Leukemia Cell Subtyping

Blood-Smear Classification and a Case Study in Data Leakage

Capstone Project Report

A computer-vision classifier that sorts a single white blood cell into a benign look-alike or one of three stages of B-cell acute lymphoblastic leukemia — and a case study in why a 99.8% test score can be the least trustworthy figure in the portfolio.

Arel · RLcapstone.ai

1. Summary

This project classifies a photograph of a single white blood cell into four classes: a benign look-alike, plus three maturation stages of B-cell acute lymphoblastic leukemia (ALL) — Early Pre-B, Pre-B, and Pro-B. The model attains **99.8% accuracy** and a **0.998 macro-F1** on the held-out test set.

> That figure is presented not as an achievement but as a **warning**. The public dataset carries no patient identifiers, so a random split almost certainly places cells from the same patient on both sides of the partition. The model may be scoring highly by recognizing per-slide staining and background rather than genuine features of leukemia.

2. Motivation

Distinguishing a benign look-alike cell (hematogones) from true leukemic blasts — and separating the leukemia stages from one another — is genuinely subtle work, even for trained hematologists, which makes it a substantive computer-vision problem. It is also the most serious subject in the portfolio, so honest reporting of the result mattered most here.

3. Dataset

The project uses a public peripheral-blood-smear dataset (Aria et al., 2021) of **3,256 images**, each showing a single cell across the four classes, split 70/15/15 into training / validation / test.

A limitation that cannot be fully remedied

The dataset provides **no patient identifiers** (file names are plain indices), so there is no way to guarantee that one patient's cells remain together. Some degree of data leakage is therefore almost certainly baked into any split of this dataset, and reporting the headline figure without this caveat would be misleading.

4. Model and training

The classifier is the same architecture used for MoleCheck — **YOLO11s-cls** (5.4M parameters), transfer-learned on the blood-cell images at 224×224. Validation accuracy reached **100%** within roughly 30 epochs, which — rather than being reassuring — is itself a signal to examine the data split more critically.

5. Results

Metric	Value
Test accuracy	99.8%
Macro-F1	0.998
Errors	1 of 489 images

On most projects this would be the headline. Here it is the red flag. Cells from a single slide share staining, lighting, focus, and background, and a high-capacity model can score near-perfectly by keying on **those** cues rather than the cell itself. This is the same failure mode the EEG project avoids through subject-level splitting; it cannot be avoided here because the identifiers are simply absent.

6. Deployment

The model was exported to ONNX for the browser demonstration and to TensorFlow Lite for the Android application (Cell Explorer), both reproducing the original model's predictions. The user can select a real cell image and classify it on-device.

7. Limitations

- Not a diagnostic tool — real leukemia diagnosis requires bone-marrow analysis, flow cytometry, and genetic testing.
- The 99.8% figure is very likely inflated by the data-leakage issue above, so it should be read as a best case, not a realistic estimate.
- A single dataset; real clinical practice presents far greater variability.

8. v2 — the honest patient-level result

The **C-NMC 2019** dataset encodes a patient identifier in every filename, enabling a strict patient-disjoint split. Re-framed as a binary task (leukemic vs normal), the model attains **80.8% accuracy** and **ROC-AUC 0.857** on 11 entirely held-out patients — a large, honest drop from v1's leakage-inflated 99.8%, and a far more meaningful figure.

9. v3 — cross-lab generalization

v2 established whether the model works on unseen patients. v3 asks the harder question: does it work on a **different lab's microscope**? Evaluated unchanged on two independent datasets (Aria, Tehran; Acevedo, Barcelona), it failed — flagging **80%** of Barcelona's normal cells as leukemic and scoring worse than chance on raw Tehran images, revealing a heavy dependence on C-NMC's background and staining rather than leukemia itself.

The remedy is multi-source training. v3 trains on **two labs at once** (C-NMC and Aria), holding out Acevedo as an unseen third lab. Generalization to that third lab more than tripled — specificity on Barcelona's normal cells rising from **20% to 67%** (and on the hard-negative lymphocytes from 89% to 36%) — while within-source accuracy also improved, to **83.3%**.

> The lesson in one number: a single-source model, even one evaluated correctly, is still a single-source model. Robustness comes from diversity in the training data, not only discipline in the split.