Molecular Classification of Leukemia

ANN project in data science working with leukemia microarray dataset

Project Members

- Hiba Mahdi Misbah
- Sara Ali Adnan
- Zeenah Shamil Kamil
- Noor-AlHuda Ayad AbdulHameed

Supervised by: Assist. Prof. Dr. Suhad Faisal Behadili

Context

Where did we get the dataset?

A famous dataset first used in T. R. Golub (1999), posted on Kaggle.

What have we been doing?

Choosing top genes and working on a suitable model

Research Tools

Platforms

Google Colab
Google Drive API
GitHub
Kaggle

AI Assistant:
DeepSeek
Gemini 2.5 Pro
Miro AI

Libraries

DATA ANALYSIS

Pandas: Data manipulation and analysis
NumPy: Numerical computations
Matplotlib & Seaborn: Data visualization

MACHINE LEARNING

Scikit-learn

- RandomForestClassifier
- StandardScaler
- SelectKBest
- ANOVA F-test
- Mutual Information
- train_test_split

DEEP LEARNING

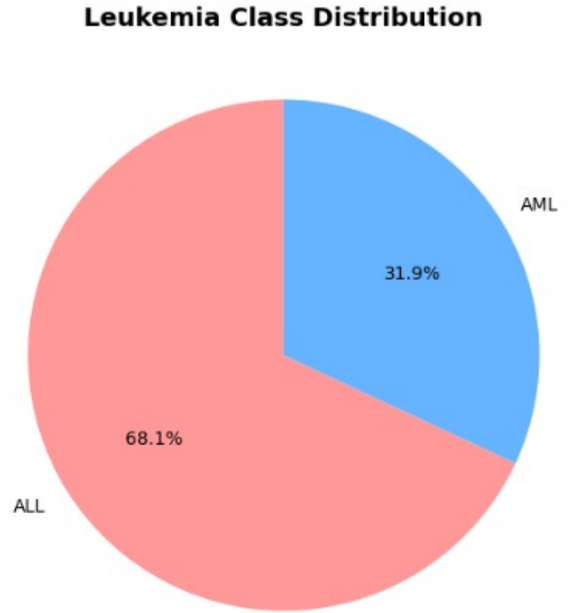
TensorFlow
Keras

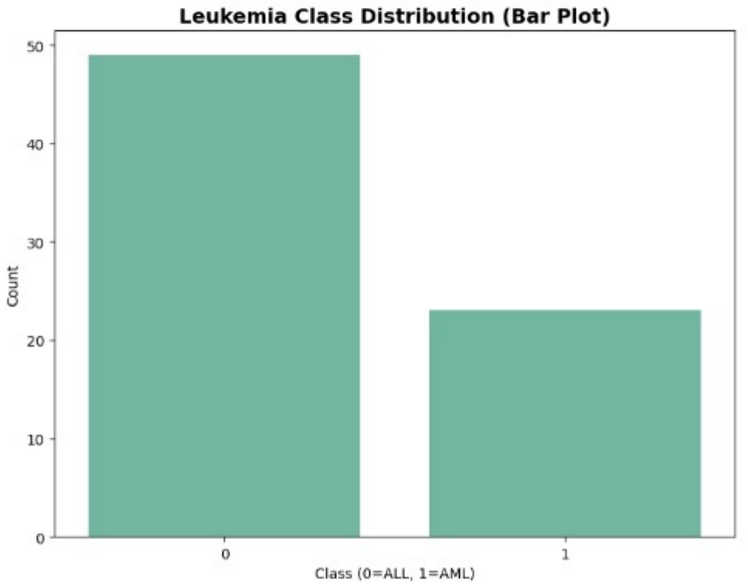
Problem Statement

Binary classification: ALL (0) vs AML (1)
ALL cases (Class 0): 49 (68.1%)
AML cases (Class 1): 23 (31.9%)

72 samples, 7,129 genes → High dimensionality challenge

Class imbalance: 68% ALL, 32% AML





Methodology Overview

Data Preprocessing & Feature Selection

- 7,129 genes → 100 most significant genes (ANOVA F-test)
- StandardScaler for normalization

Neural Network Architecture

- Input: 100 genes
- 3 Hidden Layers (10 neurons each)
- Output: Binary classification (Sigmoid)

