ensIIE

ÉCOLE NATIONALE SUPÉRIEURE D'INFORMATIQUE POUR L'INDUSTRIE ET L'ENTREPRISE

MÉTHODE DE RÉGRESSION RÉGULARISÉE

TOPIC : BREAST CANCER CLASSIFICATION

Teacher : Juhyun PARK

Sokheng DIN

Veasna RA

Tables of contents

Part I: Data & Methods

1. Dataset Description
2. Data Cleaning and EDA
3. Statistical Tests
4. Differential Expression Analysis on GeneX

Part II: Predictive methodology

5. Logistic Regression
6. Ridge & Lasso Regression
7. Adaptive Lasso Model
8. UniLasso Model
9. ElasticNet Model
10. Class Imbalance Handling with SMOTE
11. Results & Conclusions

Dataset Description

Key Information

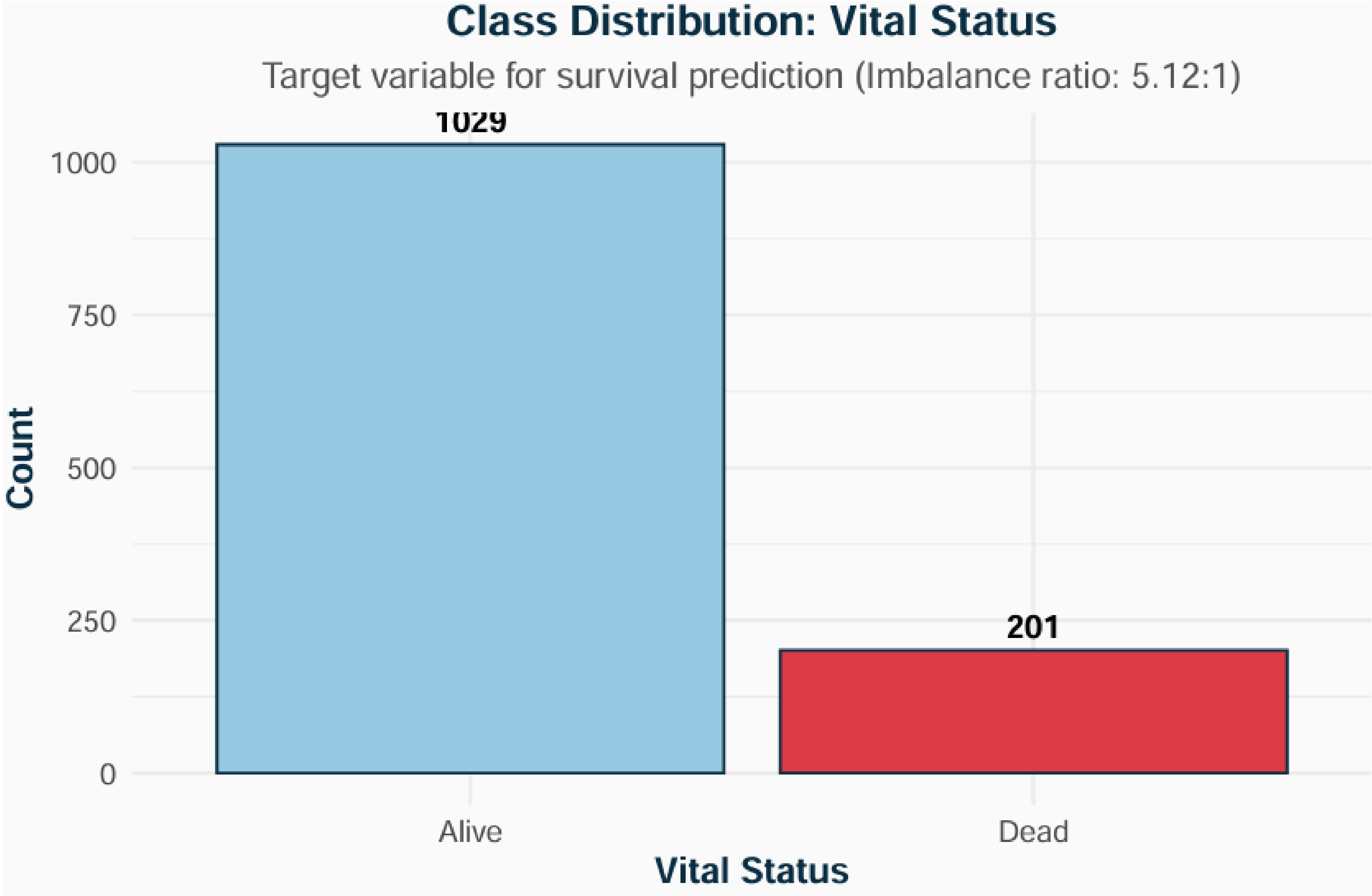
- **We have 2 mains datasets : Clinical and GeneX data**
- Clinical shape (1231 x 24)
- **GeneX shape (123 x 5000)**

Target Variables

- **Vital status : Alive & Dead**

Objectives

- **Classification: Predict vital status base on Clinical and GeneX data**

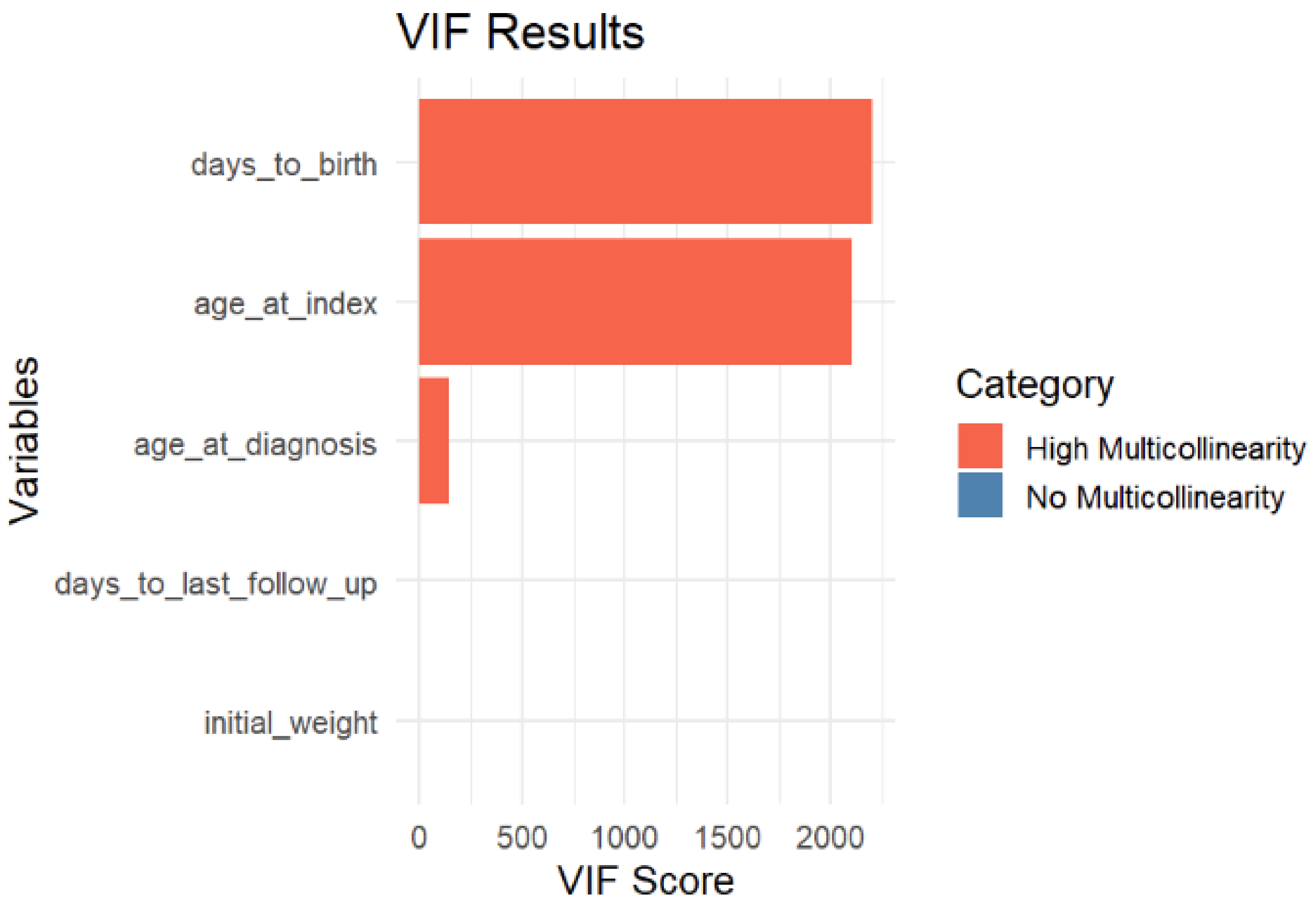


Data Cleaning & EDA

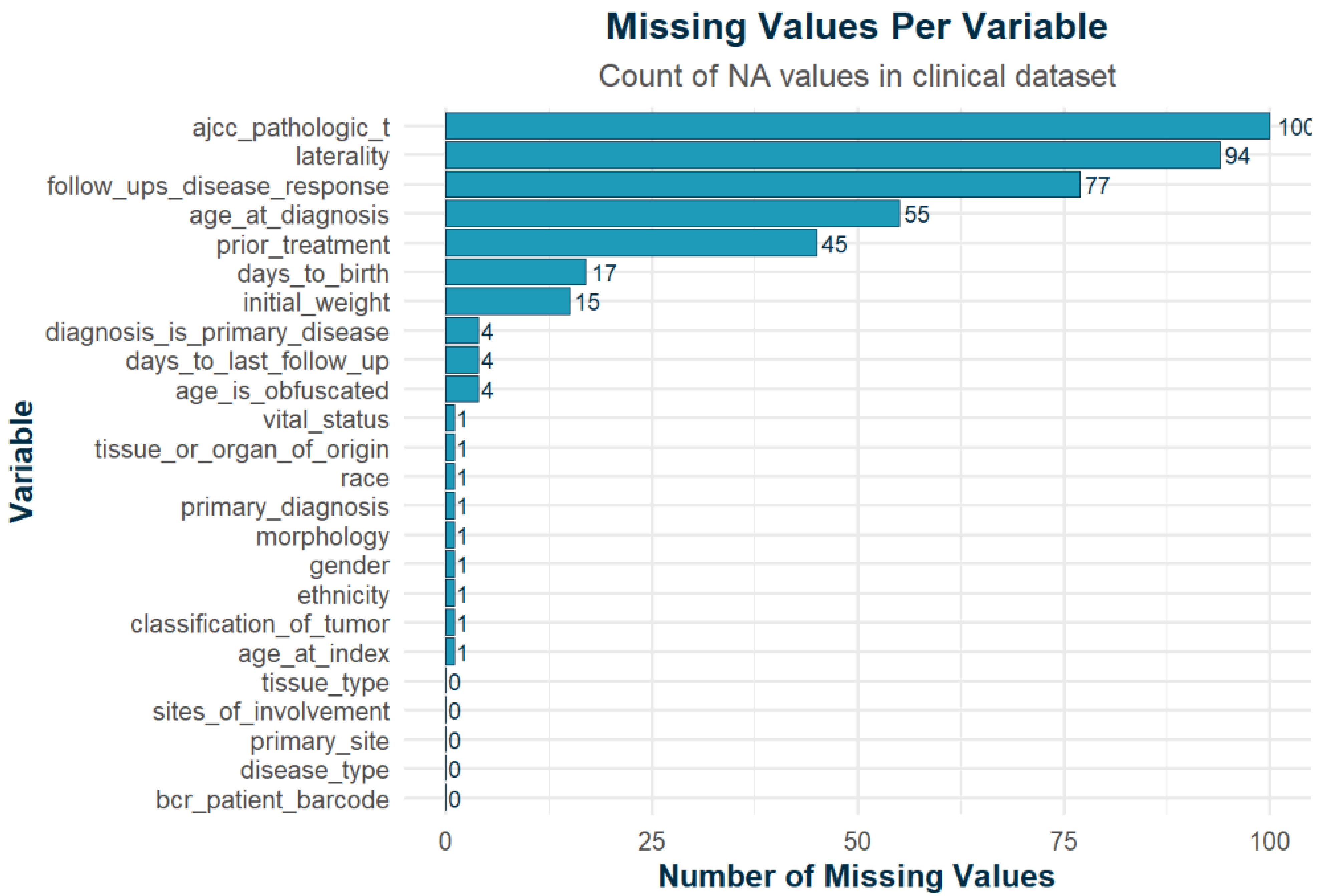
2

Handle with missing values

- **Numerical variables :** Impute by medians
- **Categorical variables :** Impute by mode



