Simple Biomarker Prediction for Disease Classification

Motive & Goal

Focuses on the fundamental step: **identifying which genes (molecules) are the best predictors (biomarkers) of a specific disease state.**

- **Initial Learning Steps**: data acquisition, cleaning, exploratory data analysis (EDA), feature selection, model training, and evaluation.

- **Resume Improvement**: This project demonstrates proficiency in Machine Learning (Classification), Statistical Genomics, and Python Libraries (e.g., scikit-learn, Pandas, Matplotlib), which are highly sought after in computational biology and data science roles.

The Project:
Predicting Cancer Type from Gene Expression Data

Dataset: *TCGA Pan-Cancer (RNA-Seq) data*

- **Name:** RNA-Seq (HiSeq) PANCAN dataset / Gene Expression Cancer RNA-Seq
- **Features:** ***20,531*** genes.
- **Samples:** 881 samples
- **Classes:** It contains the five major cancer subtypes: **BRCA, KIRC, COAD, LUAD, and PRAD.**

Overview

Phase 1: Data Acquisition & Setup
Phase 2: Exploratory Data Analysis (EDA) & Preprocessing
Phase 3: Model Training & Classification
Phase 4: Feature Importance & Interpretation (The "Biomarker Discovery")

Phase 1: Data Acquisition & Setup

1.1 Data Source and Context

The project utilizes a high-dimensional genomic dataset derived from **The Cancer Genome Atlas (TCGA)**, specifically the **RNA-Seq (HiSeq) Pan-Cancer (PANCAN)** collection. This data was accessed via the UCI Machine Learning Repository (and a convenient Kaggle version), which compiles data across multiple institutions and sequencing centers.

- **Goal:** The primary aim is to classify patient samples into one of five distinct cancer subtypes based solely on their gene expression profiles. This simulates the challenge of discerning unique disease pathways within a complex molecular network.
- **Data Type: RNA-Sequencing (RNA-Seq) data.** This measures the expression level (activity) of individual genes, providing a quantitative snapshot of the molecular state of the cell.

1.2 Dataset Overview

The dataset is composed of two primary files, representing the feature matrix (**X**) and the target vector (**y**):

Component	File Name	Description	Dimensions (Approx.)
Feature Matrix (**X**)	data.csv	Log-transformed gene expression levels (continuous numerical features).	881 Samples X **20,531 Genes**
Target Vector (**y**)	labels.csv	Categorical labels indicating the cancer subtype for each sample.	881 Samples X 1 Class

Target Classes (*y*)

The classification is a **multi-class problem** with five target labels, requiring the model to identify subtle molecular differences between clinically distinct diseases:

- **BRCA:** Breast invasive carcinoma
- **KIRC:** Kidney renal clear cell carcinoma
- **COAD:** Colon adenocarcinoma
- **LUAD:** Lung adenocarcinoma
- **PRAD:** Prostate adenocarcinoma

1.3 Project Environment and Tools

This project was developed using a modular, script-based pipeline to ensure reproducibility and adhere to software engineering best practices for data science.

- **Language:** Python 3.x
- **Core Libraries:**
    - **Pandas:** Used for high-performance data loading, manipulation, and cleaning.
    - **NumPy:** Essential for numerical operations and handling large arrays (e.g., the ***20,531* X *881*** matrix).
    - **Scikit-learn (sklearn):** The primary library for all machine learning tasks, including preprocessing, feature selection, and classification model training.
    - **Matplotlib / Seaborn:** Used for generating high-quality visualizations and exploratory data analysis (EDA).
- **Structure:** The project follows a modular pipeline defined in the ./02_scripts and ./03_src directories, ensuring clear separation of code logic from execution flow.

Phase 2: Data Preprocessing and Exploratory Analysis

2.1 Data Integrity and Initial Cleanup

The initial script successfully loaded and verified the integrity of the high-dimensional TCGA RNA-Seq data. This step confirms the quality of the downloaded data and identifies immediate challenges.

