Interpretable Machine Learning Pipeline to Identify Genomic Signatures in Age-Related Macular Degeneration

RICE | D2K Data to Knowledge Lab

Duy Ha, Patrick Yee, Qingxin Yuan, Sujitha Ravichandran, Tian Xia, Wanying Xu
Sponsor: Rinki Ratnapriya; Faculty Mentor: Arko Barman; Mentor: Maryam Khalid

Baylor College of Medicine

Background and Objective

Age-related Macular Degeneration (AMD):

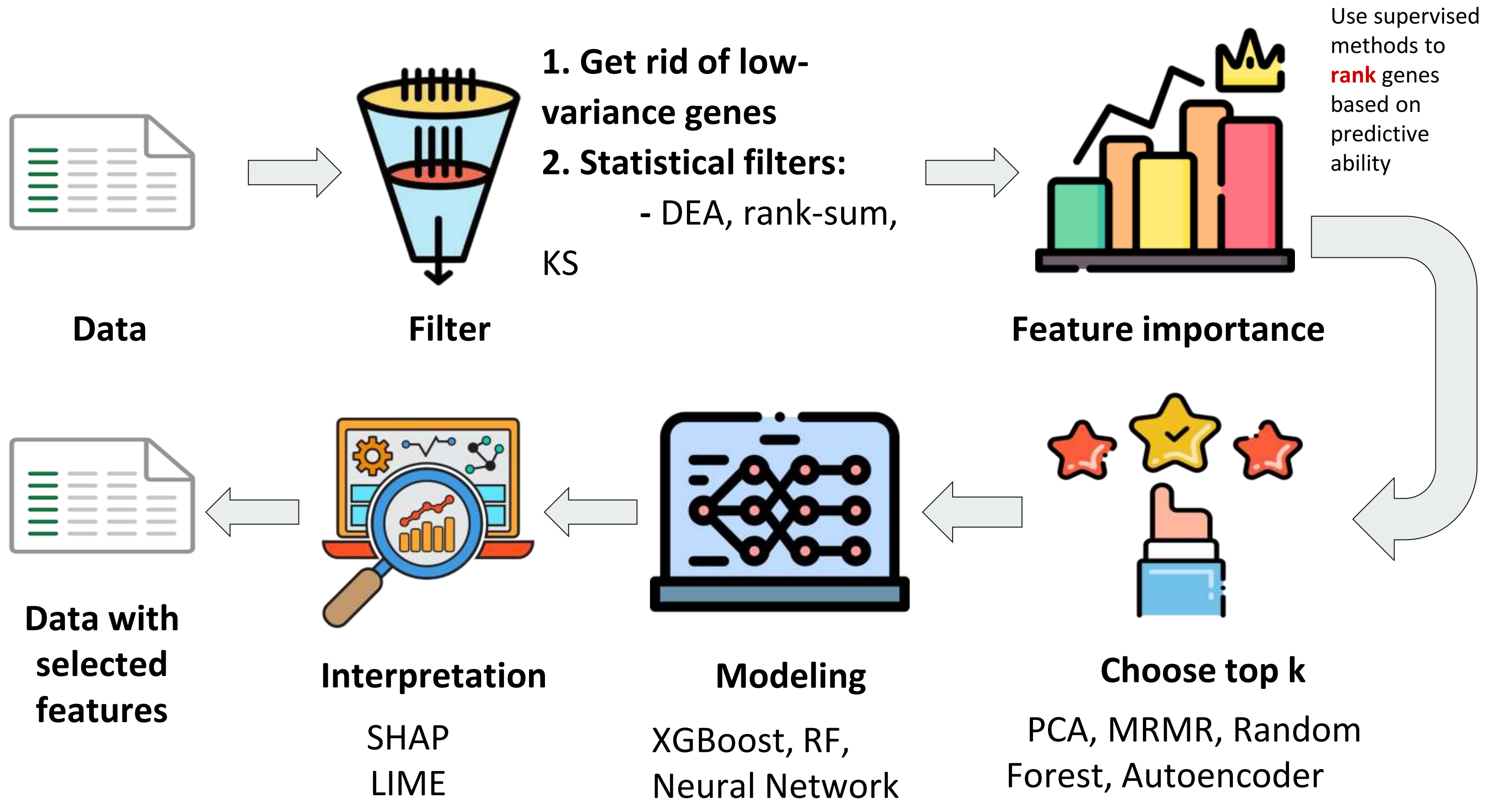
- AMD is a leading cause of vision loss in the elderly, impacting millions in the US alone.
- AMD is caused by genetic and environmental factors.
- Current Challenges: There's substantial difficulty in translating DNA variants to biological understanding, impeding progress on finding specific genes causing AMD.
- The complexity of AMD lies in its **combination of genetic variants**, **curse of dimensionality and gene processing complexity.**
- **Little is known about the risk factors of AMD.**

