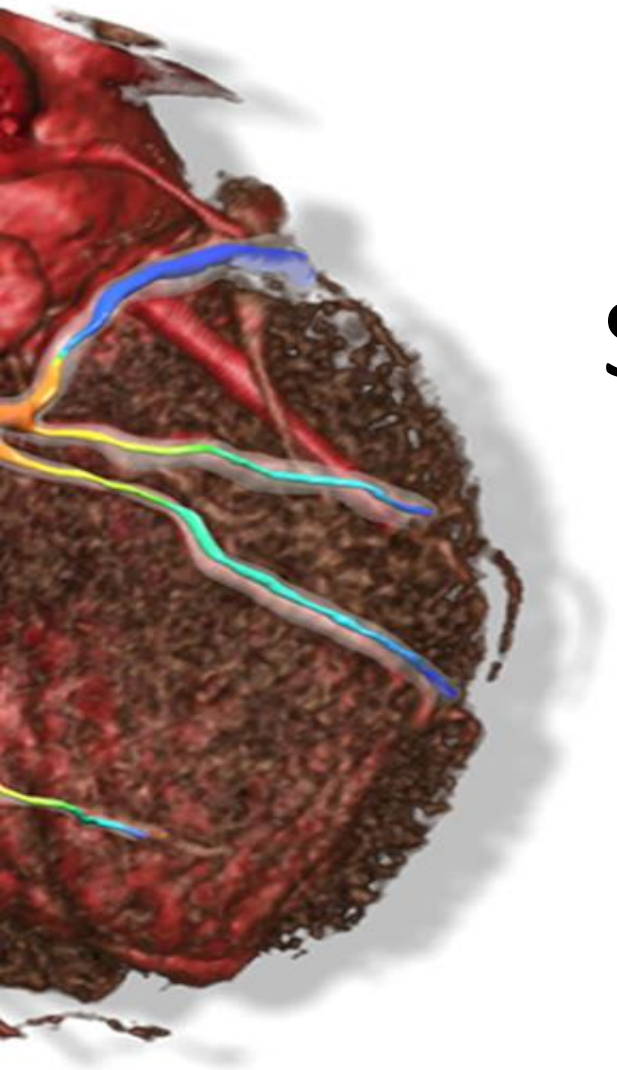
A SMARTool project workshop

CAD RISK PREDICTION AND STRATIFICATION: THE ICT APPROACH



# SMARTool clinical/molecular models

Nikolaos Tachos

Tuesday 6th November 2018

CNR Research Area Campus
Building A, Room 27
via Moruzzi, 1 Pisa - Italy

Horizon 2020
689068

# Aim

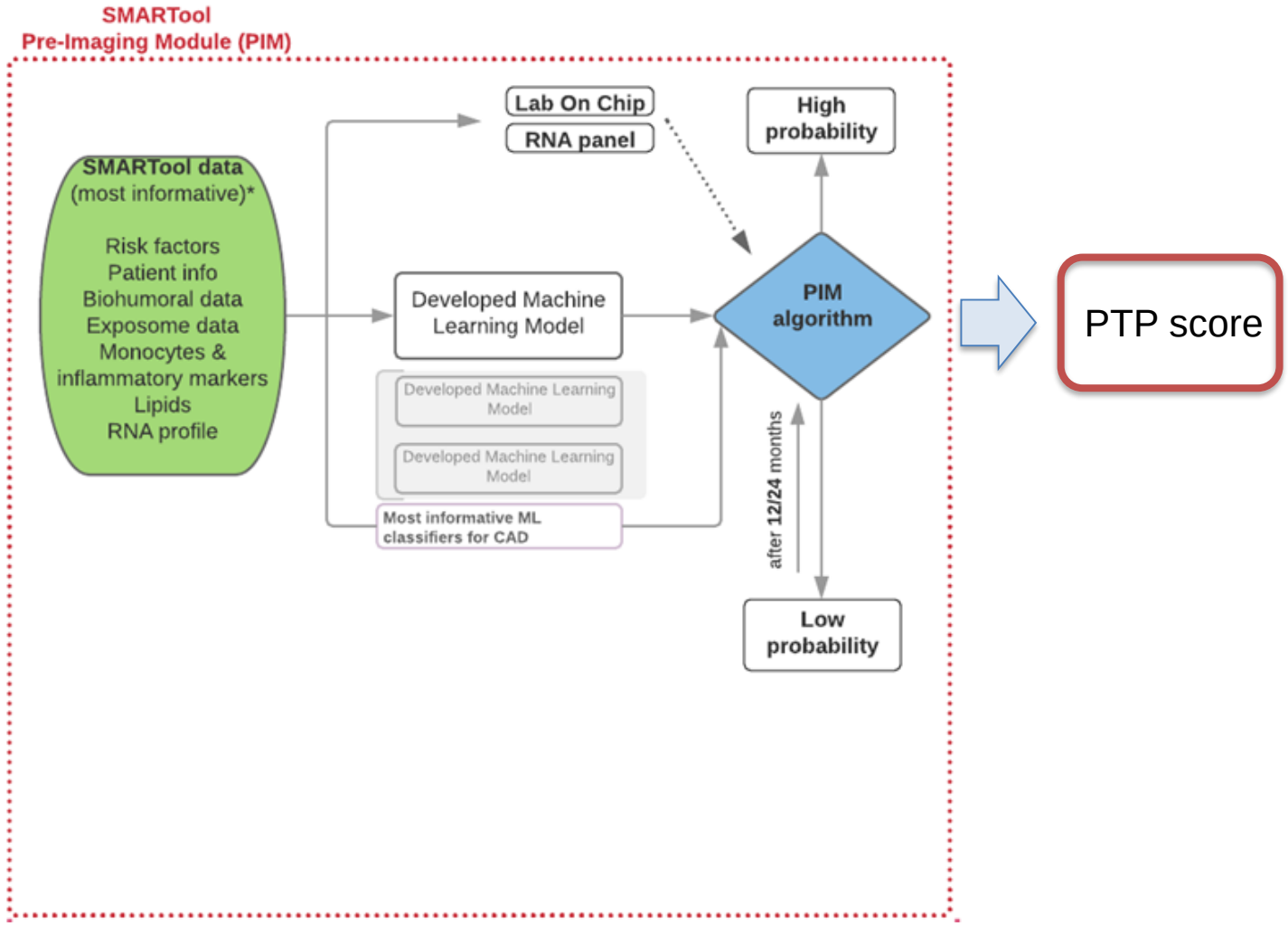
WP3, Task 3.4
Clinical/molecular ML models