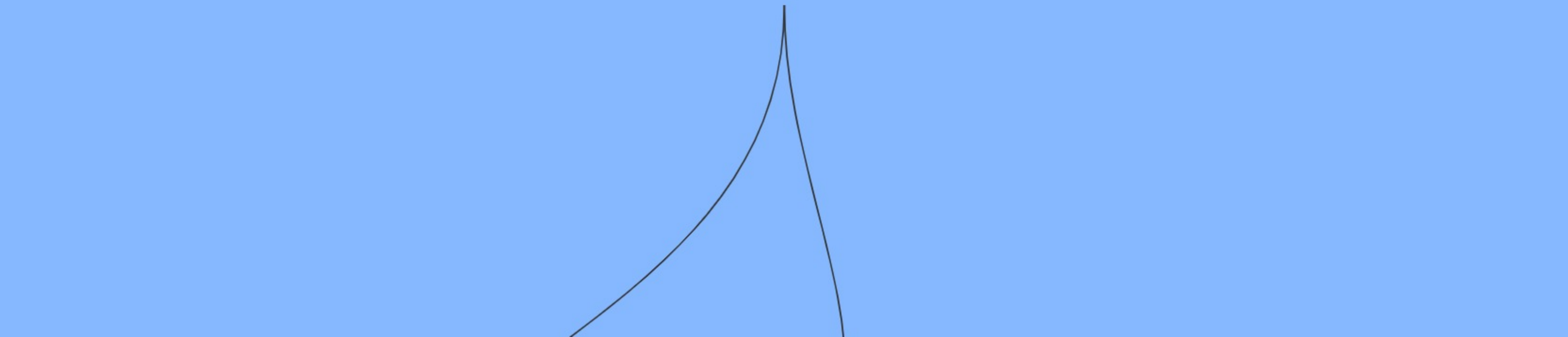
Training Strategy

- Adam Optimizer + Early Stopping
- 200 epochs maximum
- Learning rate reduction

DATA PREPROCESSING PIPELINE

NEURAL NETWORK ARCHITECTURE PIPELINE

TRAINING/TESTING RESULTS



Step 1: Initial Data Loading & Exploration

```
# Load dataset
df =pd.read_csv("/content/drive/MyDrive/leukemiamicroarray.csv", sep=';')
# Basic information about the dataset
print("Dataset Shape:", df.shape) # (72, 7130)
print("\nClass distribution:") # 49 ALL (0), 23 AML (1)
print(df['Leukemia_class'].value_counts())
```

- 72 samples, 7,129 genes + 1 target variable
- Class imbalance: 68% AML vs 32% ALL
- Data already normalized (0-1 range)

Architecture Design

- Input Layer: 100 neurons (selected genes)
- Hidden Layer 1: 10 neurons (sigmoid)
- Hidden Layer 2: 10 neurons (sigmoid)
- Hidden Layer 3: 10 neurons (sigmoid)
- Output Layer: 1 neuron (sigmoid) - Binary classification

Model Configuration

- Optimizer: Adam (adaptive learning rate)
- Loss Function: Binary Crossentropy
- Metrics: Accuracy, Precision, Recall
- Training: 200 epochs max, Batch size 32

7,129 Genes → Feature Selection → 100 Key Genes → Neural Network → AML/ALL Classification

Training Performance

- Final Training Accuracy: 100%
- Final Validation Accuracy: 100%
- Training Time: 200 epochs (no early stopping needed)
- Loss Reduction: Consistent improvement throughout training

