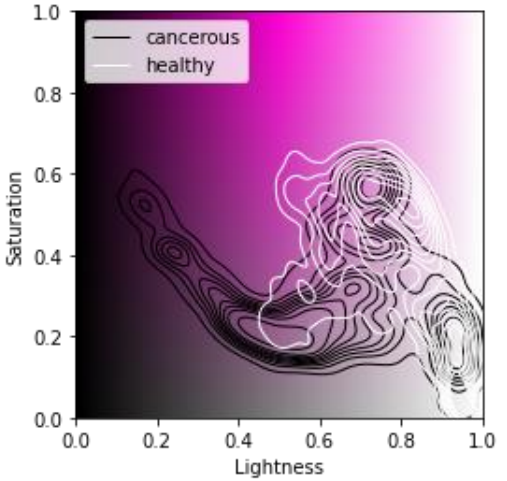
Imbalanced classes and **high variability** in #patches per patient

Reassembling breast tissue patches for PatientID 12890



Cancer **spreads to nearby tissue** such that **infected patches form a cluster**

Kernel density of tissue color for 50 random patches of each target



Cancerous tissue is **darker** on average, but **not always**

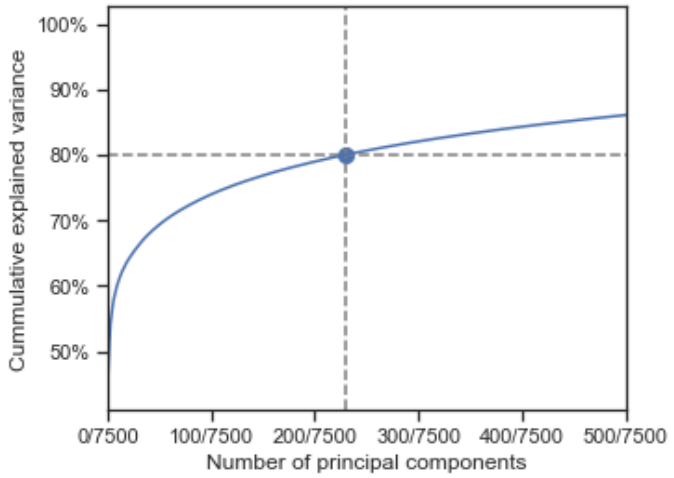
Classification for Breast Histopathology

Antonin Vidon

IV. Preprocessing task

- **Removed outliers** of size smaller than 50x50 pixels *to eliminate noise from data*
- **Extracted a balanced sample** of 2.10^4 patches (both targets in equal prop.) *to improve prediction of the minority class*
- **Performed a PCA** at 80% threshold *to avoid the curse of dimensionality*

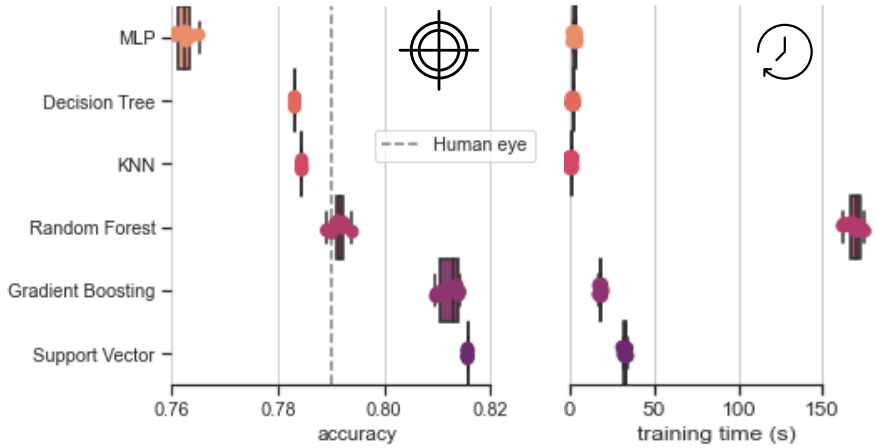
Cumulative explained variance over principal comp.



V. Tuning & Evaluation

- Classifier hyperparameters tuned using **Grid** and **Randomized Search**

Cross-validation score & training time boxplot for top 6 classifiers



- Model evaluation based on **Cross-validation score** and **computation time**

Best model

Gradient Boosting Classifier

81.4 % median C.-V. accuracy

VI. What was done well

- Assessing the issues of **class imbalance** and **curse of dimensionality** during preprocessing
- **Tuning hyperparameters to outperform diagnostic accuracy** of experienced biologists (81.4 vs. 79%)

VII. Margin for improvement

- Use **coordinates in tissue slice as features** if heatmap is promising
- Increase sample size and **oversample minority class** (e.g., rotate images)
- **Implement CNN** which have given state-of-the-art results in this domain (>90% accuracy)