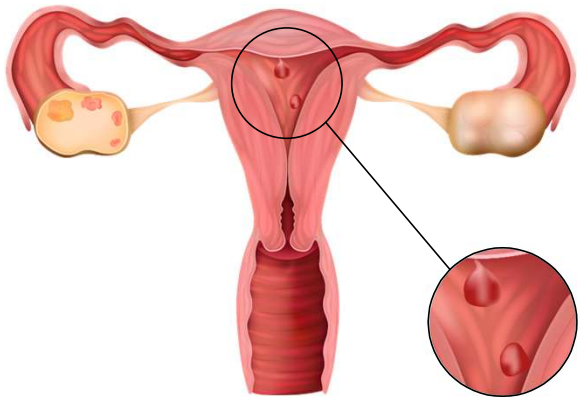
NUS National University of Singapore | Yong Loo Lin School of Medicine



A Predictive Model for Endometrial Cancer Progression Risk Using Clinical and RNA Seq Data

Group 11: Alsaidan, Almuayyad (A0336193B); Chen Tao (A0318683U); Liu Yilin (A0318483X); Qi Jingyi (A0319116J); Sheng Jie (A0054212X); Wang Xianzhe (A0318410R); Zhu Taiqi (A0318771X);

Date of Presentation: 15 Nov 2025

Table of Contents

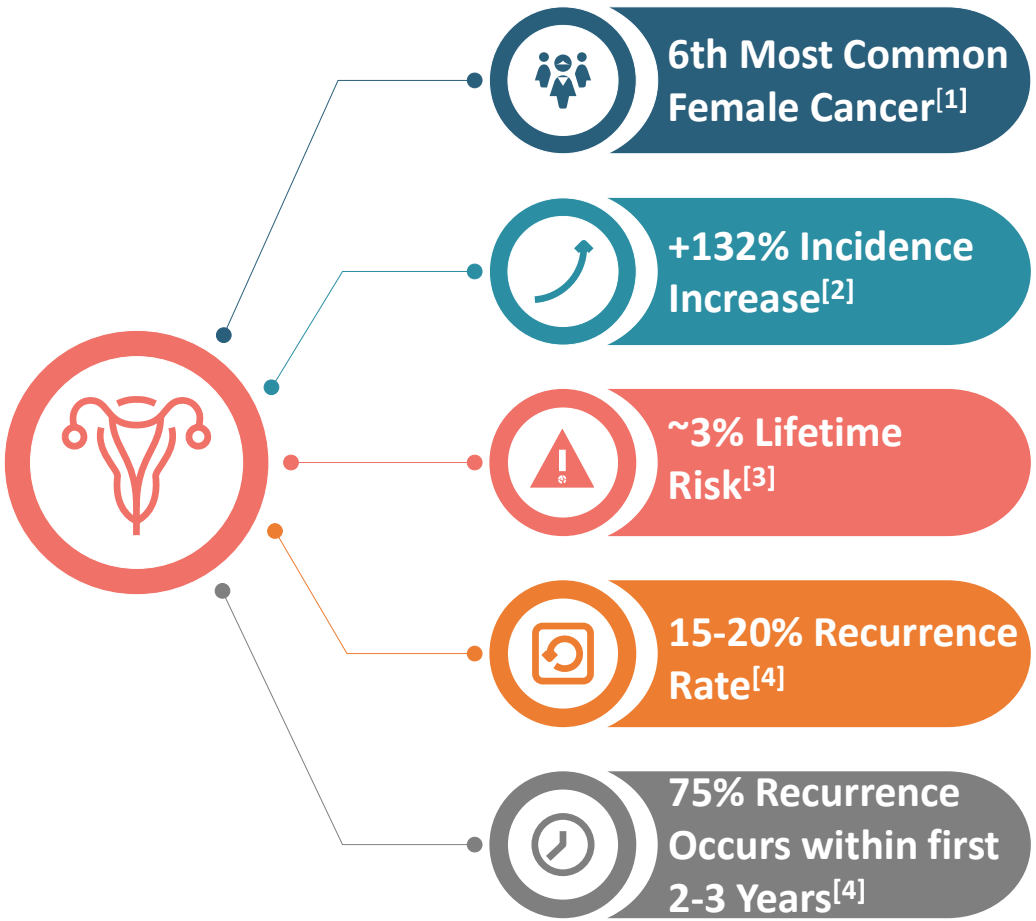
Sheng Jie

Zhu Taiqi (Judy)

Qi Jingyi

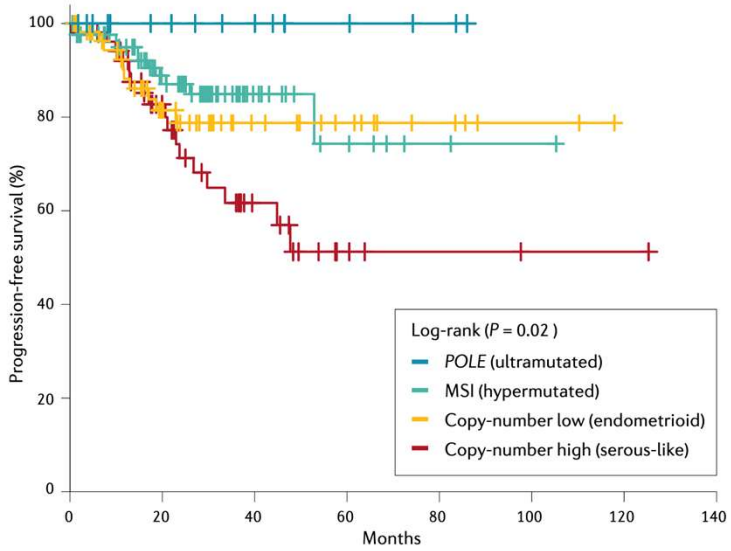
Chen Tao

Rising Burden, Unpredictable Outcome



Subtype ≠ Individual Risk

Progression-Free Survival by Molecular Subtype



Adopted figure: Makker et al., Nat Rev Dis Primers 2021[5]

Clinical Gap & Machine Learning Motivation

Current State

- **TCGA 2013:** 4 molecular subtypes with distinct prognosis
- **ESGO/ESTRO/ESP 2021:** Molecular-integrated risk groups (Low/ Intermediate/ High-Intermediate/ High/ Advanced metastatic)[6]
- **FIGO 2023:** Modified staging for POLEmut/p53abn in early disease[7]

What's Still Missing?

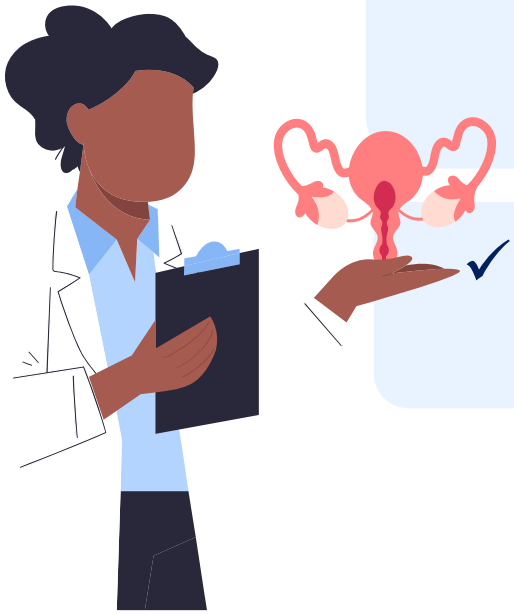
Categorical grouping vs.

individualized risk prediction

Machine Learning Opportunity

- Handle high-dimensional data
- Generate individualized risk prediction

Objectives



- ✓ Develop and validate a ML model to predict the 2-year risk of disease progression using clinical and RNA-seq data
- ✓ Identify key prognostic biomarkers (clinical and genetic) associated with patient outcomes

Dataset Overview (TCGA-UCEC[8])

Data Source: The Cancer Genome Atlas (TCGA) UCEC cohort. (N=362 patients)

- ✓ The world's largest cancer genome database
- ✓ Multi-center, standardized collection

Features:

Clinical (19) [Appendix A]

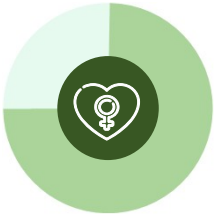
Eg: Age, FIGO stage, tumor grade, etc

RNA-Seq (879) [Appendix B]

gene expression from 7 Endometrial Cancer related signaling pathways (e.g., p53, mTOR, PI3K-Akt, etc)

Endpoint: 2-year Progression-Free Intervel (PFI) [Appendix C]

- High-risk (Label=1, 23%): PFI event ≤ 730 days
- Low-risk (Label=0, 77%): No event or event ＞ 730 days



Analysis Workflow

1. Data Preparation

- *TCGA-UCEC Raw Data → analysis-ready dataset*
- *a robust train-test split => prevent data leakage*