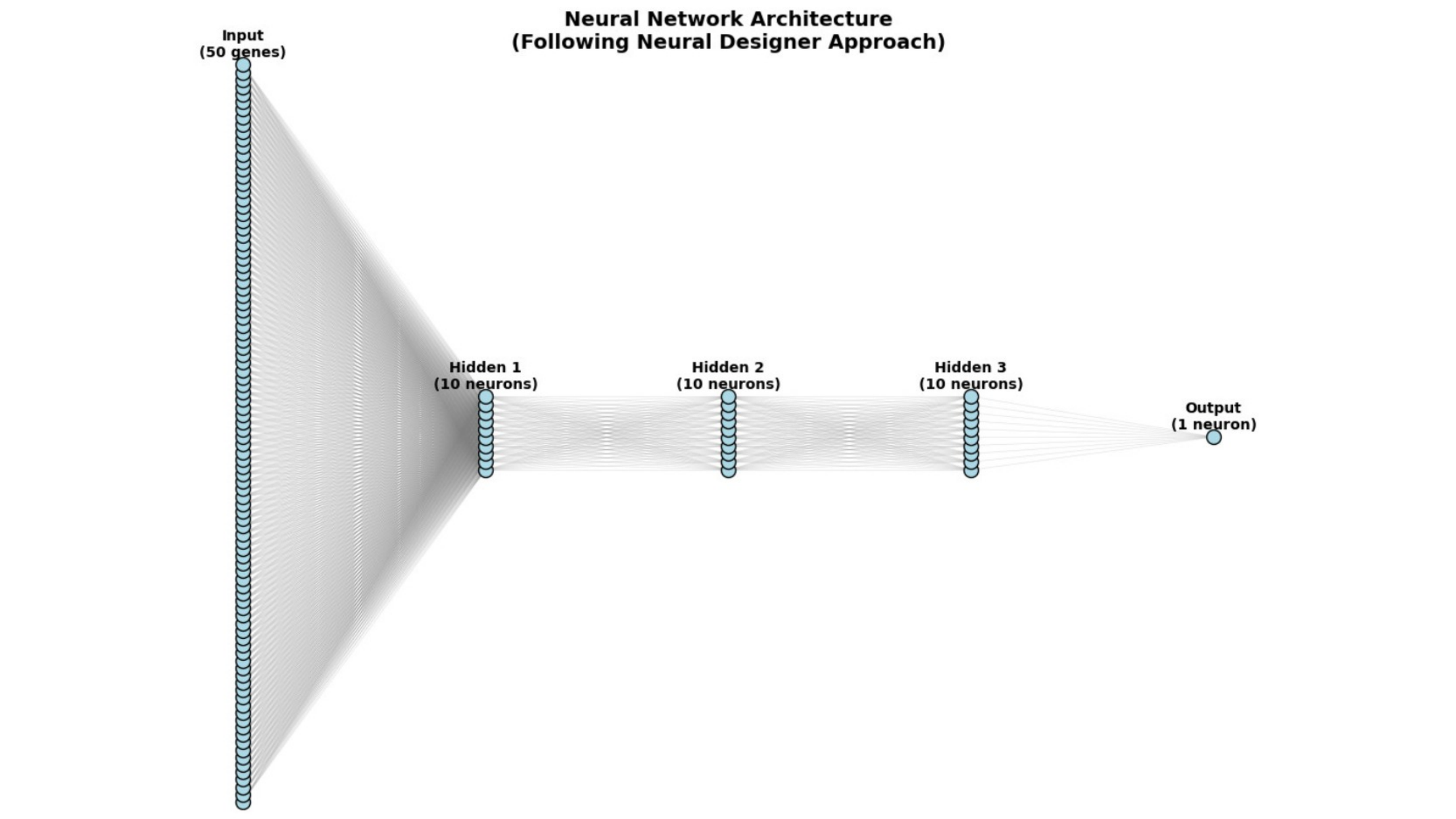
Testing Performance

- Accuracy: 100.00%
- Precision: 100.00%
- Recall: 100.00%
- F1-Score: 100.00%
- AUC-ROC: 1.0000

Step 2: Data Splitting with Stratification

```
# Prepare the data
X = df.drop(columns=['Leukemia_class'])
y = df['Leukemia_class']
# Split the data (80% train, 20% test) with stratification
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42, stratify=y
)
```

- Stratification ensures both splits have same class proportions (68% AML, 32% ALL)
- Prevents one class from being over/under-represented in splits
- 80-20 split: 57 training, 15 testing samples


Neural Network Architecture
(Following Neural Designer Approach)
Input
(50 genes)
Hidden 1
(10 neurons)
Hidden 2
(10 neurons)
Hidden 3
(10 neurons)
Output
(1 neuron)

Hidden Relationships

1. **Gene Expression Signature to Leukemia Type**: a specific subset of **genes has a strong, predictive link to the leukemia class**. The feature selection process (using ANOVA F-test) systematically uncovered that **genes are not equally important**.

2. **Non-Linear Relationships:** A simple **linear model might not capture the full complexity of the gene interactions**. The use of a neural network with non-linear activation functions (sigmoid) is designed to learn these complex, non-obvious relationships where the influence of one gene might depend on the levels of several others.

	metric	# neural designer	# our model	# improvement
1	Accuracy	94.12	100.00	5.88
2	Error Rate	5.88	0.00	-5.88
3	Sensitivity (ALL Recall)	92.86	100.00	7.14
4	Specificity (AML Recall)	95.24	100.00	4.76
5	Precision (ALL)	92.86	100.00	7.14
6	F1-Score (ALL)	92.86	100.00	7.14

