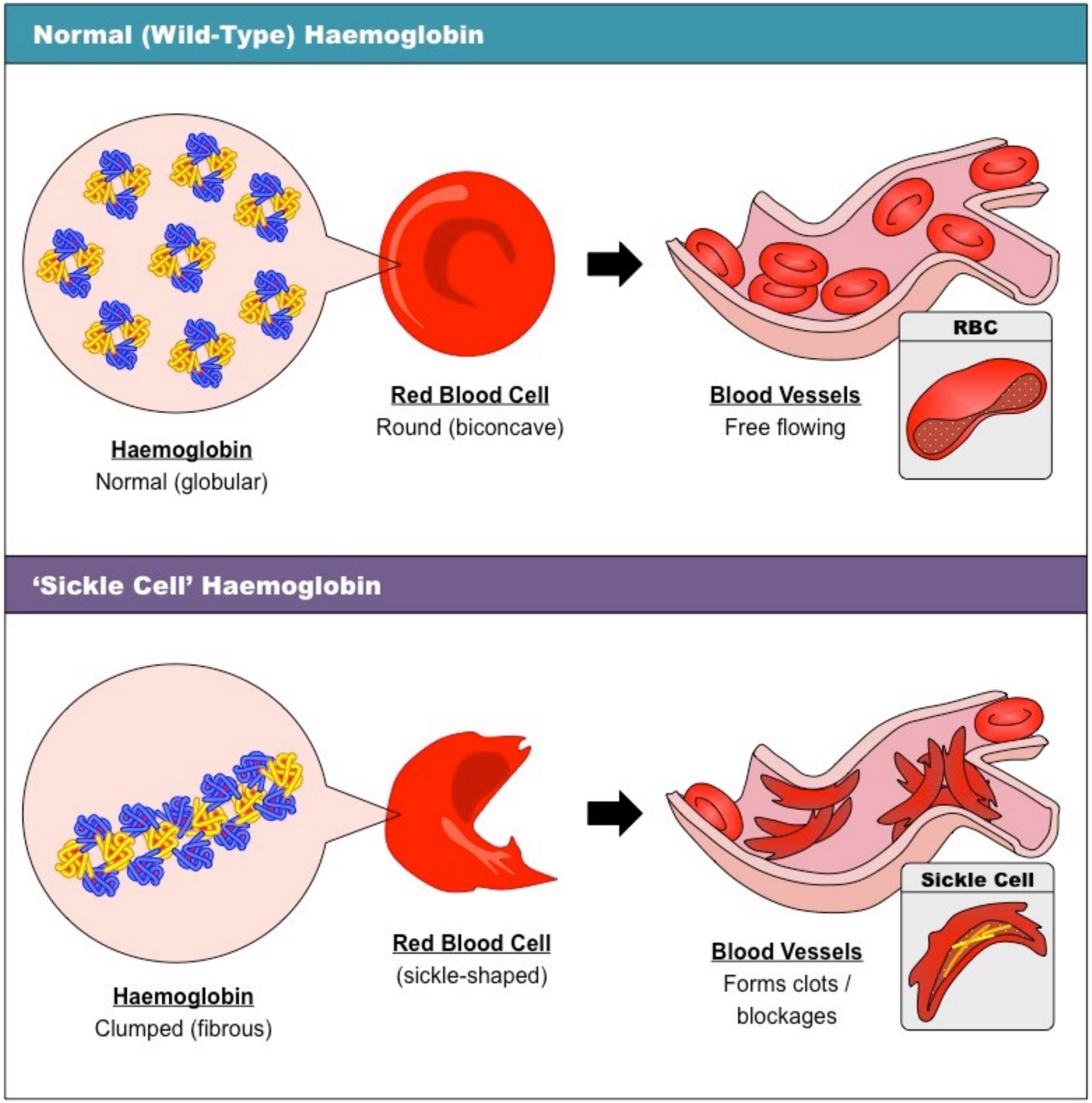
An approach to select the most appropriate machine learning method for cell morphology analysis: case study for red blood cell classification of Sickle Cell Anemia

