visiopharm
TURNING IMAGES INTO KNOWLEDGE

DTU Technical University of Denmark

Automatic Feature Selection for Digital Pathology

Dzenita Omanovic - Master Thesis Student

Introduction

Visiopharm operates in the field of quantitative digital pathology. They use image analysis to interpret and quantify microscopy images of biological samples that have been digitised. The healthcare applications of digital pathology include diagnosis and prognosis of diseases like cancer, osteoarthritis,diabetes and many more. As a preprocessing step to supervised classification of desired structures in the images, the user of their software VisiomorphDP™ needs to select a number of relevant feature images. However, the user is often a researcher within medicine and biology and lacks the needed technical expertise for choosing relevant features. In this project we look at methods for automatically selecting the most relevant features for binary classification tasks. As an extra addition we inspect whether the classifiers support vector machine and/or random forests would be a valuable addition to VisiomorphDP™.

Data: Microscopy images

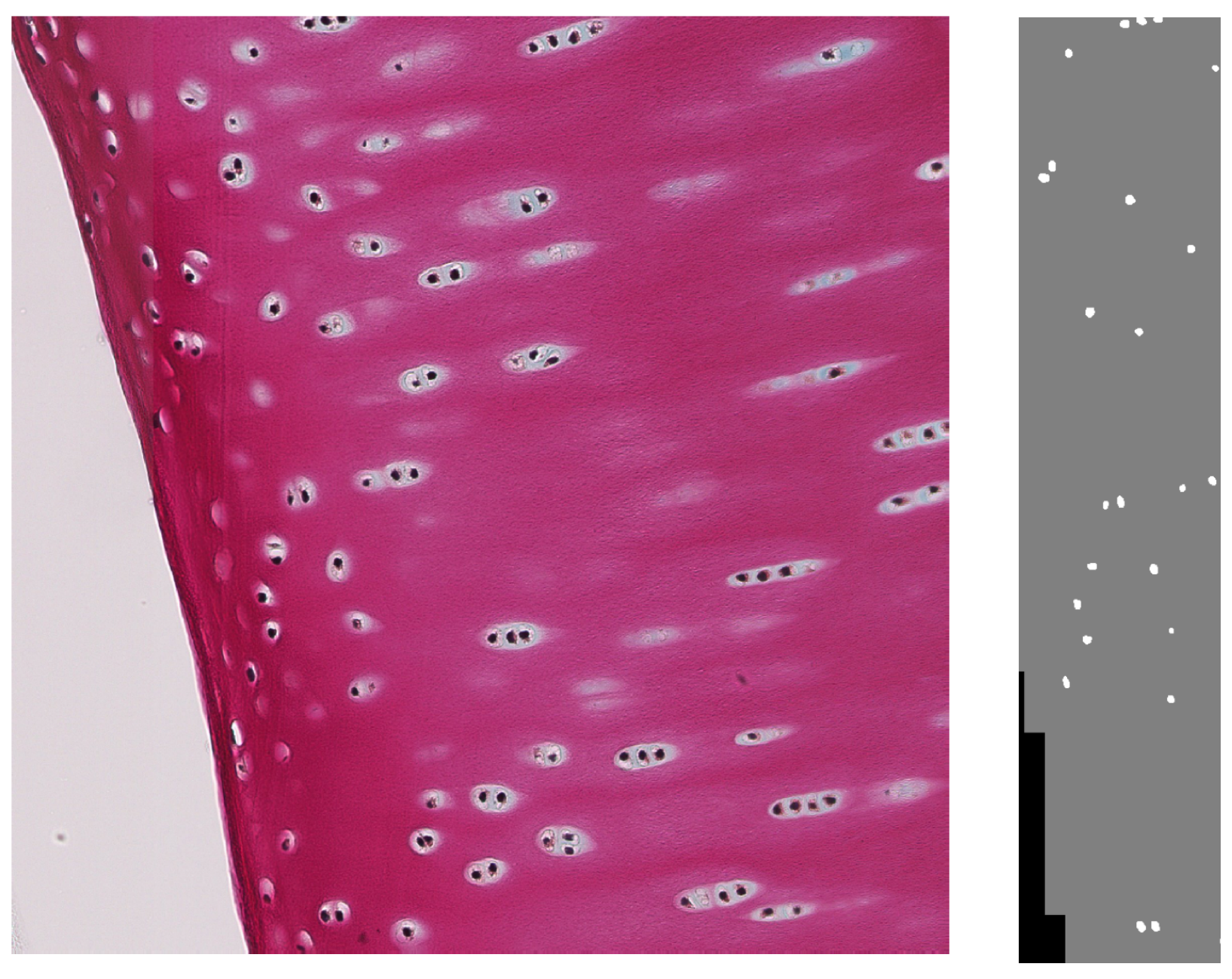


Figure 1: Image of data set 1 and segment of label image

Tissue type:
Cartilage explant stimulated to mimic diseased cartilage
Staining:
Ehrlich Triacid
Gives best differentiation between cartilage (pink) & nuclei (black)
Aim:
Identifying nuclei since a decrease in nuclei is a hallmark of Osteoarthritis.

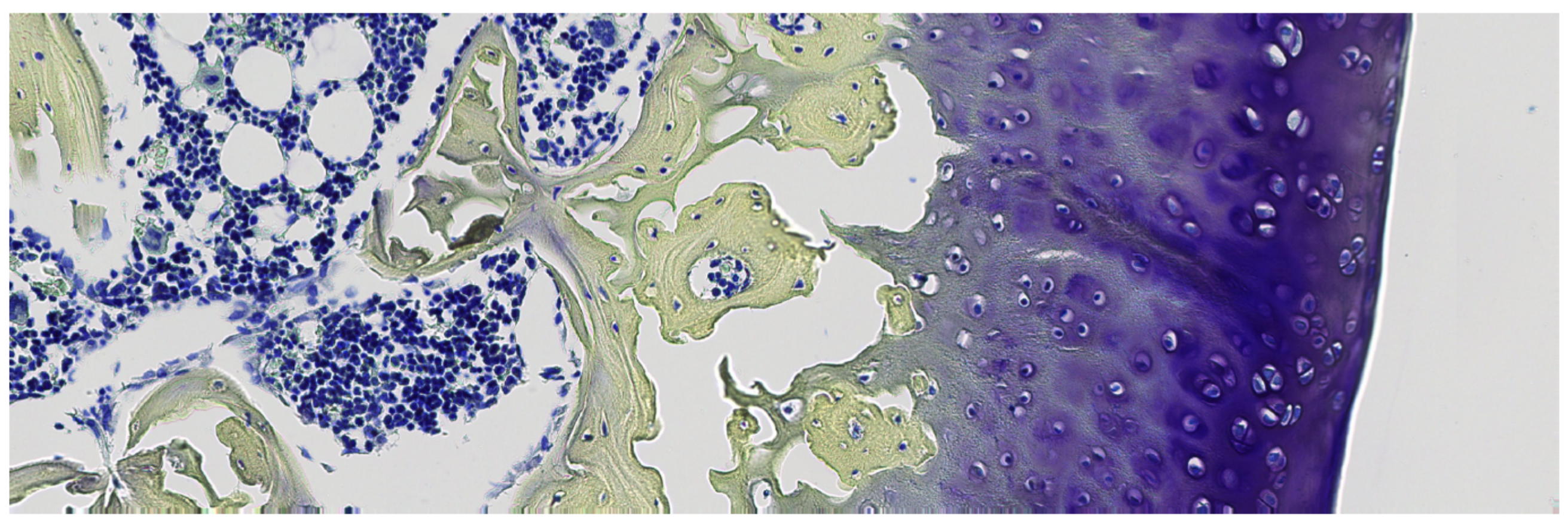


Figure 2: Image of data set 2

Figure 3: Segment of label image for data set 2

Tissue type:
Knee joint with surgically induced instability.
Staining:
Toluidine Blue - Safron du Gatinais
Best in demonstrating joint destruction
Aim:
Identifying cartilage (purple) and subchondral bone (yellow) for assessment of cartilage destruction and bone sclerosis due to Osteoarthritis.