2. Nested Cross-Validation for Model Training

- *ML Pipeline: Feature Scaling (StandardScaler) → Regularized Regression (ElasticNet)*
- *Hyperparameter tuning w/in outer loop => unbiased estimate of model's generalization*

3. Feature Importance & Interpretation

- *Identifying key prognostic biomarkers by analyzing feature stability and importance*

4. Ablation Study for Feature Validation

- *Quantify the contribution of different feature subsets and their synergistic value*

Data Preparation

1) **Data Integration:** Combined clinical and genetic features
2) **Stratified Splitting (80/20)**
   - Training: N=289; Test: N=73
   - Preserved label distribution: 23.5% high risk
3) **Data Leakage Prevention:**
   - Scaler was fitted only on the training data

   → used to transform the test data
   - Ensure unbiased evaluation

Machine Learning Pipeline

StandardScaler

↓

SelectKBest

Scoring function: **f_classif (ANOVA F test)**

Select top K features according to ANOVA F value

↓

LogisticRegression

Penalty = ElasticNet

Solver = saga

Class_weight = balanced

Max_iteration = 5000

```
param_grid = {
    'selector__k': [50, 100, 200],
    'classifier__l1_ratio': [0.2 , 0.3, 0.5, 0.7],
    'classifier__C': [0.01, 0.1, 1]
}
```

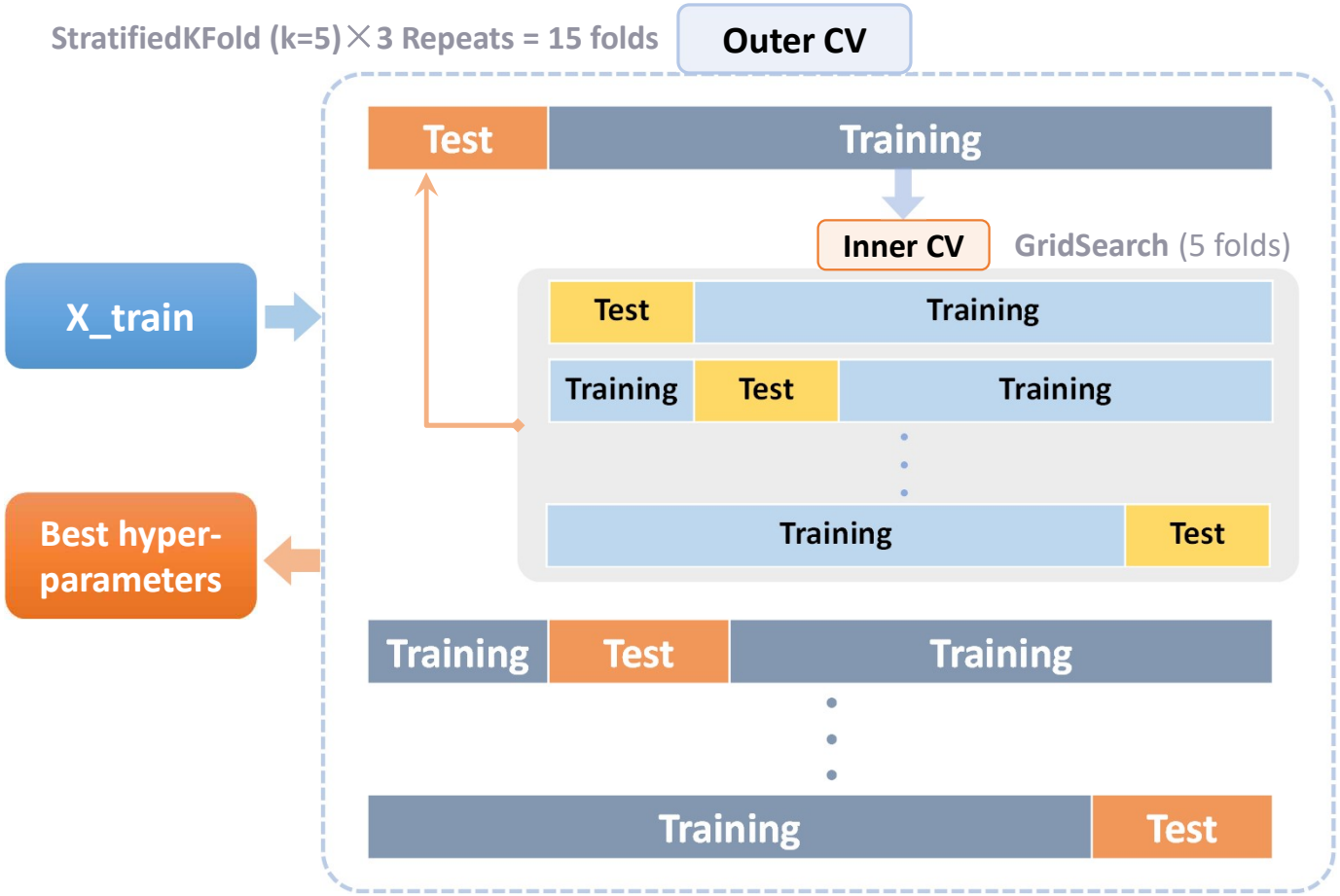
3 × 4 × 3 = 36 combinations

↓

GridSearchCV

Scoring = 'roc_auc'

Nested Cross Validation



Nested Cross Validation

- **Inner CV – Model Tuning**

  Find the best_params for each outer training set

- **Outer CV – Model Evaluation**

  Using the best hyperparameter combination to train the outer test set and get the metrics

- **Results**

  15 models generated from 15 folds

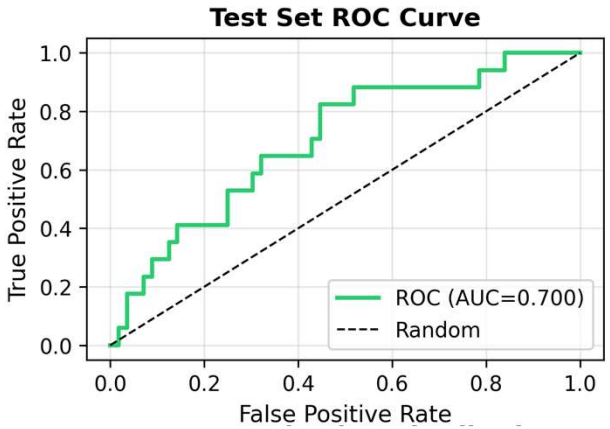
- **Strength**

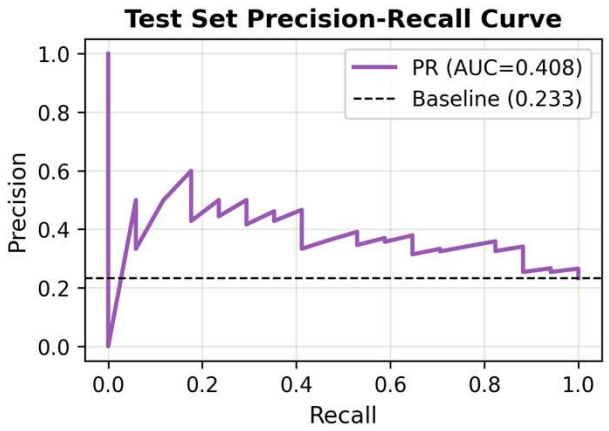
  Unbiased estimate of models' generalization performance

Model Performance on Test Set

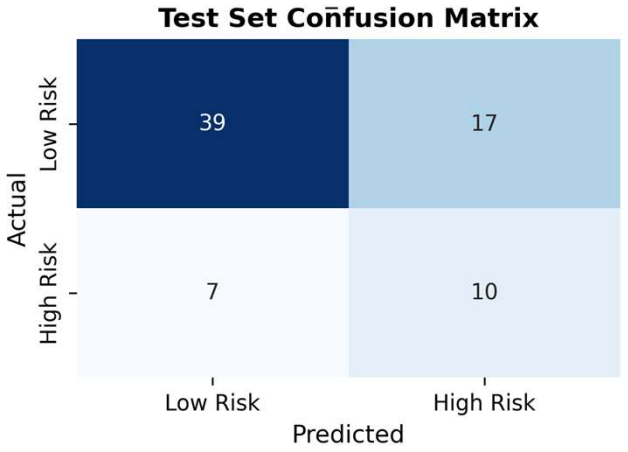
Hyperparameters selected for the final model:
(based on frequency)

- **K = 50** (8/15)
- **L1_ratio = 0.2** (7/15)
- **C = 0.1** (9/15)