VIF Results



Missing Values Per Variable

Count of NA values in clinical dataset

Data Source: Clinical Dataset

Statistical Tests on Clinical Data

3

Numerical Explanatory Features

- **t-test : Distribution is Normal**
- **Wilcoxon Rank-Sum Test : Distribution not normal**

The two-sample t-test statistic is:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$Z = \frac{W - \frac{n_1(n_1+n_2+1)}{2}}{\sqrt{\frac{n_1 n_2(n_1+n_2+1)}{12}}}$$

where the pooled variance is:

Where W = R1 be the sum of rank for group 1

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

and the degrees of freedom are:

$$df = n_1 + n_2 - 2$$

By applying the statistical tests, We can get the most signifiants variables (p_value < 0.05) such as :

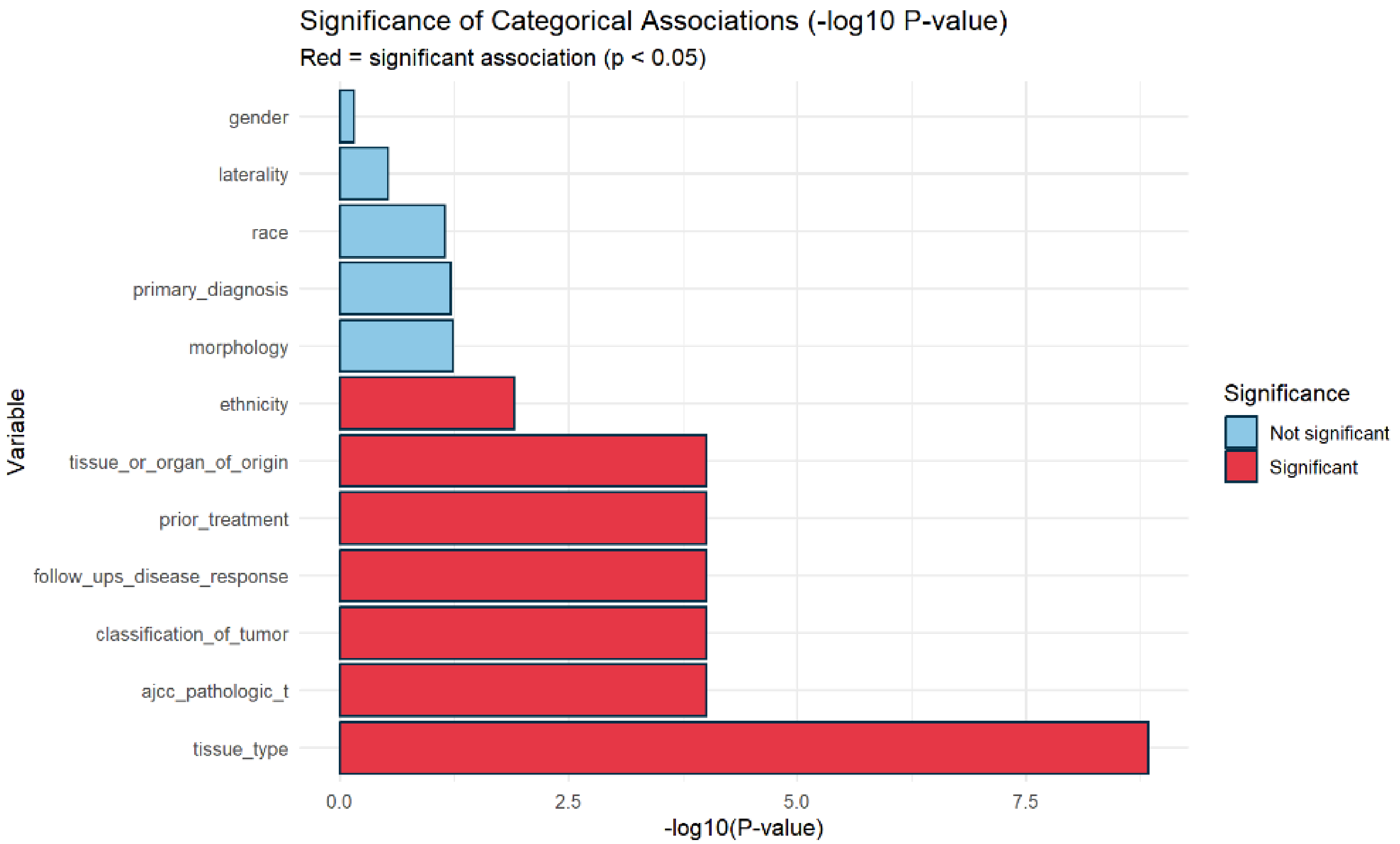
- age_at_index
- initial_weight
- days_to_last_follow_up

Categorical Variables

$$E_{ij} = \frac{(rowtotal)(columntotal)}{grandtotal}.$$

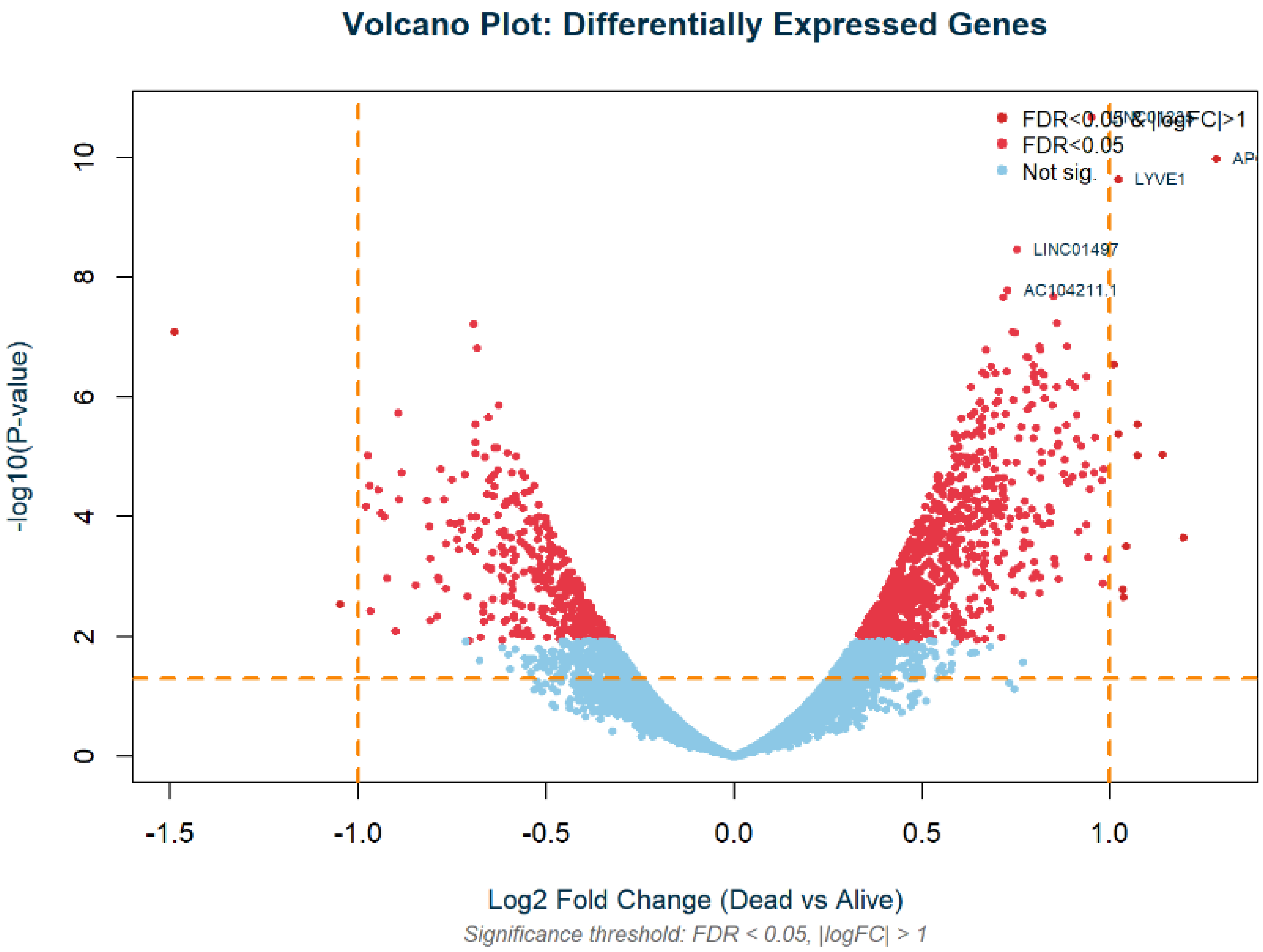
Used when all expected counts satisfy $E_{ij} \geq 5$. The test statistic is:

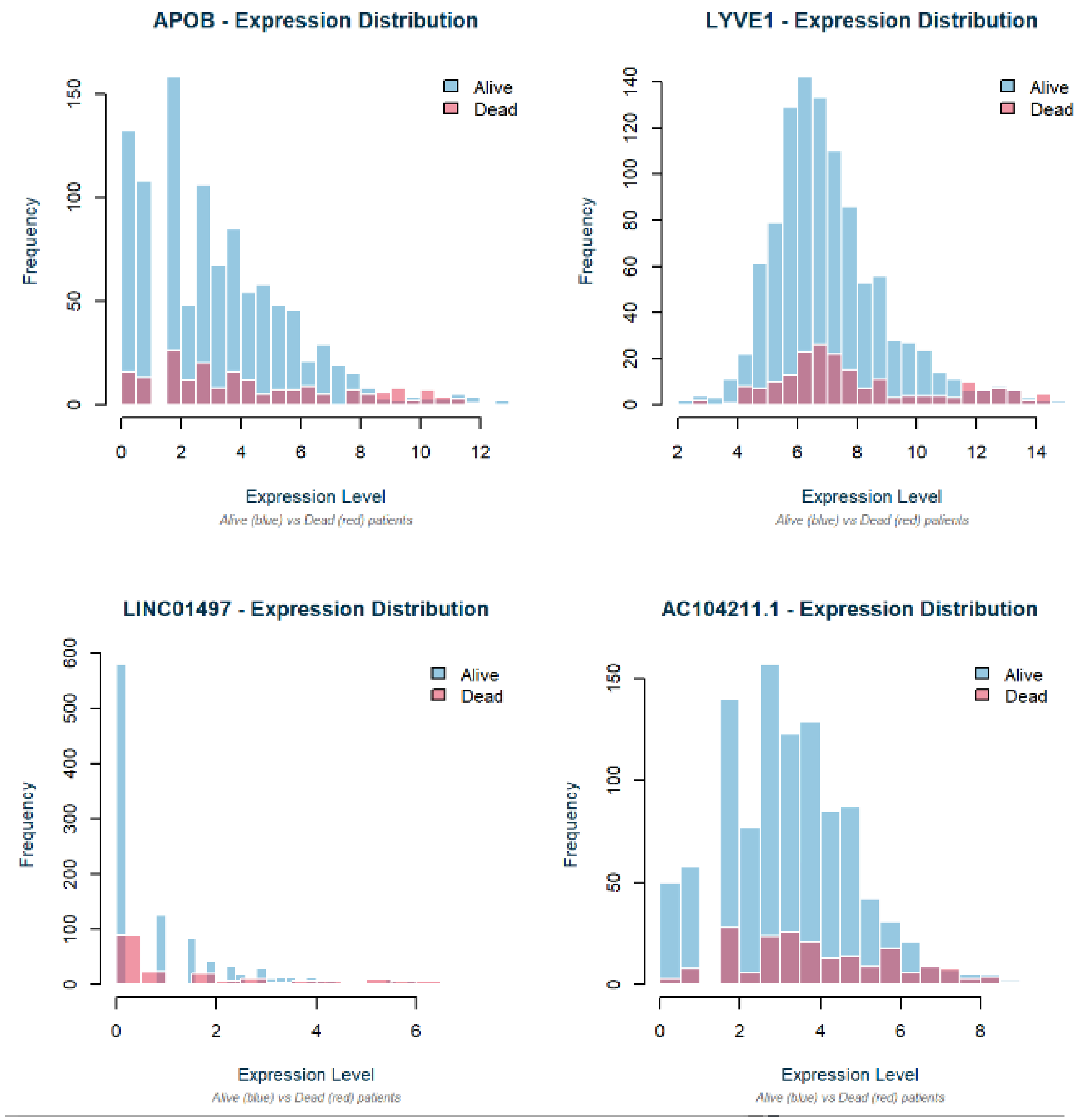
$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$



Differential Expression Analysis

4

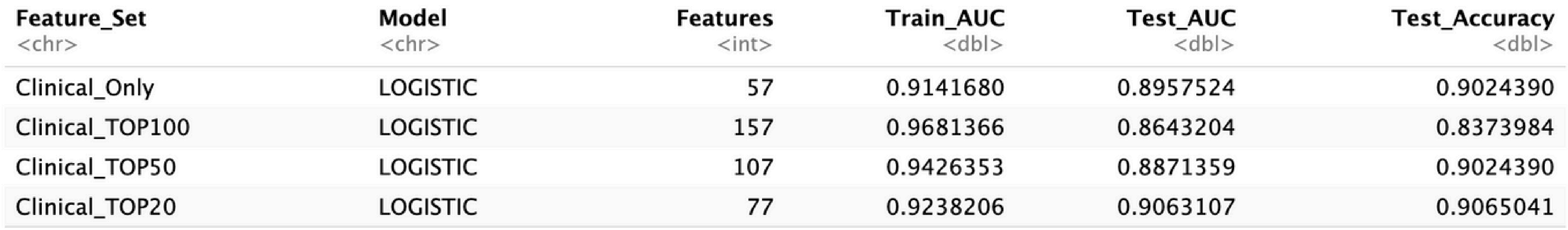




Predictive methodology 5

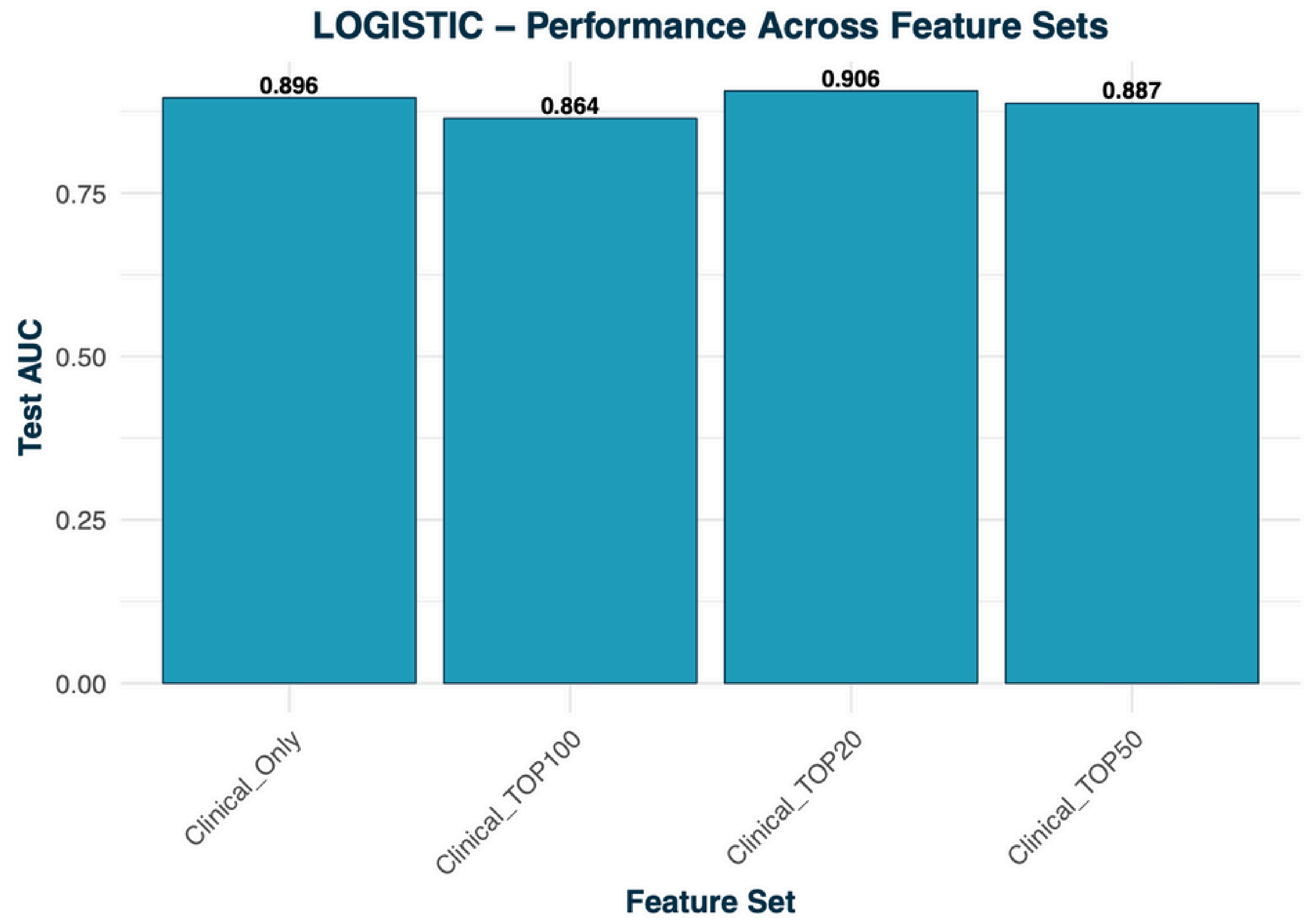
Test and Train split data

- **Train (70%) : 861 observations**
- **Test (30%) : 369 observations**

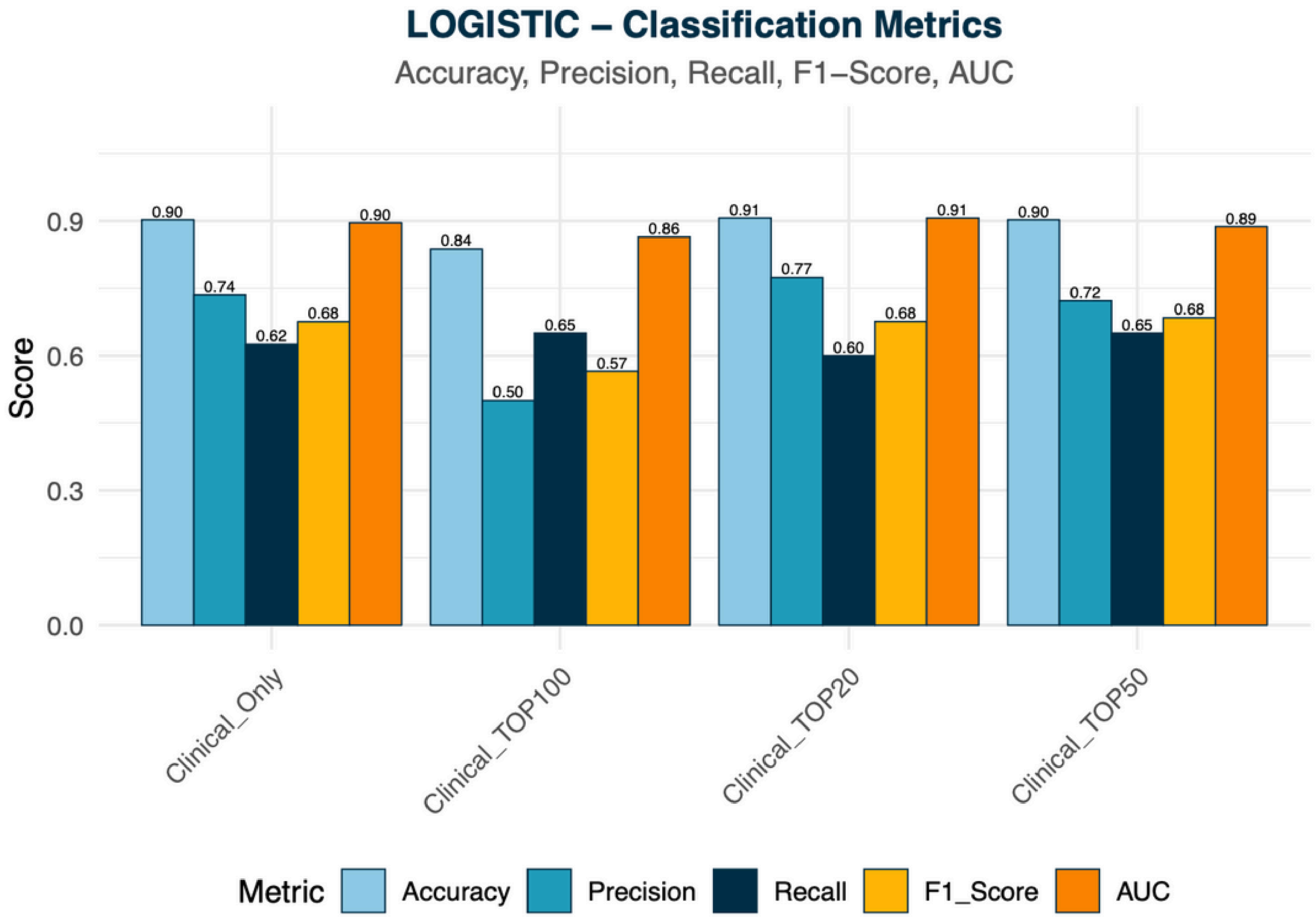


Feature_Set <chr>	Model <chr>	Features <int>	Train_AUC <dbl>	Test_AUC <dbl>	Test_Accuracy <dbl>
Clinical_Only	LOGISTIC	57	0.9141680	0.8957524	0.9024390
Clinical_TOP100	LOGISTIC	157	0.9681366	0.8643204	0.8373984
Clinical_TOP50	LOGISTIC	107	0.9426353	0.8871359	0.9024390
Clinical_TOP20	LOGISTIC	77	0.9238206	0.9063107	0.9065041

