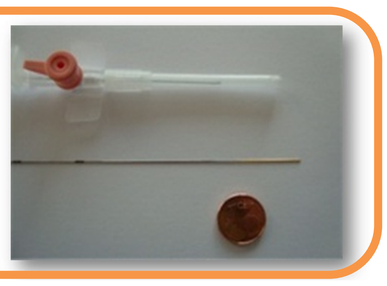
Automated enumeration of circulating tumor cells: machine learning techniques and interobserver variability

Carl-Magnus Svensson[1] and Marc Thilo Figge[1,2]
1 Applied Systems Biology, Leibniz Institute for Natural Product Research and Infection Biology – Hans Knöll Institute, Jena, Germany
2 Friedrich Schiller University, Jena, Germany

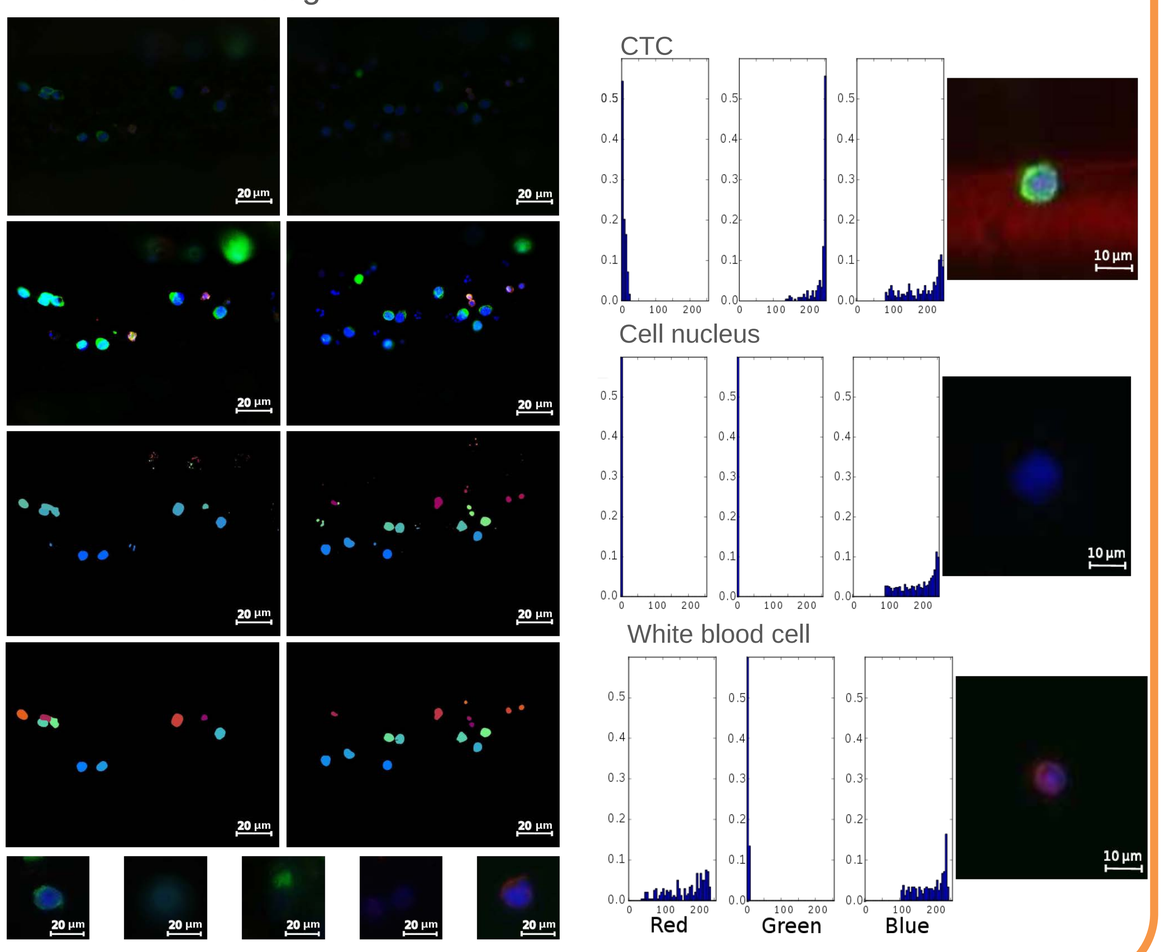
1. Circulating Tumor Cells

- Detached from primary tumors
- Metastatic seeds [1]
- Diagnostic tool
- *Ex vivo* using CellCollector from Gilupi GmbH
- Cells are fluorescently stained



2. Image analysis

- Identify regions of interest (ROIs)
