

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

C.-M. Svensson, M. T. Figge

01/09/2016

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

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

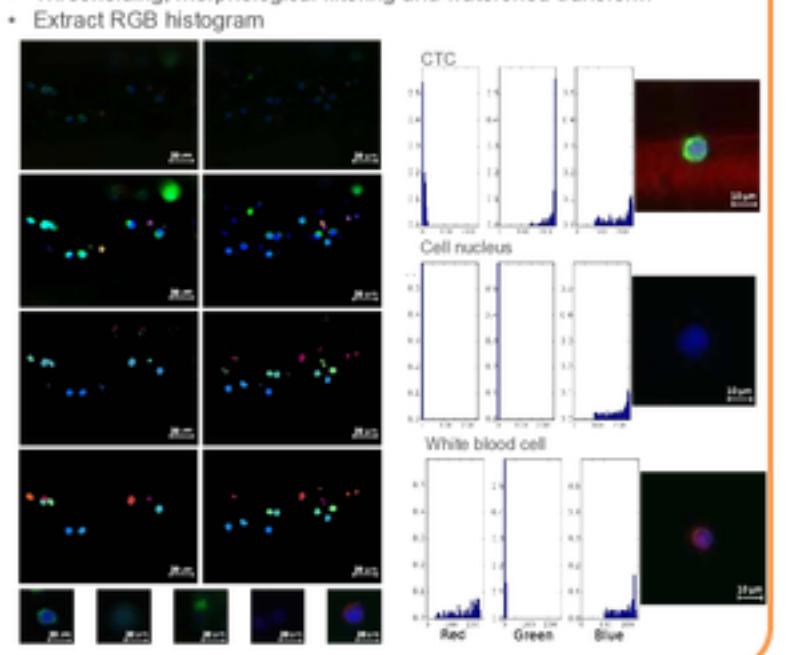
**1. Circulating Tumor Cells**

- Detached from primary tumors
- Metastatic seeds [1]
- Diagnostic tool
- Easy to collect with 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 naive 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      FN: missed CTCs  
 TN: correctly identified as not CTCs      FP: falsely identified CTCs

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

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

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

- 99% of CTCs are found at ROI identification
- Accuracy > 0.9 independent of classifier [2]
- Efficient unsupervised NBC

**4. Interobserver variability**

- N=617 ROIs, N<sub>obs</sub>=11 human observers
- Agreement: ratio of ROIs two observers annotate the same way

Observer agreement, all data



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

**5. Label noise effects**

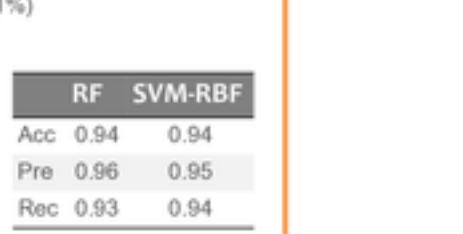
- All observers agree on 386 ROIs: Ground Truth (GT) dataset
- Remaining 252 probGT dataset
- 5 folds, 3 GT, 2 probGT
- Always test on GT

Training: Two GT folds    One GT fold and Two probGT folds    one probGT fold

	Two GT folds	One GT fold and Two probGT folds	one probGT fold
Acc:	0.98±0.00	0.96±0.01	0.94±0.02
RF:	0.98±0.00	0.94±0.02	0.89±0.05
Rec:	0.98±0.00	0.98±0.01	0.97±0.03
5:	0.96±0.00	0.92±0.01	0.81±0.04
SVM-RBF:	0.95±0.00	0.88±0.02	0.75±0.08
Rec:	0.96±0.00	0.95±0.01	0.84±0.07

**6. Consensus**

- Consensus: 95% probability that a ROI has/has not a CTC
- Consensus limit c:  $\sum_{i=1}^{11} (\frac{1}{N_{obs}})^i (0.5)^{N_{obs}-i} < 0.05$
- 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 interobserver variability on classifier training and performance. *J. of Imm. Res.*, 10, 573165.  
Email: carl-magnus.svensson@leibniz-hki.de

**HKI**  
Leibniz-Gemeinschaft