Machine Learning Approaches for Genetics-Clinical Interaction Discovery: Methods Comparison and Application using UK Biobank Data

Yi-Pin Lai, Giorgio E. M. Melloni, Sabina A. Murphy, Siona Prasad, Nicholas A. Marston, Andrea Bellavia

TIMI Study Group, Brigham and Women's Hospital, Harvard Medical School

ylai4@bwh.harvard.edu

ENAR Spring Meeting, New Orleans, March 24th, 2025

Acknowledgments:

- Dr. Marc S. Sabatine, TIMI Study Group
- Dr. Christian T. Ruff, TIMI Study Group
- Dr. Frederick K. Kamanu, TIMI Study Group

Disclaimer:

The authors declare no conflict of interest. Y-PL, GEM, SAM, NAM and AB are members of the TIMI Study Group which has received institutional research grant support through Brigham and Women's Hospital from: Abbott, Abiomed, Inc., Amgen, Anthos Therapeutics, ARCA Biopharma, Inc., AstraZeneca, Boehringer Ingelheim, Cleerly, Inc., Daiichi-Sankyo, Ionis Pharmaceuticals, Inc., Janssen Research and Development, LLC, MedImmune, Merck, Novartis, Pfizer, Softcell Medical Limited, The Medicines Company

1) Background and motivation

1) Background and motivation

- Polygenic Risk Score (PRS) quantifies an individual's genetic susceptibility to a phenotypic trait or disease relative to a population
- PRS has been utilized in various recent clinical applications to enhance risk stratification for patients

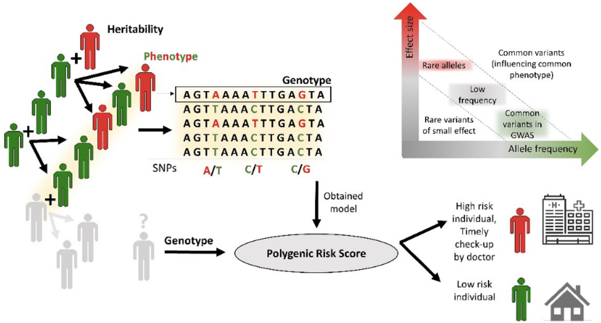
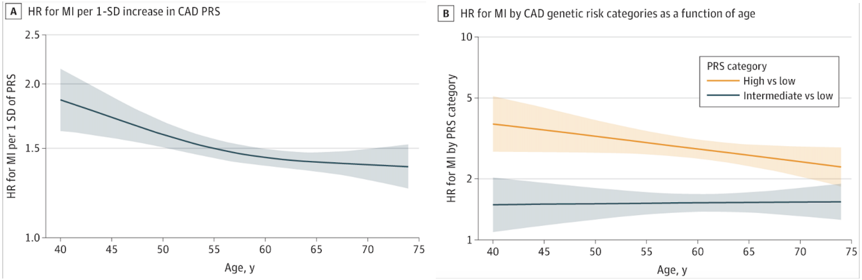


Figure 1 from Schwarzerova et al. Briefings in Bioinformatics 2024

- This is commonly achieved by assessing (potentially non-linear) interactions between PRS and clinical variables[1] defined a priori



(a) Numerical coronary artery disease (CAD) PRS x age

(b) Categorical CAD PRS x age

Figure 1 from Marston et al. JAMA Cardiology 2022

[1] E.g., demographic, physiological, medical history, medication use, behavioral/lifestyle, and biomarkers