- Thresholding, morphological filtering and watershed transform
- Extract RGB histogram



3. CTC classification

- Generative model giving naïve Bayesian classifier (NBC)
- NBC trained unsupervised and semi-supervised
- Support vector machine with radial basis function-kernel (SVM-RBF)
- Random forest (RF)
- 75% of data for training, 25% for testing
- One person annotating

TP: correctly identified CTCs FP: falsly identified CTCs

TN: correctly identified as not CTCs FN: missed CTCs

Accuracy $$\mathrm{Acc} = \frac{TP+TN}{TP+TN+FP+FN}$$

Precision $$\mathrm{Pre} = \frac{TP}{TP+FP}$$

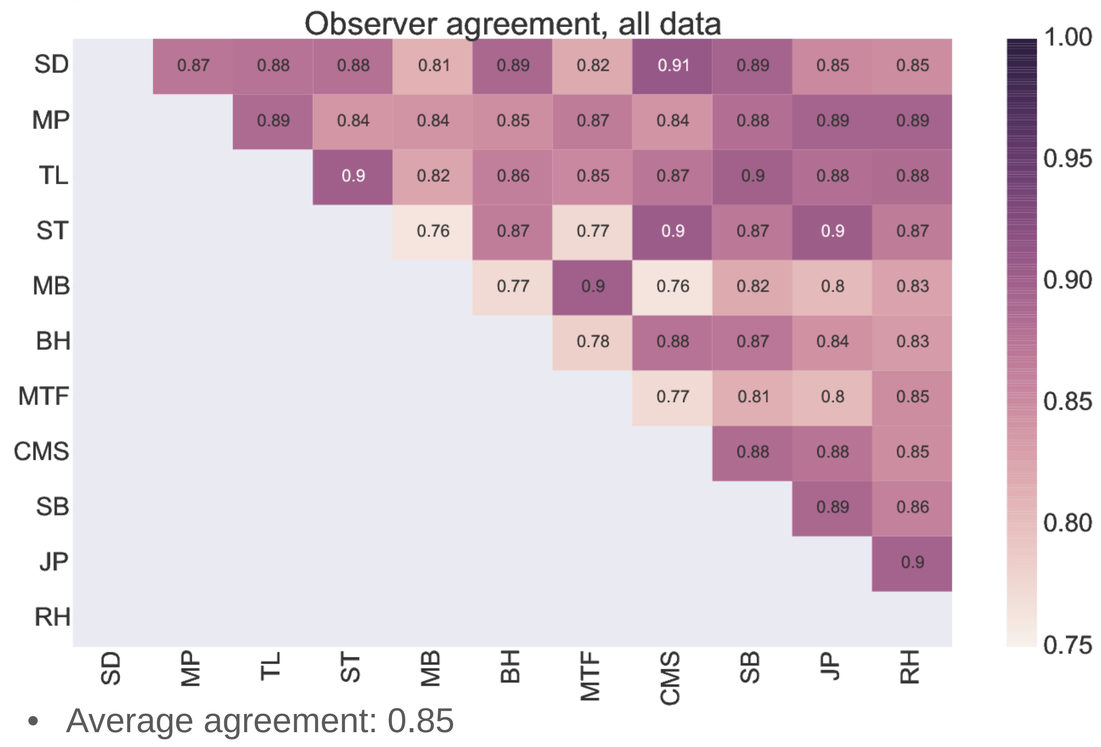
Recall $$\mathrm{Rec} = \frac{TP}{TP+FN}$$