N. Petrović, G. Moyá Alcover, A. Jaume-i-Capó,
M. González-Hidalgo

EUROWEB+

Sickle cell anemia

- Mutation of hemoglobin protein
- Affects shape of the cell
- Irregular cell lives shorter
- Depleted supply of red blood cells in the organism



https://ib.bioninja.com.au/standard-level/topic-3-genetics/31-genes/mutations.html

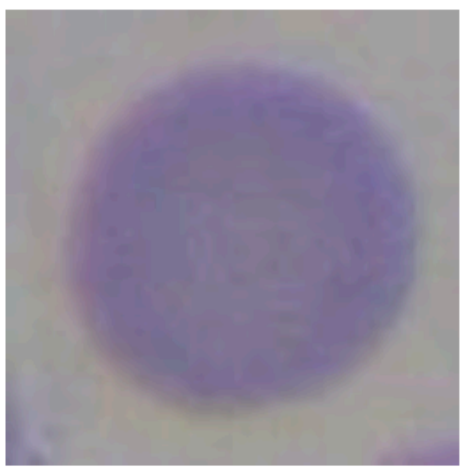
Origin of sickle cell anemia

- Mutation evolved as a beneficial adaptation
- Malaria parasites use red blood cells as incubators
- Sickle cells blocks the spreading of malaria through the blood stream
- Problem: inheritance of copies of the mutated gene from both parents

Sickle cell anemia

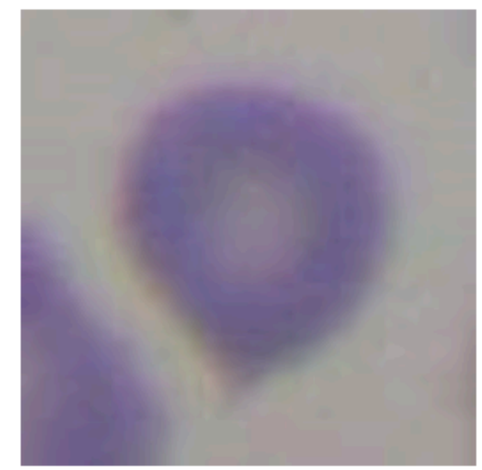
- Spread among the people with ancestors from sub-Saharan Africa, India, Saudi Arabia and Mediterranean countries
- Caused 553 000 deaths in 2016 around the world
- There is no cure, but there is treatment
- Testing includes observation of patient’s blood sample
- Excellent candidate for automation