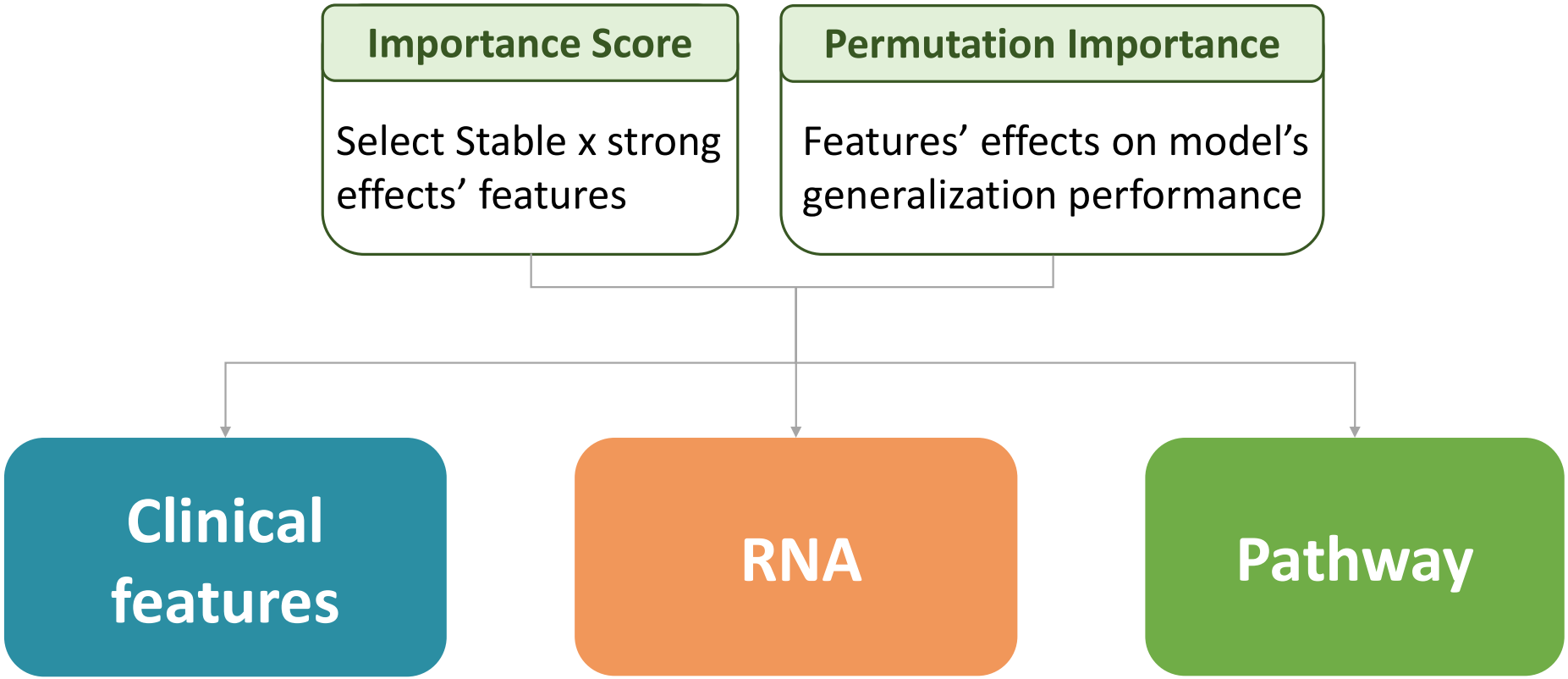




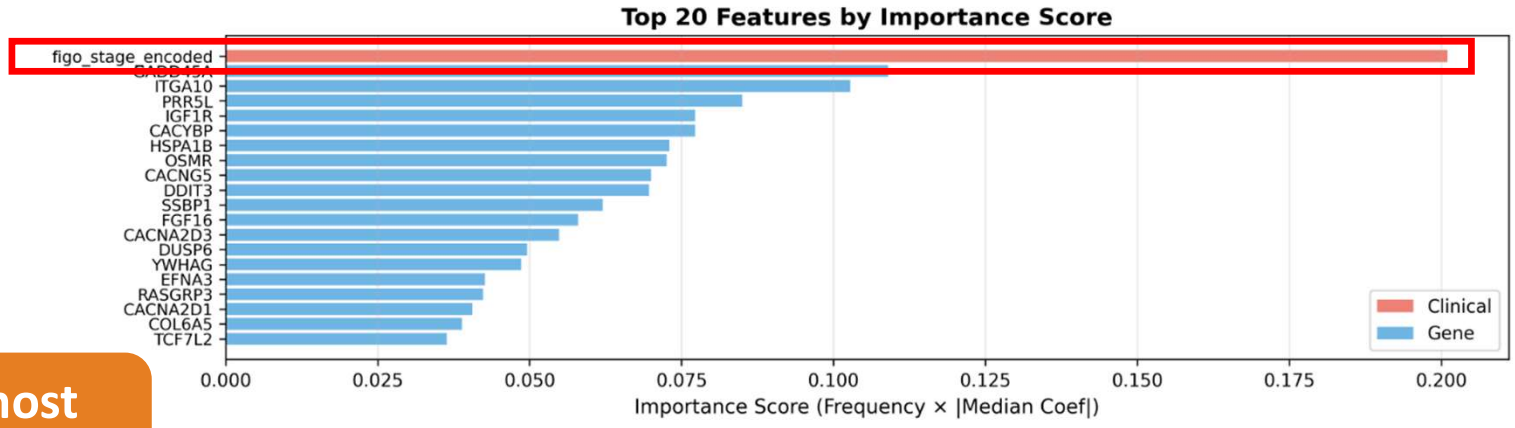
Metrics	Value	
AUROC	0.6996	Reasonably good discrimination ability
AUPRC	0.4083	Acceptable given the class imbalance
Recall	58.82%	Identify ~ 60% of high-risk patients
Precision	37.04%	
F1-Score	0.4545	
Specificity	69.64%	Identify ~ 70% of low-risk patients
Accuracy	67.12%	



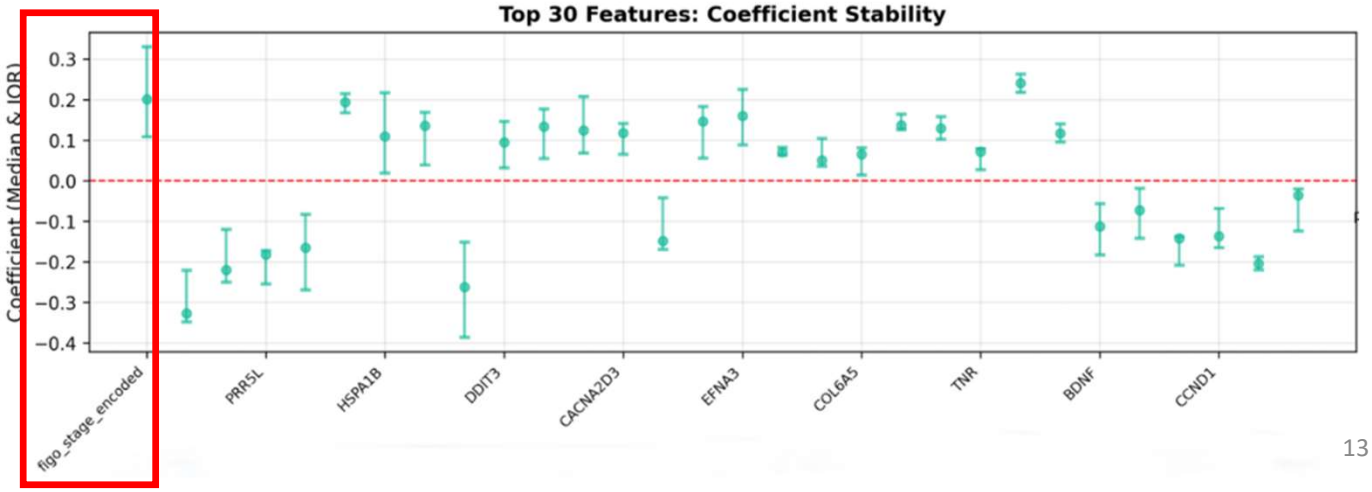
Feature Importance



Feature Importance – Clinical

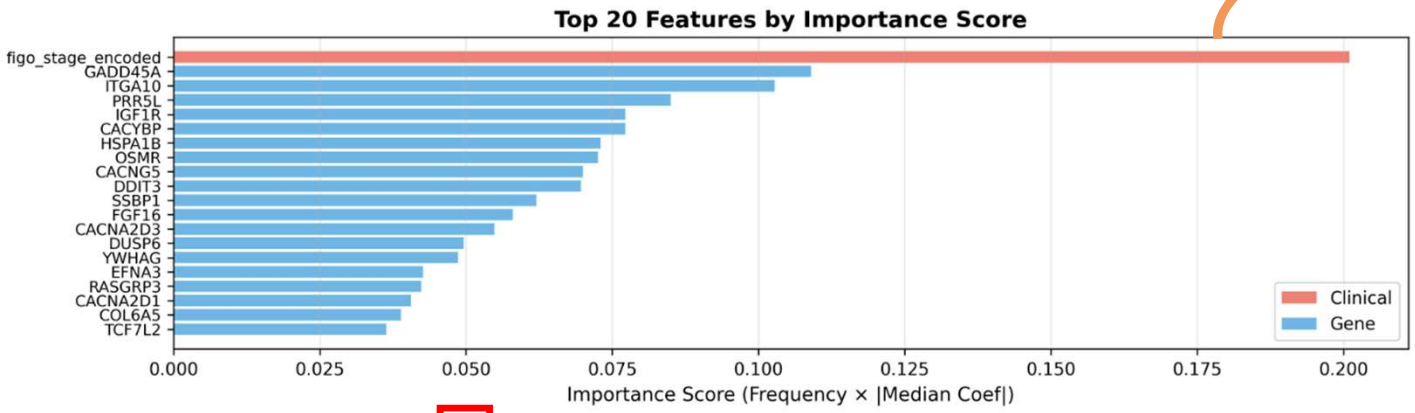


Figo_stage is the most important feature



Figo_stage has a positive correlation with high risk[9]