AMD Progression leads to vision loss

Objectives:

- **Development of an interpretable machine-learning pipeline** to predict AMD from Genomic data.
- **Discover crucial genomic signatures** that contribute to AMD
- Development of generalizable Python Package for Gene analysis for complex diseases.
- **Broader Application:** the pipeline and Python package are **generalized for other complex diseases.**

AMD Prediction & Gene Discovery Pipeline



Data

Filter

1. Get rid of low-variance genes
2. Statistical filters:
   - DEA, rank-sum, KS

Feature importance

Use supervised methods to **rank** genes based on predictive ability

Choose top k

PCA, MRMR, Random Forest, Autoencoder

Modeling

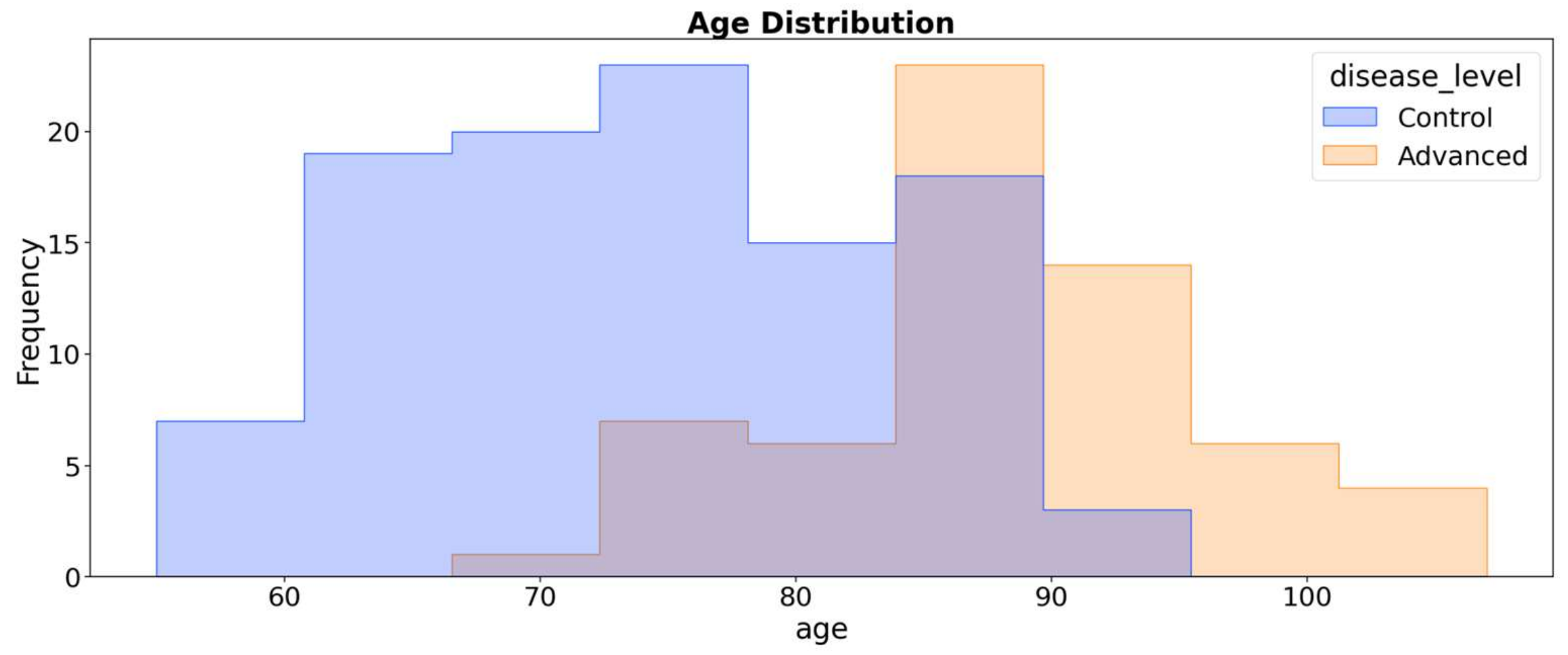
XGBoost, RF, Neural Network

Interpretation

SHAP
LIME

Data with selected features