Approaches to assessing interaction effects

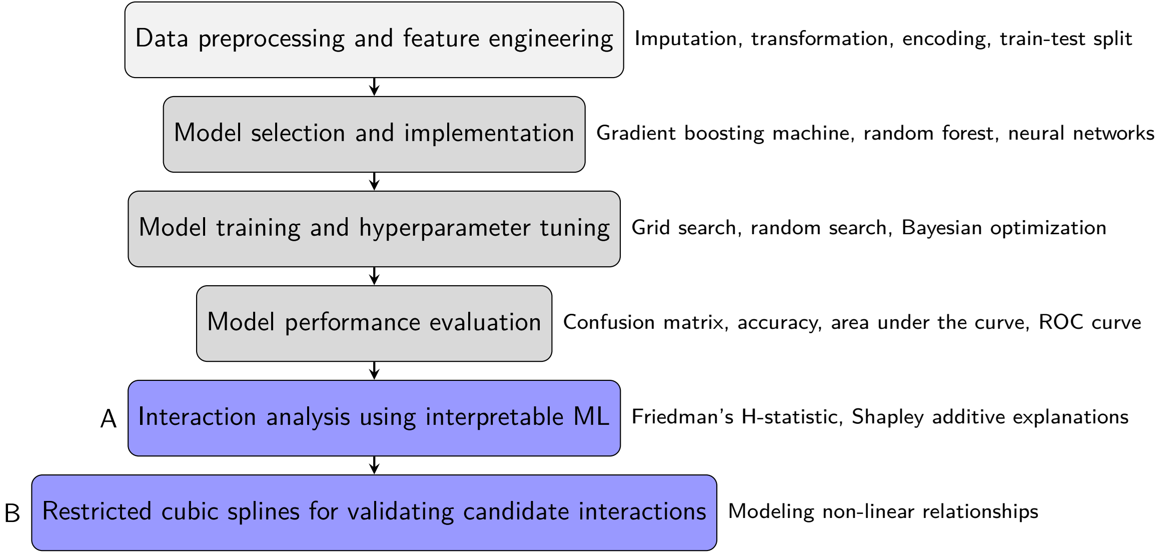
- Regression models (logistic, linear)
  - Model formulation: $log(\frac{P(Y=1)}{P(Y=0)}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \cdot X_2) + \sum \beta_i X_i + \epsilon$
    - Y: binary outcome
    - $X_1 \cdot X_2$: interaction term capturing the combined effect of two variables
    - β_3: quantifies the strength and direction of the interaction
  - Key considerations:
    - Requires pre-specification of interaction terms
    - Computationally expensive for exhaustive interaction searches in high-dimensional datasets
- Machine learning (ML)
  - Handles large-scale data and uncovers complex, non-linear interactions
  - More flexible compared to traditional regression models for interaction detection
- Challenge
  - Formal comparisons and evaluations of ML for interaction assessments with biobank-scale multimodal data have not been fully examined

Objective

- Develop a ML workflow for detecting genetic-clinical interactions in high-dimensional, large-scale datasets
- Apply the workflow to explore the relationship between genetic predisposition to an outcome and clinical risk factors
- Benchmark ML algorithms with a focus on model interpretability and clinical relevance of results

2) Study design and workflow

2) Study design and workflow



Data preprocessing and feature engineering — Imputation, transformation, encoding, train-test split

Model selection and implementation — Gradient boosting machine, random forest, neural networks

Model training and hyperparameter tuning — Grid search, random search, Bayesian optimization

Model performance evaluation — Confusion matrix, accuracy, area under the curve, ROC curve

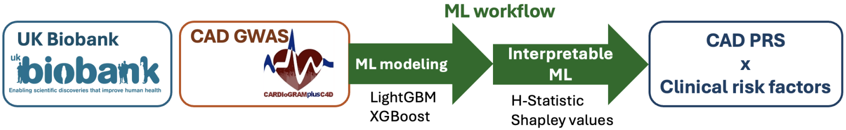
A Interaction analysis using interpretable ML — Friedman's H-statistic, Shapley additive explanations

B Restricted cubic splines for validating candidate interactions — Modeling non-linear relationships

3) Illustrative example

3) Illustrative example

- Evaluate whether the interactions between Coronary Artery Disease (CAD) PRS and clinical risk factors further explain risk for incident Myocardial Infarction (MI) using multiple ML approaches
  - Light gradient boosting machine (LightGBM), extreme gradient boosting (XGBoost), random forest (RF), symbolic regression (SR), neural networks (NNs)