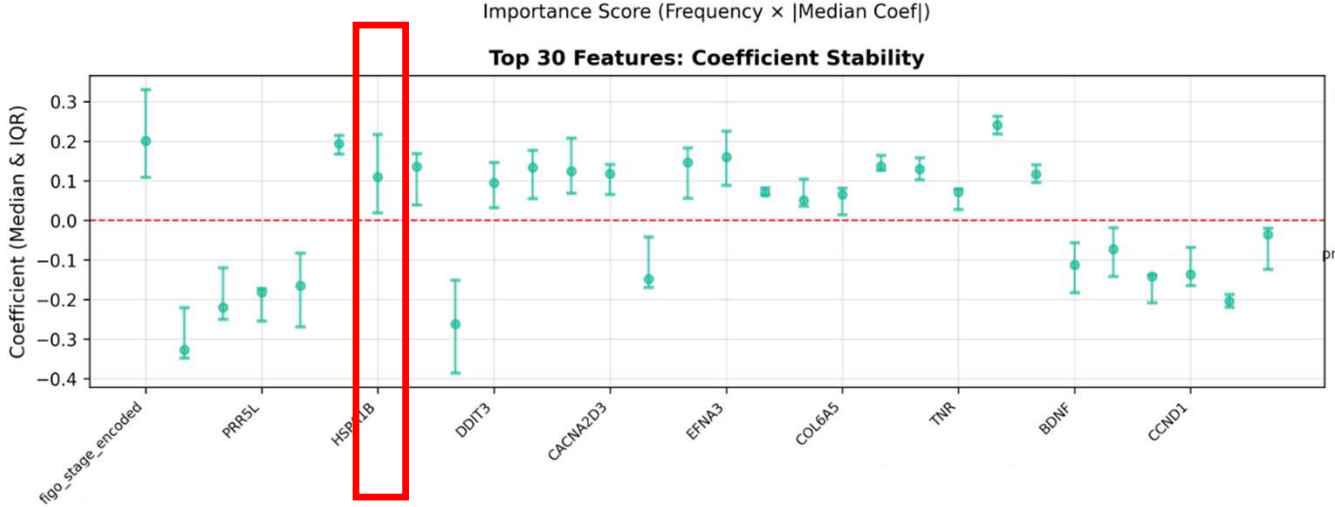
Feature Importance – RNA

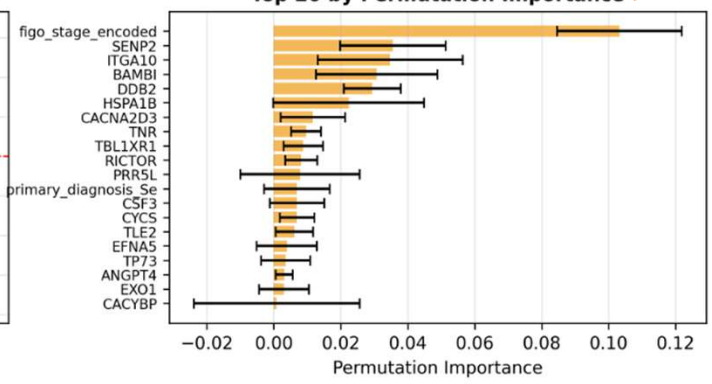


RNA important in both methods

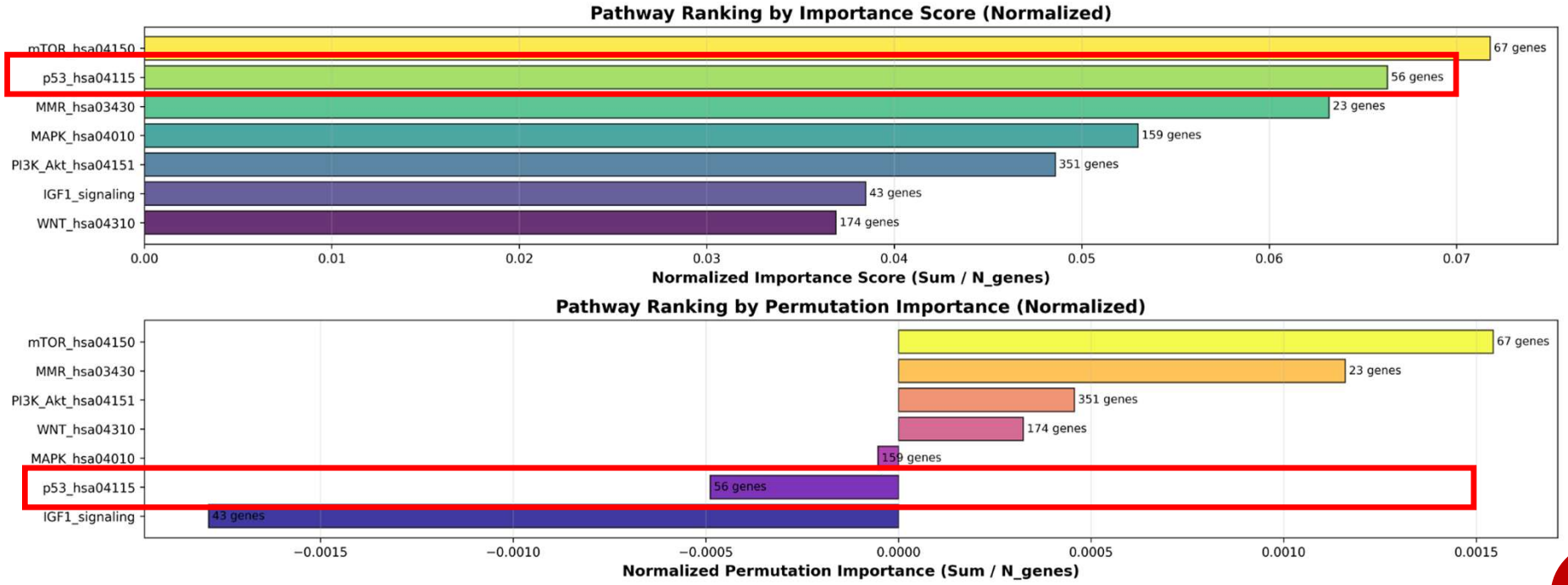
PRR5L, ITGA10, HSPA1B, CACNA2D3, CACYBP

Potential biomarker





Feature Importance – Pathway



mTOR & MMR pathways are the most important

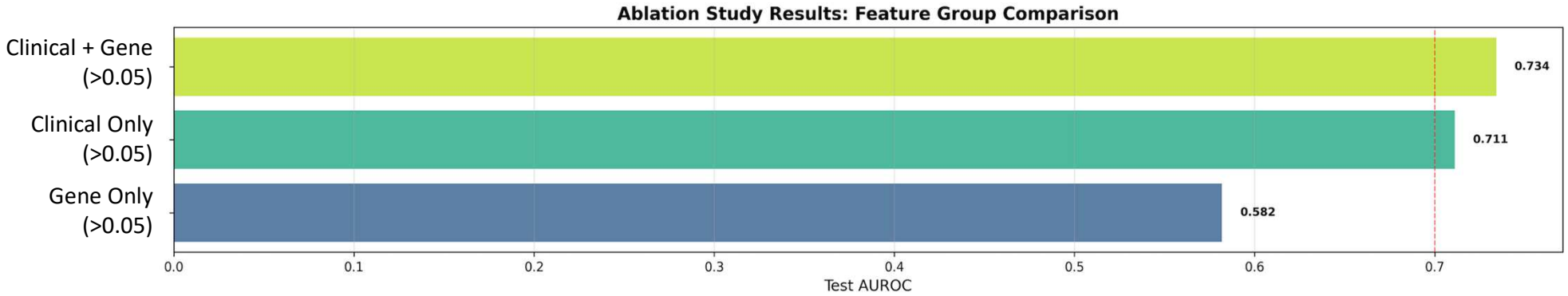
IGF1 pathway: no contribution

IGF1R affects tumor regression by activate PI3K-AKT pathway[10].

p53 pathway: opposite effect in two methods

p53 pathway highly related to tumor stage and grade[11][12] → redundancy

Ablation



Feature Group Evaluation

- 13 Features with importance score greater than 0.05
- Split these features into 3 groups: clinical only, gene only, clinical + gene
- 5-fold CV, followed by test dataset validation

Key Findings:

✓ Clinical feature alone showed high prediction power, with AUC 0.711

✓ The inclusion of RNA reading further strengthened our model's AUC to 0.734

Conclusion

Primary Achievements:

- ✓ Developed a robust model for 2-year risk prediction (AUROC＞0.70)
- ✓ **Interpretability**: Identified FIGO stage as the dominant clinical predictor and highlighted key prognostic genes in the mTOR, p53, and MMR Pathways.
- ✓ **Methodological Rigor**: Applied nested cross-validation to ensure unbiased performance estimation and fully prevent data leakage.
- ✓ **Feature Synergy**: Ablation analyses showed that combining clinical + transcriptomics features yields the strongest predictive performance.