To design and develop a ML model integrating multiple categories of biological non-imaging data towards precise risk stratification in CAD

To identify the most informative features from genomics transcriptomics, inflammatory data, lipid profile

To validate the risk stratification model on retrospective and prospective data

# SMARTool Pre-Imaging Module



# State of the Art

**Pre-test probability models of CAD based on Demographics, Risk Factors, Symptoms, ECG and conventional Biomarkers**

| STUDY | DATASET | METHODS | Acc. (%) | Sens. (%) | Spec. (%) |
|---|---|---|---|---|---|
| **Anooj, 2012** | The UCI Heart Disease Dataset<br>CAD: $n = 165$, Normal: $n = 138$<br>Demographics, Risk Factors, ECG, Symptoms | **Classification:** Automated generation of weighted fuzzy rules - Mamdani fuzzy inference system<br>**Evaluation:** Training-Test sets | 62.4 | 44.7 | 76.6 |
| **Nahar et al., 2013** | The UCI Heart Disease Dataset<br>CAD: $n = 165$, Normal: $n = 138$<br>Demographics, Risk Factors, ECG, Symptoms | **Feature Selection:** CFS, Knowledge-based feature selection<br>**Classification:** SVM<br>**Evaluation:** 10-fold cross-validation | 84.5 | 89.1 | - |
| **C. B. Fordyce, 2017** | The PROMISE Minimal-Risk Tool<br>CAD= 3388, Normal = 1243<br>Demographics, Risk Factors, Symptoms, HDL-C | **Feature Selection:** Knowledge-based feature selection<br>**Classification:** multivariable logistic regression model<br>**Evaluation:** Hosmer-Lemeshow calibration on validation set of 1544 pts | 72.6 | | |

# State of the Art

## Elashoff et al, BASED ON CATHGEN & PREDICT study

The **Corus CAD algorithm** was developed via a combination of microarray and RT-PCR gene expression data analysis, collected from age and sex-matched patients with symptoms suggestive of CAD.
The Corus CAD test incorporates **patient-specific gene expression**, **age**, and **sex data**.

- Feature Selection: Unsupervised cluster analysis and identification of meta-genes.
- Classification:
  - Age, sex, and gene expression are weighted and incorporated into the Corus CAD algorithm
  - Ridge linear regression.

Corus CAD demonstrated a **high sensitivity 85%** and **negative predictive value 83%**.

Elashoff MR, et al. *BMC Med Genomics 2011*; 4 (1):26.

## Dogan et al.,2018 BASED ON DNA AND SNP DATA

**Based on the Framingham Heart Study Data**

**Training Set ($n = 1545$)** **Test Set ($n = 142$)**

### Dataset

**Genome-wide DNA methylation and SNP data**
**Phenotype**
Age, gender, systolic blood pressure (SBP), high-density lipoprotein (HDL) cholesterol level, total cholesterol level, hemoglobin A1C (HbA1c) level, self-reported smoking status, and the use of statins.

### Model training and Testing

1. Eight Random Forest (RF) classification models were built on the eight sub-datasets using stratified 10-fold cross-validation.