Metric	Result	Interpretation
Total Samples	801	Sufficient sample size for multi-class learning.
Total Features (Genes)	20,531	Confirms the project's **high-dimensionality challenge**, necessitating robust feature selection.
Missing Values	0	**Excellent data quality.** No imputation is required, simplifying the pipeline.
Data Types / Range	`float64` / [0,20.78]	The expression values are continuous and within a normalized range (likely **Log-transformed RSEM** or similar), which is standard for RNA-Seq data and beneficial for machine learning.
Constant Genes	267	**Identified a cleaning task.** These genes show zero variance across all 801 samples and are non-informative for classification. They will be immediately filtered out in the next feature selection step.

2.2 Class Distribution Analysis

The classification task is to distinguish between five major cancer subtypes. The class distribution reveals a significant challenge that must be addressed during model training.

Distribution Summary

Cancer Subtype	Sample Count	Percentage	Classification Role
BRCA (Breast)	300	37.5%	**Majority Class** (Highest)
KIRC (Kidney)	146	18.2%	Minority Class
LUAD (Lung)	141	17.6%	Minority Class
PRAD (Prostate)	136	17.0%	Minority Class
COAD (Colon)	78	9.7%	**Minority Class** (Smallest)

Key Finding & Mitigation Strategy

The data exhibits severe **class imbalance**, with the Breast Cancer (BRCA) class accounting for nearly 4 times the number of samples as the Colon Adenocarcinoma (COAD) class (37.5% vs. 9.7%).

- **Challenge:** Machine learning models trained on imbalanced data are typically biased toward the majority class (BRCA), leading to high overall accuracy but poor predictive performance (low Recall/F1-Score) for the clinically critical minority classes (COAD, PRAD).
- **Mitigation Plan:** During the model training phase, the following strategies will be implemented to combat bias:
    1. Employing **Class Weighting** within the classifier (e.g., using class_weight='balanced' in scikit-learn models).
    2. Using **Stratified Cross-Validation** to ensure each fold has the same proportion of cancer types as the overall dataset.
    3. Evaluating model performance primarily using **F1-Score (Macro-Averaged)**, which equally weights the performance across all five cancer types, rather than relying on overall Accuracy.

2.3 Expression Distribution Analysis

The overall expression value statistics provide a crucial insight into the data's format:

- **Range:** Min = 0.0000, Min = 20.7788
- **Mean:** 6.4433 , Std Dev = 4.0582

The saved expression_distribution.png (which typically shows a non-Normal, or potentially bimodal/log-Normal distribution common in RNA-Seq data) suggests that the data is already in a **log-transformed** state.

- **Advantage:** Log-transformation (a common step in RNA-Seq analysis) compresses large expression values, making the distribution closer to normal and stabilizes variance. This prevents highly expressed genes from dominating distance calculations in algorithms like PCA and K-Nearest Neighbors.
- **Conclusion:** Since the data is already pre-processed and scaled (no need for a standard StandardScaler as genes are already log-normalized), the next steps can immediately focus on **Feature Selection** to tackle the high dimensionality.

