Team 16 – Machine Learning and Visual Analytics for Cell-based Biomarker Diagnostics

UC San Diego
JACOBS SCHOOL OF ENGINEERING
Shu Chien-Gene Lay Department of Bioengineering

Isabella Garcia, Pranav Siva
Mentor: Dr. Yu "Max" Qian, PhD

JCVI J. CRAIG VENTER INSTITUTE®
UNIVERSITY OF CALIFORNIA SAN DIEGO

Background

Leukemia Diagnosis

- Common blood cancer of white blood cells (Fig. 1, right)
- Phenotypic heterogeneity makes timely diagnosis difficult
- Manual gating of flow cytometry is both labor-intensive and subjective

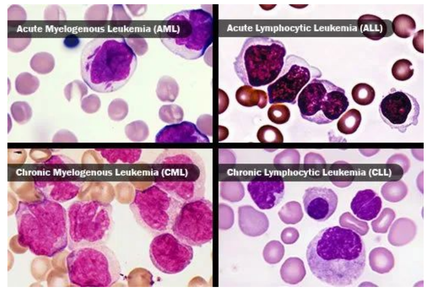

Figure 1. Picture of AML, ALL, CLL, and CML leukemia types (clockwise labeling from top left)

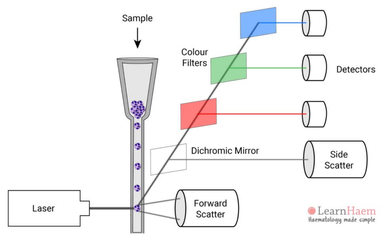

Figure 3. Illustration representing flow cytometry data collection

Flow Cytometry

- Samples are treated with unique fluorescent markers that bind selectively to surface markers of interest
- Single cells passed through a laser, scattering & emitting light for detectors to capture

Non-CLL vs. CLL Distributions on CD5 by CD19

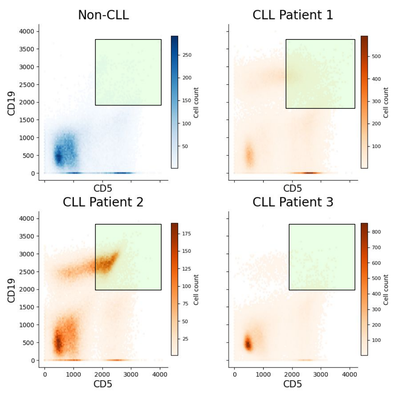

Figure 2. Distributions of non-CLL (blue) versus three CLL patients visualized on CD5 by CD19 (n=100,000 cells each)

Leukemia Variability

- Phenotypic heterogeneity exists even within the same diagnosis — see variability of orange (CLL) samples in Fig. 2, left
- A single patient may even exhibit phenotypic variability as the course of their disease progresses
- This makes reliable diagnosis challenging

Objectives

The objectives for this project were to

1. Quantitatively assess performance of existing normalization methods applied to flow cytometry data
2. Assess existing machine learning models and/or assess novel methods for leukemia classification in flow cytometry data
3. Visualize cell clustering behavior and results via UMAP and 2D plots, and determine if the UMAP embedding also aids normalization efforts.

Experimental Design

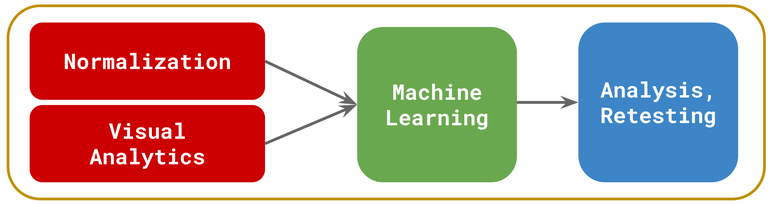

Figure 4. Workflow diagram for the project, integrating subprojects of normalization and visual analytics into the machine learning pipeline to be analyzed
We used a normalization tool called cyCombine, UMAP for visual analytics, and a novel distribution statistic-based classifier

CD45	CD22	CD5	CD19	CD79b	CD3	CD81	CD10	CD43	CD38

Table 1. Marker panel for the CLL dataset used in this study. The data also includes 6 scatter parameters (forward and side scatter Area/Height/Width)

	Train	Validate	Test
Total Samples	102	81	24
Non-CLL	42	27	7
CLL	60	54	17

Table 2. Numbers for the CLL dataset. Diagnoses labels given by CSNN data in References

