Machine learning approaches to identify predictive biomarkers for cell-cycle inhibitors in prostate cancer

Vishal Pattabiraman (31131441) , Clayton Faculty of IT
Supervisors : Lan Nguyen and Sungyoung Shin

1. Background

- Personalized medicine has emerged as a critical paradigm in oncology, owing to the considerable heterogeneity of individual tumors, even among individuals with the same tumor type.
- The development of companion biomarkers that can predict response to anti-cancer medicines and inform patient stratification (Fig 1) for clinical trials and/or therapy is a fundamental prerequisite for efficient precision oncology deployment.
- However, many anti-cancer medicines still lack accurate prognostic biomarkers, limiting their practical applicability.

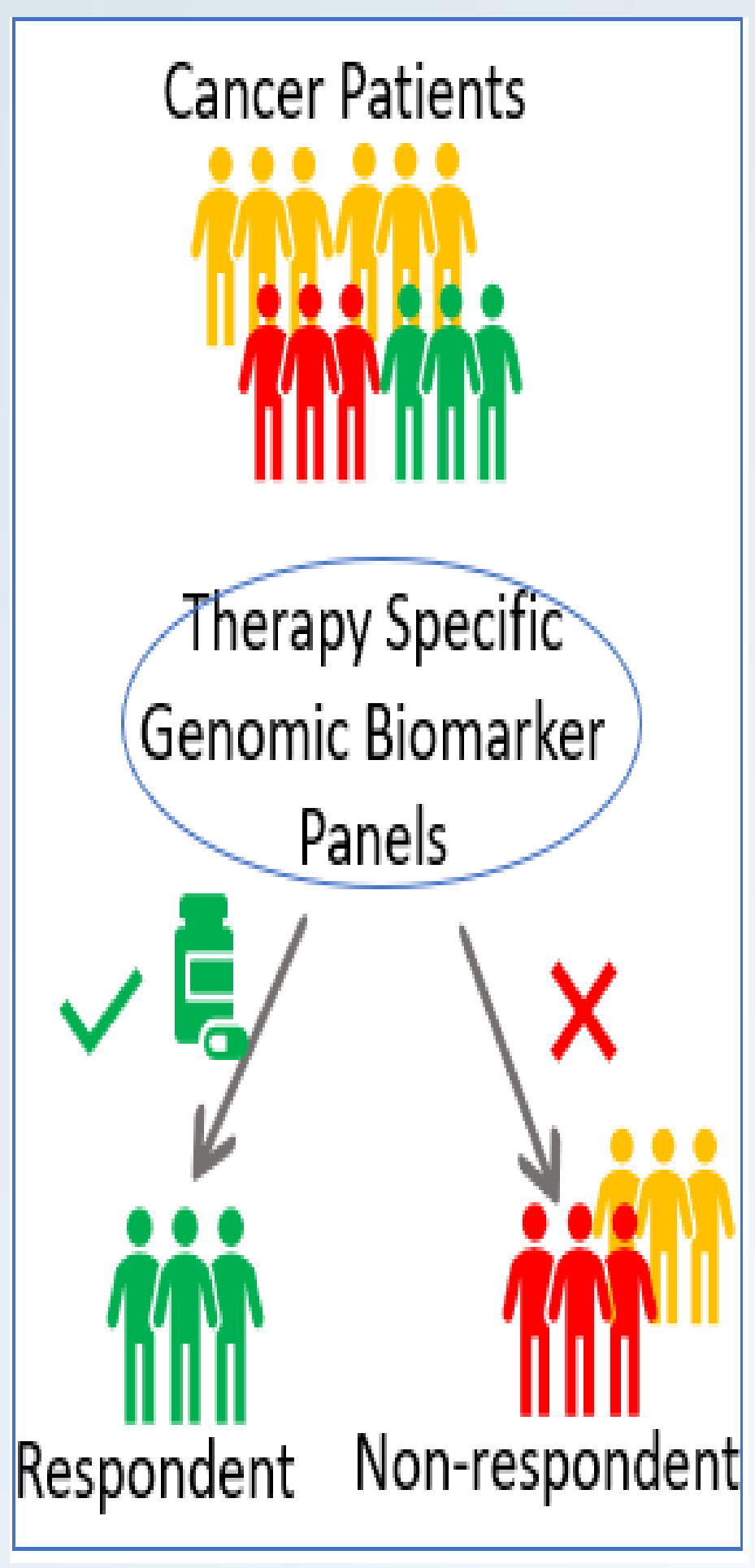


Fig 1: Patient Stratification

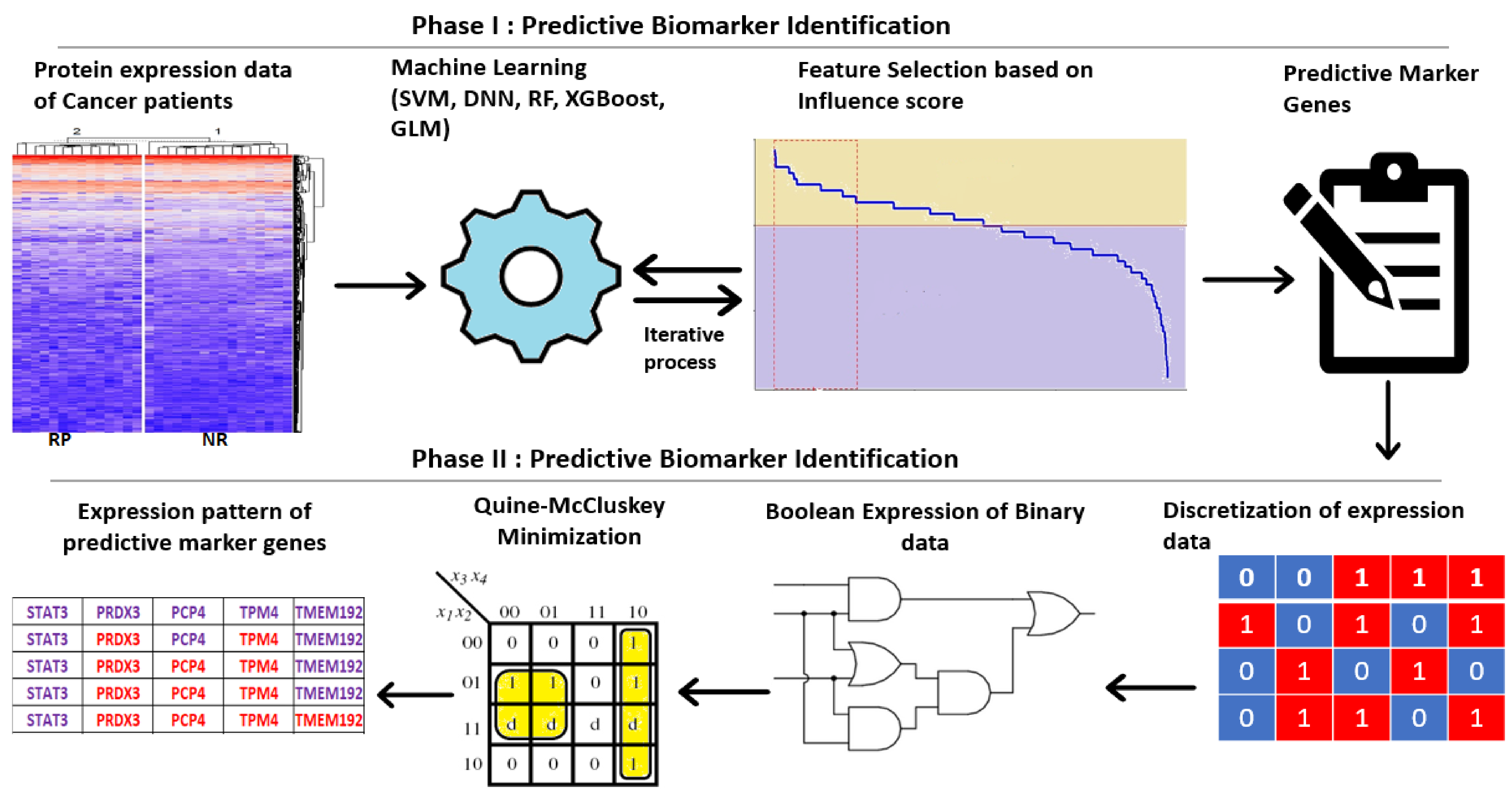
2. Aim

- To develop machine learning-based approach to derive predictive multi-protein biomarker panels and associated expression signatures that accurately predict cancer drug sensitivity

- Combine omics (gene/protein expression) data with drug-response data from pan cancer cell lines and explant samples (from public data