Dataset overview

- Dataset: UK Biobank (UKB)[2]
- Endpoint: incident Myocardial Infarction (MI) in 323,267 individuals of European ancestry, free of atherosclerotic cardiovascular disease (ASCVD)[3] and not on lipid-lowering medications at baseline
  - A total of 4,598 (1.4%) participants experienced an MI[4]
- CAD PRS: computed for each participant using 241 conditionally independent genome-wide significant SNVs identified in a recent GWAS from CARDIoGRAMplusC4D Consortium[5] (a large-scale meta-analysis with over 1 million participants)

[2] A prospective population-based study in the United Kingdom, including over half a million participants aged 40 to 69 at recruitment (2006–2010), collecting comprehensive data on environmental and lifestyle factors, genetics, biomarkers, proteomics, metabolomics, imaging, and electronic health records

[3] Prior MI, CAD diagnosis, stroke, or peripheral vascular disease

[4] Data updated to mid-2021

[5] Coronary Artery Disease Genome-Wide Replication and Meta-analysis (CARDIoGRAM) plus the Coronary Artery Disease (C4D) Genetics

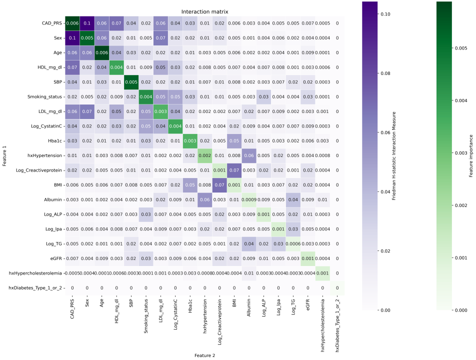
Clinical risk factors

- A comprehensive set of clinical risk factors was examined for potential interactions with CAD PRS, including:
  - Demographic: age, sex
  - Physiological: body mass index (BMI), systolic blood pressure (SBP)
  - Behavioral/lifestyle: smoking status
  - Medical history: history of hypertension, history of hypercholesterolemia, history of diabetes
  - Biomarkers: low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), triglycerides (TG), c-reactive protein (CRP), cystatin c, lipoprotein(a) (Lp(a)), albumin, alkaline phosphatase (ALP), hbA1c, eGFR
- Model training, hyperparameter tuning, and model performance evaluation were conducted (results not shown)

4) Results

Results from part A: Friedman’s H-statistic for interaction terms

- H-statistic quantifies the interaction strength between predictors by measuring the proportion of prediction variance attributed to their interaction
  - Total interaction: measures how much a predictor interacts with all other predictors
  - Pairwise interaction: measures the interaction strength between two specific predictors