Cell types



Normal

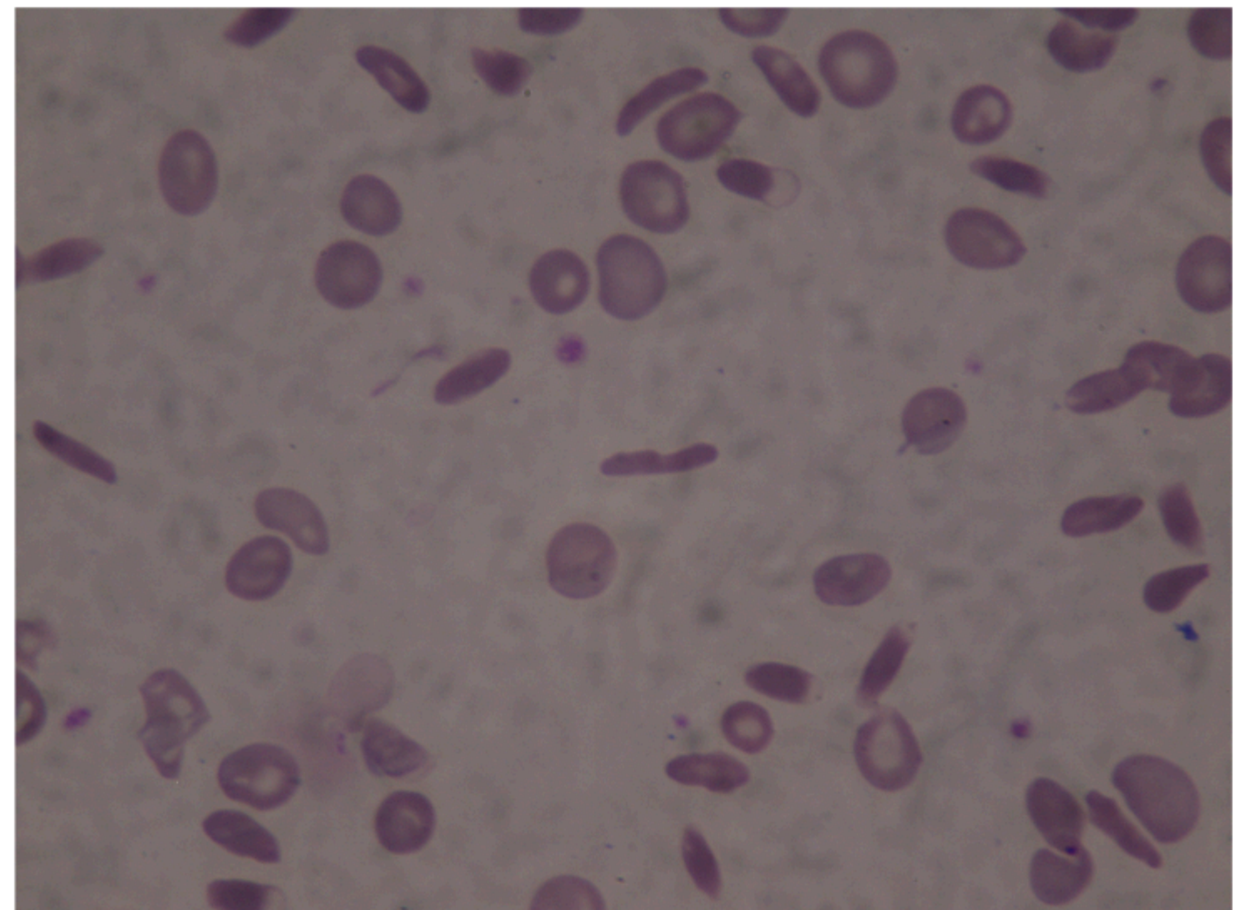
Sickle



Other

Dataset

- Microscopic images of a blood smear available at http://erythrocytesidb.uib.es/
- Total of 2550 individual cells labeled by medical experts from “Dr. Juan Bruno Zayas” General Hospital in Santiago de Cuba
- Imbalanced dataset - 1575 normal, 657 sickle, 318 other cells