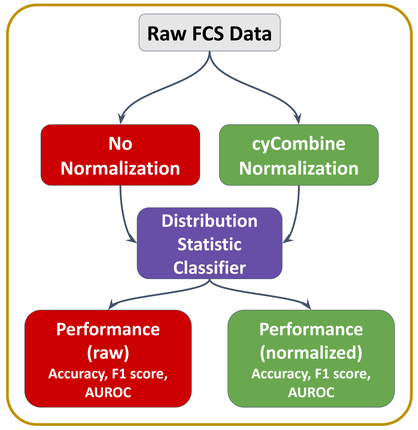

Figure 5. Normalization assessment workflow

Distribution Statistic Classifier

- Non-CLL samples have relatively consistent distributions, whereas CLL samples tend to have high variability (see Fig. 6)
- Variability shown in markers associated with B cells (e.g. CD19)
- If we use distribution statistics as input, little to no data preprocessing is required

Normalization by cyCombine

- Isolated normalization performance is hard to assess, i.e. distribution-level statistics have no inherent meaning
- We test normalization by comparing downstream performance, i.e. A/B testing
- Due to computational constraints, 5,000 cells per sample were subsampled for normalization; downstream analyses were performed on this subsampled dataset

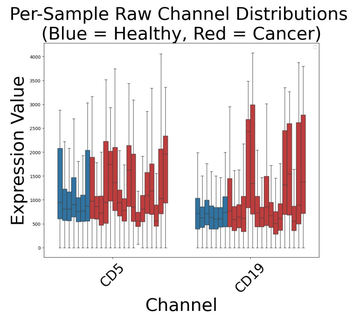

Figure 6. Distributions of FCM samples in the CD5 and CD19 dimensions, with blue non-CLL samples (n=7) and red CLL samples (n=17)

Expected Outcomes

We hypothesized that

1. Normalization will improve classifier accuracy
2. A model trained on simple distribution statistics could achieve better than random classification based on known biological signals,
3. UMAPs can be used to visually identify whether a sample is cancerous

Results

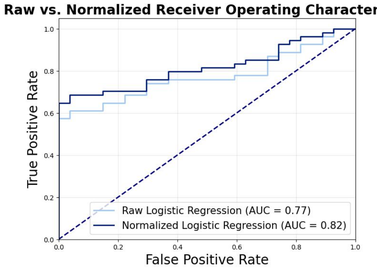

Figure 7. Raw vs. Normalized ROC curve.
Normalization appears to improve classifier accuracy, even when subsampled.
Macro avg F1-score increases marginally from 0.75 to 0.77, indicating insignificant change

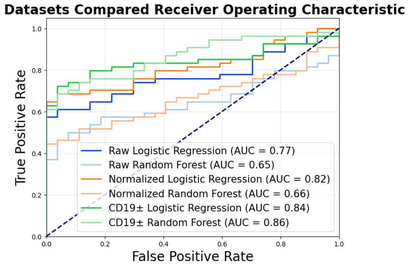

Figure 8. All ML model ROC curves.
The best models are generally the normalized variants, and logistic regression.
CD19± split LR has macro avg F1-score 0.8

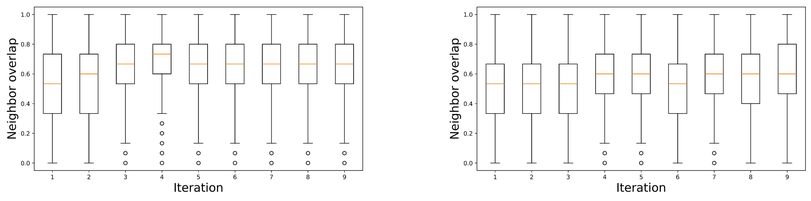

Figure 9. Healthy vs. Diseased Clustering.
Healthy samples see higher rates of cells staying with the same clusters across UMAP runs of the same sample as well as less variance amongst individual clusters

Conclusions & Future Work

In conclusion,

- It is inconclusive whether the normalization tool cyCombine helps classification
- A classifier trained on distribution statistics can achieve relatively good performance
- Heterogeneity can be visualized using UMAP and cluster stability

Future work could:

- assess other common normalization tools
- expand analysis to more complex leukemias (e.g. AML)
- improve feature selection
- expand analysis to other panels/tubes of the CLL dataset, e.g. analyze Igκ and Igλ

Acknowledgements & References

Thank you to our mentor, Dr. Yu "Max" Qian (JCVI), Senior Design Teaching Professor Dr. Taylor, and our TA Yashwin for support and guidance during our senior design project.