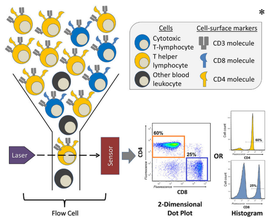
Pulmonary Department and Critical Care, Northwestern University

Machine learning classifier for automated and scalable analysis of clinical flow cytometry samples

Stanislav Bratchikov[1], Cathy A. Gao[1], Suchitra Swaminathan[1], Yuliana Sokolenko[1], Estefani Diaz[1], Emmy Jonasson[1], Lucy Luo[1], Zhan Yu[1], Ankit Agrawal[2], Richard G. Wudnerink[1,3], Alexander V. Misharin[1,3]

1 Division of Pulmonary and Critical Care, Department of Medicine, Northwestern University, 2 Department of Electrical and Computer Engineering, Northwestern University, 3 Simpson Querrey Lung Institute for Translational Science

Introduction



Multiparameter flow cytometry is crucial for translational research, providing in-depth immunophenotyping of clinical samples. Traditional manual analysis by experts faces biases and scalability challenges. We hypothesize that machine learning classifier can be used to achieve expert-level accuracy, handle technical variations, and analyze hundreds of clinical samples efficiently.

Figure 1. Flow cytometry concept

Methods

We have generated a diverse set of 209 expert-annotated clinical flow cytometry samples generated from bronchoalveolar lavage fluid from patients with lung diseases, including severe pneumonia, respiratory PASC, samples from lung transplant patients, and samples from healthy volunteers. We split this dataset with a 7:3 train-validation ratio to train and optimize a model based on a gradient boosting LightGBM classifier. We have validated our model using previously annotated external dataset of hundreds of clinical samples from patients with lung diseases.

Flow cytometry experiment statistics

STUDY	Total number of experiments	Number of exp. used for training
CLAD	761	50
SCRIPT	504	107
PASC	42	36
Duke ozone	16	12
Total	**1323**	**205**

Figure 2. Including experiments from multiple study allows model to learn biological variability

Expert-Level Knowledge Transfer: Feature Importance Analysis

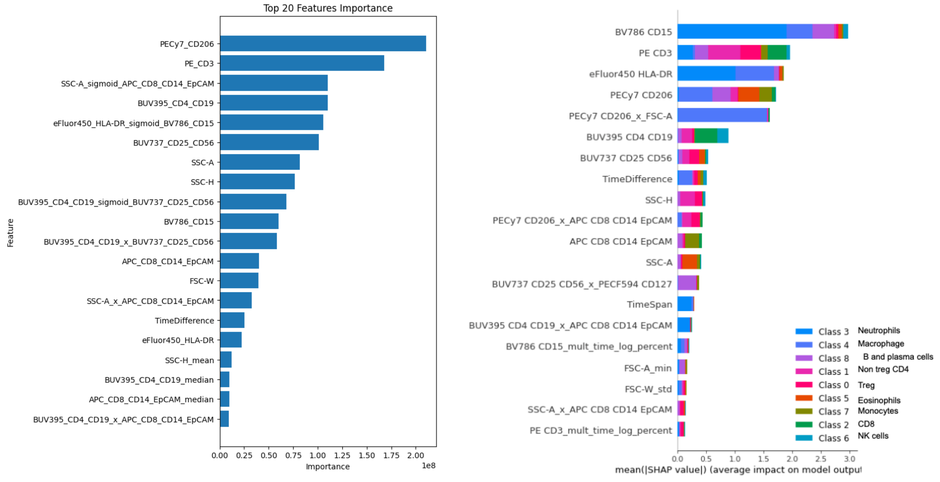


Figure 4. Model gained understanding of biological variation between different populations, which is supported by features it focused on during training

Heterogenous structure of flow cytometry data

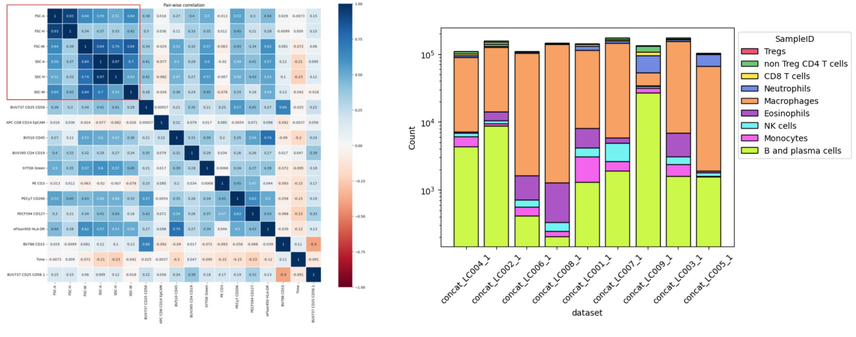


Figure 3. Model uses the fact that biological properties are specific to respective cell populations. Flow cytometry data is represented by populations of varying sizes, which needs to be accounted for during training.

Evaluating Model Precision with Normalized Matrices

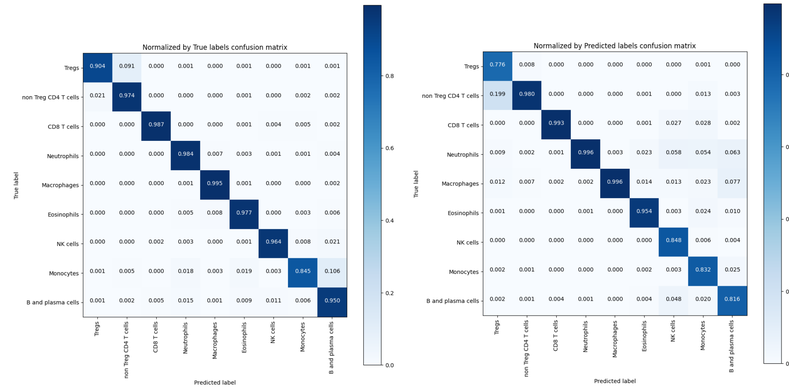


Figure 5. Model is highly specific and accurate for all nine cell populations

Scalability of model uncovers biological insights

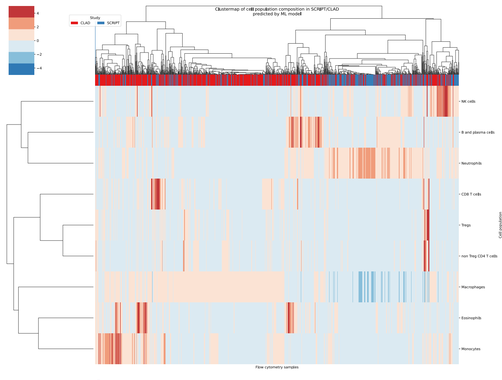


Figure 6. Application of model to 1108 flow cytometry experiments (CLAD-711, SCRIPT-911) reveals structural differences between pneumonia and non pneumonia studies

Conclusions

- Unbiased machine learning applied to flow cytometry data allows scalable and robust identification of different sized cell populations across multiple health conditions.
- Patients with known or suspected pneumonia have both increased population of neutrophils and decreased population of neutrophils when compared to patients that underwent lung transplantation.
- Scalability of this model allows unbiased analysis of hundreds of samples in one hour instead of months of work required from human expert.

* Verschoor, Chris & Lelic, Alina & Bramson, Jonathan & Bowdish, Dawn. (2015). An Introduction to Automated Flow Cytometry Gating Tools and Their Implementation. Frontiers in immunology. 6. 380. 10.3389/fimmu.2015.00380.

Northwestern Medicine®
Feinberg School of Medicine