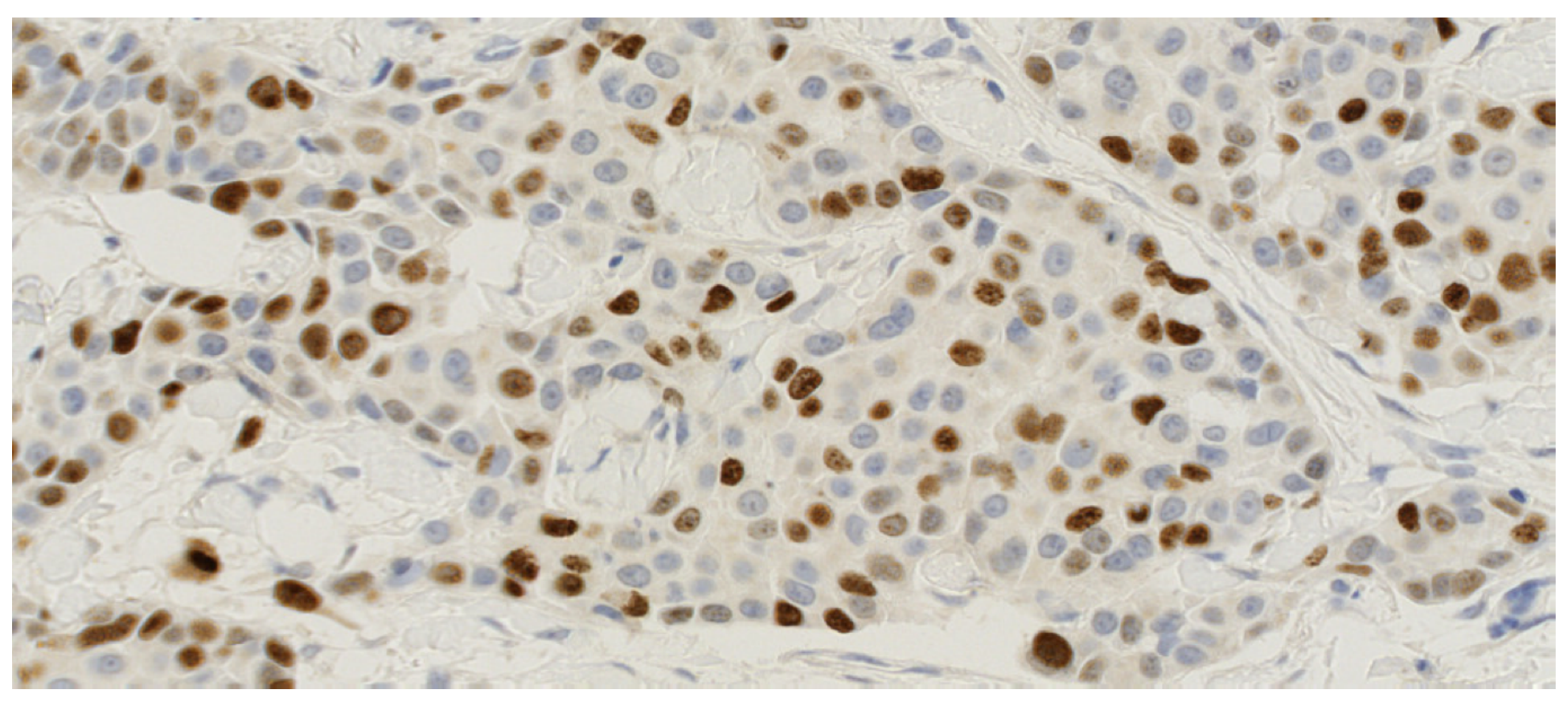


Figure 3: Image of data set 3

Figure 4: Segment of label image for data set 3

Tissue type:
Cancerous breast tissue
Staining:
Ki-67 Immunohistochemical (IHC) staining.
Cell-proliferation can be assessed w. Ki-67 IHC.
Aim:
Identifying Ki-67 positive (brown) and Ki-67 (blue) negative nuclei that can be correlated to the tumor grade and clinical course of the disease.

Methods

We create a set of 88 feature images for each of the images seen above. We do feature selection via the methods presented below, and finally we train a classifier on a training set using only the selected features and validate on a validation set. We use 5 fold cross-validation.