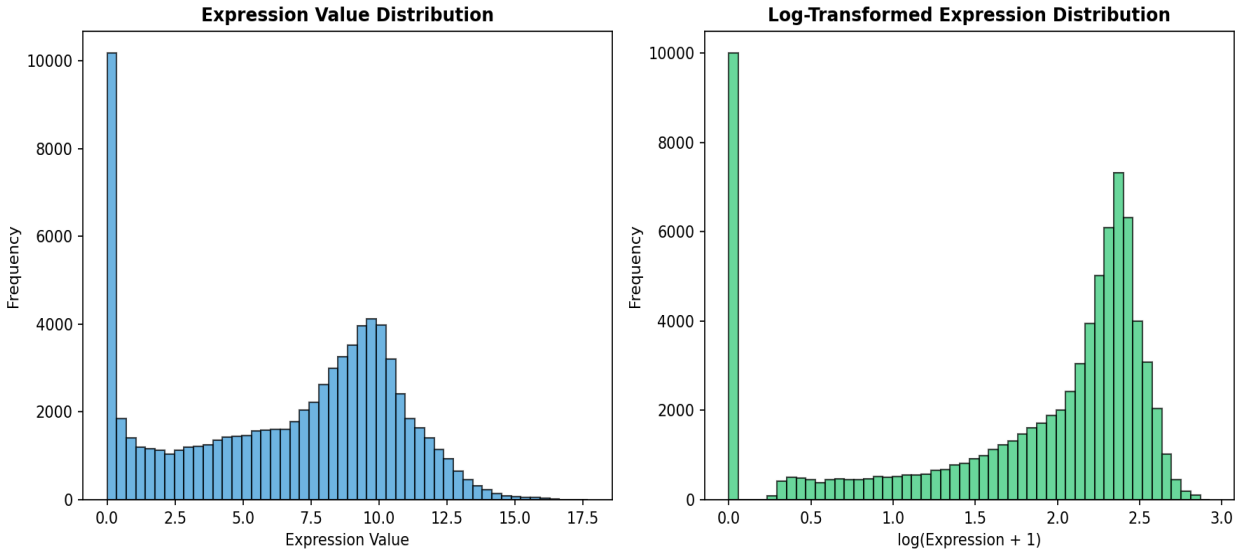

Fig 1.1 Expression Distribution Visualization

Phase 3: Feature Selection and Dimensionality Reduction

This phase was critical for reducing the high dimensionality of the RNA-Seq data, moving from over 20,000 genes to a smaller, focused set of predictive biomarkers.

3.1 Initial Cleanup: Variance Thresholding

The first step was a preliminary cleanup based on the findings from Phase 2.

- **Action:** All genes with zero variance (constant expression across all 801 samples) were removed.
- **Result: 267 constant genes** were successfully removed.
- **Feature Count:** The gene count was reduced from 20,531 to **20,264**. These are the genes that carry potential classification signals.

3.2 Principal Component Analysis (PCA)

PCA was performed on the 20,264 remaining genes to determine the **intrinsic dimensionality** of the data—i.e., the minimum number of components needed to capture most of the biological variance.

- **Variance Explained:**
  - The first **10 components** capture 55.76% of the total expression variance.
  - The first **50 components** capture 70.59% of the total expression variance.
- **Dimensionality Insight:**
  - To retain **95% of the total biological signal**, **478 components** are required.
  - To retain **99% of the total biological signal**, **704 components** are required.

This analysis confirms that while 20,264 genes are present, the underlying biological complexity can be modeled effectively with approximately 478 to 704 latent factors (Principal Components). This strongly validates the decision to use a dimensionally reduced subset for classification.