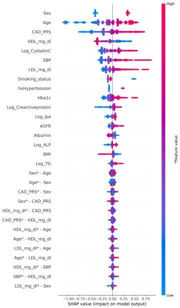
	CAD_PRS	Sex	Age	HDL_mg_dl	SBP	Smoking_status	LDL_mg_dl	Log_CystatinC	Hba1c	hxHypertension	Log_Creactiveprotein	BMI	Albumin	Log_ALP	Log_lpa	Log_TG	eGFR	hxHypercholesterolemia	hxDiabetes_Type_1_or_2
CAD_PRS	0.006	0.1	0.06	0.07	0.04	0.02	0.06	0.04	0.03	0.01	0.02	0.006	0.003	0.004	0.005	0.005	0.007	0.0005	0
Sex	0.1	0.005	0.06	0.02	0.01	0.005	0.07	0.02	0.02	0.01	0.02	0.005	0.001	0.004	0.006	0.002	0.004	0.0004	0
Age	0.06	0.06	0.006	0.04	0.03	0.02	0.02	0.02	0.01	0.003	0.005	0.006	0.002	0.002	0.004	0.001	0.009	0.0001	0
HDL_mg_dl	0.07	0.02	0.04	0.004	0.01	0.009	0.05	0.03	0.02	0.008	0.005	0.007	0.004	0.007	0.008	0.001	0.003	0.0006	0
SBP	0.04	0.01	0.03	0.01	0.005	0.02	0.02	0.02	0.01	0.007	0.01	0.008	0.002	0.003	0.007	0.02	0.02	0.0003	0
Smoking_status	0.02	0.005	0.02	0.009	0.02	0.004	0.05	0.05	0.03	0.01	0.01	0.005	0.003	0.03	0.004	0.004	0.03	0.0001	0
LDL_mg_dl	0.06	0.07	0.02	0.05	0.02	0.05	0.003	0.04	0.02	0.002	0.002	0.007	0.02	0.007	0.009	0.002	0.003	0.001	0
Log_CystatinC	0.04	0.02	0.02	0.03	0.02	0.05	0.04	0.004	0.01	0.002	0.004	0.02	0.009	0.004	0.007	0.007	0.009	0.0003	0
Hba1c	0.03	0.02	0.01	0.02	0.01	0.03	0.02	0.01	0.003	0.02	0.005	0.05	0.01	0.005	0.003	0.002	0.006	0.003	0
hxHypertension	0.01	0.01	0.003	0.008	0.007	0.01	0.002	0.002	0.02	0.002	0.01	0.008	0.06	0.005	0.02	0.005	0.004	0.0008	0
Log_Creactiveprotein	0.02	0.02	0.005	0.005	0.01	0.01	0.002	0.004	0.005	0.01	0.001	0.07	0.003	0.02	0.02	0.001	0.02	0.0004	0
BMI	0.006	0.005	0.006	0.007	0.008	0.005	0.007	0.02	0.05	0.008	0.07	0.001	0.01	0.003	0.007	0.02	0.02	0.0004	0
Albumin	0.003	0.001	0.002	0.004	0.002	0.003	0.02	0.009	0.01	0.06	0.003	0.01	0.0009	0.009	0.005	0.04	0.009	0.01	0
Log_ALP	0.004	0.004	0.002	0.007	0.003	0.03	0.007	0.004	0.005	0.005	0.02	0.003	0.009	0.001	0.005	0.02	0.01	0.0003	0
Log_lpa	0.005	0.006	0.004	0.008	0.007	0.004	0.009	0.007	0.003	0.02	0.02	0.007	0.005	0.005	0.001	0.03	0.005	0.0004	0
Log_TG	0.005	0.002	0.001	0.001	0.02	0.004	0.002	0.007	0.002	0.005	0.001	0.02	0.04	0.02	0.03	0.0006	0.003	0.0003	0
eGFR	0.007	0.004	0.009	0.003	0.02	0.03	0.003	0.009	0.006	0.004	0.02	0.02	0.009	0.01	0.005	0.003	0.001	0.0004	0
hxHypercholesterolemia	0.0005	0.0004	0.0001	0.0006	0.0003	0.0001	0.001	0.0003	0.003	0.0008	0.0004	0.0004	0.01	0.0003	0.0004	0.0003	0.0004	0.001	0
hxDiabetes_Type_1_or_2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

XGBoost

- Green cells: total interactions
- Purple cells: pairwise interactions
- Interaction strength increases with color intensity

Results from part A: Shapley additive explanations (SHAP) interaction values

- SHAP is a game-theory-based method for explaining ML model outputs by assigning an importance value to each predictor for a specific prediction
    - The contribution of each predictor can be further decomposed into main effects and pairwise interaction effects



XGBoost

- X-axis: represents the SHAP values for each predictor
- Y-axis: lists the predictors included in the model, arranged vertically by importance (high to low)
- Color gradient: shows the predictor's value, where darker red correspond to higher values

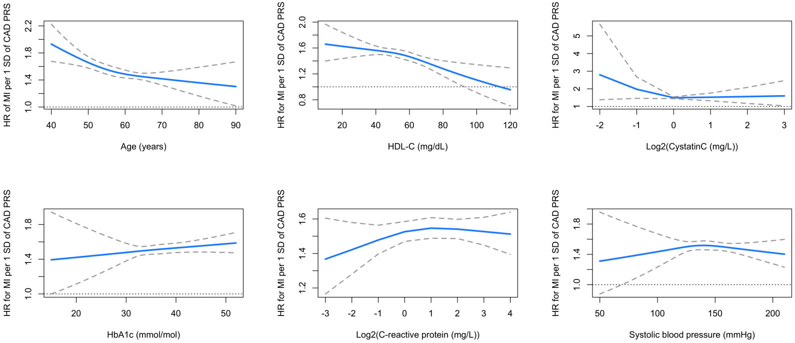
Results from part A: Concordance of PRS-clinical interactions across ML models