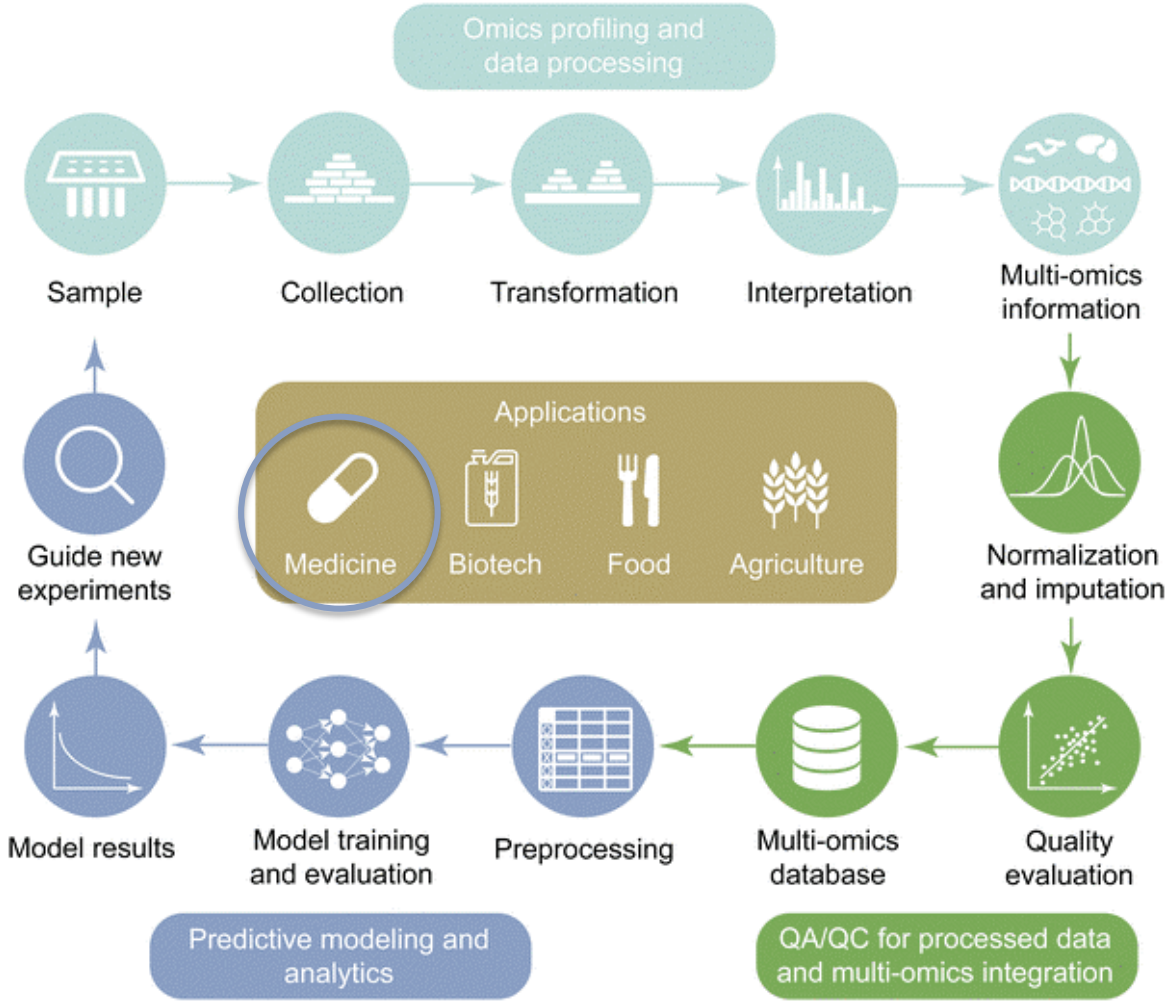
| | Acc. | Sens. | Sp. |
|---|---|---|---|
| Integrative model | 77.5% | 0.75 | 0.80 |
| Conventional CHD risk factor model | 65.4% | 0.42 | 0.89 |

Dogan MV et al., *PLOS ONE*, 2018

# Predictive Modeling through Machine Learning

## End-to-end pipeline of predictive analytics over multi-omics data

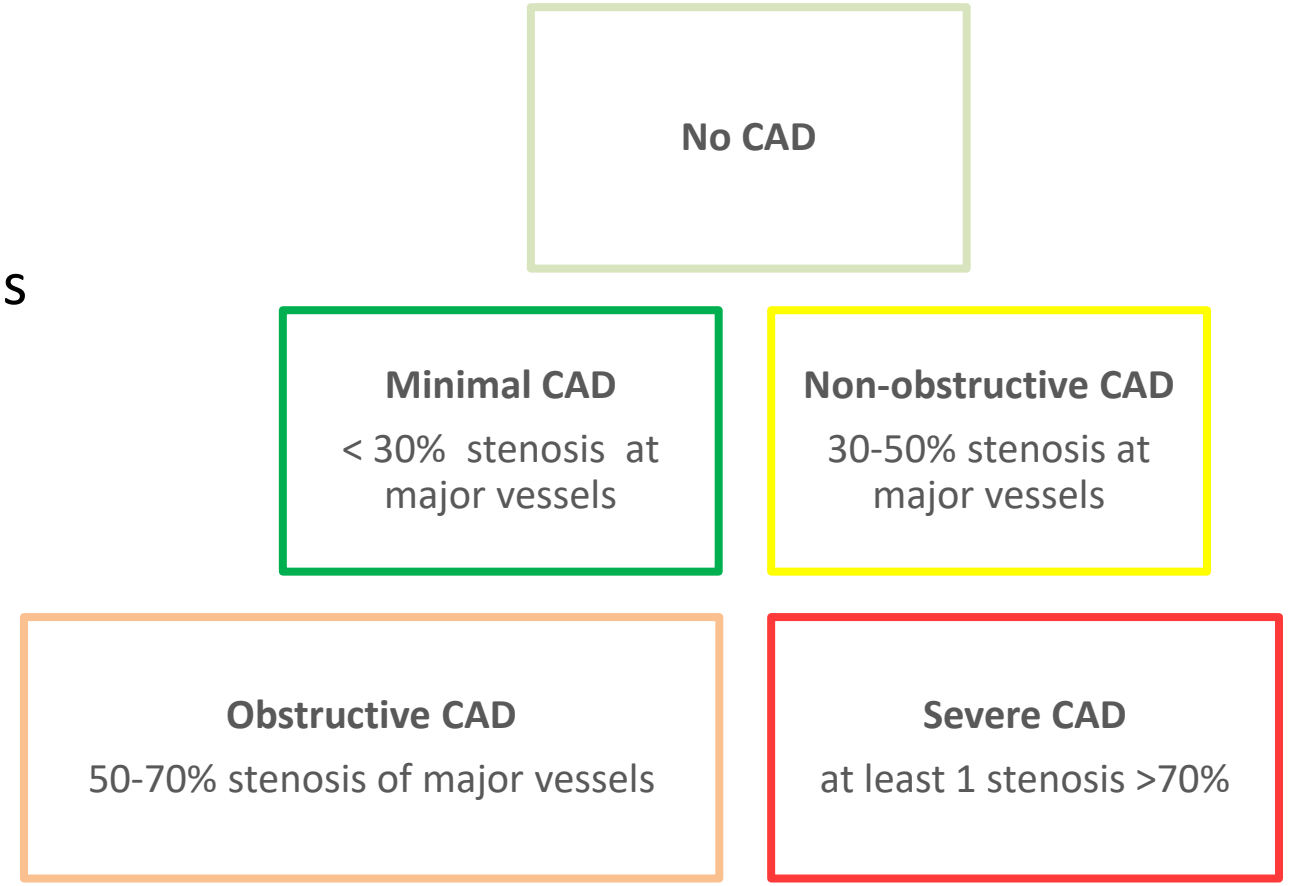
1. **Data Acquisition**
   - transformation, interpretation
2. **Multi-omics Integration**
   - normalization, imputation, quality control
   - integration within a single-omics type or across multi-omics-types
3. **Predictive Modeling**
   - feature selection, dimensionality reduction
   - unsupervised or supervised machine learning