Classifier	Acc	Pre	Rec
ROI identification	0.99	0.51	0.96
NBC unsupervised	0.87±0.02	0.85±0.04	0.92±0.03
NBC semisupervised	0.88±0.02	0.85±0.03	0.93±0.02
SVM-RBF	0.89±0.03	0.87±0.05	0.93±0.08
RF	0.88±0.04	0.84±0.06	0.90±0.08

- 96% of CTCs are found at ROI identification
- Accuracy < 0.9 independent of classifier [2]
- Efficient unsupervised NBC

4. Interobserver variability

- N=617 ROIs, N_{obs}=11 human observers
- Agreement: ratio of ROIs two observers annotate the same way



- Average agreement: 0.85
- Classifiers not worse than the average human observers

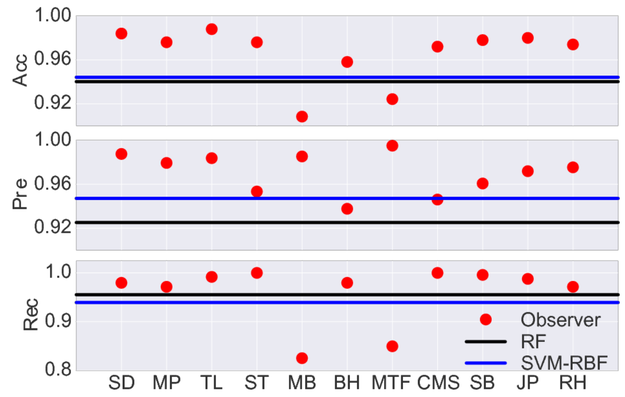
5. Label noise effects

- All observers agree on 365 ROIs: Ground Truth (GT) dataset
- Remaining 252: probGT dataset
- 5 folds, 3 GT, 2 probGT
- Always test on GT

Training	Two GT folds	One GT fold and one probGT fold	Two probGT folds
RF	Acc: 0.98±0.00 Pre: 0.98±0.00 Rec: 0.98±0.00	Acc: 0.96±0.01 Pre: 0.94±0.02 Rec: 0.98±0.01	Acc: 0.94±0.02 Pre: 0.89±0.05 Rec: 0.97±0.03
SVM-RBF	Acc: 0.96±0.00 Pre: 0.95±0.00 Rec: 0.96±0.00	Acc: 0.92±0.01 Pre: 0.88±0.02 Rec: 0.95±0.01	Acc: 0.81±0.04 Pre: 0.75±0.08 Rec: 0.84±0.07

6. Consensus

- Consensus: 95% probability that a ROI has/has not a CTC
- Consensus limit c: $\sum_{i=1}^{c}\binom{N_{obs}}{i}(0.5)^{N_{obs}}<0.05$
- Observers reach consensus on 502 ROIs (81%)



	RF	SVM-RBF
Acc	0.94	0.94
Pre	0.96	0.95
Rec	0.93	0.94

7. Conclusions

- Computer vision methods are extremely useful for CTC identification
- Classification is predominantly limited by data quality
- SVM-RBF is sensitive to label noise during training
- RF is robust against label noise during training
- Evaluate classifiers against consensus data if possible

References
[1] Hayes *et al.*, (2006) Circulating tumor cells at each follow-up time point during therapy of metastatic breast cancer patients predict progression-free and overall survival. *Clin. Cancer Res.*, 12(14):4218–4224.
[2] Svensson *et al.*, (2014) Automated detection of circulating tumor cells with naive Bayesian classifiers. *Cytometry A*, 85(6):501-511.
[3] Svensson *et al.*, (2015) Automated classification of circulating tumor cells and the impact of interobsever variability on classifier training and performance. *J. of Imm. Res.*, ID 573165.

Email: carl-magnus.svensson@leibniz-hki.de