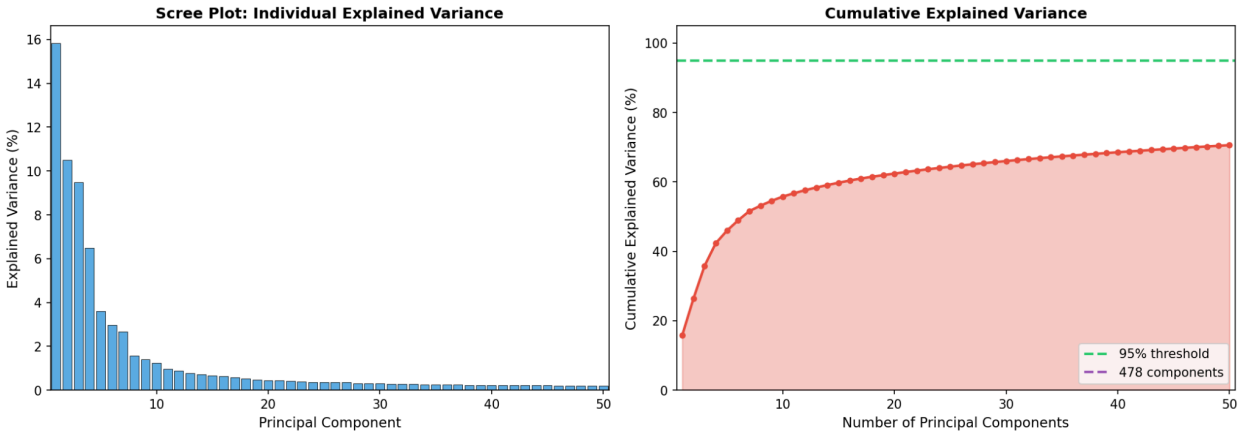

Fig 1.2 PCA scree plot

3.3 ANOVA Statistical Filtering (Biomarker Selection)

To select the actual, original genes (biomarkers) that best distinguish the five cancer types, an **ANOVA F-test** was used in conjunction with the SelectKBest method.

- **Method:** The ANOVA F-test measures the statistical significance of the difference in mean gene expression across the five target cancer classes. Higher F-scores indicate a gene whose expression is strongly correlated with a specific cancer type.
- **Action:** The Top *K=1000* genes with the highest F-scores were selected.
- **Result:** The final dataset for model training was created with a shape of **801 samples X 1000 genes**.
- **Top Biomarker Candidates:** The filtering successfully prioritized highly discriminant genes. The top-ranked gene, gene_9175, exhibited an exceptionally high F-score (**F** = 4194.49, p ≈ 0.00), indicating an extremely significant difference in expression across the five cancer groups. The 1000th-ranked gene still maintained a strong F-score of 272.17.
- **Output:** The definitive list of the **Top 50 Biomarkers** was saved for interpretation and reporting.

