DETECTING RARE CELL POPULATIONS IN FLOW CYTOMETRY DATA USING UMAP

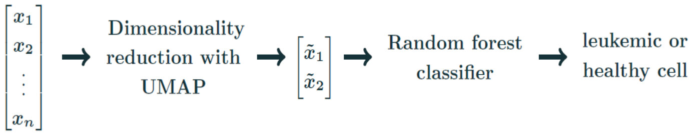
Lisa Weijler[1], Markus Diem[1], Michael Reiter[1], Margarita Maurer-Granofszky[2], Angela Schumich[2], and Michael Dworzak[2]

[1]TU WIEN, [2]CCRI

BACKGROUND

We present an approach based on unsupervised Uniform Manifold Approximation[1] (UMAP) and supervised classification to detecting small cell populations in flow cytometry (FCM) samples with a focus on minimal residual disease (MRD) quantification to monitor Acute Lymphoblastic Leukemia (ALL) treatment response. A common issue within automated FCM data analysis is the lack of (labeled) training data; an approach operating on least possible number of training samples is of considerable value.

METHODS


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

Using **UMAP** to embed FCM features into a latent 2D space prior to classification by **Random Forest**[2] (RF)

Comparison with a state-of-the-art method[3] based on Gaussian mixture manifolds (GMM) on varying training set sizes

Comparison with alternative embedding dimensionality and feature transform methods,

t-SNE[4], PCA[5], and LLE[6]

RESULTS

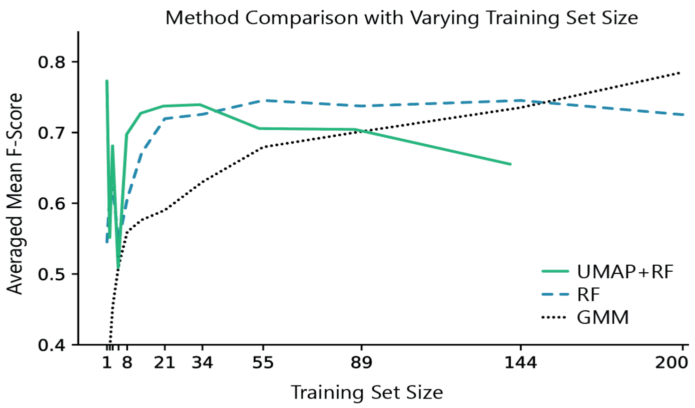


Figure 1: Comparison of Gaussian Mixture Manifolds (GMM), Random Forest (RF) and feature transformation with UMAP before RF classification (UMAP+RF) in terms of average F₁-score (vertical axis) for different training set size N (horizontal axis).

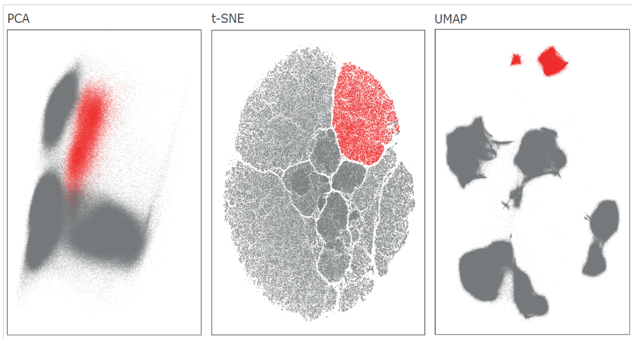


Figure 2: 8 training samples embedded to a 2D space with PCA (left), t-SNE (middle) and UMAP (right). Leukemic cells are shown in red.

Table 1: 8 samples randomly chosen are used for training. Precision (p), recall (r), average F₁-score (avg F₁), and median F₁-score (med F₁) are calculated per single cell.

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

CONCLUSION

- proposed method allows for a training set size reduction of more than 90% (N=144 to N= 8) in the problem setting discussed
- performance of the standard classifier can be improved significantly by combining it with a preceding unsupervised learning step involving UMAP
- UMAP proves superior to other dimension reduction methods in terms of run time performance and F₁-score

REFERENCES

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

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

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

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

[5] 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.

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

MyeFLOW

vienna business agency