Method

Image preprocessing

↓

Cell extraction

↓

Feature extraction

↓

Run classifiers with all features

↓

Feature selection/reduction

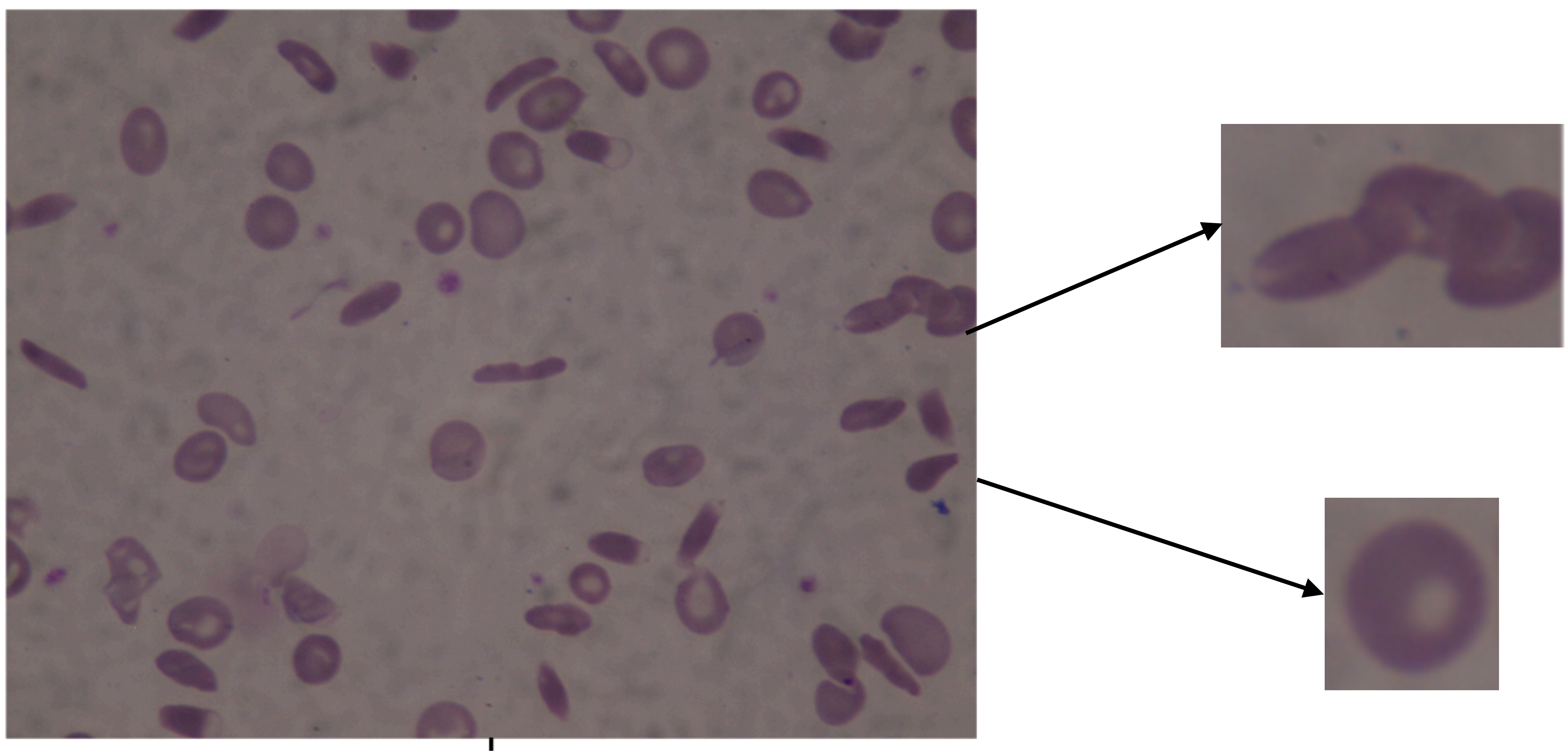
↓

Comparison of results

Development Tools

- Python 3.5
- OpenCV
- Scikit-learn

Image preprocessing



Feature extraction

- Shape features - based on cell contour, 33 features
- Texture features - based on GLCM, 60 features
- Color features - mean and standard deviation of color channels of different color spaces, 18 features
- In total 111 features

