PhD in Intelligenza Artificiale in medicina e innovazione nella ricerca clinica e metodologica

Coordinatore: Prof. Domenico Russo

Identifying of novel gene expression signatures in Acute Myeloid Leukemia (AML) patients: comparison of different methodological approaches

Dottorandi XXXIX ciclo

Adriana Blanda

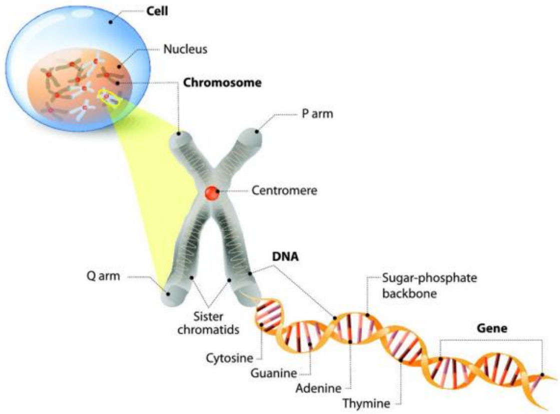
Supervisor: Prof. Russo Domenico

UNIVERSITÀ
DEGLI STUDI
DI BRESCIA

FONDAZIONE IRCCS
ISTITUTO NAZIONALE
DEI TUMORI

OUTLINE:

I. Background
II. Aims
III. Methods
IV. Results
V. Next Steps



I. Background

Acute myeloid leukaemia (AML) is a type of blood cancer. AML starts from the fast and uncontrolled growth of early myeloid blood cells in the bone marrow. The bone marrow is the soft inner part of the bones, where new blood cells are made.

AML is a biologically and clinically heterogeneous clonal disorder of hematopoietic progenitor cells, driven by a complex interplay of **cytogenetic** and molecular aberrations.

Risk stratification in AML is primarily based on cytogenetic abnormalities and recurrent gene mutations, as outlined by **ELN** and **WHO** classifications. These genetic lesions influence leukemogenesis, treatment response, and overall prognosis.

Emerging evidence suggests that the combinatorial effects of specific gene mutations—such as those in NPM1, FLT3, DNMT3A, and others—can modulate disease behavior and therapeutic sensitivity.

Integrative genomic profiling holds the potential to refine prognostic models and identify molecular signatures associated with improved survival outcomes, paving

I. Background

As of 2025, the estimated 5-year survival rate for AML is approximately **29.8% to 32%**

→ AGE

For individuals ***under 60 years*** of age, the 5-year survival rate ranges from **30% to 40%**

For patients ***over 60***, the 5-year survival rate decreases to **less than 20%**

In paediatric cases, particularly those ***under 15 years old***, survival rates can be significantly higher, reaching up to **67%**

II. Aims

1. **To identify and validate one or more gene signatures related to outcome in terms of survival. (Survival Cox Regression)**

2. To identify and validate one or more gene signatures **related to the cytogenetic risk**. (Logistic Regression)

3. To characterize and compare the above identify signatures especially in terms of involved biological pathways by gene ontology and enrichment analysis.

III. Methods

DISCOVERY - TRAINING

Gene Expression Omnibus (GEO) is a public repository maintained by the National Center for Biotechnology Information (NCBI), specifically designed to store and provide open access to gene expression, transcriptomic, and other genomic data.

Dataset: GSE6891*
#pts: 457[15-60y]
#genes: 20888
# event: 290 dead
Median FU: 205 [IQR: 45-272]

*Verhaak RG, Wouters BJ, Erpelinck CA, Abbas S, Beverloo HB, Lugthart S, Löwenberg B, Delwel R, Valk PJ. Prediction of molecular subtypes in acute myeloid leukemia based on gene expression profiling. Haematologica. 2009 Jan;94(1):131-4. doi: 10.3324/haematol.13299. Epub 2008 Oct 6. PMID: 18838472; PMCID: PMC2625407.