Algorithms	LightGBM		XGBoost	
Interactions[a]	H-statistic	SHAP	H-statistic	SHAP
PRS × Sex	1[b]	1	1	1
PRS × HbA1c	2	8	7	7
PRS × HDL-C	3	4	2	2
PRS × SBP	4	3	6	4
PRS × Smoking	5	5	9	
PRS × Age	6	2	3	3
PRS × LDL-C	7	7	4	6
PRS × CRP	8	9	8	9
PRS × CystatinC	9	10	5	5
PRS × hxHTN	10	6		
PRS × eGFR				8

[a] Top-ranked interactions based on importance were evaluated and compared

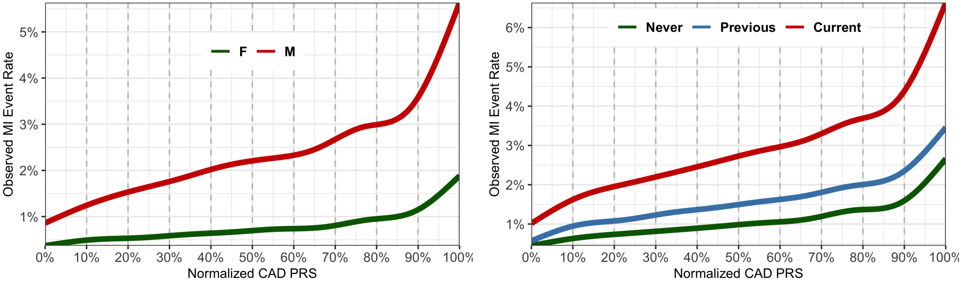
[b] Ranks of interactions within each model

Results from part B: Restricted cubic splines for key interactions between CAD PRS and continuous variables



- Negative interactions were observed between CAD PRS and increased age, HDL-C, and Cystatin C whereas high CAD PRS yielded joint positive associations with HbA1c

Results from part B: Event rate of MI across CAD PRS stratified by categorical variables



(a) Sex

(b) Smoking status

- Joint risk increases were observed in males and current smokers with a high CAD PRS

5) Summary and discussion

5) Summary and discussion

- Most PRS-clinical interactions identified by the ML models for predicting myocardial infarction risk were consistent and further assessed using restricted cubic splines to validate non-linear relationships
- ML-driven screening allowed identifying and validating interactions that had not been defined a priori
- This study demonstrated the benefits of using ML to detect genetic-clinical interactions, enhancing both hypothesis generation and patient risk stratification

Open source ML algorithms and resources for interaction identification

- ML algorithms
  - Gradient boosting machine: LightGBM v4.5.0 (https://github.com/microsoft/LightGBM)
  - Extreme gradient boosting: XGBoost v2.1.4 (https://github.com/dmlc/xgboost)
  - Symbolic regression: Feyn (QLattice algorithm) v3.4.0 (https://github.com/abzu-ai/QLattice-clinical-omics)
- Interpretable ML
  - Friedman's H-statistic: artemis v0.1.5 (https://github.com/pyartemis/artemis)
  - Shapley additive explanations: SHAP v0.46.0 (https://github.com/shap/shap)
  - Restricted cubic splines: interactionRCS v0.1.1 (https://github.com/cran/interactionRCS)
- Others
  - Python modules for ML: scikit-learn v1.5.2 (https://github.com/scikit-learn/scikit-learn)

References

- Schwarzerova J, Hurta M, Barton V, et al. A perspective on genetic and polygenic risk scores—advances and limitations and overview of associated tools. Briefings in Bioinformatics. 2024. doi:10.1093/bib/bbae240
- Marston NA, Pirruccello JP, Melloni GEM, et al. Predictive Utility of a Coronary Artery Disease Polygenic Risk Score in Primary Prevention. JAMA Cardiol. 2023;8(2):130–137. doi:10.1001/jamacardio.2022.4466
- Ke G, Wang S, Yang Q, et al. LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems (Vol. 30). 2017. doi:10.1109/NIPS.2017.5590
- Chen T and Guestrin C. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 785–794. 2016. doi:10.1145/2939672.2939785
- Friedman JH and Popescu BE. Predictive learning via rule ensembles. The Annals of Applied Statistics. JSTOR, 916–54. 2008. doi:10.1214/07-AOAS148
- Lundberg SM and Lee SI. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems (Vol. 30). 2017. doi:10.48550/arXiv.1705.07874

Thanks for your attention!