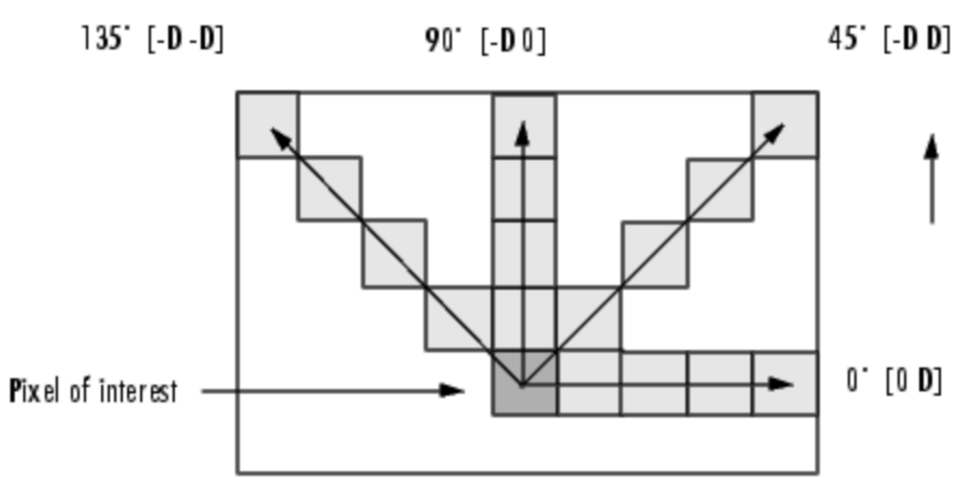
Shape features

- Based on cell contour: Perimeter, Area, Max feret, Min feret, Elongation, Solidity, Shape, Circularity, Modification ratio, Hu moments…
- Based on fitted ellipse contour: Major axis, Minor axis, Aspect ratio, Compactness, Eccentricity..
- Total 33 features

Texture features

- Based on GLCM
- 4 angles and 3 distances
- Total of 60 features

Texture feature	Description	Equation
Contrast	measure of the intensity contrast between pixels	$\sum_{i,j=0}^{levels-1} P_{i,j}(i-j)^2$
Dissimilarity	belongs to contrast group of features and its weights are linear	$\sum_{i,j=0}^{levels-1} P_{i,j}\lvert i-j\rvert$
Homogeneity	measures closeness of distribution of the elements in the GLCM to the GLCM diagonal	$\sum_{i,j=0}^{levels-1} \frac{P_{i,j}}{1+(i-j)^2}$
Energy	measures the textural uniformity and detects disorders in textures	$\sqrt{\sum_{i,j=0}^{levels-1} P_{i,j}^2}$
Correlation	gray level linear dependency of neighbor pixels	$\sum_{i,j=0}^{levels-1} P_{i,j} \frac{(i-\mu_i)(j-\mu_j)}{\sqrt{\sigma_i^2\sigma_j^2}}$



http://matlab.izmiran.ru/help/toolbox/images/enhanc15.html

Color features

- Three color spaces: RGB, HSV and CIE L*a*b*
- Mean and standard deviation values from the channels
- 18 features

Classifiers

- SVM - finds optimal decision boundary that maximizes the distance from nearest data points of all classes
- Decision Tree - uses simple decision rules to predict output values
- Random Forest - uses fully grown DTs with low bias and high variance
- Extra Trees - similar as RF, difference in testing random splits over fraction of features
- Gradient Booster - uses DT stumps with high bias and low variance
- kNN - computes output values from majority of the nearest neighbors of each point
- MLP - type of neural network, consists of at least three layers, each node is a neuron that uses nonlinear activation function

Preprocessing data

- Different ranges of data can affect performance of SVM and MLP
- Solution: Standardization of the data: $z = \frac{x - \mu}{\sigma}$

Metrics

- F1-measure: $\text{F1} = 2 \times \frac{Precision*Recall}{Precision+Recall}$

- SDS-score (Sickle cell disease Diagnosis Support score):

$$SDS\text{-}score = \frac{\sum_{i=1}^{3} n_{ii} + n_{23} + n_{32}}{\sum_{i=1}^{3} \sum_{j=1}^{3} n_{ij}}$$

Experiments

- 10-fold cross-validation
- Fine tuning of classifiers
- 1. Baseline experiment with fine tuning
- 2. Feature selection/reduction
- 3. Comparison with other algorithms

First experiment

- Running all classifiers with default parameters on all the features
- Searching for best parameter using RandomizedSearch with cross-validation
- Selecting 2 best performing classifiers

First experiment results

	SVM	DT	RF	ET	GB	kNN	MLP
Baseline F1	87.40%	85.07%	88.95%	87.75%	89.38%	82.54%	88.1%
Fine tuning F1	88.72%	88.49%	**90.26%**	90.05%	**90.14%**	83.77%	89.84%
Baseline SDS	89.53%	88.12%	91.31%	90.71%	91.41%	84.78%	90.75%
Fine tuning SDS	90.51%	91.45%	**92.59%**	92.16%	**92.35%**	85.98%	91.76%