Data Description

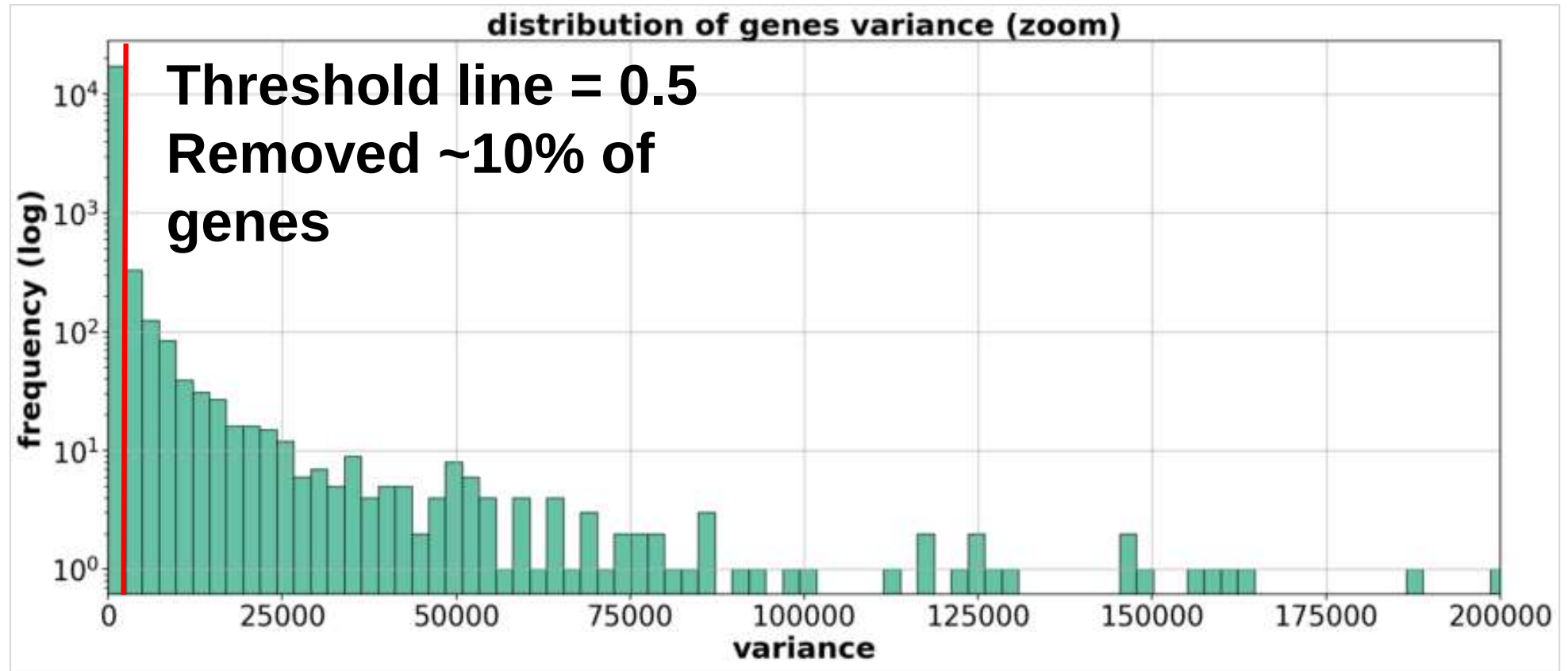


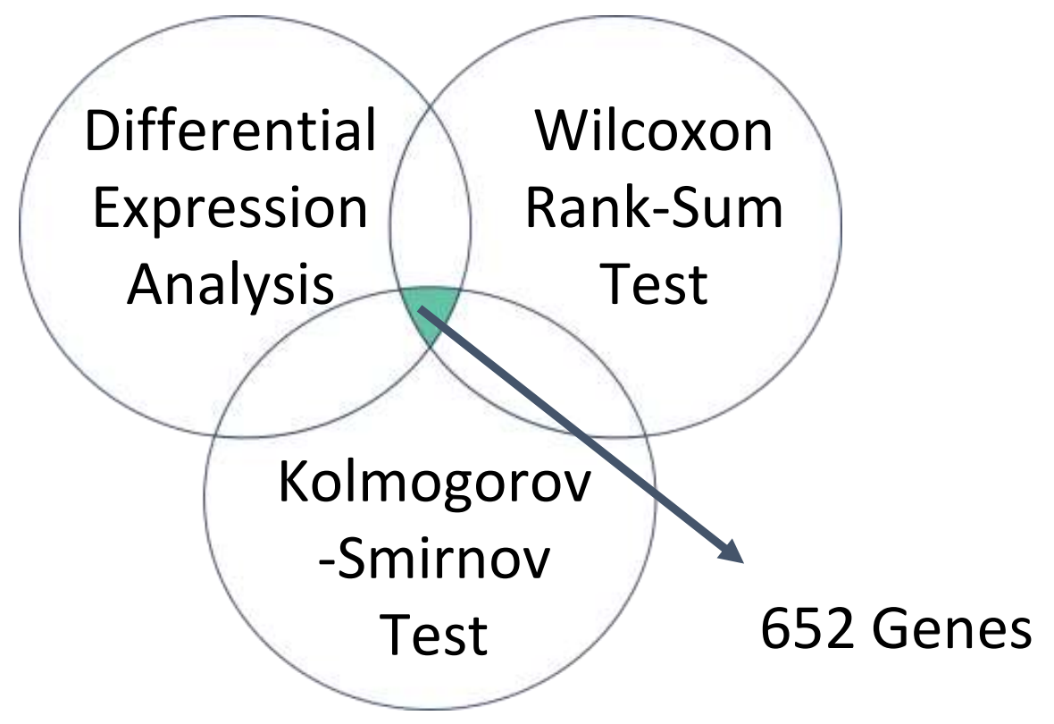
Genomic Data:
18056 Genes from **453** patients

- 105 samples with normal AMD levels
- 61 samples with advanced AMD levels

System Modeling & Results

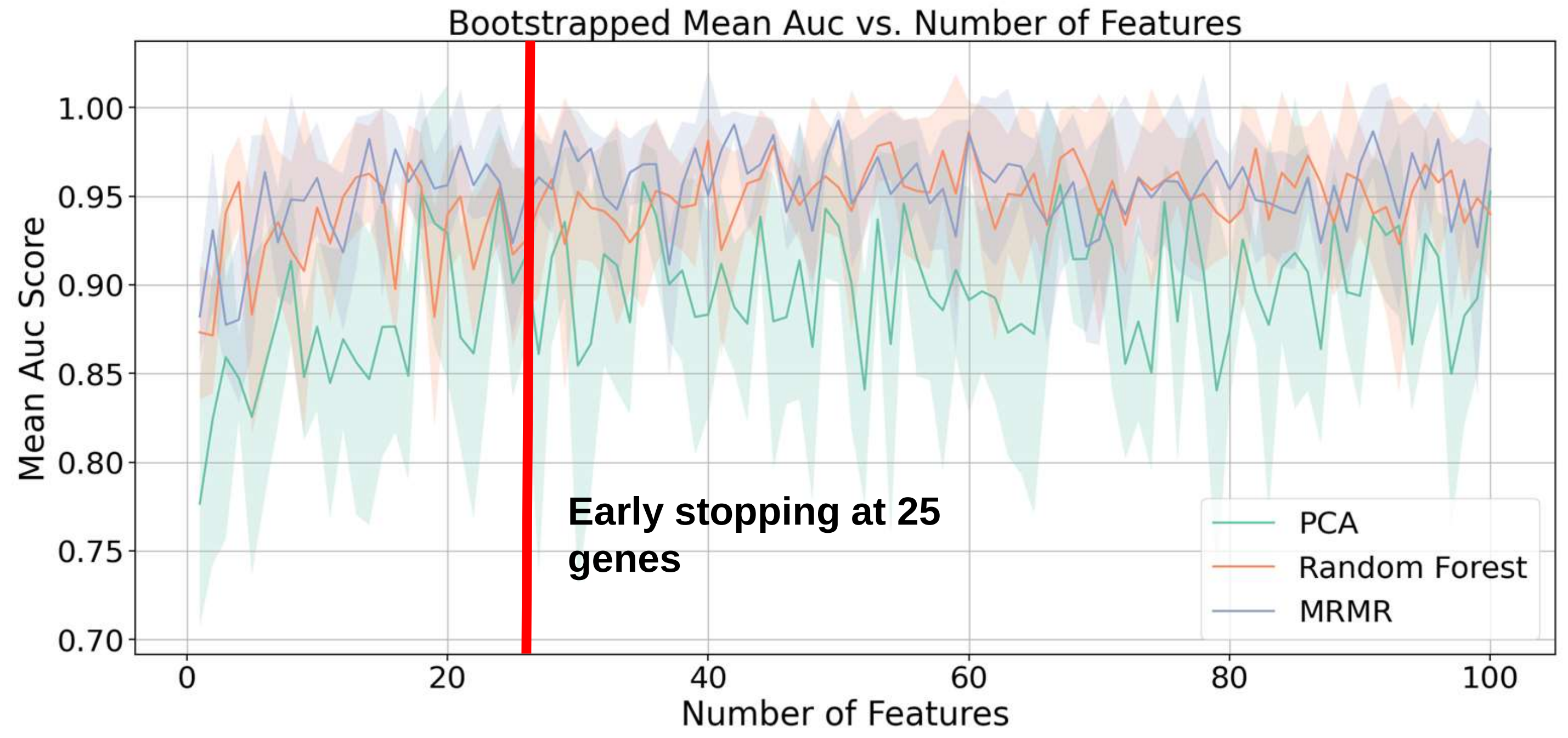
Step 1: Statistical filtering: filtering features using gene variance and statistical methods