collaborators) as inputs into machine learning models for predicting cancer treatment sensitivity using predictive multi-protein biomarker panels and related expression signatures.

- Validate biomarkers on independent datasets from cell lines and prostate cancer patients.

3. Research Design



Phase I : Predictive Biomarker Identification

Protein expression data of Cancer patients

Machine Learning (SVM, DNN, RF, XGBoost, GLM)

Iterative process

Feature Selection based on Influence score

Predictive Marker Genes

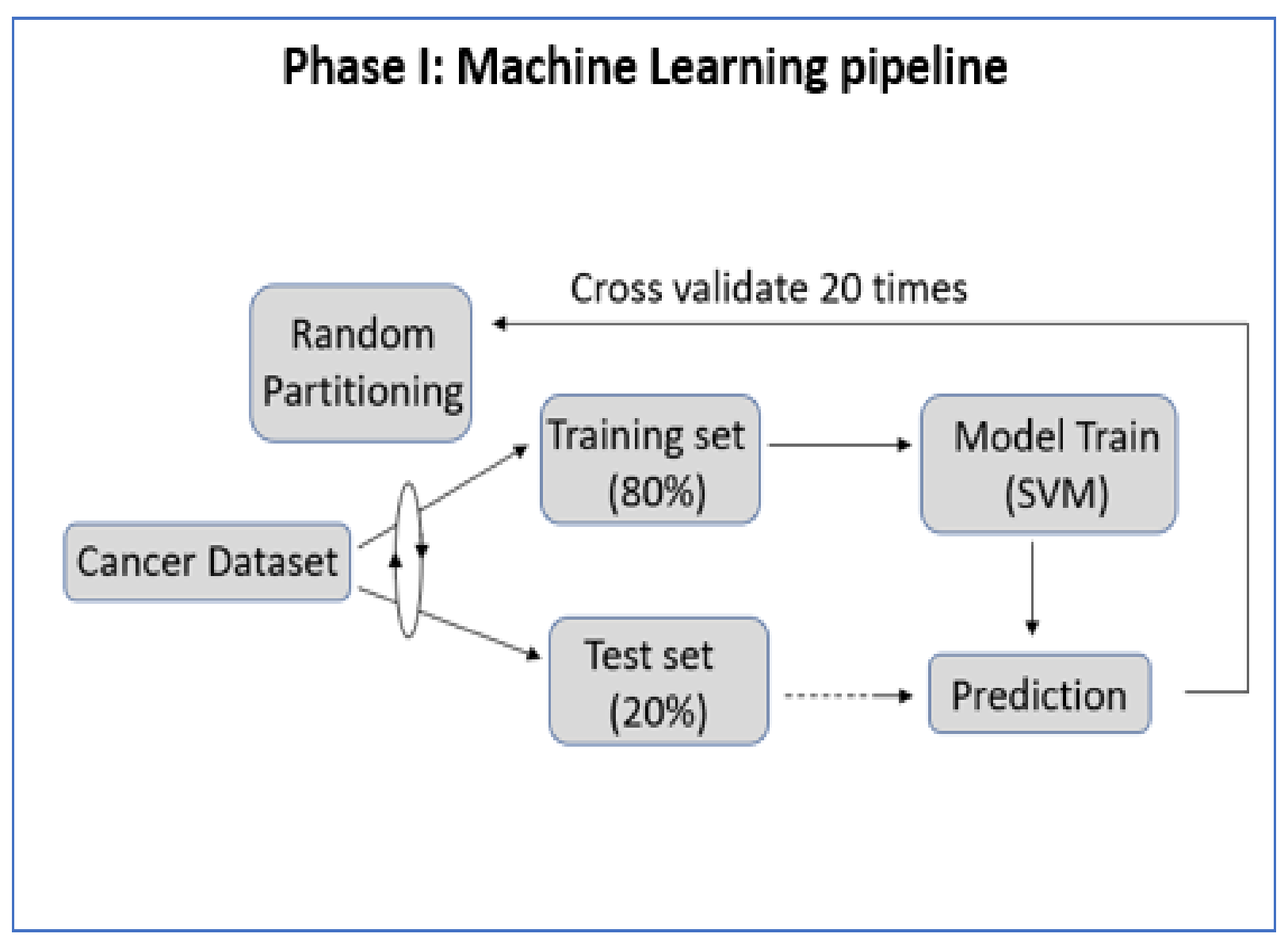
Phase II : Predictive Biomarker Identification