Three paths are pursued:

1) Feature subset selection: select a subset of features among all input features.
We use **Stepwise feature selection** and **Lasso**

2) Dimension reduction: construct new features using linear combinations of all original input features.
We use **Principal Component Analysis (PCA)**

3) Hybrid method: combine feature selection and dimension reduction.
We use **Sparse Linear Discriminant Analysis (SLDA)**

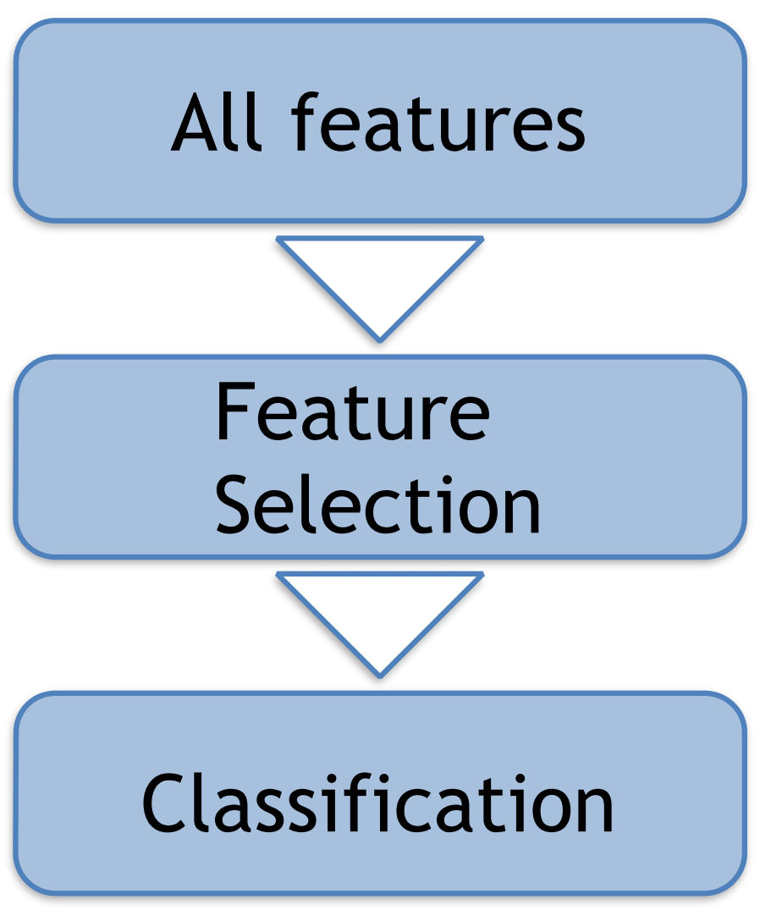
Four classifiers are tested:

1) Linear Bayesian classifier (LDA)
included in VisiomorphDP™

2) Quadratic Bayesian classifier (QDA)
included in VisiomorphDP™

3) Support Vector Machine (SVM)
not included in VisiomorphDP™ - new addition

4) Random Forests (RF)
not included in VisiomorphDP™ - new addition



All features

Feature Selection

Classification

Preliminary Results

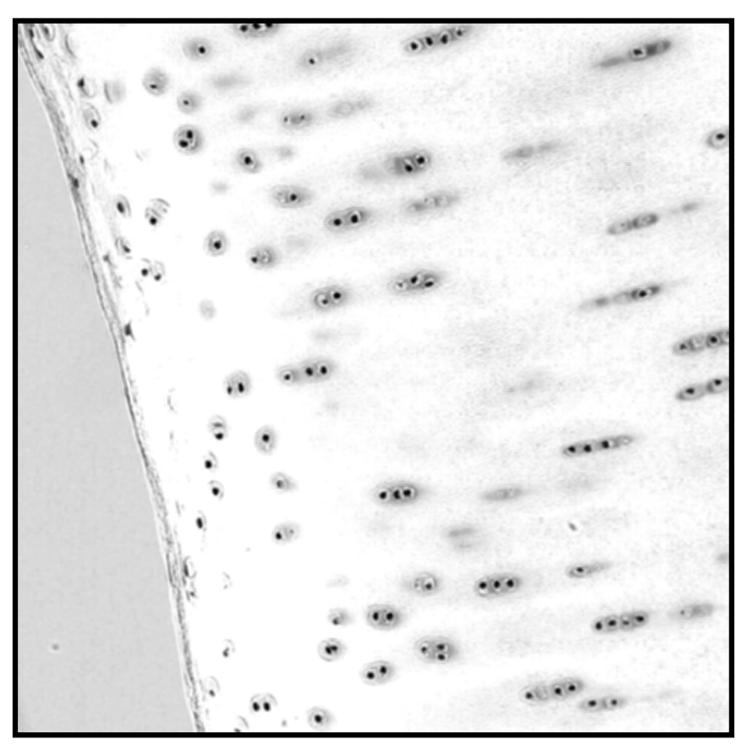
For each data set we show the most "cooked down" feature set for all the selection-methods and the lowest obtained misclassification error among the 4 classifiers.