Step 2: Feature selection

- **Principal Component Analysis** (PCA)
  - Ranks genes by the number of times they appear at top principal components with certain cutoff
- **Random Forest**
  - Selects genes using the average permutation-based feature importance over 100 bootstrapped resamples
- **Minimum-Redundancy Maximum-Relevance** (MRMR)
  - Identifies features that maximizes the algorithm's predictive power using an f-statistic relevance score and minimized redundancy using Pearson correlation



Step 3: Modeling using XGBoost (eXtreme Gradient Boosting)

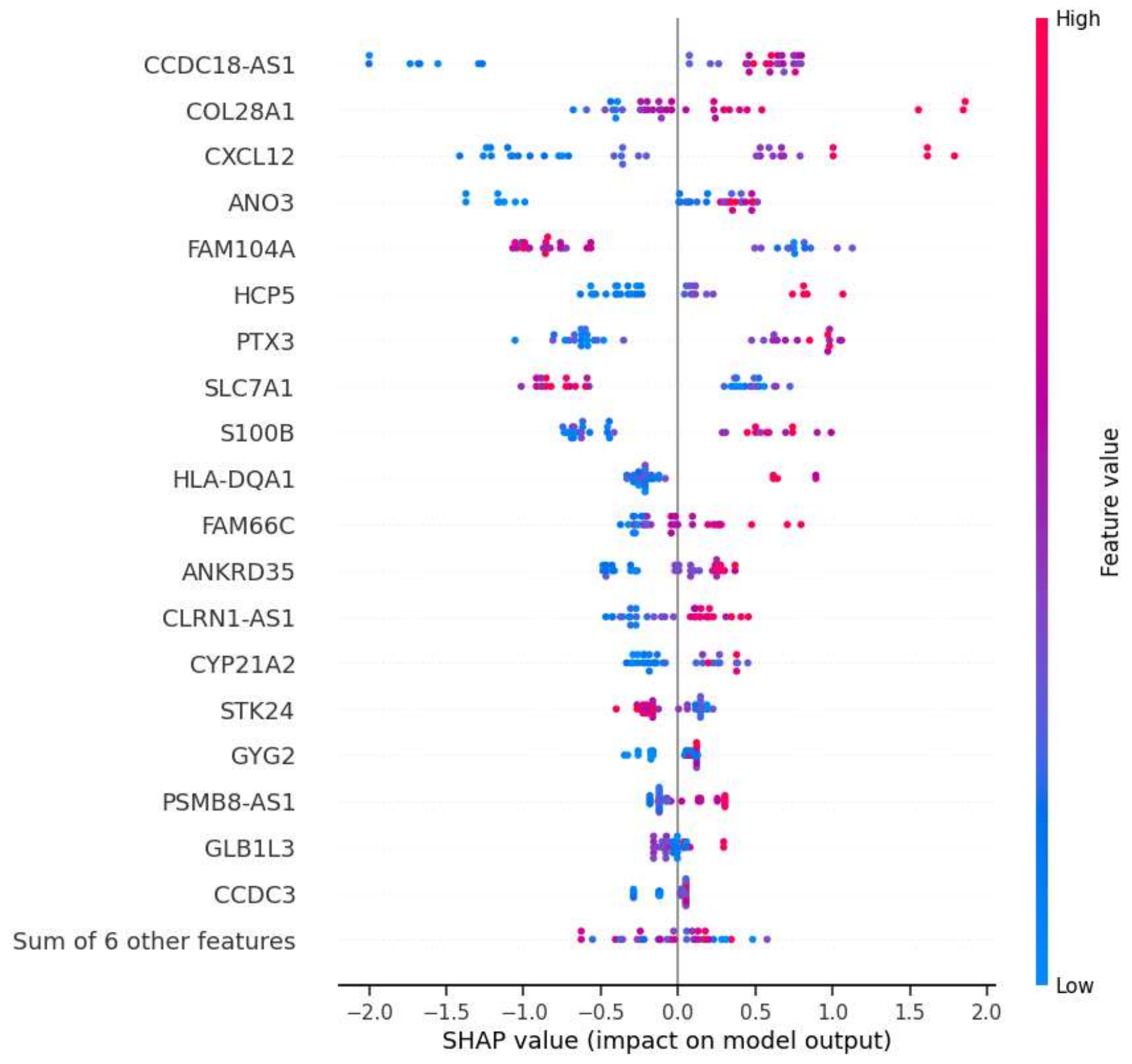
- Used bootstrapping and hyperparameter tuning on selected Top 25 genes