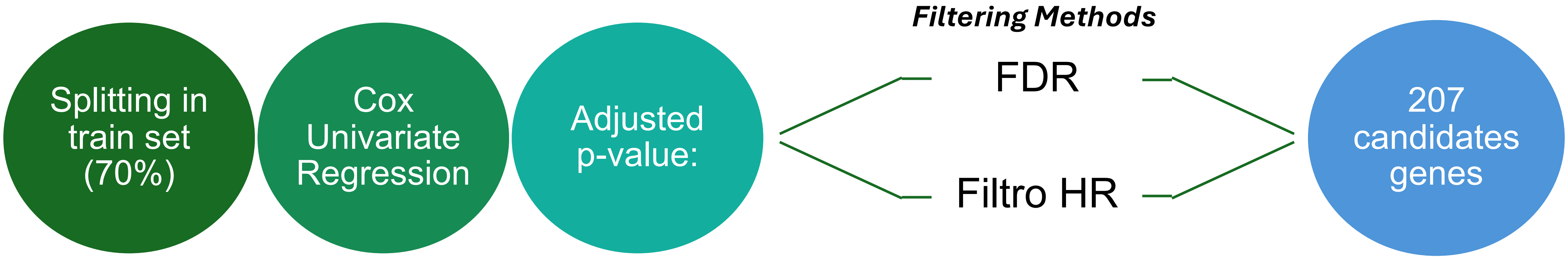
DISCOVERY DATASET GSE6891

Clinic variables		N	%
Gender			
F		228	50%
M		229	50%
Age,median[Q1-Q3]		43[33-53]	
Score			
FABM0		16	4%
FABM1		94	21%
FABM2		104	23%
FABM3		24	5%
FABM4		79	17%
FABM4E		5	1%
FABM5		103	23%
FABM6		6	1%
FABMX		1	0%
FABUNK		8	2%
RAEB		4	1%
RAEB-T		13	3%
Risk			
good		97	21%
intermediate		259	57%
poor		91	20%
unknown		10	2%

Mutation			
npm1			
neg		318	70%
pos		139	30%
evi1			
neg		426	93%
pos		31	7%
n_ras			
neg		411	90%
pos		45	10%
1 NA			
Flt3_tkd			
neg		407	89%
pos		50	11%
Cebpa			
neg		422	92%
pos		31	7%
4 NA			
Flt3_itd			
neg		333	73%
pos		124	27%
K_ras			
neg		453	99%
pos		4	1%
Idh1			
neg		420	92%
pos		34	7%
3 NA			
Idh2			
neg		417	91%
pos		37	8%
3 NA			

III. Methods

GSE6891



Splitting in train set (70%)

Cox Univariate Regression

Adjusted p-value:

Filtering Methods

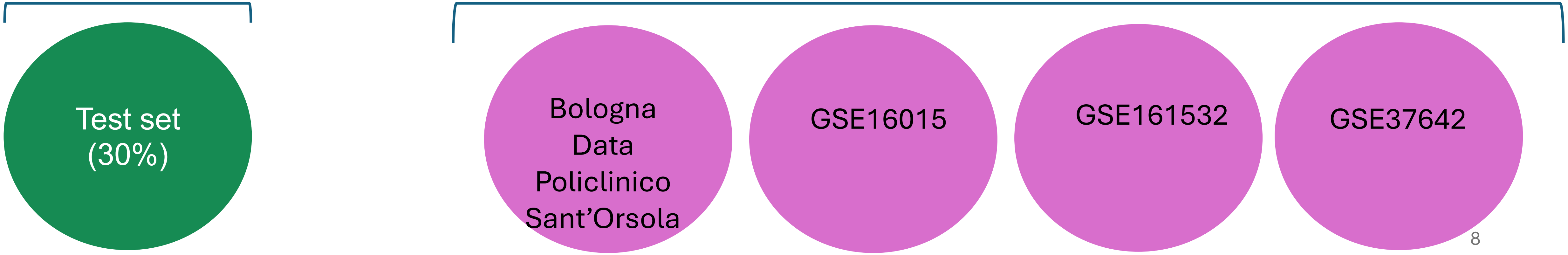
FDR

Filtro HR

207 candidates genes

Internal dataset

External dataset



Test set (30%)