Step 3: Feature Scaling

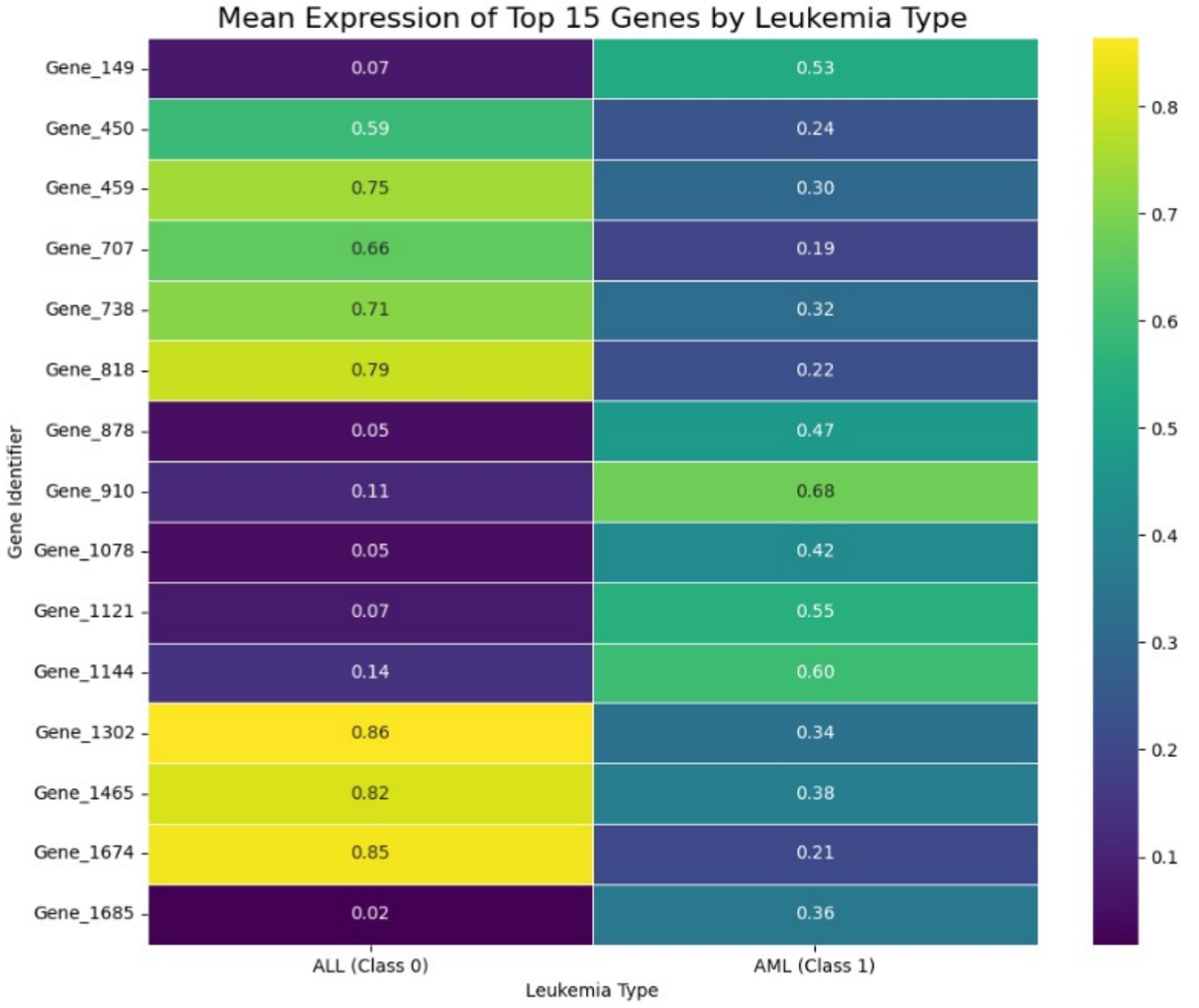
```
# Scale the features
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

StandardScaler Effect:

- Centers data to mean=0, standard deviation=1
- Formula: (x - mean) / std
- Prevents features with large ranges from dominating
- Fit on train only to avoid data leakage

Analysis of Top 15 Gene Expression Patterns

- High Expression (Bright Yellow): Indicates the gene is more active.
- Low Expression (Dark Purple): Indicates the gene is less active.



Step 4: Three Feature Selection Methods Compared

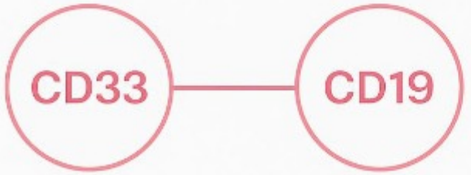
ANOVA F-test **Mutual Information**

Random Forest Importance