Logistic Regression



LOGISTIC – Performance Across Feature Sets



LOGISTIC – Classification Metrics

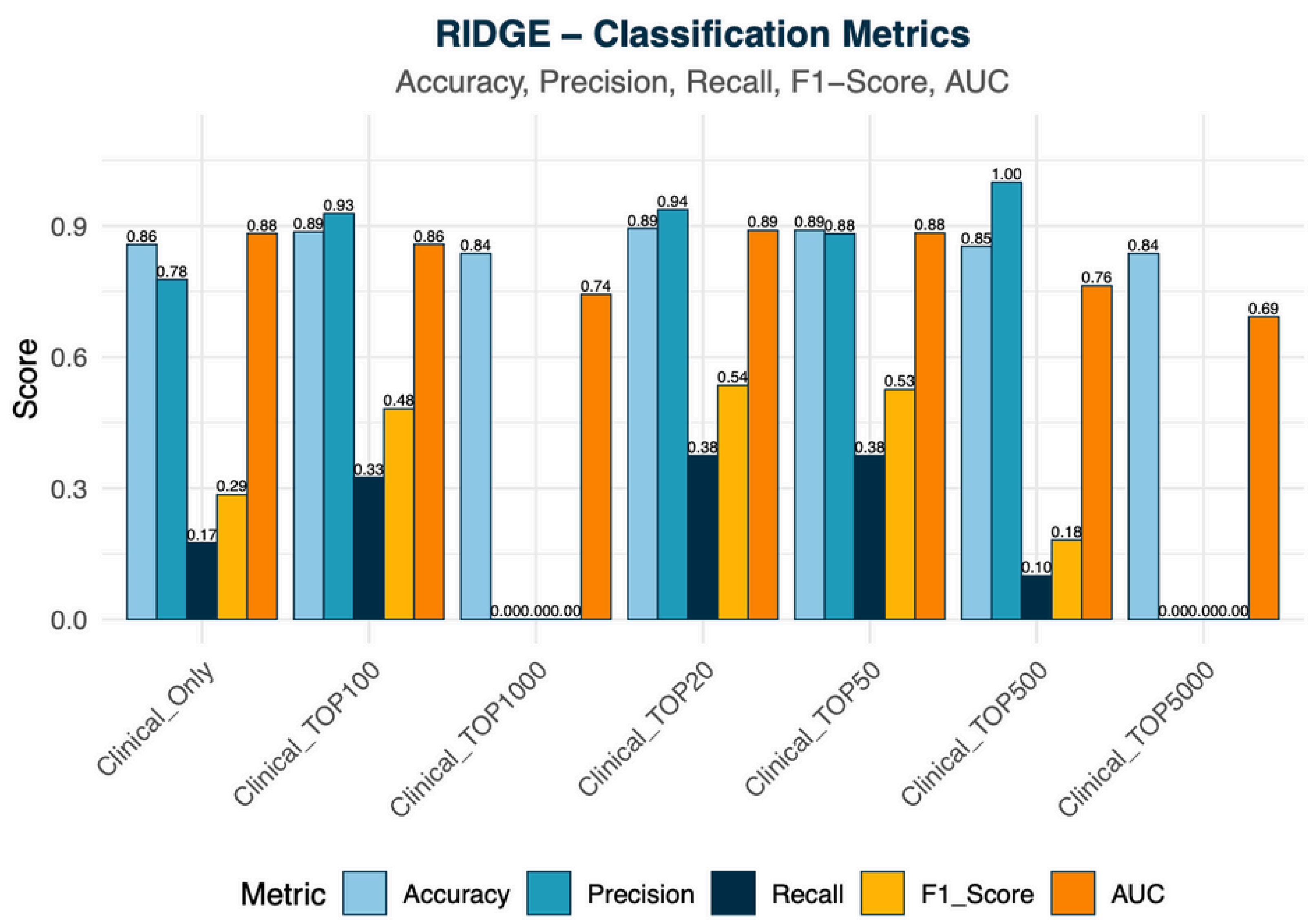
Accuracy, Precision, Recall, F1–Score, AUC

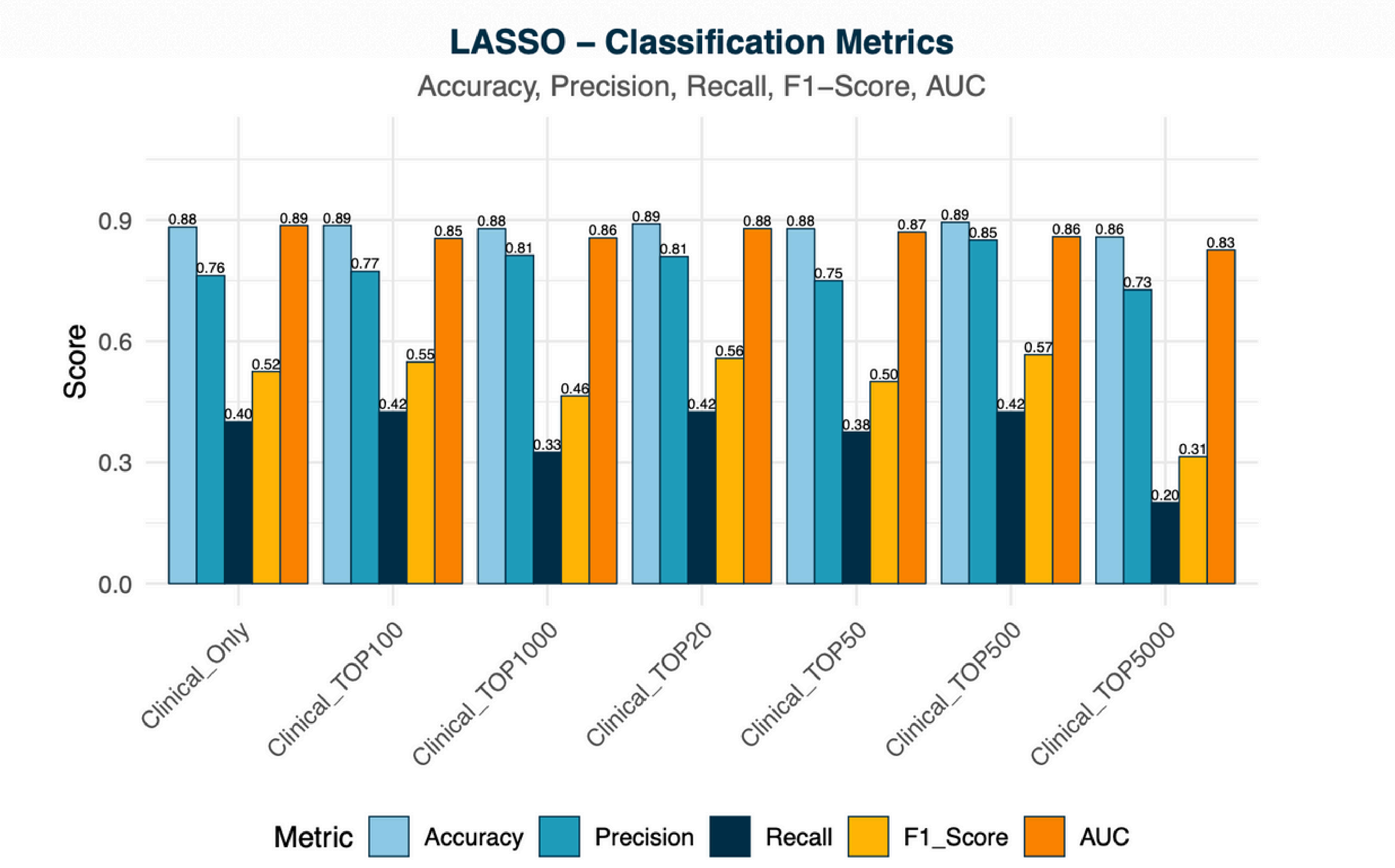
Ridge & Lasso Regression

6

$$\hat{\beta}^{\text{ridge}} = argmin_{\beta}\{-l(\beta) + \lambda\|\beta\|_2^2\}$$

$$\hat{\beta}^{\text{lasso}} = argmin_{\beta}\{-l(\beta) + \lambda\|\beta\|_1\}$$