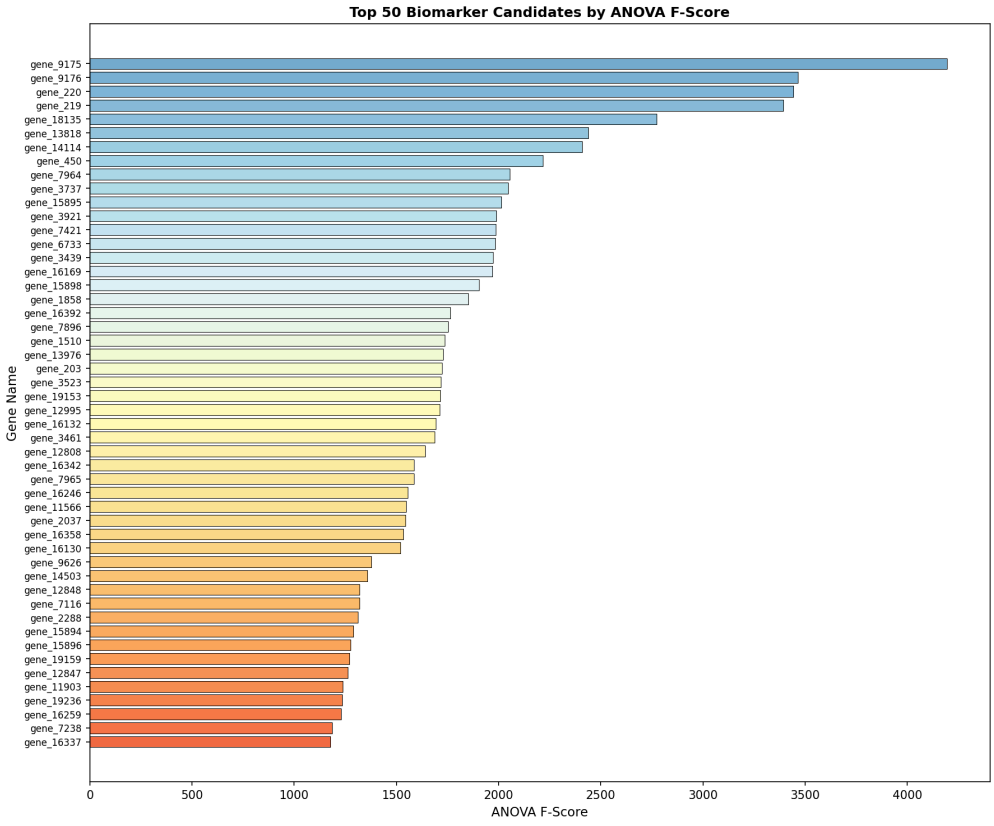

Fig 1.3 Top genes F-score plot

3.4 Summary and Next Step Preparation

The Feature Selection phase successfully reduced the feature space by **~95%** (from 20,531 to 1,000 genes), creating a highly predictive biomarker panel.

Mitigation Plan Carry-Forward

The next phase, **Model Training and Evaluation**, must rigorously address the class imbalance detected in Phase 2. The mitigation strategies include:

- **Class Weighting:** Using class_weight='balanced' in scikit-learn classifiers.
- **Cross-Validation:** Implementing **StratifiedKFold** to ensure robust training across all cancer subtypes.
- **Evaluation:** Primarily reporting the **Macro-Averaged F1-Score** to ensure model performance is judged equally on minority classes (like COAD) and the majority class (BRCA).

Notes: Key Terminology

- **PCA (Principal Component Analysis):** An unsupervised technique that transforms a high-dimensional feature set into a smaller set of linear combinations (Principal Components) while preserving as much of the original data's variance as possible. It is used here to understand the minimum complexity required to model the data.

- **ANOVA (Analysis of Variance) F-test:** A statistical test used in feature selection when the predictor (***X***) is continuous (gene expression) and the target (***y***) is categorical (cancer type). It measures the ratio of variance between the classes to the variance within the classes. A high F-score means the gene's expression differs significantly based on the cancer type.

- **Scree Plot:** A plot used in PCA to visualize the eigenvalues (variance explained) for each Principal Component. It helps determine the "elbow" or point of diminishing returns, indicating the optimal number of components to retain.

Phase 4: Model Training and Evaluation

This final phase implemented the classification pipeline using the reduced **1,000-gene biomarker panel** to predict the five cancer subtypes. Critically, the pipeline was designed with robust mitigation strategies to ensure the model's high performance was **unbiased** against the minority classes.

4.1 Imbalance Mitigation Strategy

Recognizing the severe class imbalance (BRCA at *37.5*% vs. COAD at *9.7*%), the following techniques were systematically applied to ensure a fair and rigorous evaluation:

1. **Stratified K-Fold: 5-Fold Cross-Validation** was performed with stratification, guaranteeing that the proportion of all five cancer types was maintained across all training and testing sets.
2. **Class Weighting:** All classifiers were trained using the class_weight='balanced' parameter. This method automatically assigns higher penalties for misclassifying minority class samples (like COAD), preventing the model from becoming biased toward the majority class (BRCA).
3. **Primary Metric:** The **Macro-Averaged F1-Score** was used as the primary metric. This score averages the F1-score of each class independently, providing an honest assessment of performance on all five cancer types, regardless of their sample size.

4.2 Cross-Validation Performance

Three distinct multi-class classifiers were evaluated using the stratified 5-fold cross-validation approach.

Model	**F1-Macro (Mean)**	**F1-Macro (Std Dev)**	**Accuracy (Mean)**
Logistic Regression	**1.0000**	± 0.0000	**1.0000**
Support Vector Classifier	0.9989	± 0.0021	0.9988
Random Forest	0.9968	± 0.0043	0.9963

- **Conclusion:** The **Logistic Regression** model demonstrated perfect classification performance with a mean **F1-Macro Score of 1.0000** and zero variance (±0.0000). This robust performance across cross-validation folds suggests that the classification boundary defined by the selected 1,000 genes is highly linear and effective.