```
methods = {
    'ANOVA F-test': (X_train_scaled, y_train),
    'Mutual Information': (X_train_scaled, y_train),
    'Random Forest': (X_train_scaled, y_train)
}
```

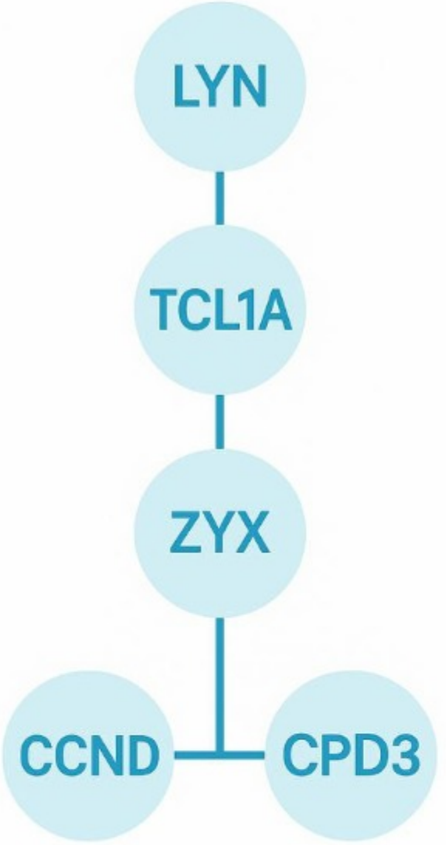
ANOVA F-test was selected as the best method

Receptors & Cell Recognition



Immune & Inflammatory Genes

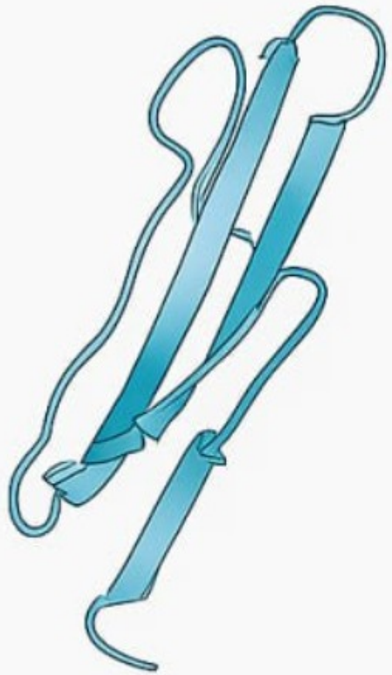
Signaling & Structural Regulatior



Enzymes & Metabolism

CD33 (Gene_818)

Transmembranc receptor on myeloid cells, Overexpressed in Acute Myeloid Leukemia (AML) and is a therapeutic target.

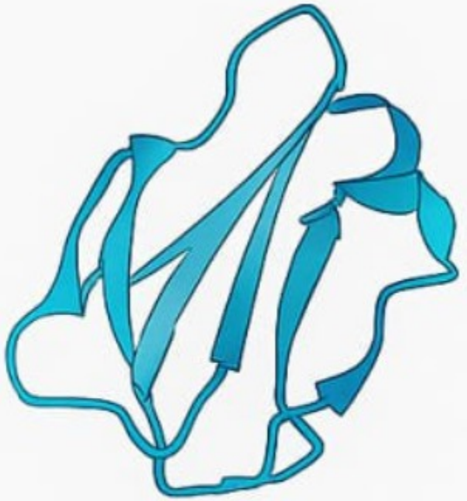


(Source: RCSB PDB – 51HB)

CST3 (Gene_149)

Cystatin C is a cysteine protease inhibitor regulating cathepsins.

Dysregulation in leukemia alters extracellular proteolysis, promoting tumor invasion.



(Source: RCSB PDB – 3GAX)

ZYX (Gene_910)

LIM-domain cytoskeletal protein controlling cell adhesion and migration.

Abnormal signaling may enhance leukemia cell motility and disrupt bone marrow interactions.

(Source: AlphaFold - AF-Q625233-F1)

Thank you!

	Placeholder ID	Gene Symbol	Brief Functional...
1	Gene_818	CD33	A surface protein on myeloid cells. A well-known biomarker and major therapeutic target for AML.
2	Gene_149	CST3	Cystatin C. Involved in protein degradation; its altered expression is linked to various cancers.
3	Gene_910	ZYX	Zyxin. A protein involved in the cell's structural skeleton and adhesion; often implicated in cancer metastasis.
4	Gene_1674	PTGS2	Prostaglandin-Endoperoxide Synthase 2 (also known as COX-2). An enzyme involved in inflammation.
5	Gene_1078	CTSD	Cathepsin D. An enzyme that degrades proteins; plays a role in tumor progression and invasion.