Data set 1	No.of features /components	Misclassification rate (%)	Classifier
Stepwise	3	0.06	SVM
Lasso	4	0	SVM/RF
PCA	14	0.04±0.02	SVM
SLDA	10	1.5±1.1	SVM

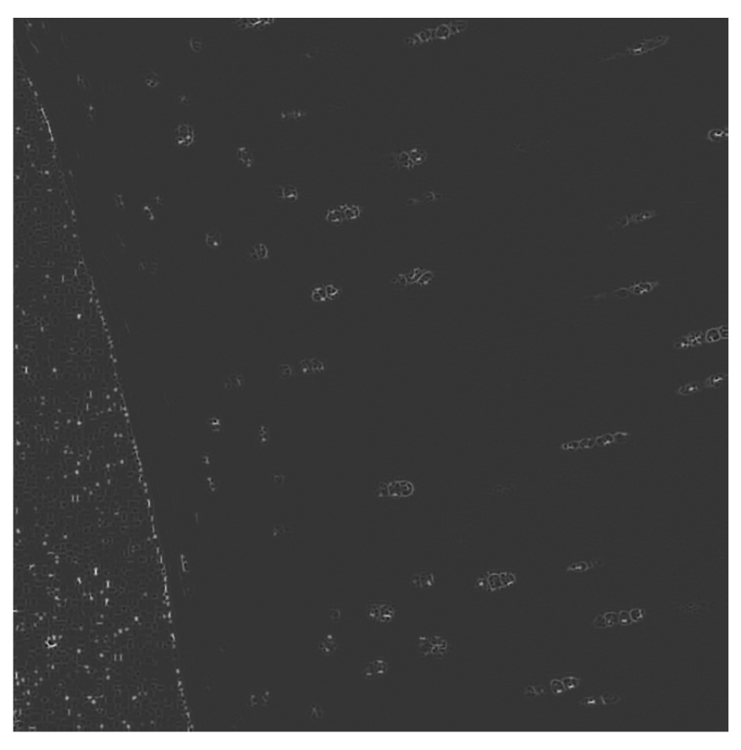
Data set 2	No.of features /components	Misclassification rate (%)	Classifier
Stepwise	22	1.1	SVM/RF
Lasso	31	0.89	SVM
PCA	19	0.95±0.29	SVM
SLDA	6	1.9±0.31	SVM

Data set 3	No.of features /components	Misclassification rate (%)	Classifier
Stepwise	27	4.59	RF
Lasso	23	6.1	SVM
PCA	15	6.3±1.0	SVM
SLDA	12	8.2±0.64	LDA

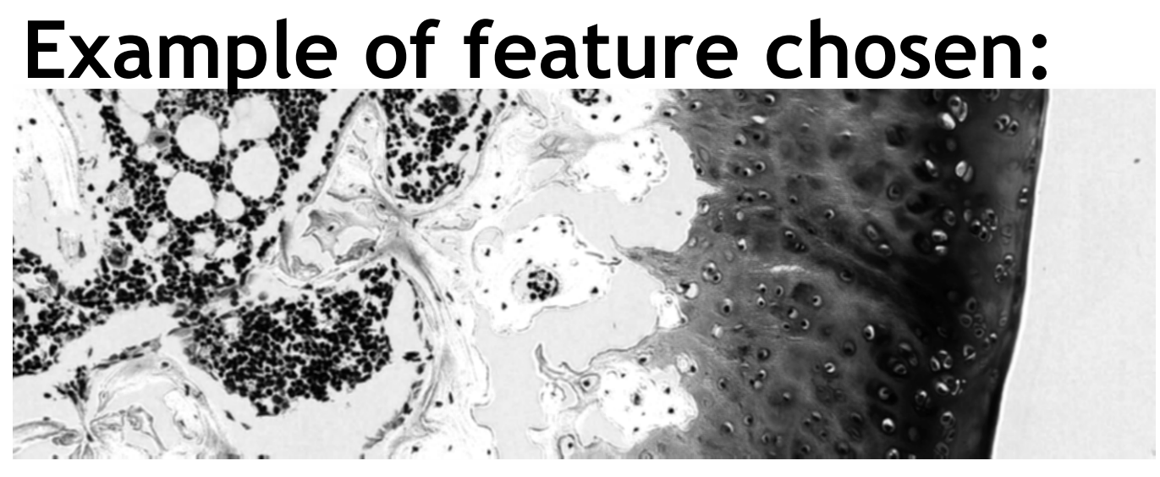
Example of feature chosen:



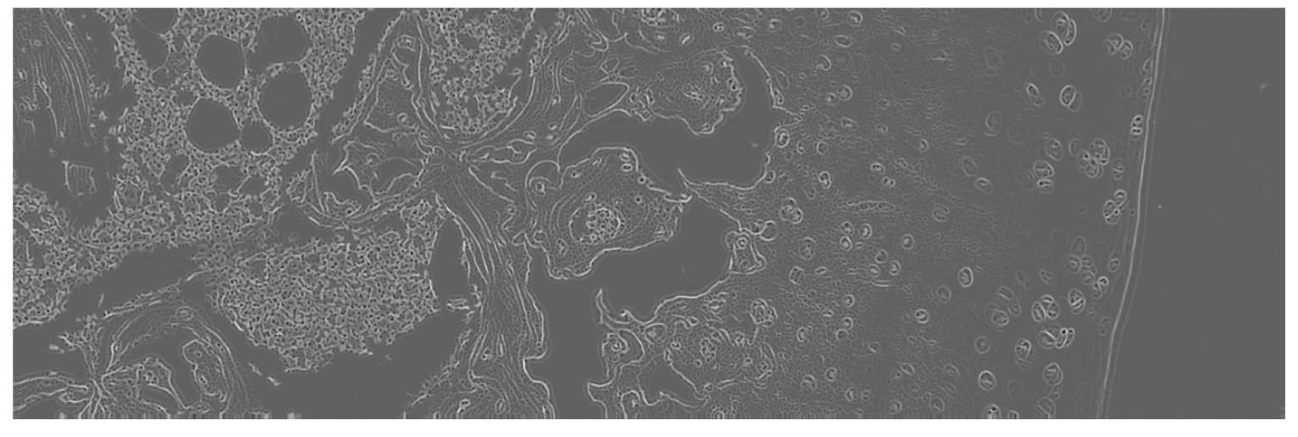
Example of feature NOT chosen:



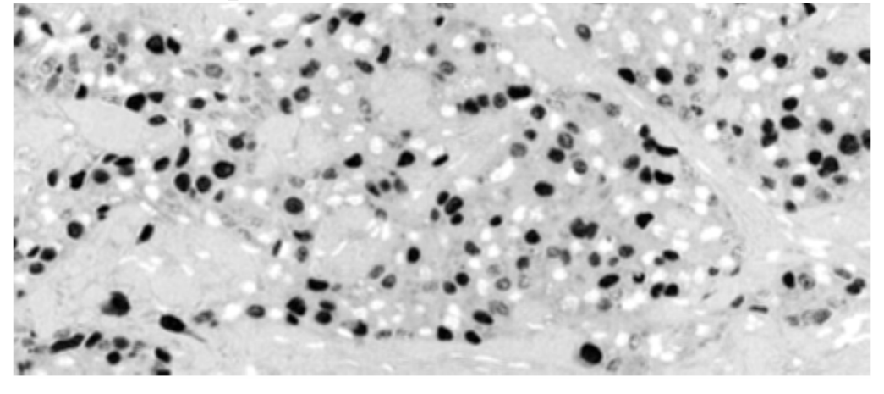
Example of feature chosen:



Example of feature NOT chosen:



Example of feature chosen:



Example of feature NOT chosen:

Conclusions

- For all the methods the general observation is that the set of 88 features is at least cut down to about 30.
- The classifier giving the lowest misclassification error is the SVM followed by RF.
- Feature selection via SLDA is observed to give the highest misclassification error.
- Lasso and SLDA have the advantage that the user can define the number of features desired as output.
- All feature selection methods have been implemented to work independently from the classifiers. This is viewed as a big advantage by Visiopharm.