Contact:

ylai4@bwh.harvard.edu

timi.org/biostatistics/

@TimiStudyGroup

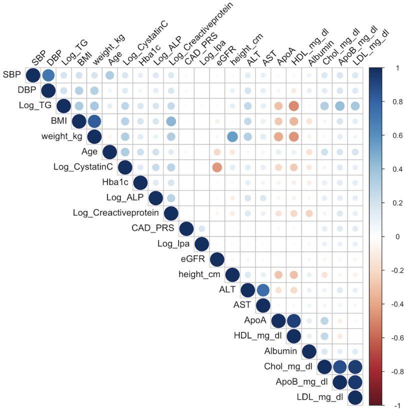
Appendix

PRS estimation

- Polygenic Risk Score (PRS) quantifies an individual's genetic predisposition to a specific trait or disease based on the cumulative effect of multiple genetic variants within a population
- A PRS of an individual j is calculated as a weighted sum of risk alleles across independent genome-wide statistically significant single-nucleotide variants (SNVs):
  - $PRS_j = \sum_{i=1}^{N} \beta_i G_{ij}$
  - where N is the total number of SNVs identified from genome-wide association studies (GWAS), β_i represents the effect size of SNV_i, and G_{ij} denotes the number of risk alleles of SNV_i that individual j carries

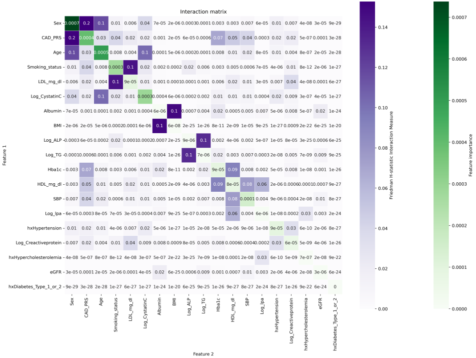
Results: Correlations of pairwise variables

- Variables with a correlation coefficient ≥ 0.7 were removed prior to modeling to reduce multicollinearity

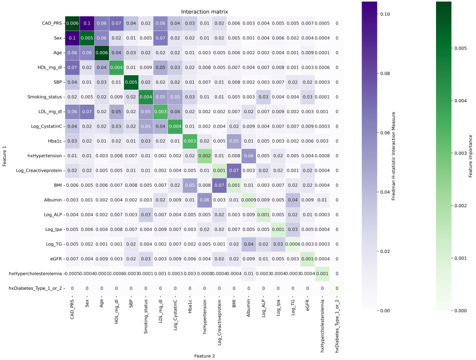


Results: Friedman's H-statistic for interaction terms

- H-statistic quantifies the interaction strength between a pair of predictors by assessing the proportion of prediction variance attributed to their interaction



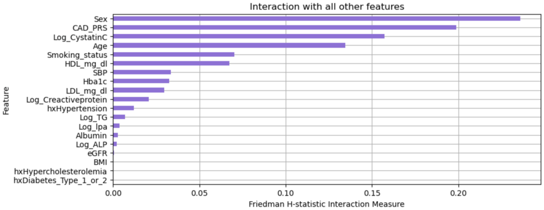
(a) LightGBM



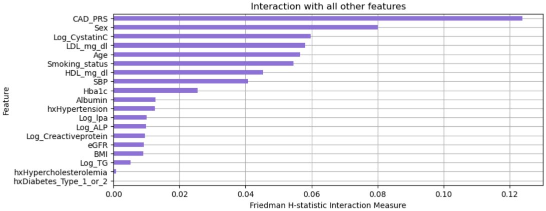
(b) XGBoost

Results: Friedman’s H-statistic for total interactions

- The total interaction measure quantifies the extent to which a predictor interacts with all other predictors in the model