Kim, Minseung, and Ilias Tagkopoulos. "Data integration and predictive modeling methods for multi-omics datasets." *Molecular omics* 14.1 (2018): 8-25.

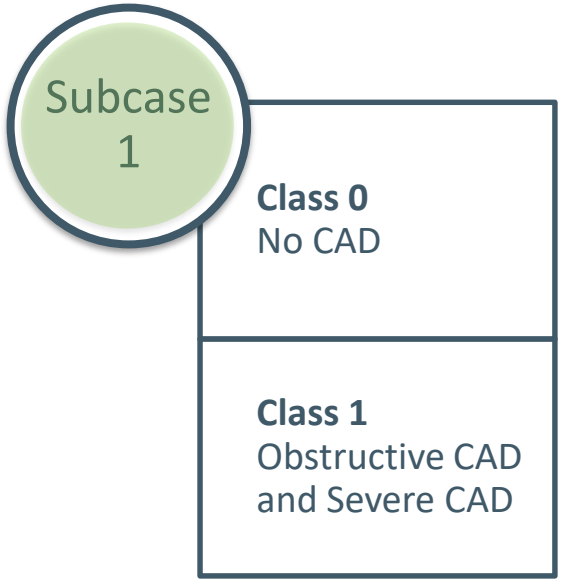
# Problem formulation

- In the PIM module, the CAD risk stratification is formulated as a multiclass classification problem.
- The severity of the disease is represented as a nonlinear parametric function of a confined set of features $f(x) = C_i, x = [x_1, \dots, x_d],\ i = 1, \dots k.$
- Five dominant classes $C_i, i = 1, \dots 5$ have been defined by the **SMARTool experts** based on stenosis severity, as assessed by computed tomography coronary angiography.

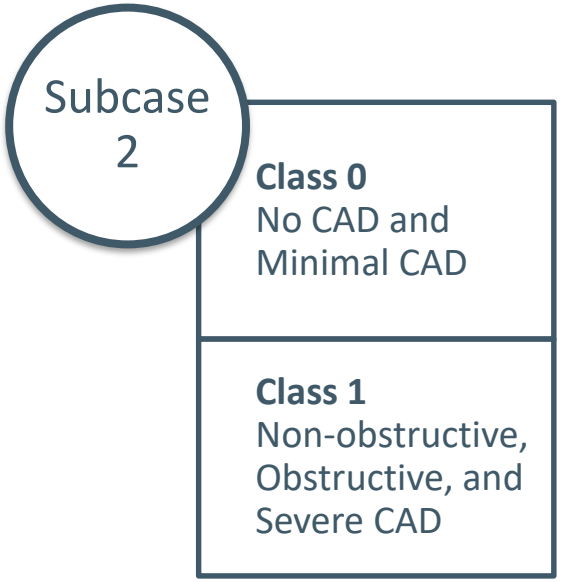


# Problem formulation

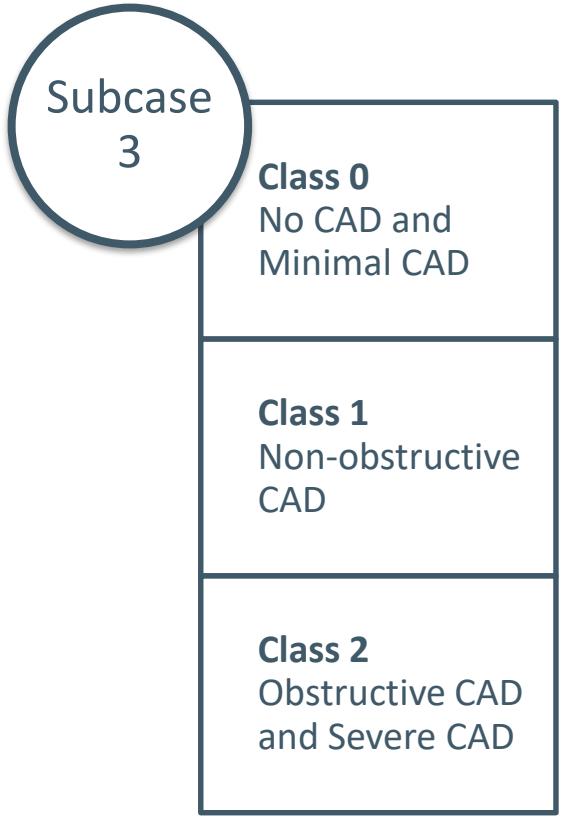
## DEFINITION OF SUBCASES




*2-class problem*




*2-class problem*




*3-class problem*

# Coronary Artery Disease Risk Stratification

## PROBLEM FORMULATION of subcase 1

The binary classification problem is addressed based on stenosis severity of major vessels, as assessed by computed tomography coronary angiography (CCTA).