Discretization of expression data

Boolean Expression of Binary data

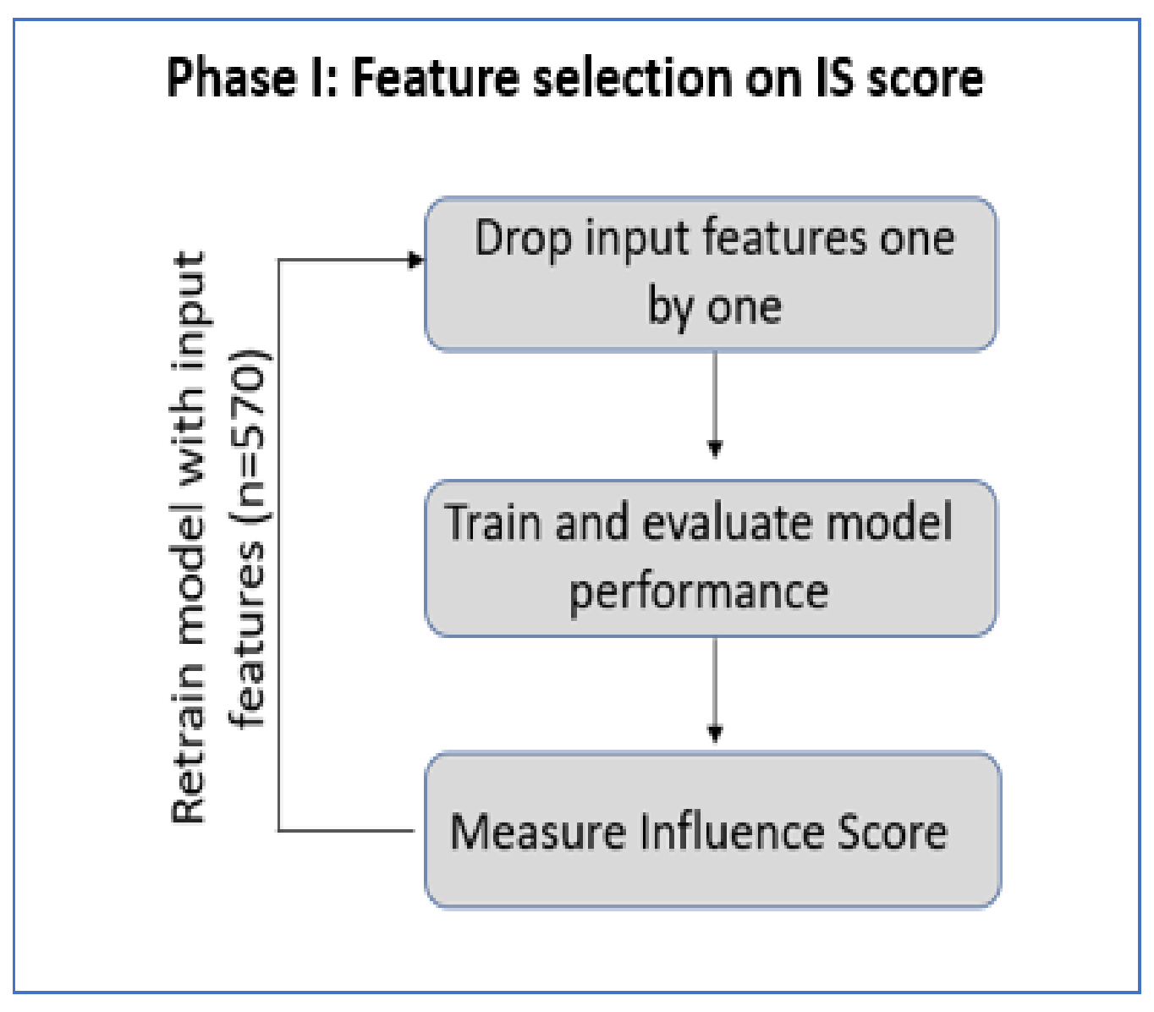
Quine-McCluskey Minimization

Expression pattern of predictive marker genes