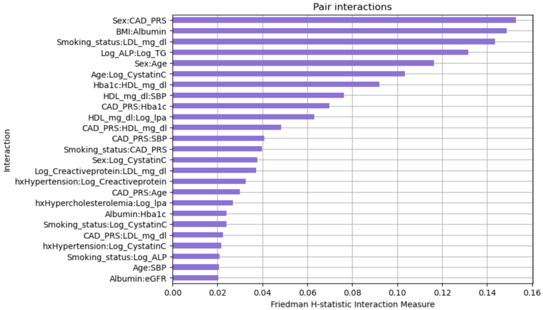
(a) LightGBM



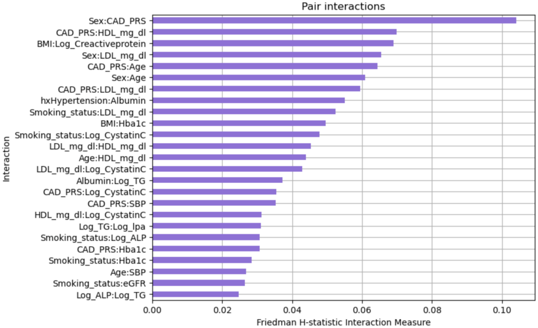
(b) XGBoost

Results: Friedman’s H-statistic for pairwise interactions

- A pairwise interaction measure evaluates the presence and magnitude of interaction between two specific predictors within the model

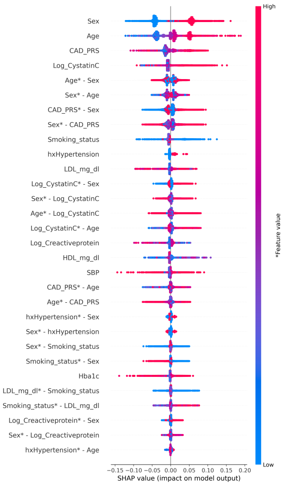


(a) LightGBM

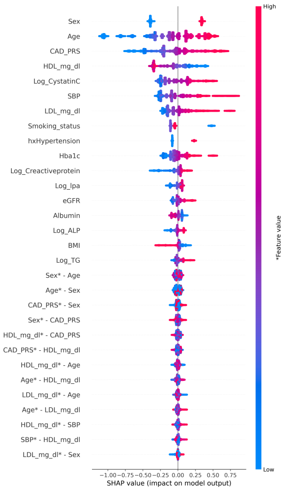


(b) XGBoost

Results: Shapley additive explanations (SHAP) interaction values



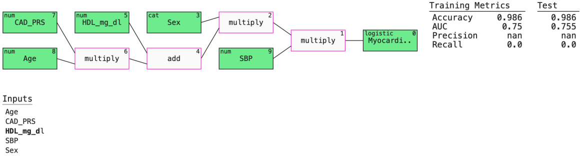
(a) LightGBM



(b) XGBoost

Results: Symbolic regression

- Symbolic regression is an evolutionary algorithm-based technique that searches for the optimal mathematical expression to describe a given dataset by combining mathematical operators, variables, and constants, without assuming a predefined model structure

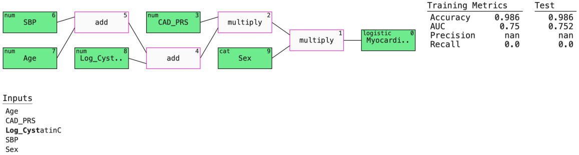


Training Metrics		Test
Accuracy	0.986	0.986
AUC	0.75	0.755
Precision	nan	nan
Recall	0.0	0.0

Inputs

Age
CAD_PRS
HDL_mg_dl
SBP
Sex

(a) Best model from SR



Training Metrics		Test
Accuracy	0.986	0.986
AUC	0.75	0.752
Precision	nan	nan
Recall	0.0	0.0

Inputs

Age
CAD_PRS
Log_CystatinC
SBP
Sex

(b) Second best model from SR