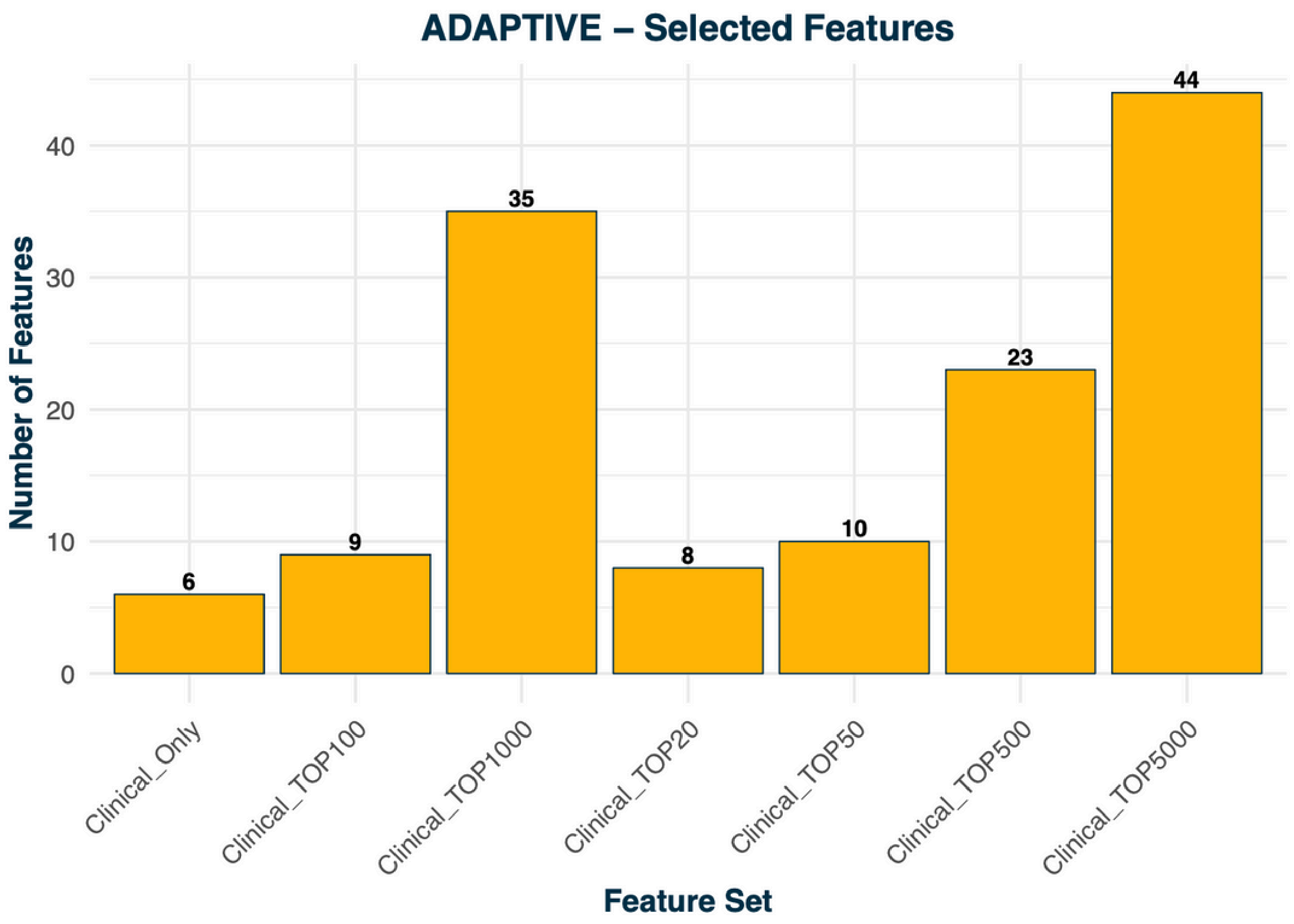


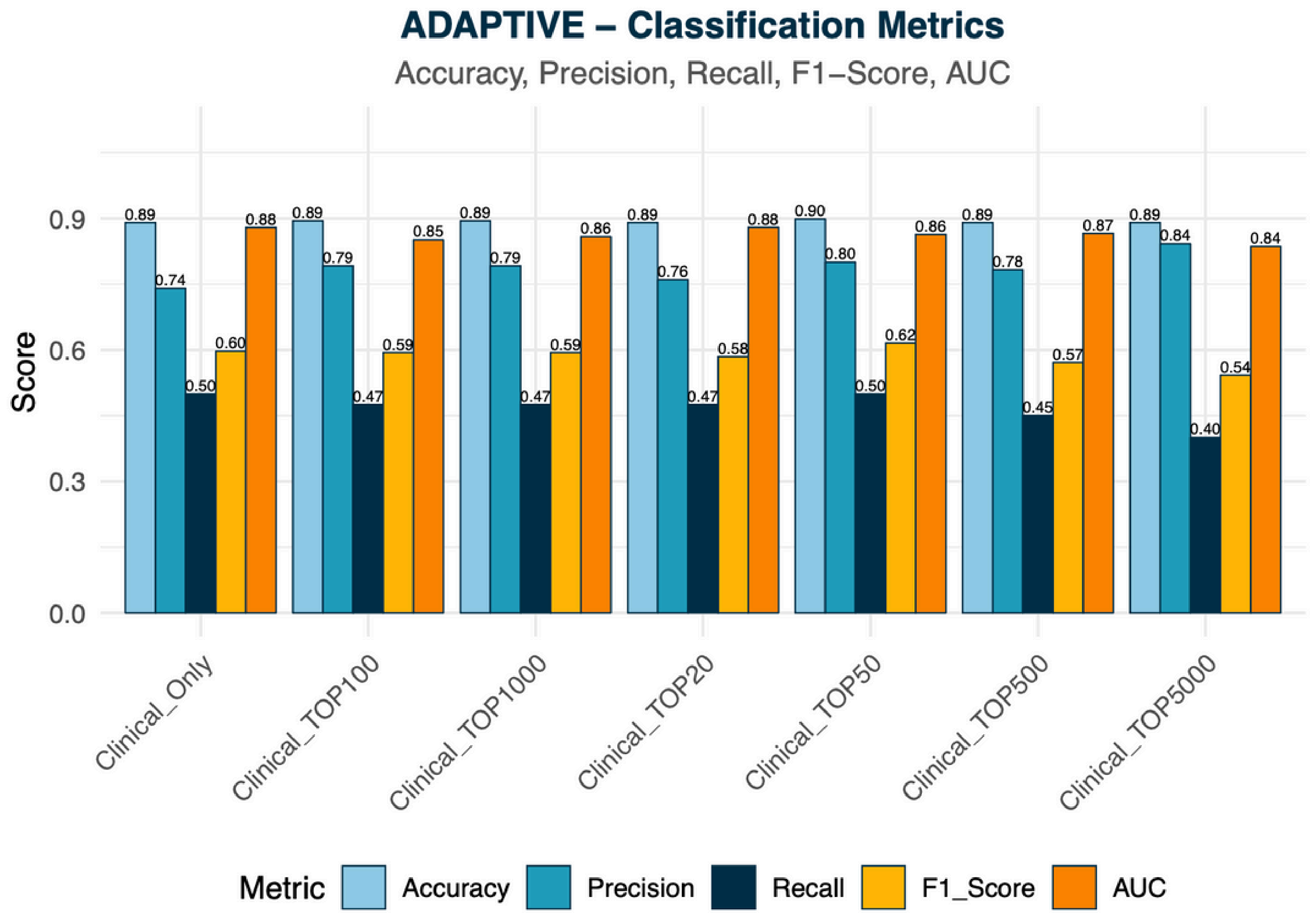


Adaptive Lasso Model

- **The Adaptive Lasso is introduced by Hui Zou (2006), "The Adaptive Lasso and Its Oracle Properties", one of the papers provided. This method modifies the standard Lasso by applying individual penalty weights to each coefficient, allowing the model to penalize weak predictors more strongly while preserving important ones.**

$$\hat{\beta}^{AL} = argmin_{\beta}\left\{-l(\beta) + \lambda \sum_{j=1}^{p} w_j|\beta_j|\right\} \qquad w_j = \frac{1}{|\hat{\beta}^{initial}|^{\gamma}}, \quad \gamma > 0$$