Phase I: Machine Learning pipeline

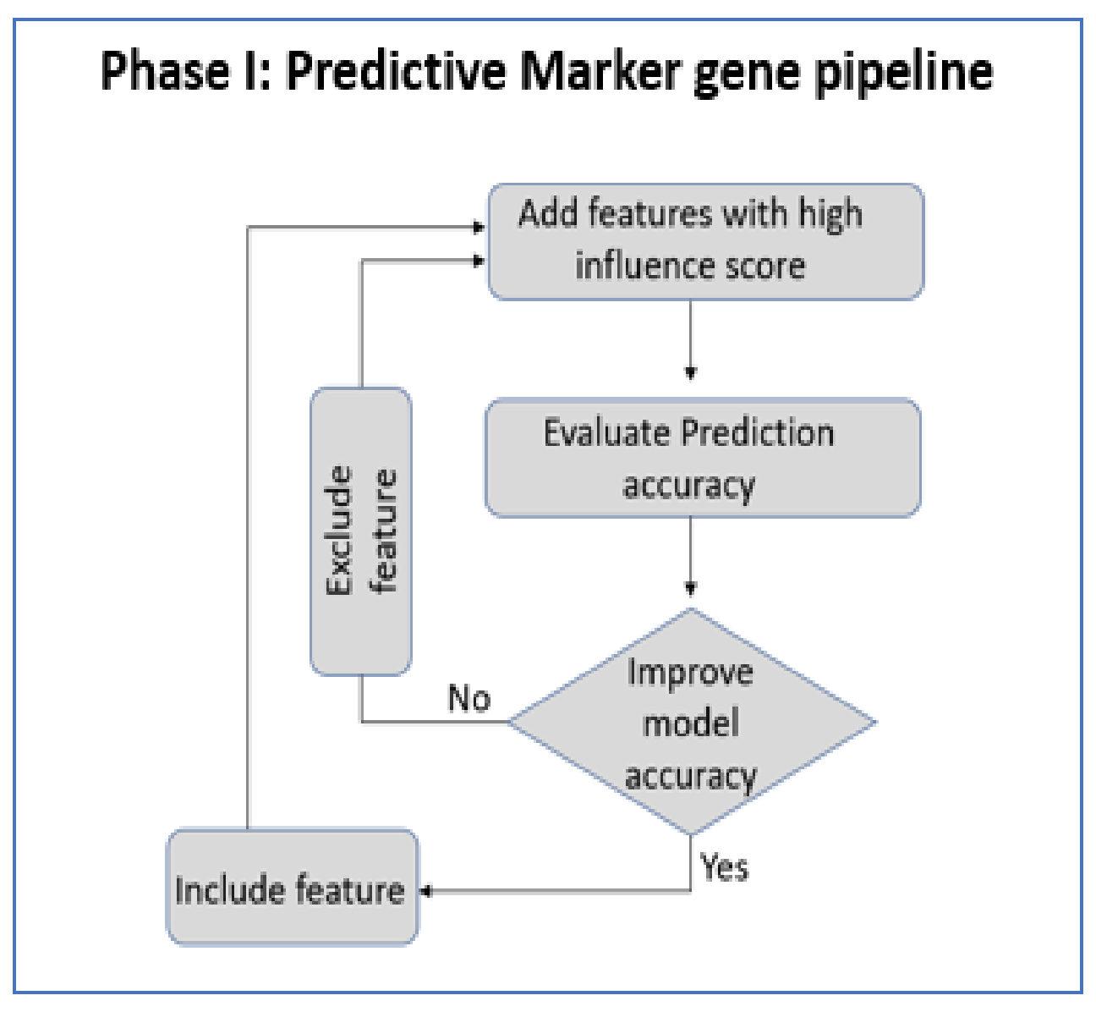
Cancer Dataset → Random Partitioning → Training set (80%) → Model Train (SVM) → Prediction; Test set (20%) → Prediction; Cross validate 20 times