GB baseline	c	e	o
c	1476	33	66
e	40	592	25
o	80	25	213

GB max F1	c	e	o
c	1497	30	48
e	33	610	14
o	84	29	205

GB max SDS	c	e	o
c	1497	30	48
e	33	610	14
o	84	29	205

RF baseline	c	e	o
c	1485	35	55
e	30	620	7
o	101	42	175

RF max F1	c	e	o
c	1502	28	45
e	28	615	14
o	90	35	193

RF max SDS	c	e	o
c	1498	29	48
e	28	610	19
o	84	35	199

Feature selection/reduction

- RF most important features - 22 features: HU2, Eccentricity, R factor, Modification ratio, HU1, HU4, Circularity, Aspect ratio, Shape, Roundness, HU3, Blue mean…

- GB most important features - 20 features: HU2, Eccentricity, HU3, HU1, Modification ratio, Aspect ratio, Shape, Roundness, Circularity, HU4, Blue mean, Max Feret…

- 16 features overlapping

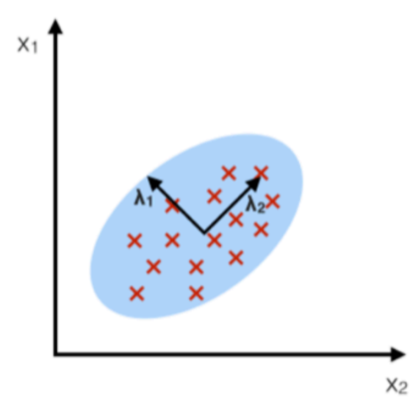
- No texture features!

Feature reduction

- Principal Component Analysis (PCA) - 13 components, 95% of variance explained

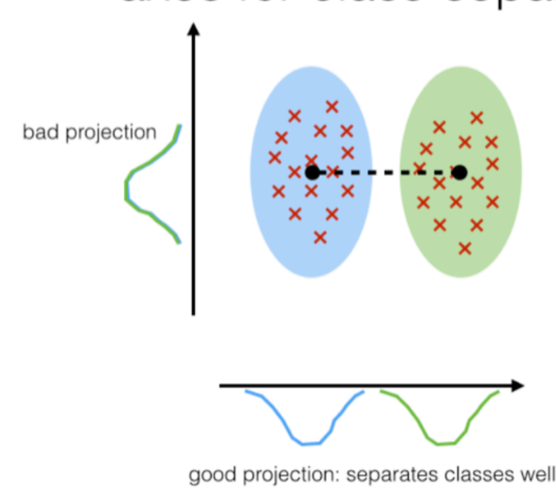
- Linear Discriminant Analysis (LDA) - 2 components

PCA:
component axes that
maximize the variance



https://sebastianraschka.com/Articles/2014_python_lda.html

LDA:
maximizing the component
axes for class-separation



Second experiment results

	First experiment	Feature selection	PCA	LDA
GB F-measure	90.14%	91.50%	86.58%	92.45%
GB SDS	92.35%	93.82%	88.40%	94.04%
RF F-measure	90.51%	91.15%	85.62%	**94.16%**
RF SDS	92.51%	93.65%	87.53%	**94.16%**

Second experiment results

	SVM	DT	RF	ET	GB	kNN	MLP
Baseline SDS	90.40%	91.45%	92.59%	92.16%	92.35%	85.96%	91.76%
LDA SDS	94.12%	**94.48%**	94.16%	94.20%	94.04%	94,00%	94.47%
Union SDS	90.98%	91.21%	92.94%	92.94%	92.90%	91.33%	91.29%
Intersection SDS	92.47%	91.37%	93.06%	93.73%	92.90%	92.43%	92.47%
Baseline F-measure	88.72%	87.59%	90.26%	89.11%	90.14%	83.77%	89.62%
LDA F-measure	92.71%	92.42%	92.65%	92.68%	92.45%	92.50%	**93.04%**
Union F-measure	89.19%	85.66%	90.52%	90.55%	90.98%	88.63%	89.71%
Intersection F-measure	89.27%	88.24%	90.13%	91.08%	90.16%	88.79%	89.04%

Future work

- Run ensemble of models
- Focus only on shape features
- Evaluate best models on bigger dataset

Thank you!