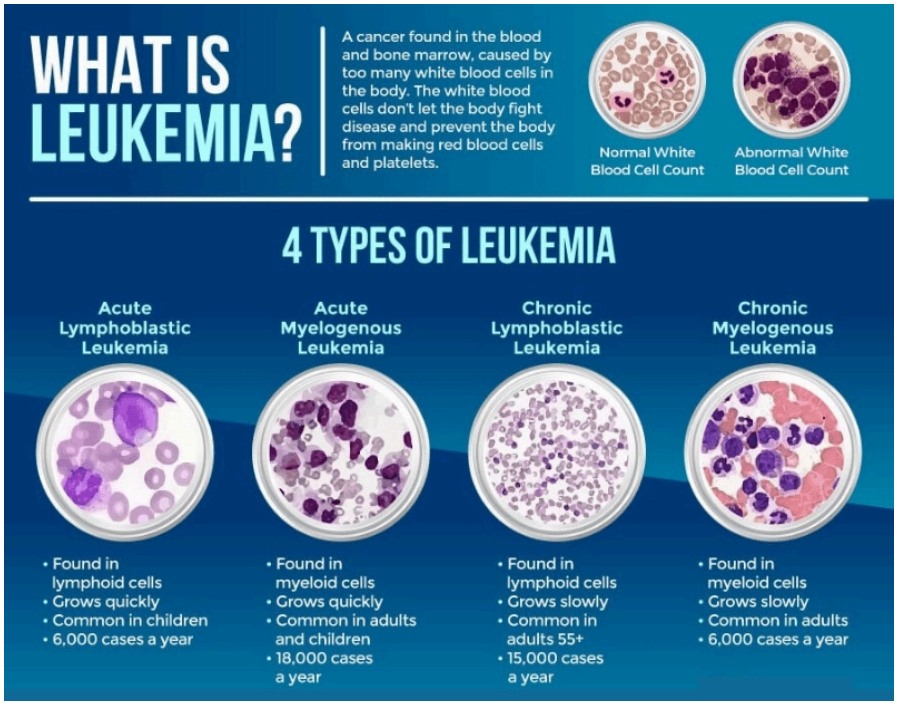
The Computational Quest for Precision in Leukemia Diagnosis

The diagnosis of leukemia is not a single event, but a determination of a precise disease subtype. A patient with Acute Myeloid Leukemia (AML) demands a drastically different treatment strategy than one with Chronic Lymphocytic Leukemia (CLL) or Acute Lymphoblastic Leukemia (ALL). The current diagnostic process, relying on complex lab work and cytogenetics, can be time-consuming and often lacks the objective certainty needed at the point of care.



The project tackles this critical diagnostic challenge by developing a machine learning system capable of rapid, objective classification of the **four major leukemia types**: ALL, AML, CLL, and CML. The goal is to transform thousands of raw biological signals into a clear, confident clinical decision.

The key to this precision lies in the gene expression profile—the genetic fingerprint of the cancer cells. This project is using a massive public resource, the MILE study data (**GSE13164**), which provides over **1,100 patient samples** and the robust genomic evidence required to train our intelligent system.

The focus of this project is a structured approach designed to distill the raw biological data into a high-performance predictive model. The process begins with rigorous **data wrangling**. The data, measured in technical "probe IDs," must be purified and accurately mapped to official Gene Symbols for biological interpretability.

Next, the focus will be on the **discovery of biomarkers**. Training a model on the full ~20,000 genes introduces noise and the risk of overfitting. To overcome this, ANOVA (Analysis

of Variance) will be employed, to pinpoint the 500 most differentially expressed genes. These selected features are the true, critical biomarkers that distinguish the four leukemia subtypes, allowing our model to focus on the strongest biological signals.

Finally, in **model development** where a multi-class classifier will be built using algorithms highly effective for genomic data: Ridge Regression for its ability to manage high dimensionality and Random Forest for its robustness and interpretability. Using Grid Search and K-Fold Cross-Validation, we will systematically optimize these models. The ultimate benchmark for success is stringent: achieving an F1-score and Accuracy exceeding 95% across all four classes, thus validating the system as a reliable, data-driven "second opinion" for diagnosis.