- **Class 0:** Control subjects
- **Class I:** Obstructive CAD (≥50% stenosis at major vessels)

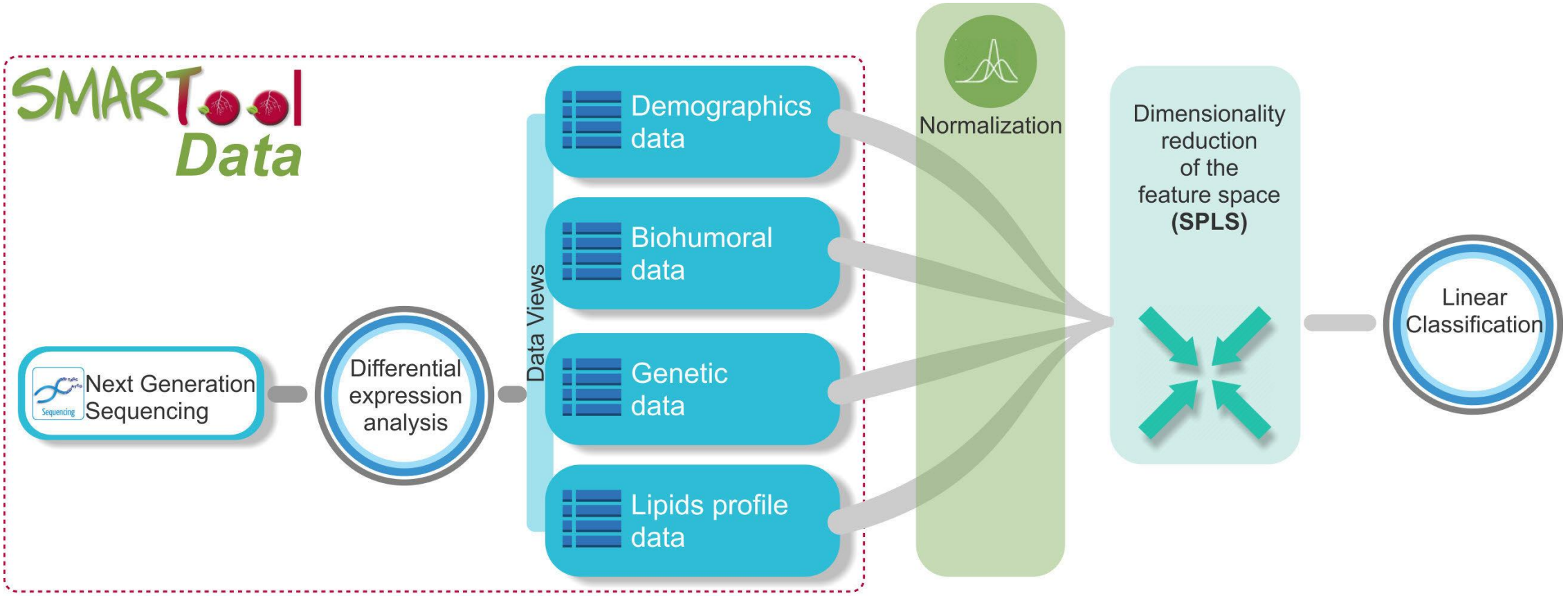
## SMARTool dataset at follow-up

- The total number of annotated patients in follow-up with gene expression is 210pts
- The dataset is reduced to 87pts for subcase 1 problem
  - N=35 control subjects
  - N= 52 cases

# Feature Set Description

| | |
|---|---|
| **Demographics** | Age, Gender |
| **Risk Factors** | Family History of CAD, Hypertension, Diabetes, Dyslipidaemia, Smoking, Obesity, Metabolic Syndrome |
| **Biohumoral data** | Creatinine, Erythrocytes, Glucose, Fibrinogen, HCT, HDL, Haemoglobin, INR, LDL, Leukocytes, MCH, MCV, Platelets, Total Cholesterol, Triglycerides, Uric Acid, aPTT, Alanine Aminotransferase, AlkalinePhosphatase, Aspartate Aminotransferase, Gamma Glutamyl Transferase, High-Sensitivity C-Reactive Protein, Interleukin-6, Leptin |
| **Inflammatory and Monocyte Markers** | ICAM1, VCAM1, CCR2, CCR5, CD11b, CD11b, CD14(++/+), CD14++/CD16+/CCR2+, CD14++/CD16-/CCR2+, CD14+/CD16++/CCR2-, CD163, CD16, CD18, CX3CR1, CXCR4, HLA-DR, MONOCYTE COUNT |
| **Omics Data** | Gene Expression Data, Lipidomics |
| **Symptoms data** | Typical Angina, Atypical Angina, Non Angina Chest Pain, Other Symp-toms, No Symptoms |
| **Exposome data** | Alcohol Consumption, Vegetable Consumption, Physical Activity, Home Environment, Exposition to Pollutants |