Methods +XGBoost	Precision	Sensitivity	Specificity	F1 Score	AUC
PCA	0.840 ± 0.067	0.829 ± 0.068	0.924 ± 0.085	0.824 ± 0.068	0.835 ± 0.055
MRMR	0.900 ± 0.052	0.894 ± 0.054	0.922 ± 0.065	0.893 ± 0.054	0.946 ± 0.056
Random Forest	0.873 ± 0.021	0.870 ± 0.023	0.931 ± 0.028	0.869 ± 0.023	0.930 ± 0.037

- MRMR feature selection method + XGBoost model perform the best for AMD patient classification.
- Random Forest feature selection method + XGBoost model has the least standard deviation and thus it is the most stable model.

Model Interpretation

Interpretation of XGBoost with mRMR feature selection method
Shapley Additive Explanations (SHAP)



SHAP applies game theoretic approach to explain the output of machine learning model and breaks down a prediction to show the impact of each feature.

Highly expressed genes in AMD subjects	Highly expressed genes in Control subjects
CCDC18-AS1, COL28A1, CXCL12	FAM104A, SLC7A1

DREAMR Package

Dimensionality **R**eduction, (feature) **E**xtraction, **a**nd **M**odeling for **R**NA

We developed a deployable and generalizable genomic analytics package for medical professionals!

Modules	Details & Functionalities
Preprocessing Tools	Loading, Merging, Filtering by Variance, Normalization Techniques (Z-score, Min-Max), Converting ENSG ID to Gene Name, etc.
EDA Tools	Visualization of Feature Distributions, Visualizing Variance and Standard Deviation, Summarizing Data, Finding Missing Values, etc.
Feature Filtering	Univariate Generalized Linear Model Filter, KS-test, Wilcoxon Rank Sum, Differential Expression Analysis
Feature Select	PCA, MRMR, Optimal Transport, Random Forest, XGBoost
Feature Scoring	AIC, BIC, Testing Agreement in Gene Rankings
Modeling Classes	XGBoost, Random Forest, Neural Network, Autoencoder
Modeling Tools	Bootstrapping Class, Hyperparameter Tuning, Evaluation Functionalities and Plotting
Interpretation Tools	SHAP, LIME
Utilities	Saving and Loading Trained Models, Data Visualization Wrappers

Conclusion

DREAMR: Developed a **generalizable genomic analytics package**, featuring comprehensive **data preprocessing, feature filtering and selection, modeling, and evaluation** techniques and functionalities.

Using DREAMR's Pipeline for AMD Findings

- Models **classified AMD patients with high performance (~0.98 AUC).**
- Interpretation (SHAP and LIME) **identified potential disease-causing genes.**
- Best performing **models may not show all** significant genes
  - e.g. MRMR's inherent nature to exclude redundant genes
- DREAMR provides a **variety of models** for users to choose according to their needs.

Future Improvements:

- Plans to **extend DREAMR's capabilities** to **handle multiclass classification problems.**
- Potential usage example: classify AMD patients for **multiple disease stages.**