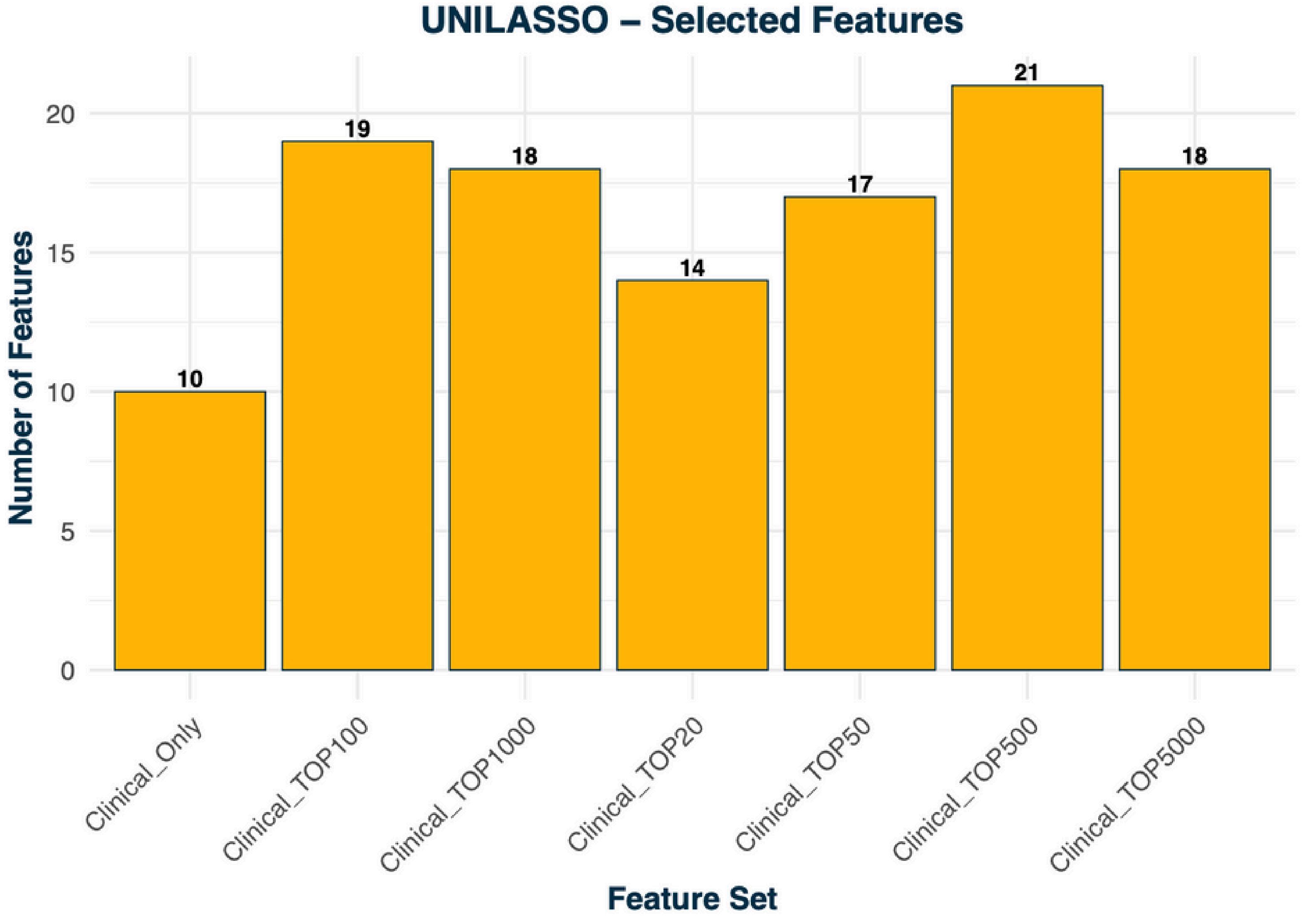


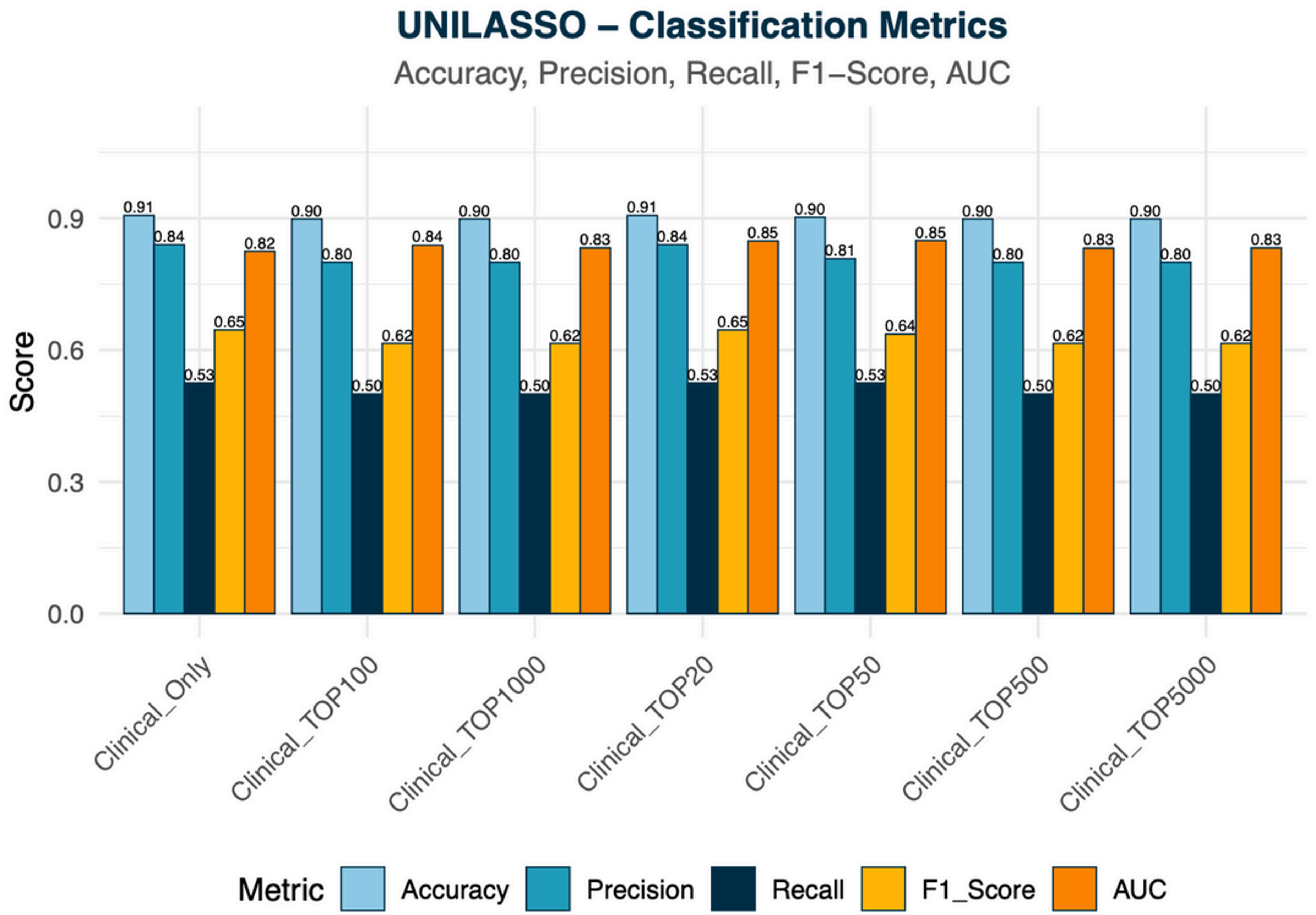
uniLasso Model

8

- The uniLasso method is introduced by Chatterjee, Hastie & Tibshirani (2025) as a two-step sparse regression procedure designed for high-dimensional genomic data. The key idea is to guide multivariate Lasso using univariate signal, improving stability and reducing the chance of selecting false genes.

$$\hat{\theta} = argmin_{\theta\geq 0}\left\{-l(\theta)+\lambda\sum_{j=1}^{p}\theta_j\right\} \qquad \tilde{\gamma}_j = \hat{\beta}_j^{univ}\hat{\theta}_j$$





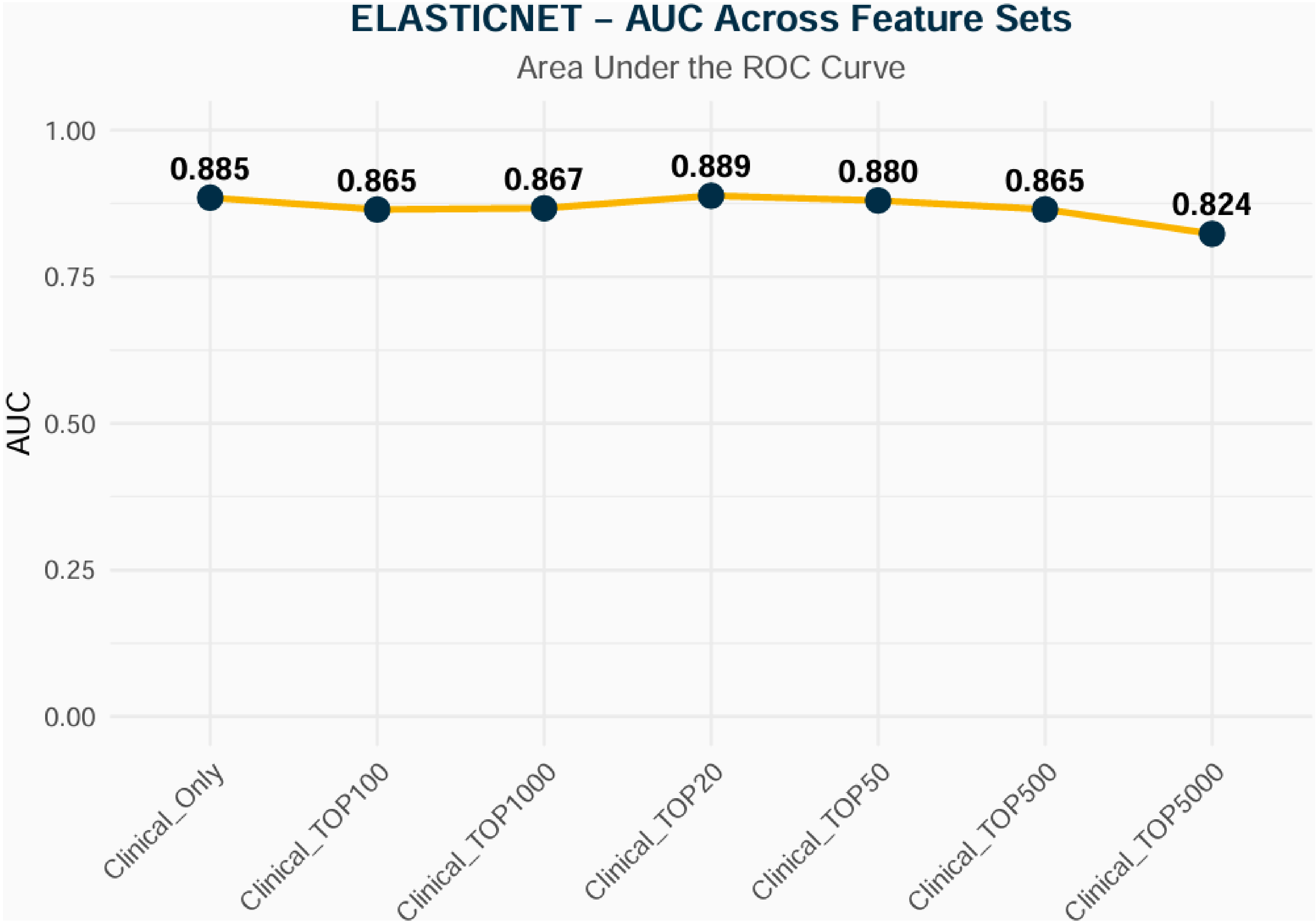
ElasticNet Model

- **Elastic Net combines the strengths of both Ridge (L2) and Lasso (L1) penalties.**

$$\hat{\beta} = argmin_{\beta}\{-l(\beta) + \lambda(\alpha\|\beta\|_1) + (1-\alpha)\|\beta\|_2^2\}$$

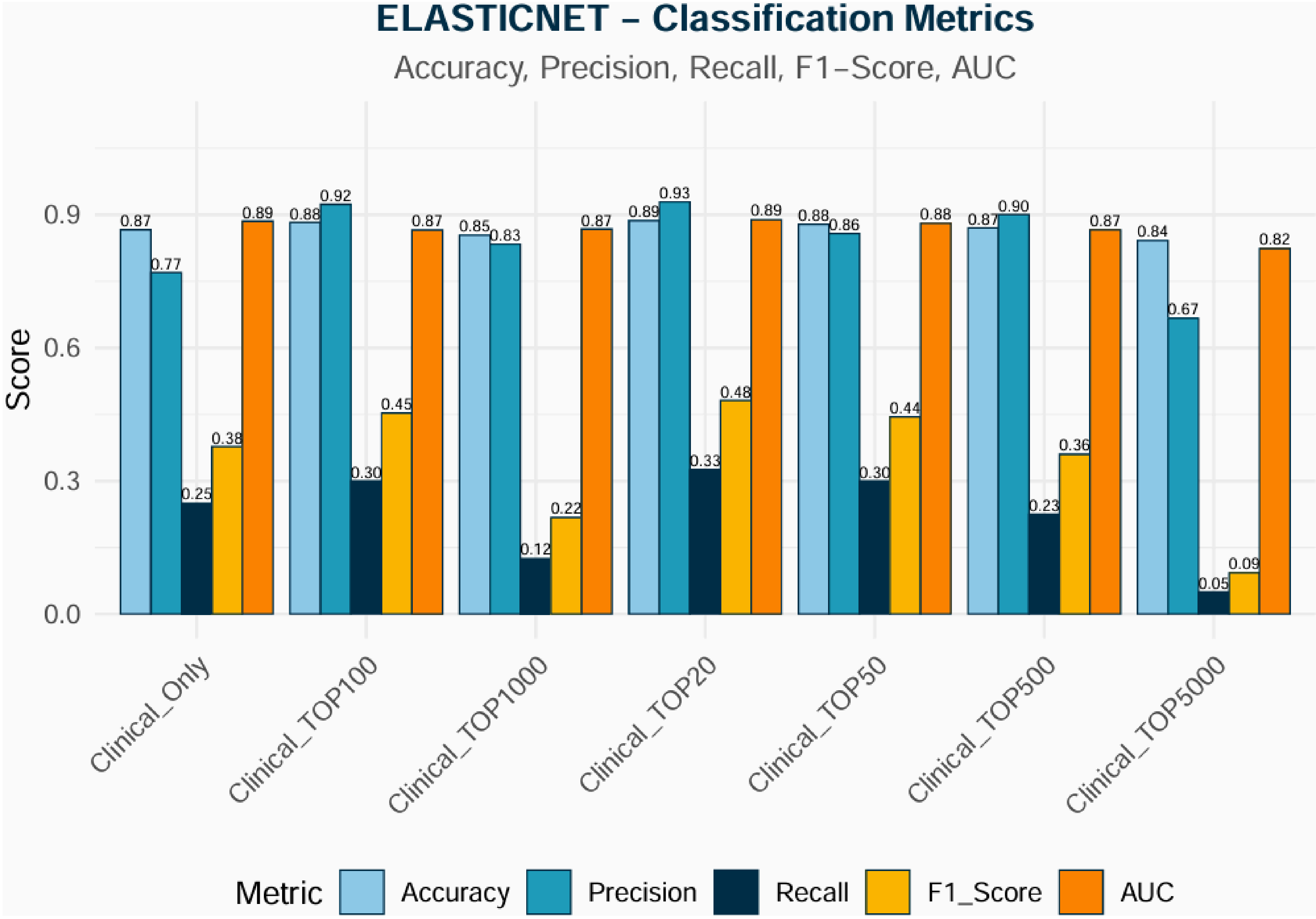
Where

- $\alpha = 1$ is Lasso
- $\alpha = 0$ is Ridge
- $0 < \alpha < 1$ is Elastic Mixed model