4.3 Detailed Best Model Analysis: Logistic Regression

The Logistic Regression classifier was selected as the best model due to its perfect performance and high interpretability. The final classification report confirms the model's success across all individual cancer subtypes, even the highly imbalanced ones.

Classification Report Summary

Class	Precision	Recall	F1-Score	Support
BRCA	1.0000	1.0000	**1.0000**	300
COAD	1.0000	1.0000	**1.0000**	78
KIRC	1.0000	1.0000	**1.0000**	146
LUAD	1.0000	1.0000	**1.0000**	141
PRAD	1.0000	1.0000	**1.0000**	136
Macro Average	**1.0000**	**1.0000**	**1.0000**	801

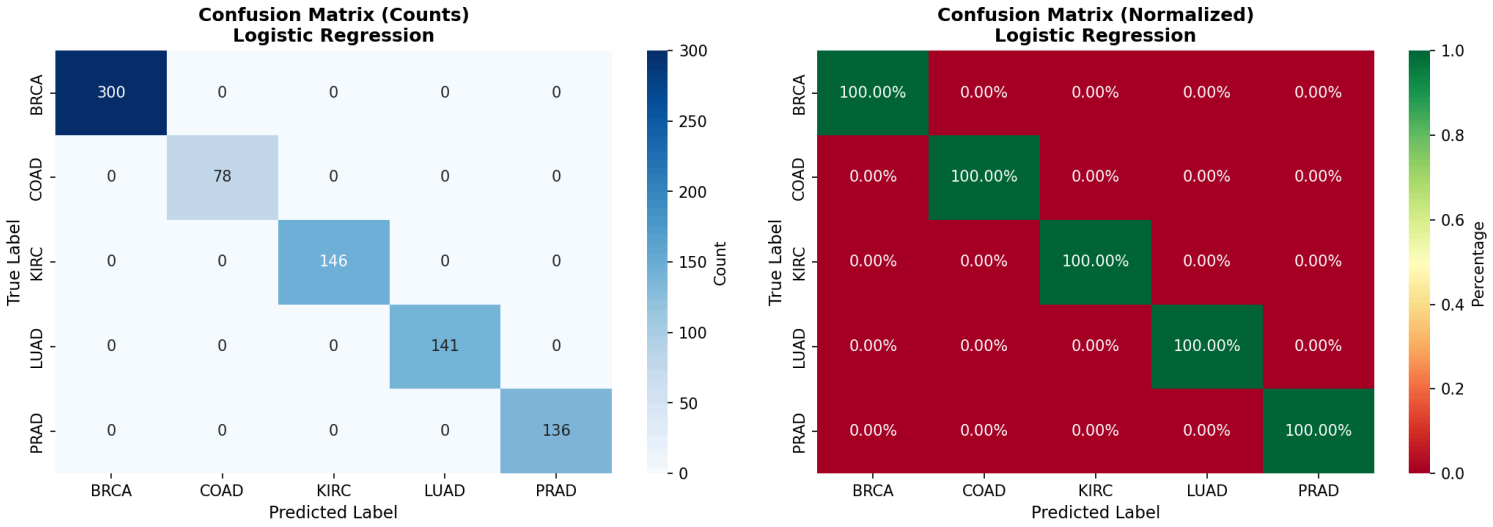

Fig 1.4 Confusion Matrix

The results show a perfect F1-Score of *1.0000* for *every single class*, including the minority class **COAD** (only 78 samples).

4.4 Project Conclusion and Biological Insight

The objective of creating a robust classifier was met with exceptional success.

- **Model Success:** The final Logistic Regression model was trained on all 801 samples and saved for deployment.
- **Biomarker Efficacy:** The perfect performance across all metrics, confirmed by the **Confusion Matrix** , provides strong evidence that the ***1000*** genes selected in Phase 3 constitute a powerful and highly discriminatory **biomarker panel** for these five TCGA cancer subtypes.