Phase I: Feature selection on IS score

Drop input features one by one → Train and evaluate model performance → Measure Influence Score

Retrain model with input features (n=570)



Phase I: Predictive Marker gene pipeline

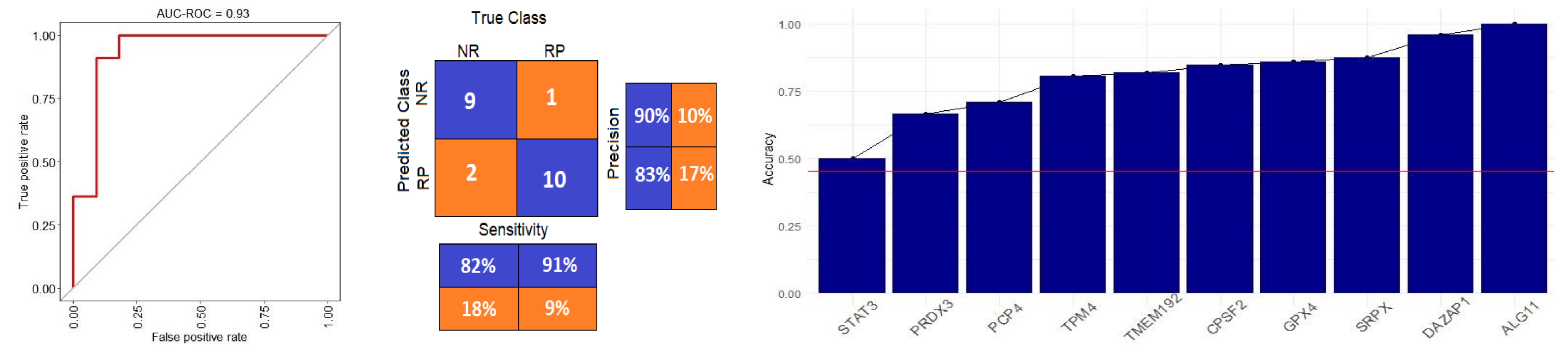
Add features with high influence score → Evaluate Prediction accuracy → Improve model accuracy