ELASTICNET – AUC Across Feature Sets

Area Under the ROC Curve

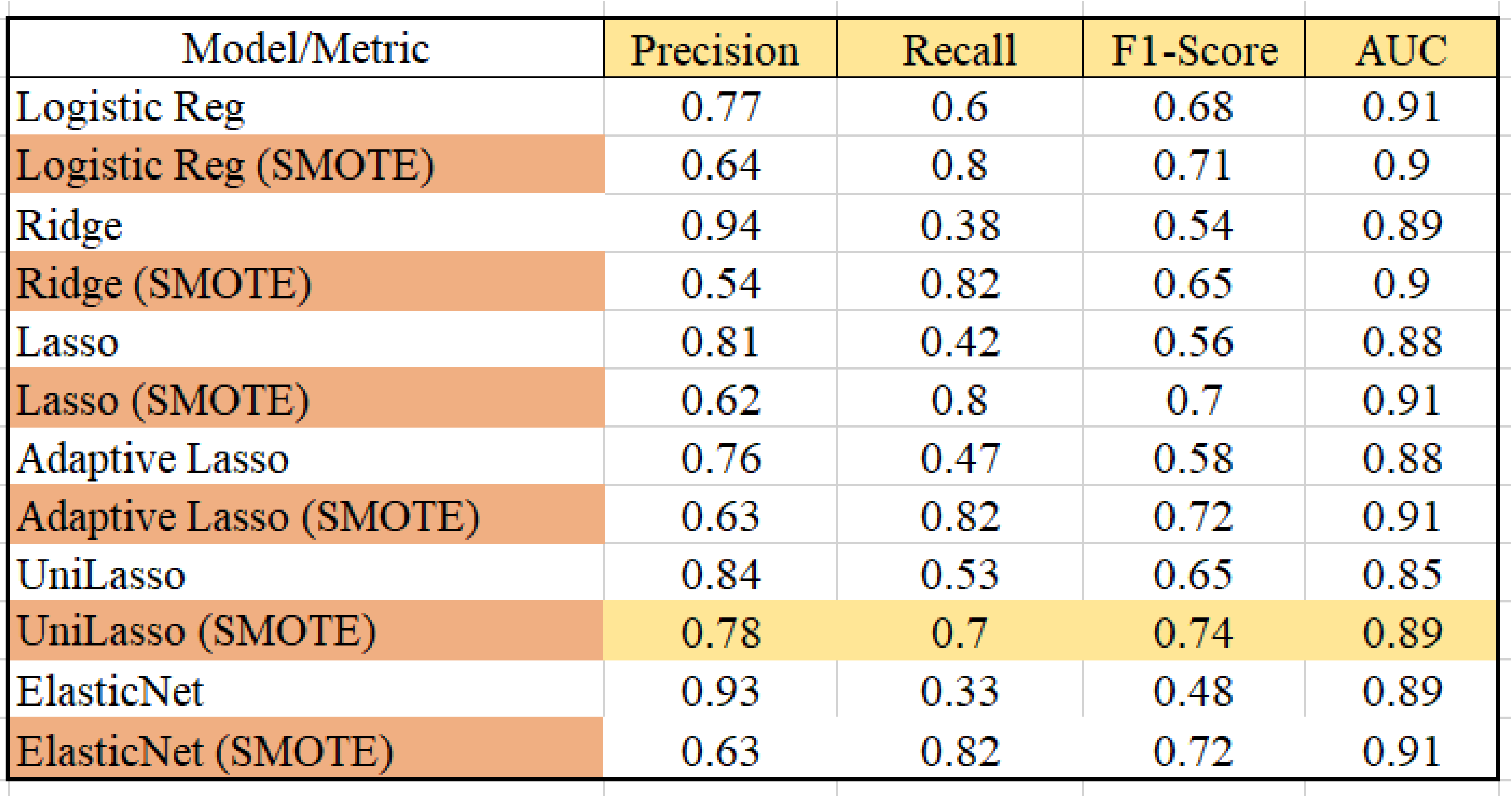


ELASTICNET – Classification Metrics

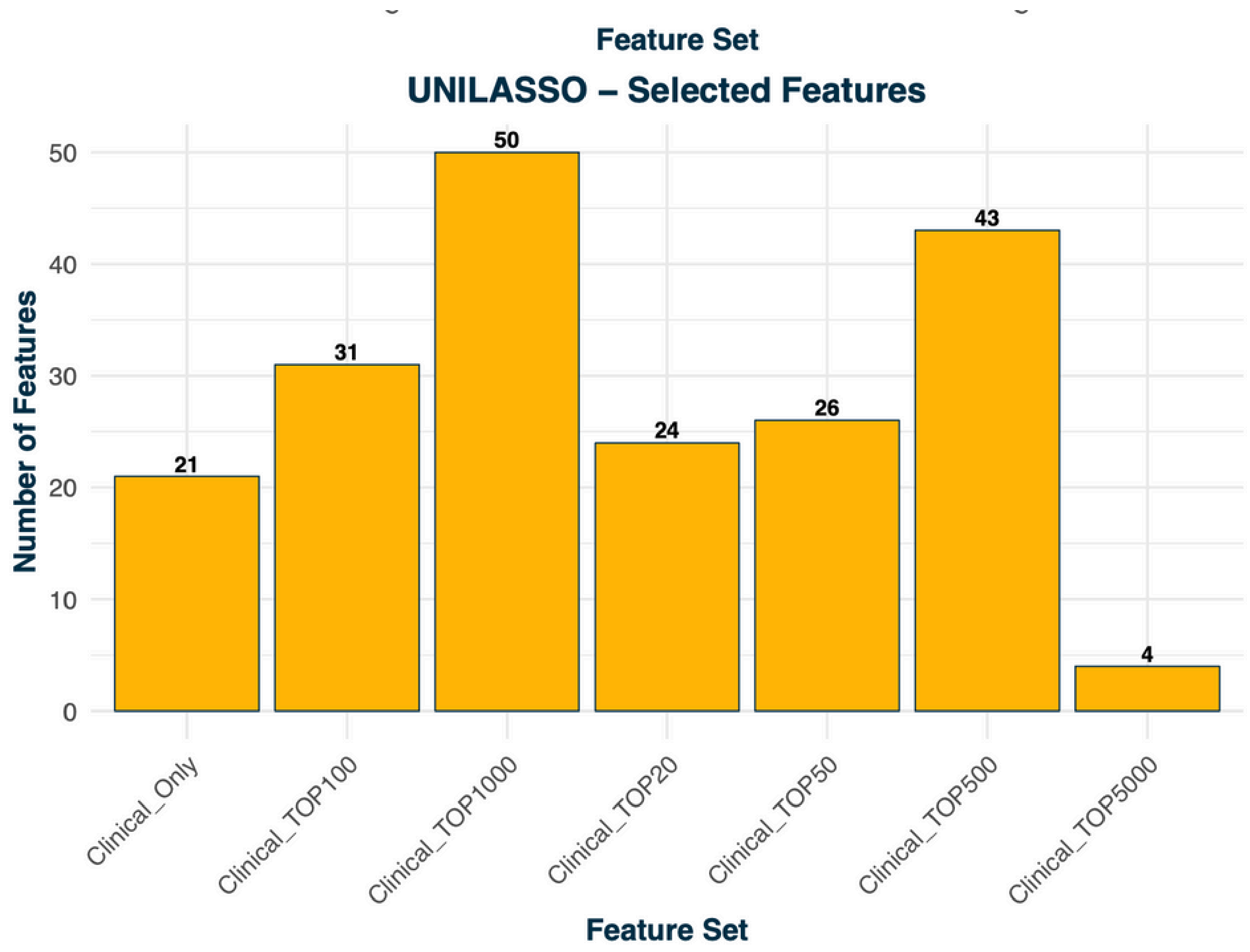
Accuracy, Precision, Recall, F1–Score, AUC

Class Imbalance Handling with SMOTE 10

- **Current imbalance ratio: 5.12:1 (Alive:Dead)**
- **Why SMOTE: - Creates synthetic minority class samples (Dead patients) - Balances training data to ~1:1 ratio - Forces models to learn Dead patient patterns - No data loss (vs downsampling) - Prevents overfitting (vs simple upsampling)**