Limitations

1) **Single-Cohort Study**: Requires external validation on independent datasets.
2) **Moderate Sample Size (N = 362):** May limit generalizability **=> Cross comparison**
3) **Model Scope**: Future work should evaluate additional ML methods and deep learning architectures to capture more complex biological patterns.

Cross Comparison

Machine Learning Task	Best ML Model	Type of Features	Number of Samples	Best Model Performance	Year	Ref.
Recurrence Prediction	Random Forest	**FIGO stage,** DNA (molecular subtypes), and RNA (immune markers)	230 (TCGA UCEC, filtered)	Accuracy 0.537 (w/o RNA) Accuracy 0.686 (w RNA) AUC not reported	2023	[13]
Overall Survival Prediction	Cox Regression	Clinical (Age, **FIGO stage,** Tumor grade, BMI) and RNA (PCD & DEG)	507 (TCGA UCEC, filtered)	AUC 0.708 (1 years) AUC 0.702 (3 years) AUC 0.737 (5 years)	2025	[14]
Our Research	Logistic Regression	**FIGO stage** and RNA (6 pathways)	362 (TCGA UCEC, filtered)	AUC 0.734 (2 years)		
Recurrence and Metastasis Prediction	Random Forest ★	lncRNA (Best performance compared to miRNA, CNV & mRNA)	238 (TCGA UCEC, filtered)	AUC 0.763 ★	2022	[15]

Next Step: Other Model

Metrics	Logistic Regression	Random Forest	XGBoost
Model Setting	Selected 13 features K = 50, L1_ratio = 0.2, C = 0.100	Not optimized yet	Not optimized yet
AUROC	0.734	0.652	0.508
AUPRC	0.402	0.344	0.333
Specificity	0.661	0.980	0.910
Accuracy	0.685	0.753	0.739

Logistic Regression provided superior overall predictive power (higher AUC), proving more robust on our imbalanced dataset than models with higher accuracy scores alone.

Acknowledgement

NUS National University of Singapore | Yong Loo Lin School of Medicine

Thank you for attention!

Wang Xianzhe

Alsaidan Almuayyad

Liu Yilin

Reference

[1] Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*. 2021;71(3):209-249. doi:10.3322/caac.21660

[2] Gu B, Shang X, Yan M, et al. Variations in incidence and mortality rates of endometrial cancer at the global, regional, and national levels, 1990–2019. Gynecologic Oncology. 2021;161(2):573-580. doi:10.1016/j.ygyno.2021.01.036

[3] Crosbie EJ, Kitson SJ, McAlpine JN, Mukhopadhyay A, Powell ME, Singh N. Endometrial cancer. *The Lancet*. 2022;399(10333):1412-1428. doi:10.1016/S0140-6736(22)00323-3

[4] Beavis AL, Fader AN. Surveillance Strategies in Endometrial Cancer Care: Why Less Represents Progress. *J Clin Oncol*. 2022;40(33):3790-3795. doi:10.1200/JCO.22.01551

[5] Makker V, MacKay H, Ray-Coquard I, et al. Endometrial cancer. *Nat Rev Dis Primers*. 2021;7(1):88. doi:10.1038/s41572-021-00324-8

[6] Concin N, Matias-Guiu X, Vergote I, et al. ESGO/ESTRO/ESP guidelines for the management of patients with endometrial carcinoma. Int J Gynecol Cancer. 2021;31(1):12-39. doi:10.1136/ijgc-2020-002230

[7] Berek JS, Matias-Guiu X, Creutzberg C, et al. FIGO staging of endometrial cancer: 2023. International Journal of Gynecology & Obstetrics. 2023;162(2):383-394. doi:10.1002/ijgo.14923

[8] GDC. portal.gdc.cancer.gov. https://portal.gdc.cancer.gov/projects/TCGA-UCEC

[9] Miller, K.D., Nogueira, L., Devasia, T., Mariotto, A.B., Yabroff, K.R., Jemal, A., Kramer, J. and Siegel, R.L. (2022), Cancer treatment and survivorship statistics, 2022. CA A Cancer J Clin, 72: 409-436. https://doi.org/10.3322/caac.21731

[10] Wade T. Iams, Christine M. Lovly; Molecular Pathways: Clinical Applications and Future Direction of Insulin-like Growth Factor-1 Receptor Pathway Blockade. Clin Cancer Res 1 October 2015; 21 (19): 4270–4277. https://doi.org/10.1158/1078-0432.CCR-14-2518

[11] Chang YW, Kuo HL, Chen TC, Chen J, Lim L, Wang KL, Chen JR. Abnormal p53 expression is associated with poor outcomes in grade I or II, stage I, endometrioid carcinoma: a retrospective single-institute study. J Gynecol Oncol. 2024 Mar;35(6):e78. https://doi.org/10.3802/jgo.2024.35.e78

[12] Singh, N., Piskorz, A.M., Bosse, T., Jimenez-Linan, M., Rous, B., Brenton, J.D., Gilks, C.B. and Köbel, M. (2020), p53 immunohistochemistry is an accurate surrogate for TP53 mutational analysis in endometrial carcinoma biopsies. J. Pathol., 250: 336-345. https://doi.org/10.1002/path.5375

[13] Bruno V, Betti M, D'Ambrosio L, et al. Machine learning endometrial cancer risk prediction model: integrating guidelines of European Society for Medical Oncology with the tumor immune framework. *Int J Gynecol Cancer*. 2023;33(11):1708-1714. doi:10.1136/ijgc-2023-004671

[14] Pan W, Cheng J, Lin S, et al. Construction of a prognostic model for endometrial cancer related to programmed cell death using WGCNA and machine learning algorithms. *Front Immunol*. 2025;16:1564407. doi:10.3389/fimmu.2025.1564407

[15] Li L, Qiu W, Lin L, Liu J, Shi X, Shi Y. Predicting recurrence and metastasis risk of endometrial carcinoma via prognostic signatures identified from multi-omics data. *Front Oncol*. 2022;12. doi:10.3389/fonc.2022.982452

[16] Qi X. Artificial intelligence-assisted magnetic resonance imaging technology in the differential diagnosis and prognosis prediction of endometrial cancer. *Sci Rep*. 2024;14(1):26878. doi:10.1038/s41598-024-78081-3

