Detecting rare cell populations in flow cytometry data using UMAP

Lisa Weijler[1], Markus Diem[1], Michael Reiter[1],
Margarita Maurer-Granofszky[2],
Angela Schumich[2], and Michael Dworzak[2]

25th International Conference on Pattern Recognition
Milan, January 10th – 15th, 2021

1 Computer Vision Lab, TU Wien, Austria
2 Children Cancer Research Institute, Vienna, Austria

Introduction

Acute Lymphoblastic Leukemia (ALL) & Minimal Residual Disease (MRD)

- development of large numbers of immature lymphocytes
- evaluation of response to treatment with MRD assessment
- assessment of MRD via bone-marrow puncture and flow cytometer

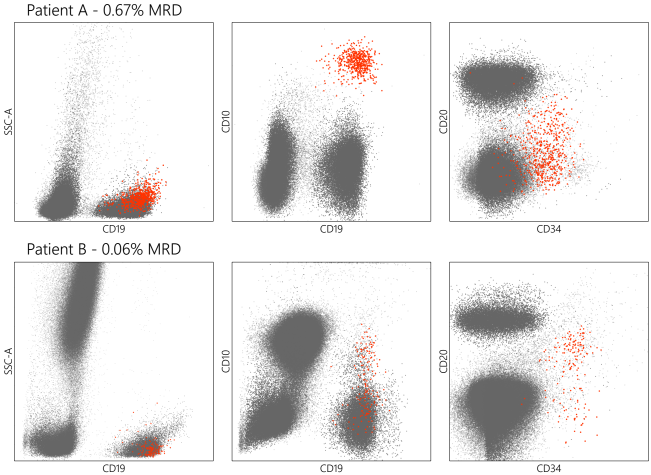
Flow cytometry (FCM) to transform cell characteristics into feature vectors

A bone marrow biopsy. Picture taken from
http://www.news.navy.mil/view_single.asp?id=4416
Description from that page: 021204-N-0696M-180 Georgetown University Hospital, Washington, D.C. (Dec. 4, 2002)

bone marrow sample → flow cytometry → find leukemic cells (MRD)

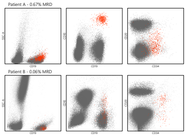
Challenges involved

What's the difficulty of automation of MRD assessment?



Challenges involved

What's the difficulty of automation of MRD assessment?



- variation of cell populations

- cell population to be detected as low as 0.01%

- limited availability of training data

holistic approach:
flexible decision regions

vs.

event-based approach:
fixed decision regions

Methods

- Random forest[1] (RF) for classification

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \rightarrow$$

Random forest
classifier

→

leukemic or
healthy cell

Methods

- Random forest[1] (RF) for classification
- Transforming FCM features into 2D embedding before classification with RF

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \rightarrow \textbf{Dimensionality reduction} \rightarrow \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{bmatrix} \rightarrow \text{Random forest classifier} \rightarrow \text{leukemic or healthy cell}$$

Uniform Manifold Approximation & Projection[2] (UMAP)

t-distributed Stochastic Neighbor Embedding[3] (t-SNE)

Principal Component Analysis[4] (PCA)

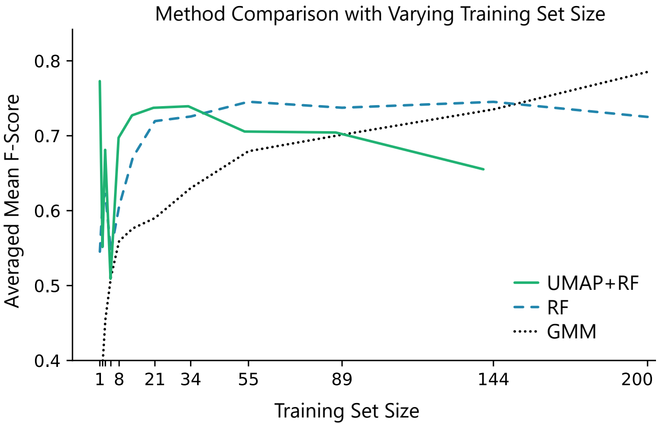
Locally Linear Embeddings[5] (LLE)

Basic idea:

- Reduce redundance and noise
- Mapping to latent space as a calibration step

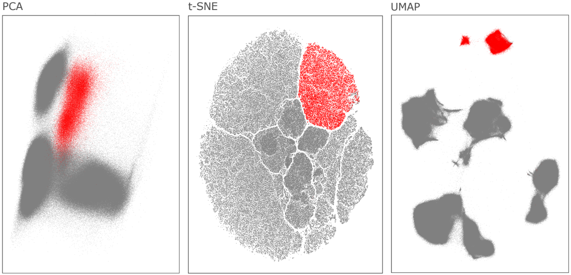
Results 1/2

Training set: 200 samples from Vienna - **Test set:** 72 samples from Berlin



GMM refers to approach [**6**]

Alternative feature transform & embedding dimensionality



Method	N	p	r	avg F_1	med F_1
PCA+RF	8	0.362	0.656	0.377	0.324
t-SNE+RF	8	0.696	0.747	0.615	0.693
LLE+RF	8	0.725	0.729	0.621	0.730
UMAP+RF (6D)	8	0.823	0.687	0.688	0.785
UMAP+RF	8	0.790	0.758	**0.697**	0.833

Key findings

- sufficient training samples → holistic approach[6] superior
- event-based method RF has a glass ceiling (F_1-score $= 0.73$)
- dimensionality reduction with UMAP prior to classification with RF allows for a **training set size reduction of more than 90%** ($N = 144$ to $N = 8$) without performance loss in the problem setting discussed
- choice of dimensionality reduction method and embedding dimensionality matters → **UMAP proves superior** in terms of run time performance and F_1-score

the performance of the standard classifier can be improved significantly by combining it with a preceding unsupervised learning step involving UMAP

References

[1] L. Breiman Random Forests. Machine learning, 45(1): 5-32, 2001.

[2] L. McInnes et al. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. `https://arxiv.org/abs/1802.03426`

[3] L. van der Maaten Accelerating t-sne using tree-based algorithms. Journal of Machine Learning Research, 15(1): 3221–3245, 2014.

[4] K. Pearson F.R.S. LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2(11): 559–72, 1901.

[5] S. T. Roweis et al. Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science, 290(5500): 2323–2326, 2000.

[6] M. Reiter et al. Automated Flow Cytometric MRD Assessment in Childhood Acute B- Lymphoblastic Leukemia Using Supervised Machine Learning. Cytometry, 95: 966-975, 2019.

Table 1: Comparison of random forest (RF), preprocessing with UMAP and Gaussian Mixture Manifolds trained with all available training data VIE (N) and tested against BLN (M). Precision (p), recall (r), average F-Score (avg F_1), and median F-Score (med F_1) are calculated per single cell.

Method	N	M	p	r	avg F_1	med F_1
GMM [**6**]	165	72	0.854	0.808	**0.785**	0.912
RF	200	72	0.872	0.750	0.725	0.886
UMAP+RF	200	72	0.596	0.974	0.657	0.801

Table 2: Comparison of RF, UMAP+RF and GMMs if 8 training samples are randomly chosen from VIE (N) and tested against BLN (M).

Method	N	M	p	r	avg F_1	med F_1
GMM [**6**]	8	72	0.756	0.613	0.559	0.637
RF	8	72	0.758	0.726	0.606	0.832
UMAP+RF	8	72	0.790	0.758	**0.697**	0.833