No → Exclude feature

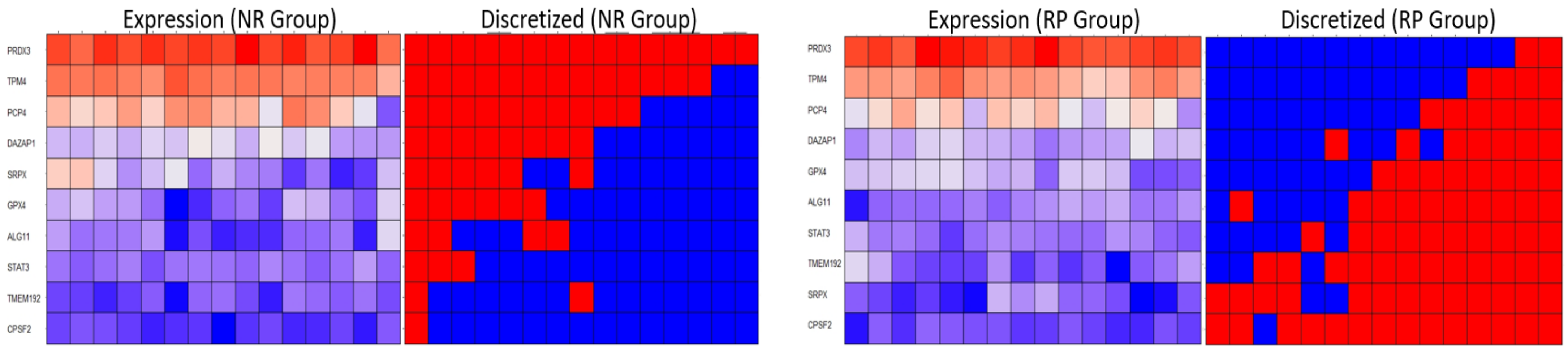
Yes → Include feature

4. Results

Prostate cancer predictive marker genes from Model computation



Gene Expression and discretization



Boolean Expression minimization using Quine McCluskey Algorithm

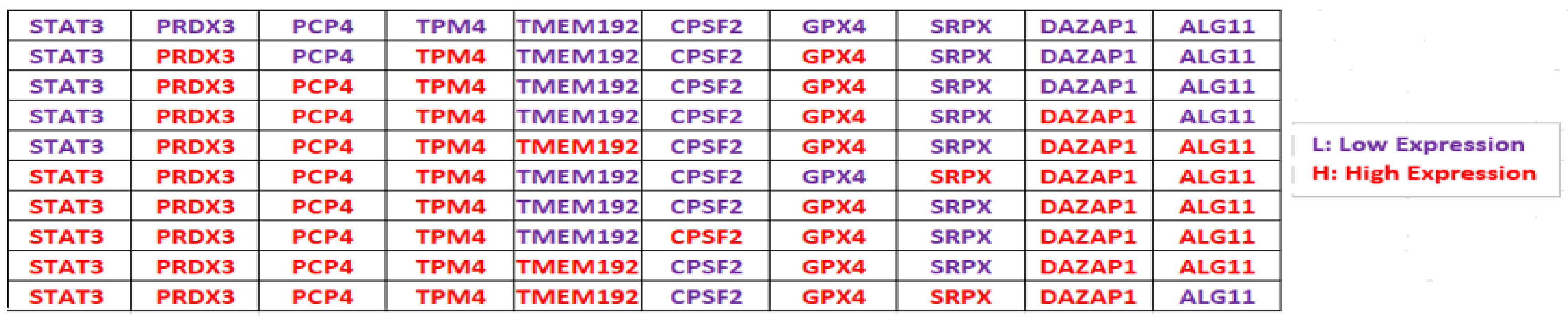
Expression signature of NR predictive marker genes

STAT3	PRDX3	PCP4	TPM4	TMEM192	CPSF2	GPX4	SRPX	DAZAP1	ALG11	Count (%)
L	H	L	L	L	L	L	L	L	L	4 (22%)
L	H	L	H	L	L	L	L	L	L	4 (22%)
L	H	H	H	L	L	L	L	L	L	3 (16%)
L	H	H	H	L	L	H	H	H	L	2 (10%)
L	H	H	H	L	L	H	H	H	H	1 (5%)
L	H	H	H	H	L	H	L	H	L	1 (5%)
H	H	H	H	L	L	L	L	H	L	1 (5%)
H	H	H	H	L	L	L	H	H	L	1 (5%)
H	H	H	H	L	L	H	H	H	H	1 (5%)
H	H	H	H	H	H	H	H	H	H	1 (5%)