# SMARTool Machine Learning pipeline



# CLASSIFICATION PERFORMANCE

**Sparse PLS of demographics and gene expression data**

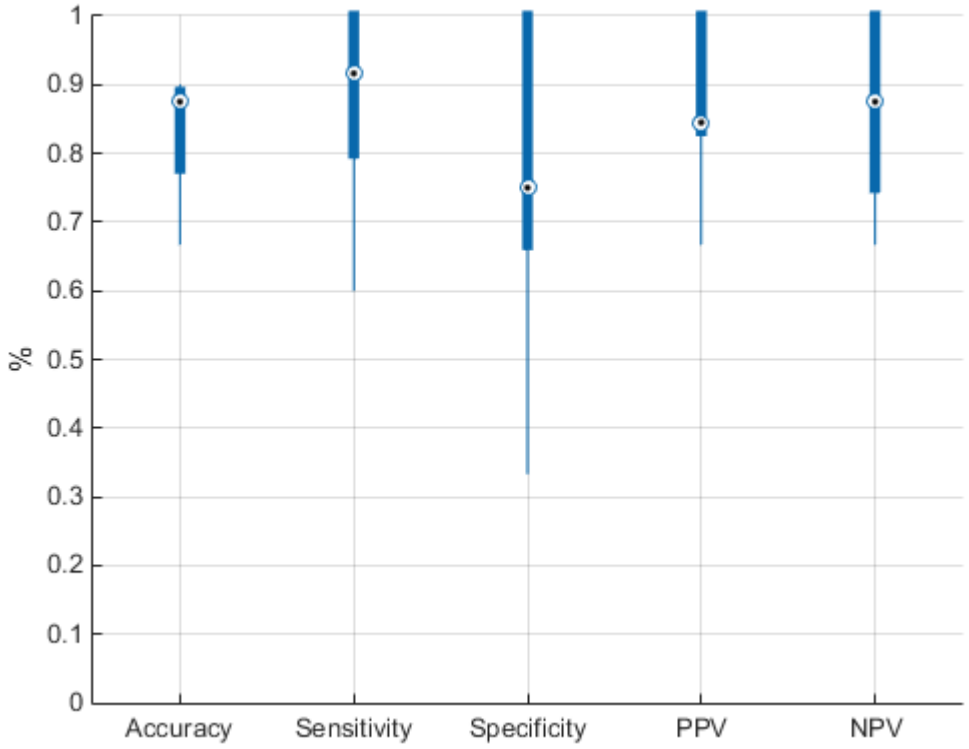
**Evaluation Procedure:** 10-fold cross validation accompanied by an internal 10-fold cross-validation for hyper-parameter tuning .

### Performance Metrics

| | |
|---|---|
| Accuracy | 0.85±0.14 |
| Sensitivity | 0.90±0.14 |
| Specificity | 0.77±0.33 |
| Positive Predictive Value | 0.88±0.16 |
| Negative Predictive Value | 0.87±0.19 |

### Confusion Matrix

| | | Predicted | |
|---|---|---|---|
| | | Class 0 | Class I |
| Actual | Class 0 | 26 | 8 |
| | Class I | 6 | 46 |



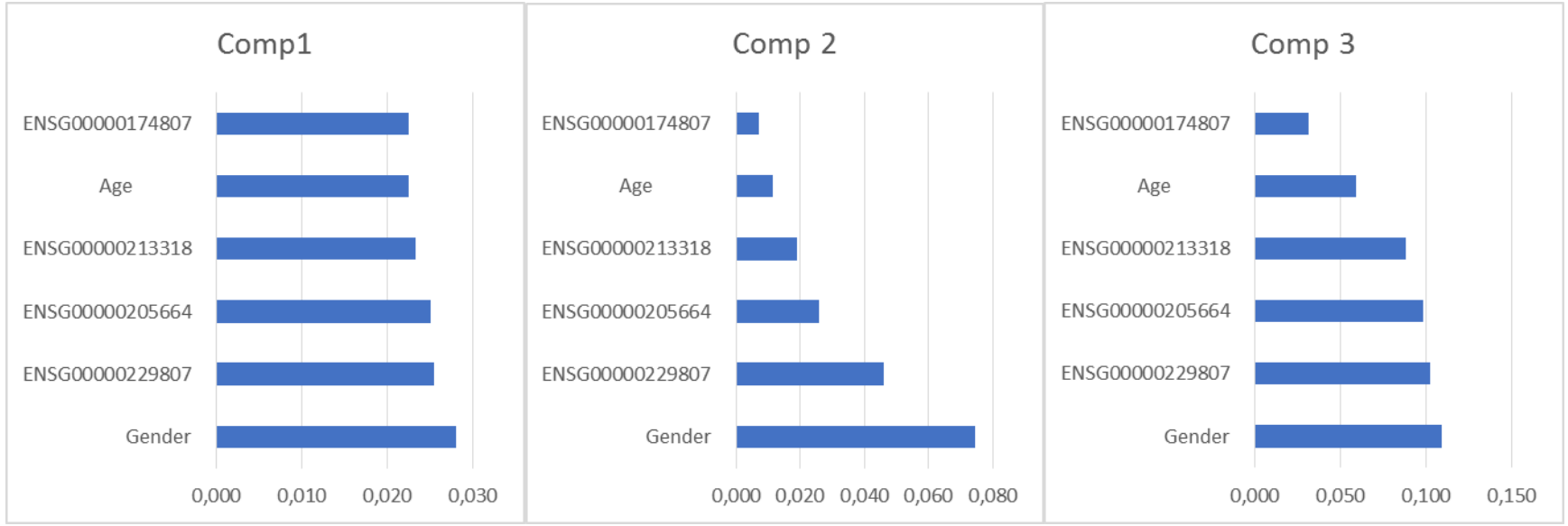
# CLASSIFICATION PERFORMANCE

**Sparse PLS of demographics and gene expression data**

**Evaluation Procedure:** 10-fold cross validation accompanied by an internal 10-fold cross-validation for hyper-parameter tuning .