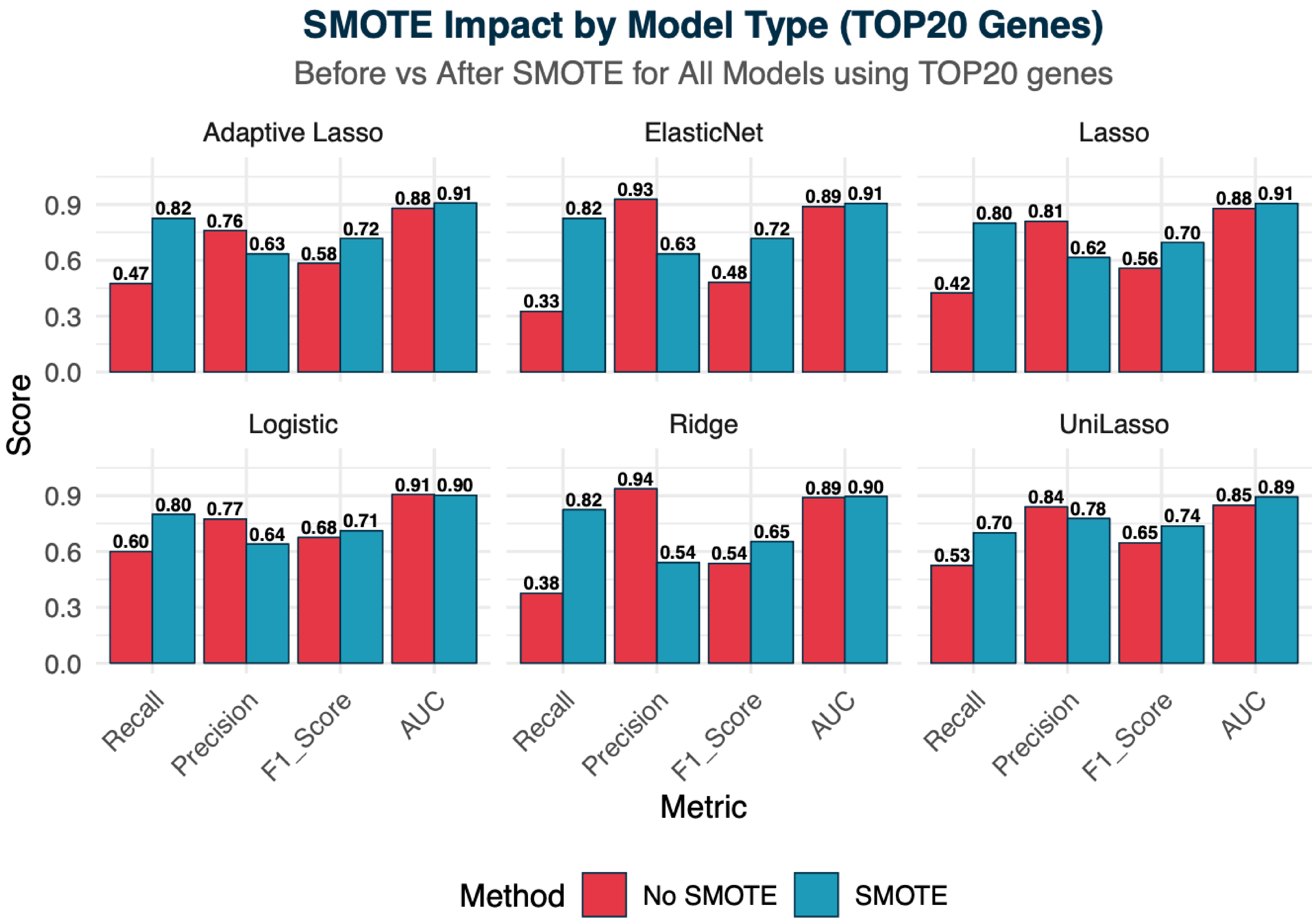


Model/Metric	Precision	Recall	F1-Score	AUC
Logistic Reg	0.77	0.6	0.68	0.91
Logistic Reg (SMOTE)	0.64	0.8	0.71	0.9
Ridge	0.94	0.38	0.54	0.89
Ridge (SMOTE)	0.54	0.82	0.65	0.9
Lasso	0.81	0.42	0.56	0.88
Lasso (SMOTE)	0.62	0.8	0.7	0.91
Adaptive Lasso	0.76	0.47	0.58	0.88
Adaptive Lasso (SMOTE)	0.63	0.82	0.72	0.91
UniLasso	0.84	0.53	0.65	0.85
UniLasso (SMOTE)	0.78	0.7	0.74	0.89
ElasticNet	0.93	0.33	0.48	0.89
ElasticNet (SMOTE)	0.63	0.82	0.72	0.91



Class Imbalance Handling with SMOTE

10



SMOTE Impact by Model Type (TOP20 Genes)

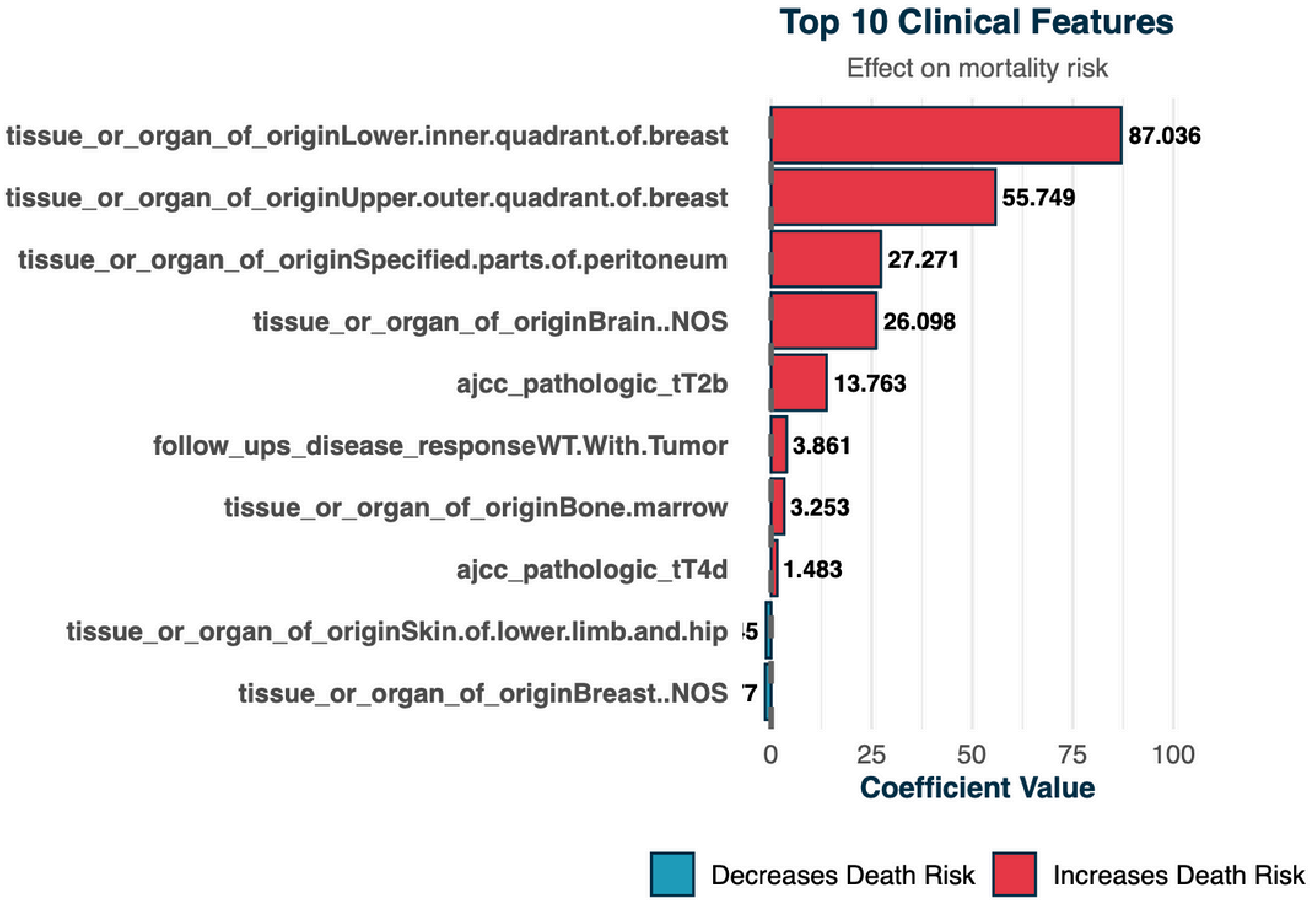
Before vs After SMOTE for All Models using TOP20 genes

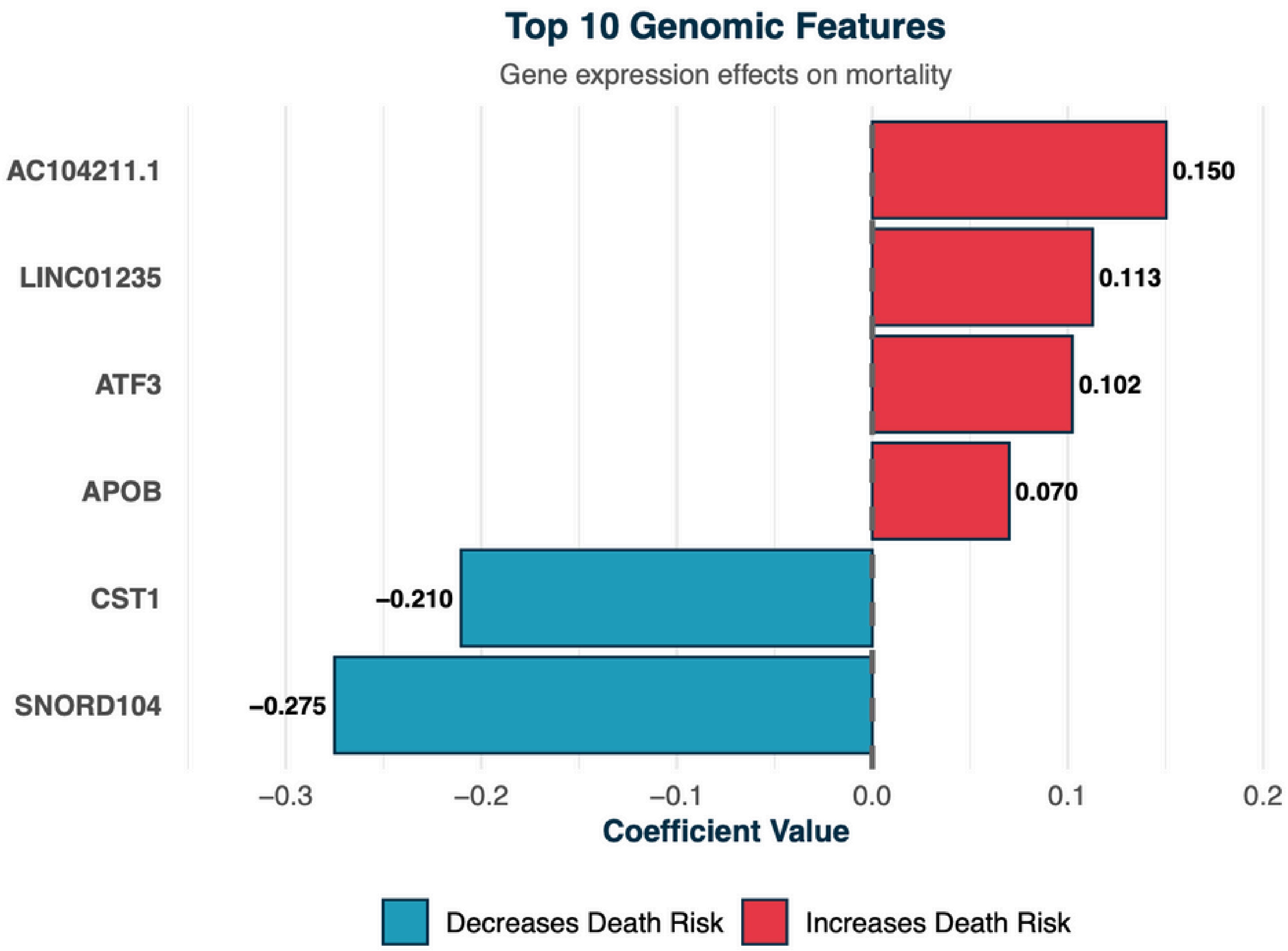
Results & Conclusions

11

Best Models

- UniLasso (SMOTE) with Clininal + Top20 GeneX data:
    - ROC-AUC: 0.89
    - Recall: 70% (catches dieds)
    - F1-Score: 74%





THANK YOU !