Expression signature of RP predictive marker genes

STAT3	PRDX3	PCP4	TPM4	TMEM192	CPSF2	GPX4	SRPX	DAZAP1	ALG11	Count (%)
L	L	L	L	L	L	L	L	L	L	2 (19%)
L	H	L	H	L	L	H	L	L	L	1 (9%)
L	H	H	H	L	L	H	L	L	L	1 (9%)
L	H	H	H	L	L	H	L	H	L	1 (9%)
L	H	H	H	H	L	H	L	H	H	1 (9%)
H	H	H	H	L	L	L	H	H	H	1 (9%)
H	H	H	H	L	L	H	L	H	H	1 (9%)
H	H	H	H	L	H	H	L	H	H	1 (9%)
H	H	H	H	H	L	H	L	H	H	1 (9%)
H	H	H	H	H	L	H	H	H	L	1 (9%)

Expression pattern of predictive marker gene panel of Respondent(RP) Group



STAT3	PRDX3	PCP4	TPM4	TMEM192	CPSF2	GPX4	SRPX	DAZAP1	ALG11
STAT3	PRDX3	PCP4	TPM4	TMEM192	CPSF2	GPX4	SRPX	DAZAP1	ALG11
STAT3	PRDX3	PCP4	TPM4	TMEM192	CPSF2	GPX4	SRPX	DAZAP1	ALG11
STAT3	PRDX3	PCP4	TPM4	TMEM192	CPSF2	GPX4	SRPX	DAZAP1	ALG11
STAT3	PRDX3	PCP4	TPM4	TMEM192	CPSF2	GPX4	SRPX	DAZAP1	ALG11
STAT3	PRDX3	PCP4	TPM4	TMEM192	CPSF2	GPX4	SRPX	DAZAP1	ALG11
STAT3	PRDX3	PCP4	TPM4	TMEM192	CPSF2	GPX4	SRPX	DAZAP1	ALG11
STAT3	PRDX3	PCP4	TPM4	TMEM192	CPSF2	GPX4	SRPX	DAZAP1	ALG11
STAT3	PRDX3	PCP4	TPM4	TMEM192	CPSF2	GPX4	SRPX	DAZAP1	ALG11
STAT3	PRDX3	PCP4	TPM4	TMEM192	CPSF2	GPX4	SRPX	DAZAP1	ALG11

L: Low Expression
H: High Expression

5. Discussion

- On the Prostate Cancer dataset, we discovered a highly predictive 10-protein panel with an overall prediction accuracy of 93 percent. This panel outperforms any individual Differential Expressed Proteins (DEPs) as well as a panel containing all 570 DEPs in terms of predictive capability.
- Notably, STAT3, TPM4, TMEM192, CPSF2, GPX4 and SRPX are found to be associated with Cancer cell-cycle[1].
- On Multiomics Pan-cancer Dataset we identified 8 Panel Marker genes with 70 percent overall prediction accuracy.
- Notably, CDKN2A, CASZ1, AGBL4 and PDXDC1 are associated with Pan-cancer cell cycle.

6. Conclusion and Future work

- In comparison to existing techniques, this pipeline aids in the identification of most predictive biomarkers by comprehensive examination of Influence scores corresponding to putative features that contribute positively to overall predictive performance.
- This proposed technique also allows for the development of expression signature patterns for proteins in the identified panel that correspond to specific responses (Responsive or Resistant).
- We established the machine learning approach utility by focusing on prostate cancer; nevertheless, it has a broad applicability and can be used in future research with various drugs and cancer types.

Reference : 1. UniProt ConsortiumEuropean Bioinformatics InstituteProtein Information ResourceSIB Swiss Institute of Bioinformatics. (n.d.). UniProt Consortium. Retrieved from https://www.uniprot.org/