**Selected variables in each of the 3 components ($K = 3$)**

ENSG00000174807
ENSG00000205664
ENSG00000213318
ENSG00000229807
Age
Gender



# CLASSIFICATION PERFORMANCE

**Logistic regression of demographics and biohumoral data**

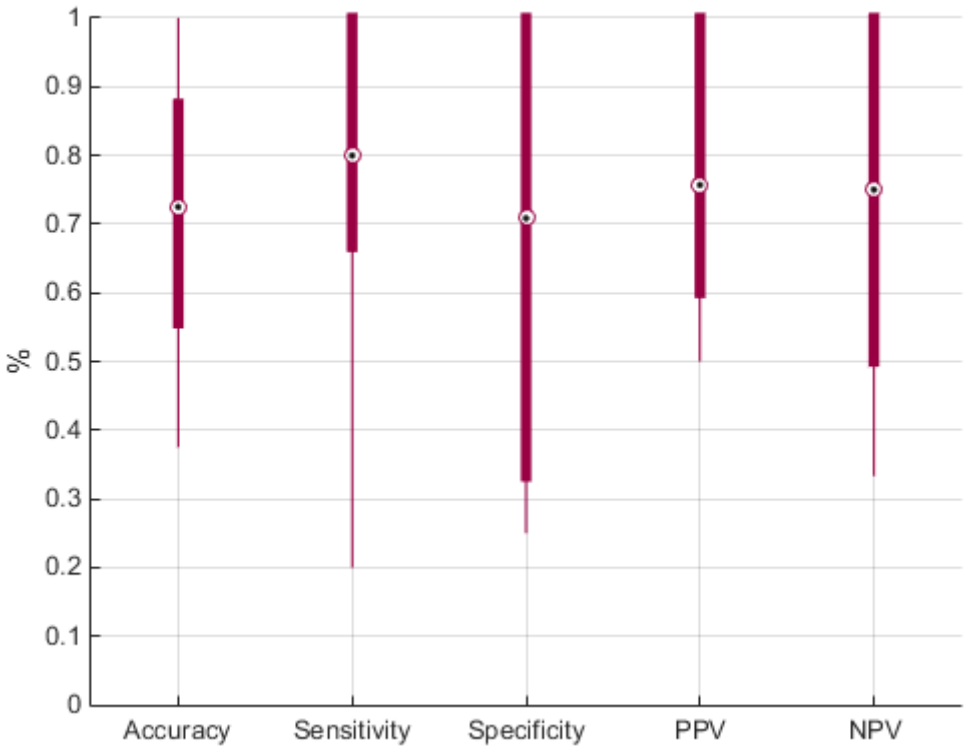
**Evaluation Procedure:** 10-fold cross validation accompanied by an internal 10-fold cross-validation for hyper-parameter tuning .

### Performance Metrics

| | |
|---|---|
| Accuracy | 0.71±0.19 |
| Sensitivity | 0.77±0.24 |
| Specificity | 0.63±0.32 |
| Positive Predictive Value | 0.76±0.19 |
| Negative Predictive Value | 0.70±0.26 |

### Confusion Matrix

| | | Predicted | |
|---|---|---|---|
| | | Class 0 | Class I |
| Actual | Class 0 | 22 | 13 |
| | Class I | 12 | 40 |



# CLASSIFICATION PERFORMANCE

**Linear discriminant analysis (LDA) of demographics and biohumoral data**

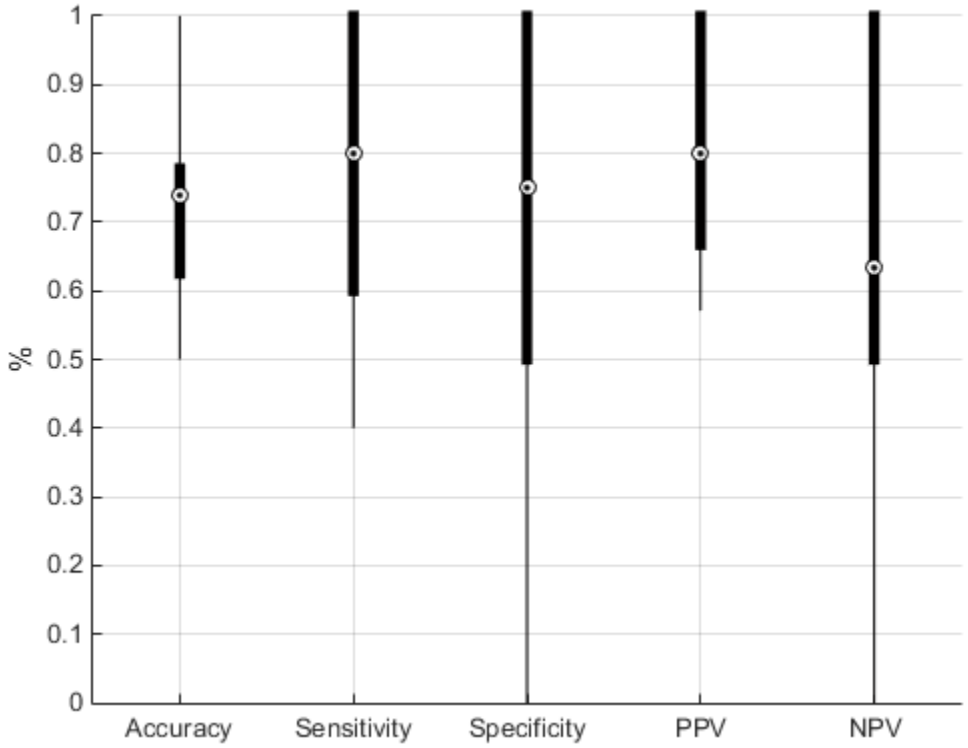
**Evaluation Procedure:** 10-fold cross validation accompanied by an internal 10-fold cross-validation for hyper-parameter tuning .