Chat history with AI

[1] https://www.genspark.ai/agents?id=d8f61c59-99b2-4756-a4aa-1db2cb1fef95

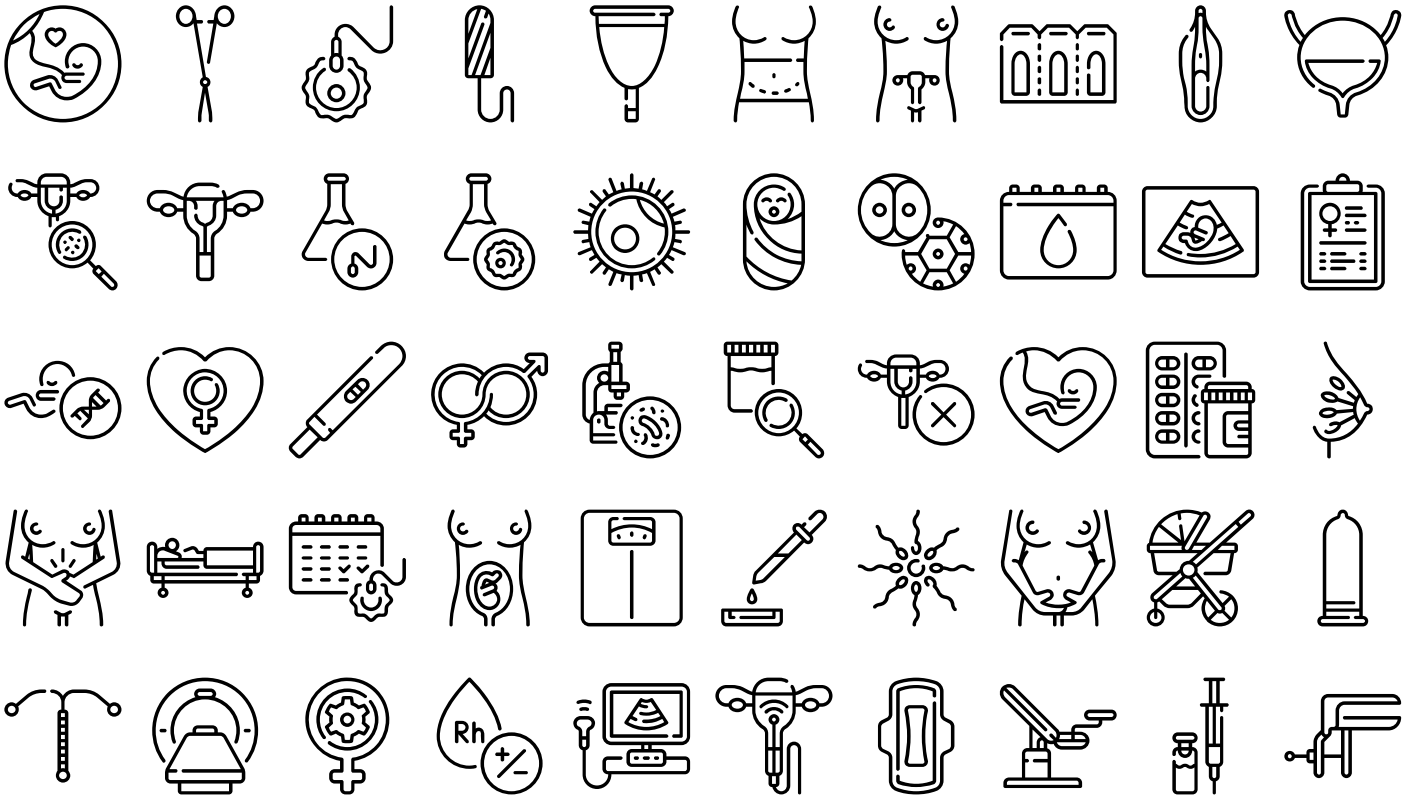
[2] UCEC median survival time | Claude

[3] https://www.genspark.ai/agents?id=a67abd93-1b0e-4e4c-adb6-29b67dba97c2

[4] https://github.com/copilot/share/422c5398-4044-84a3-a003-5c4d44e56826

[5] https://www.genspark.ai/agents?id=edaa2a6b-7da4-4417-995b-b4d373e12ffa

Icon pack



Contributions

	Alsaidan, Almuayyad	Chen Tao	Liu Yilin	Qi Jingyi	Wang Xianzhe	Sheng Jie	Zhu Taiqi
Idea proposals/Sourcing datasets			×		×	×	×
Background/literature search	×	×		×		×	
Project management			×			×	×
Data preprocessing/cleaning			×				×
Analyzing data/writing code		×	×	×	×		×
Making diagrams/ graphs/tables		×		×	×	×	×
Finding references	×	×			×	×	
Making slides		×		×	×	×	×
Prepare script + presentation		×		×		×	×

Reflection

S

Appendix List

Appendix A – Clinical Dataset (preprocessing)

Appendix B – RNA Dataset

- *Literature supported on why transcriptomic (RNA seq)*
- *Dataset filtering*

Appendix C – Endpoint Justification (2yr PFI)

Appendix D: Feature Share by Label

Appendix E: Model Performance on Outer CV

Appendix F: Hyperparameter tuning

Appendix A : Clinical Dataset, preprocessed

1. For multiple visits of the same patient, only the baseline (initial) diagnosis information was retained [A1]
2. Select features that were relevant, excluding:
    1. Those with >90% missing values
    2. Non-biologically relevant features

Clinical Features	Definition	Effects
age_at_diagnosis	Age when diagnosis at the first time	Higher → poorer survival
disease_type	➢ Type I → Endometrioid adenocarcinoma ➢ Type II → serous carcinomas and mixed cell	Type II → poorer prognoses, higher death rates
tumor_grade	➢ Grade 1: Less than 5% solid growth. ➢ Grade 2: 6% to 50% solid growth. ➢ Grade 3: More than 50% solid growth.	Higher grade → poorer prognosis, higher death rates
figo_stage	FIGO staging system: ➢ Stage I: The cancer is confined to the organ where it originated. ➢ Stage II: The cancer has spread to surrounding tissues or organs. ➢ Stage III: The cancer has spread to lymph nodes or other parts of the pelvis. ➢ Stage IV: The cancer has metastasized to distant organs.	advanced stages → poor prognosis, with a five-year survival rate falling of less than 20%.

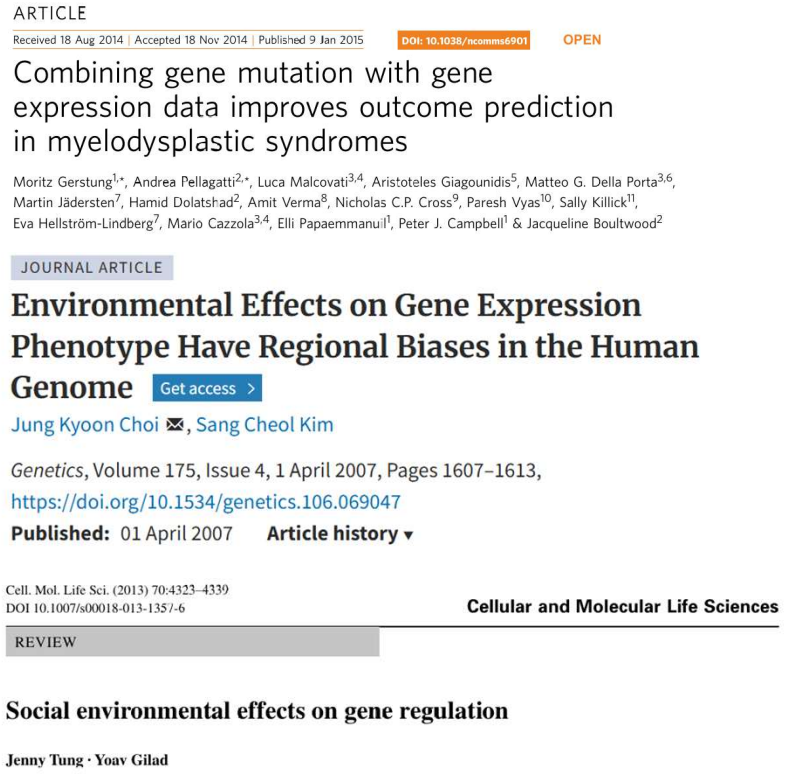
[A1] Moons KGM, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. Annals of Internal Medicine. 2015;162(1):W1. doi:https://doi.org/10.7326/m14-0698

Appendix B1: Why transcriptomic (RNA seq)

Patients with the same cancer and driver mutations often exhibit markedly different clinical features and treatment responses.

Transcriptome: A Promising Core Indicator of Tumor Heterogeneity

- Highest prognostic value in certain tumor types(1).
- Integrates both genetic and environmental information(2,3).
- Gene expression represents the critical link between genotype and phenotype.



ARTICLE

Received 18 Aug 2014 | Accepted 18 Nov 2014 | Published 9 Jan 2015 DOI: 10.1038/ncomms6901 OPEN

Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes

Moritz Gerstung[1,*], Andrea Pellagatti[2,*], Luca Malcovati[3,4], Aristoteles Giagounidis[5], Matteo G. Della Porta[3,6], Martin Jädersten[7], Hamid Dolatshad[2], Amit Verma[8], Nicholas C.P. Cross[9], Paresh Vyas[10], Sally Killick[11], Eva Hellström-Lindberg[7], Mario Cazzola[3,4], Elli Papaemmanuil[1], Peter J. Campbell[1] & Jacqueline Boultwood[2]

JOURNAL ARTICLE

Environmental Effects on Gene Expression Phenotype Have Regional Biases in the Human Genome Get access >

Jung Kyoon Choi ✉, Sang Cheol Kim

Genetics, Volume 175, Issue 4, 1 April 2007, Pages 1607–1613,
https://doi.org/10.1534/genetics.106.069047

Published: 01 April 2007 **Article history** ▾

Cell. Mol. Life Sci. (2013) 70:4323–4339
DOI 10.1007/s00018-013-1357-6

Cellular and Molecular Life Sciences

REVIEW

Social environmental effects on gene regulation

Jenny Tung · Yoav Gilad

Appendix B2: RNA Dataset filtering

60000+ genes → 7literature supported pathways → 800+ genes

Pathway	Definition and Function	Effects
PI3K-Akt/mTOR signaling pathway	regulates cell growth, survival, and metabolism.	The most common mutated pathway in EC. Overactivation Lead to Type I EC.
MAPK/ERK signaling pathway	regulating cell proliferation, differentiation	Overactivation due to mutations is a common feature in many cancers.
IGF1 signaling pathway	insulin-like growth factor-1, related to obesity(which is highly risk factor of EC)	stimulate the proliferation and migration of EC cells and to inhibit their apoptosis
MMR pathway	DNA mismatch repair: correcting errors in DNA	MMR deficiency leads to microsatellite instability-high (MSI-H) EC.
p53 signaling pathway	regulates a cell's response to stress	Loss/mutation leads to Type II EC
Wnt/β-catenin Pathway	regulates key cellular processes, stem cell self-renewal	Around 40% of ECs show irregularities in the pathway