### Performance Metrics

| | |
|---|---|
| Accuracy | 0.73±0.17 |
| Sensitivity | 0.77±0.20 |
| Specificity | 0.68±0.34 |
| Positive Predictive Value | 0.82±0.17 |
| Negative Predictive Value | 0.65±0.31 |

### Confusion Matrix

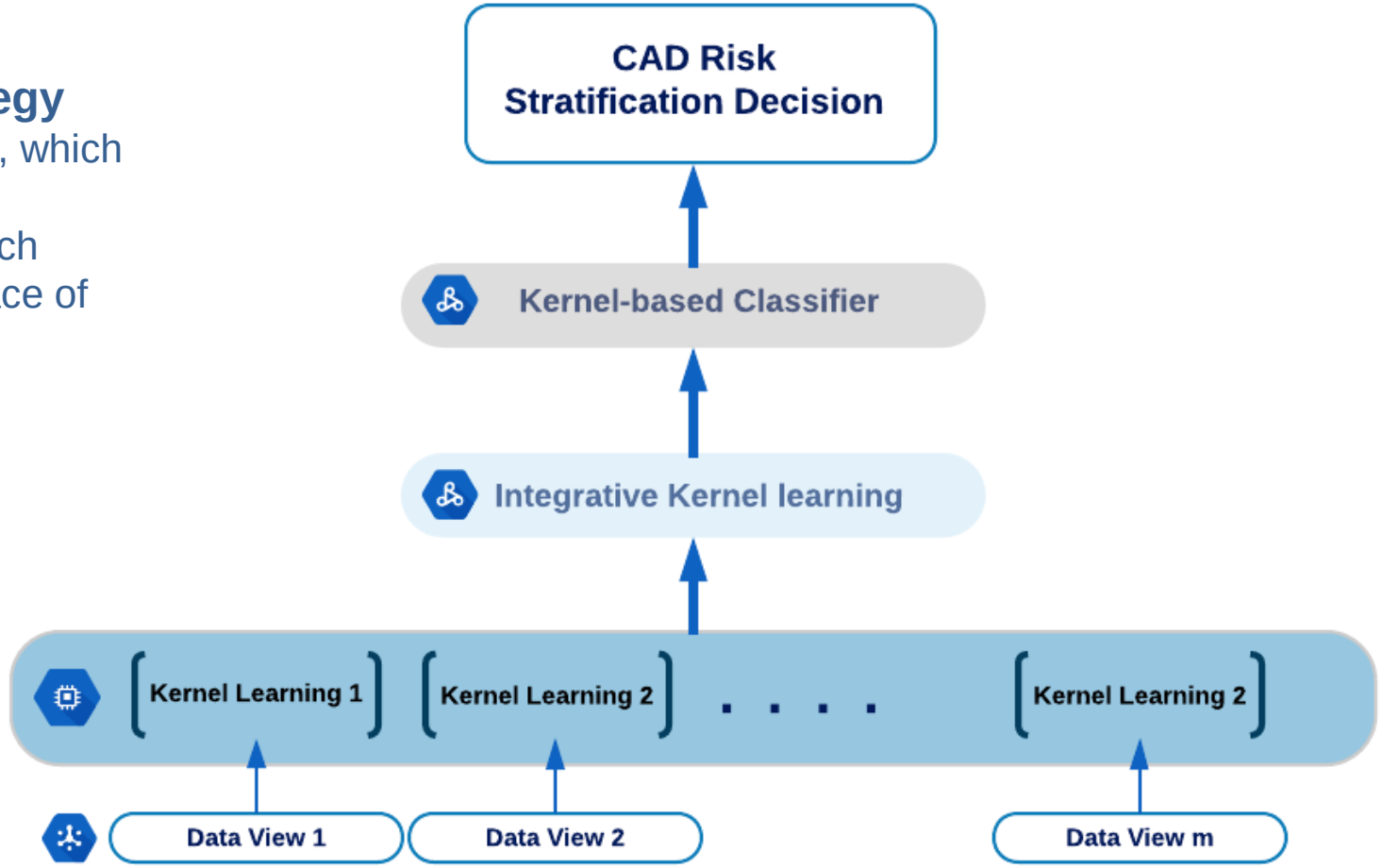
| | | Predicted | |
|---|---|---|---|
| | | Class 0 | Class I |
| Actual | Class 0 | 24 | 11 |
| | Class I | 12 | 40 |



# FUTURE WORK: INTEGRATIVE MACHINE-LEARNING MODEL

**Intermediate data integration strategy**

- a purely nonlinear multi view approach, which is based on multiple kernel learning
- instead of dimensionality reduction, each data view is projected on a feature space of higher dimension




[1]Y. Li, *et al.*, *Briefings in Bioinformatics*, 2016
[2]D. Arneson, *et al.*, *Frontiers in Cardiovascular Medicine,* 2017
[3]S. Min, *et al.*, *Briefings in Bioinformatics*, 2017

# CONCLUSIONS

- A multimodal pipeline has been presented relying on sparse dimensionality reduction techniques and linear classification.
- The model can stratify patients with a high accuracy when demographics and genes are integrated using the SPLS framework.
- The feature set comprised of biohumoral and demographics data produces a lower classification performance.
- A higher-level integration of all data views requires a more sophisticated dimensionality reduction approach which is under development.
- Non-linear data integrative models are also examined for the definition of multiclass problems.