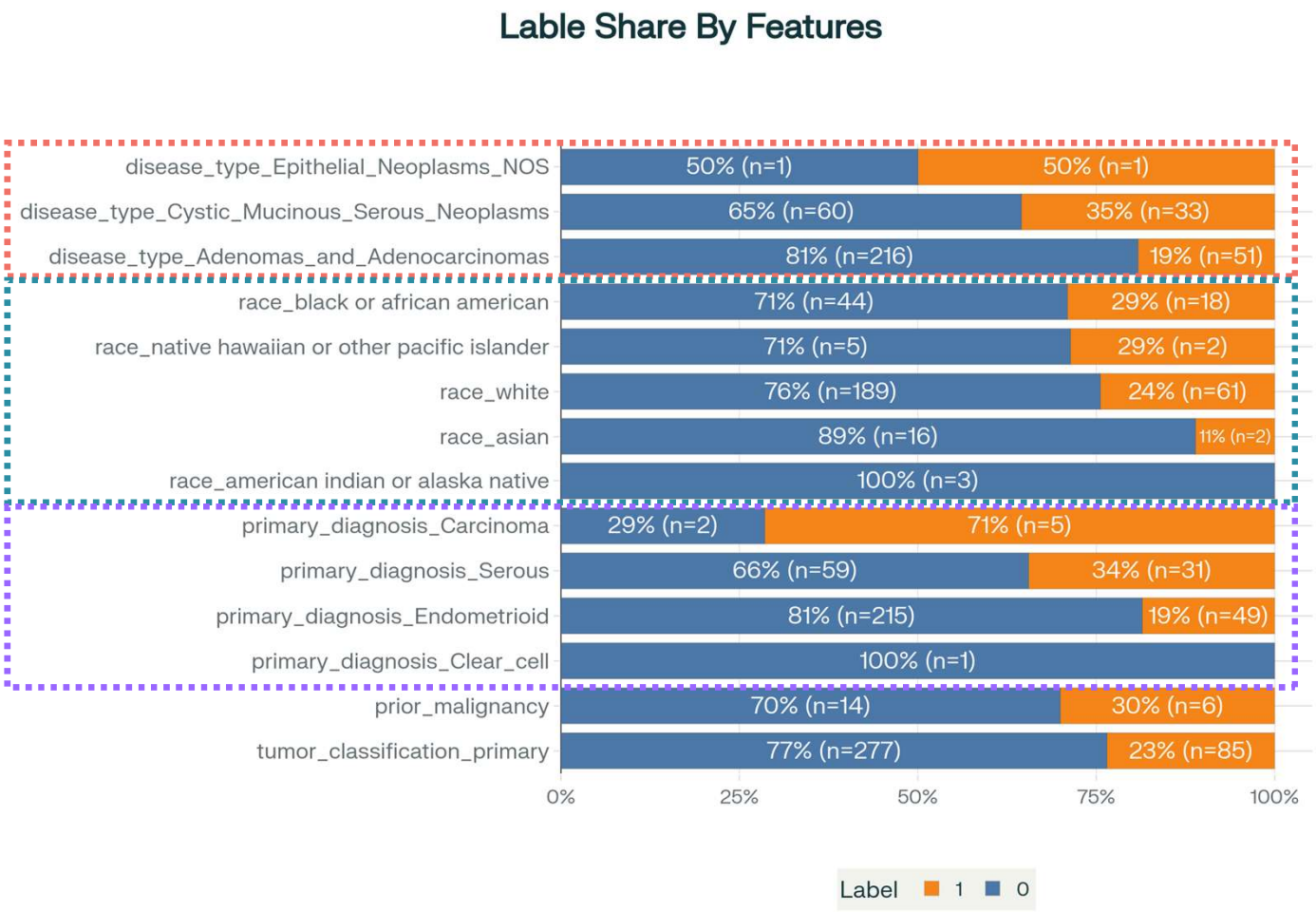
Appendix C: End-Pt Justification (2yr PFI)

Endpoint Definition	Usable Samples (Estimated)	High-Risk Event Rate	Pros (with Literature)	Cons
OS (Overall Survival)	368	≈ 13.0%	Traditional clinical endpoint	Low event rate → fewer events for ML; delayed maturity
PFI / 2-Year PFI (Recommended)	371	≈ 24.3%	- Good event rate for modeling- Clinically meaningful window (relapse happens early) (LWW Journals) - Supported in genomic/TCGA studies using PFI (BioMed Central)	Requires robust PFI data; need to ensure follow-up completeness
Composite (PFI or OS)	386	≈ 27.2%	Maximizes event capture	Harder to interpret clinically – mixing death and progression

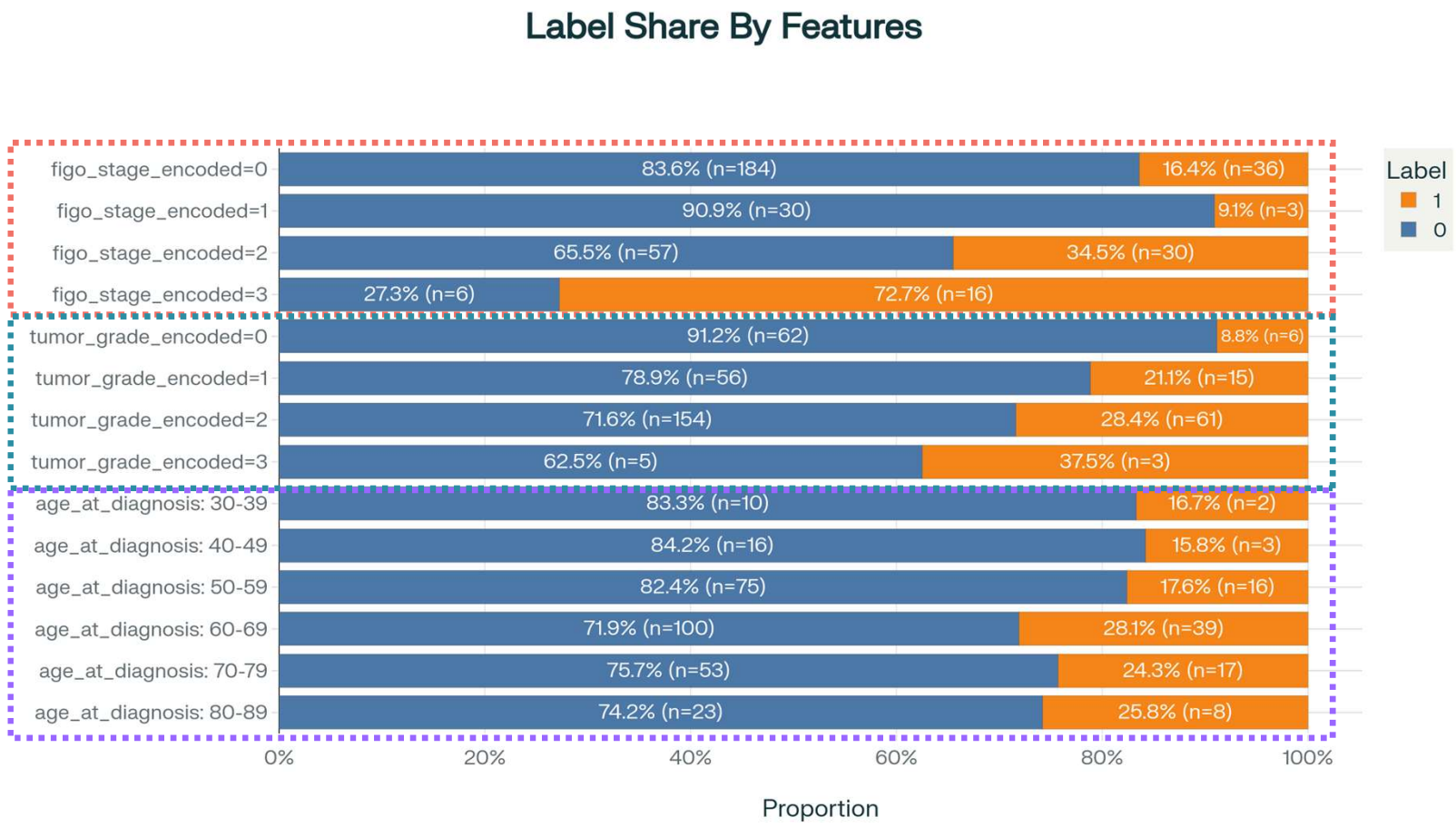
PFI as a valid endpoint in bioinformatics / ML studies

- In computational prognostic modeling, many studies (e.g., in TCGA data) use **PFI** (progression-free interval) rather than just OS. BioMed Central
- Recent ML-based prognostic models for UCEC (e.g., using programmed cell-death genes) also validate risk stratification based on survival intervals [C7]

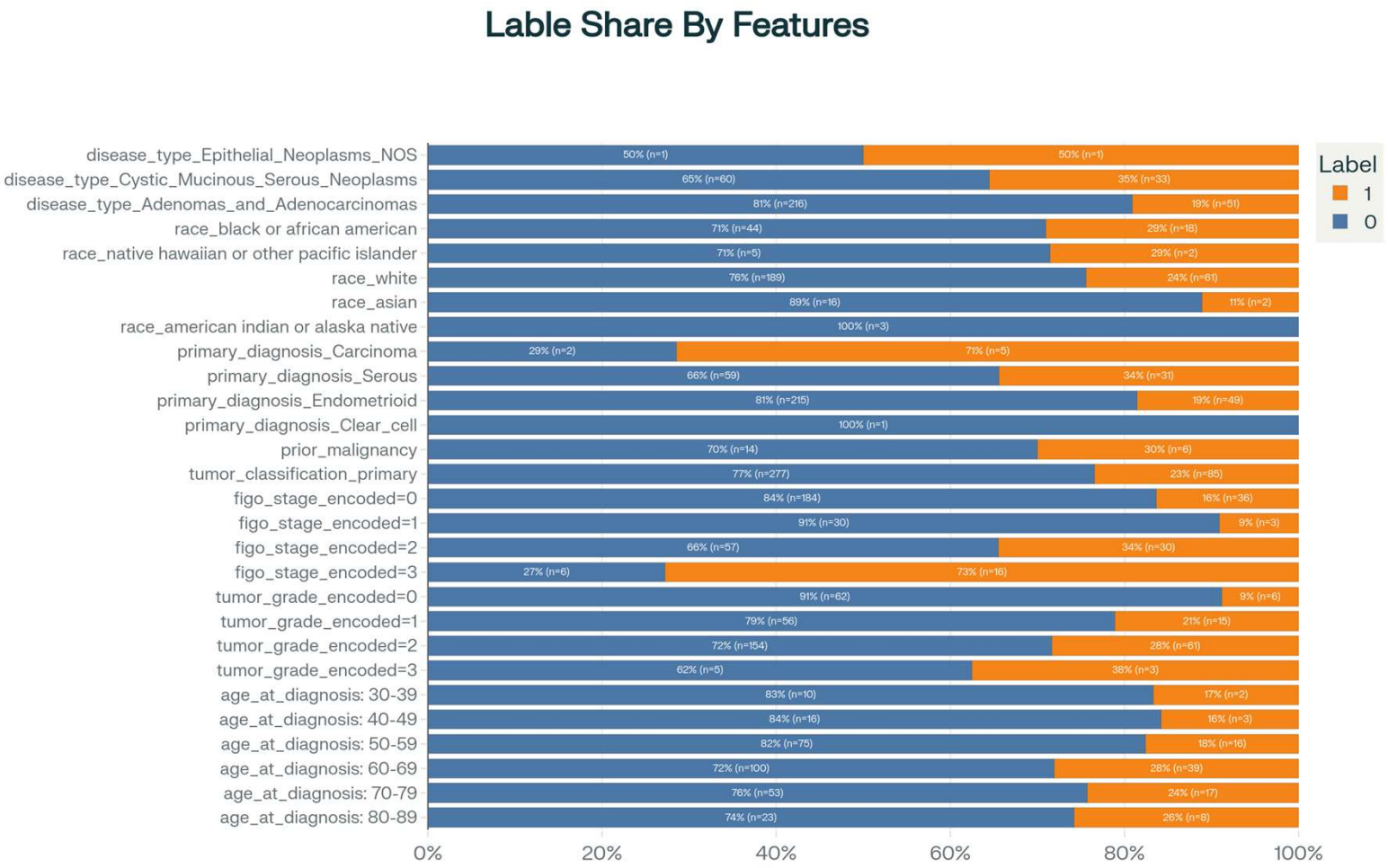
Appendix D1: Feature Share by Label_a



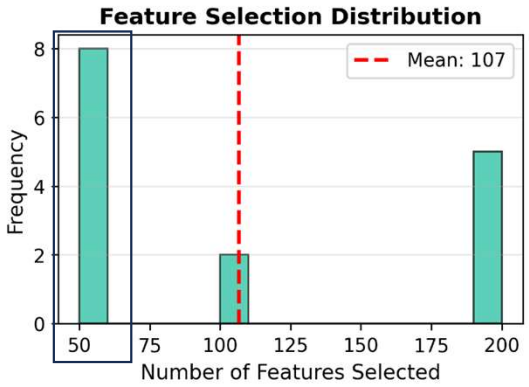
Appendix D2: Feature Share by Label_b

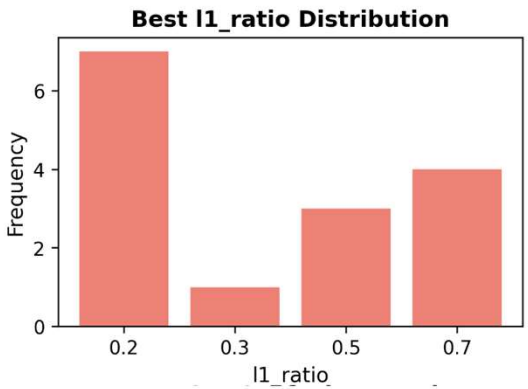


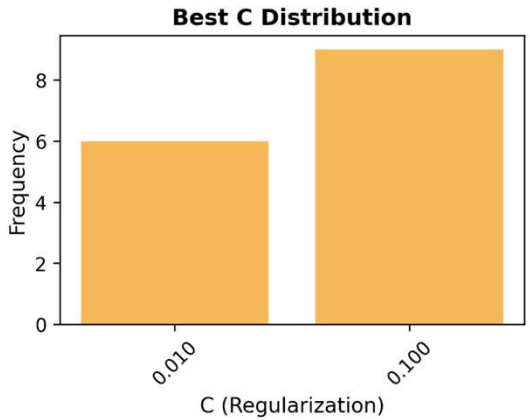
Appendix D3: Feature Share by Label Full

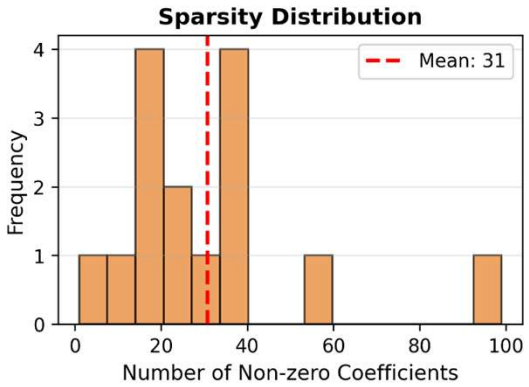


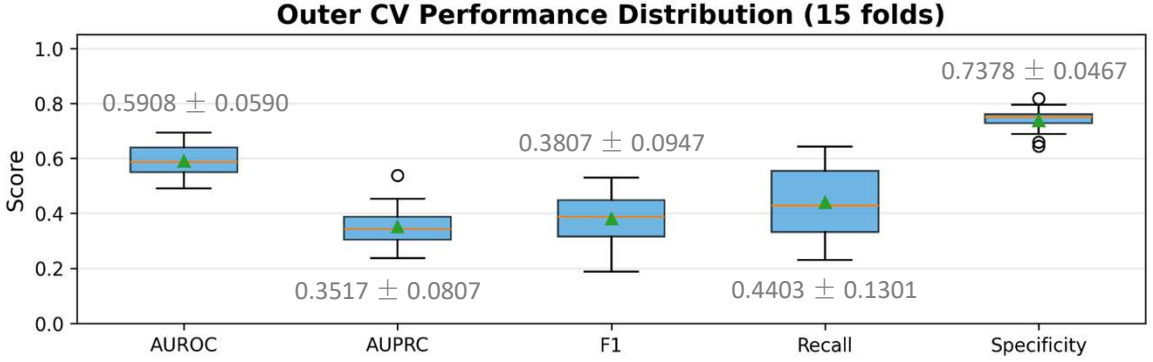
Appendix E: Model Performance on Outer CV











Appendix F: Hyperparameter tuning

ElasticNet LR

Inner CV
5-Fold
GridSearchCV（AUROC）

Total parameter Combinations
36
k × l1_ratio × C

Regularization
ElasticNet
Balances sparsity and coefficient stability

Parameter	Search Range	Description
selector__k	300 500 700 all	Univariate feature selection (Top-k). Larger k increases model flexibility; smaller k encourages sparsity and mitigates overfitting in high-dimensional settings.
classifier__l1_ratio	0.2 0.5 0.8	L1/L2 mixing ratio. Higher L1 increases sparsity and feature selection pressure; higher L2 stabilizes coefficients.
classifier__C	0.001 0.01 0.1 1 10 100	Inverse regularization strength. Larger C weakens regularization; smaller C imposes stronger penalization, balancing bias–variance trade-offs.