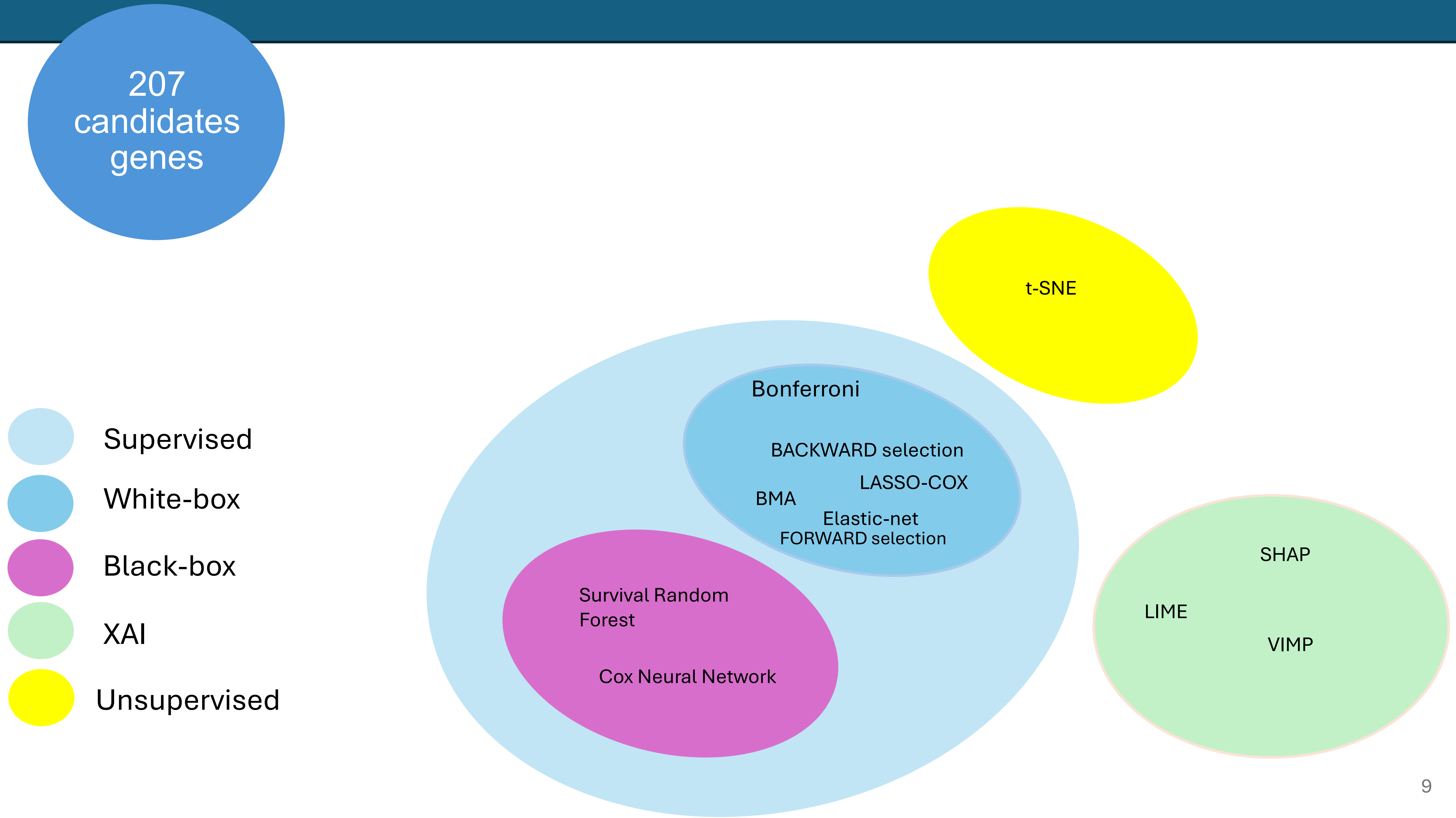
Bologna Data Policlinico Sant’Orsola

GSE16015

GSE161532

GSE37642

III. Methods



207 candidates genes

Supervised

White-box

Black-box

XAI

Unsupervised

t-SNE

Bonferroni

BACKWARD selection

LASSO-COX

BMA

Elastic-net

FORWARD selection

Survival Random Forest

Cox Neural Network

SHAP

LIME

VIMP

III. Methods

Evaluation metrics

C-Index (Concordance Index) measures the discriminative ability of a survival model.

$$\text{C-index} = \frac{\sum_{i,j} 1_{T_j<T_i} \cdot 1_{\eta_j>\eta_i} \cdot \delta_j}{\sum_{i,j} 1_{T_j<T_i} \cdot \delta_j}$$

η_i, the risk score of a unit i

$1_{T_j<T_i} = 1$ if $T_j < T_i$ else 0

$1_{\eta_j>\eta_i} = 1$ if $\eta_j > \eta_i$ else 0

Integrated Brier Score (IBS) evaluates the overall predictive accuracy over time.

$$\text{IBS}(t_{\max}) = \frac{1}{t_{\max}} \int_0^{t_{\max}} BS(t)dt$$

$$BS(t) = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{\left(0 - \hat{S}(t, \vec{x}_i)\right)^2 \cdot 1_{T_i \leq t, \delta_i = 1}}{\hat{G}(T_i^-)} + \frac{\left(1 - \hat{S}(t, \vec{x}_i)\right)^2 \cdot 1_{T_i > t}}{\hat{G}(t)} \right)$$

IV. Results

Bonferroni	LASSO	Elastic-net	Forward	Backward	SRF	BMA
CA13	CA13	CA13	CA13	BCHE	CA13	CA13
CDCP1	CDCP1	CDCP1	TEDC2_AS1	KRT8	CDCP1	TMEM79
	TEDC2_AS1		CDCP1	MADCAM1	TEDC2_AS1	ZGLP1
	TMEM79		PRRG1	PROCR		ZNF311
			ZGLP1	PTHLH		TEDC2_AS1
				EDC2_AS1		CDCP1
				CDCP1		PRRG1
				HEY2		
				TMEM79		

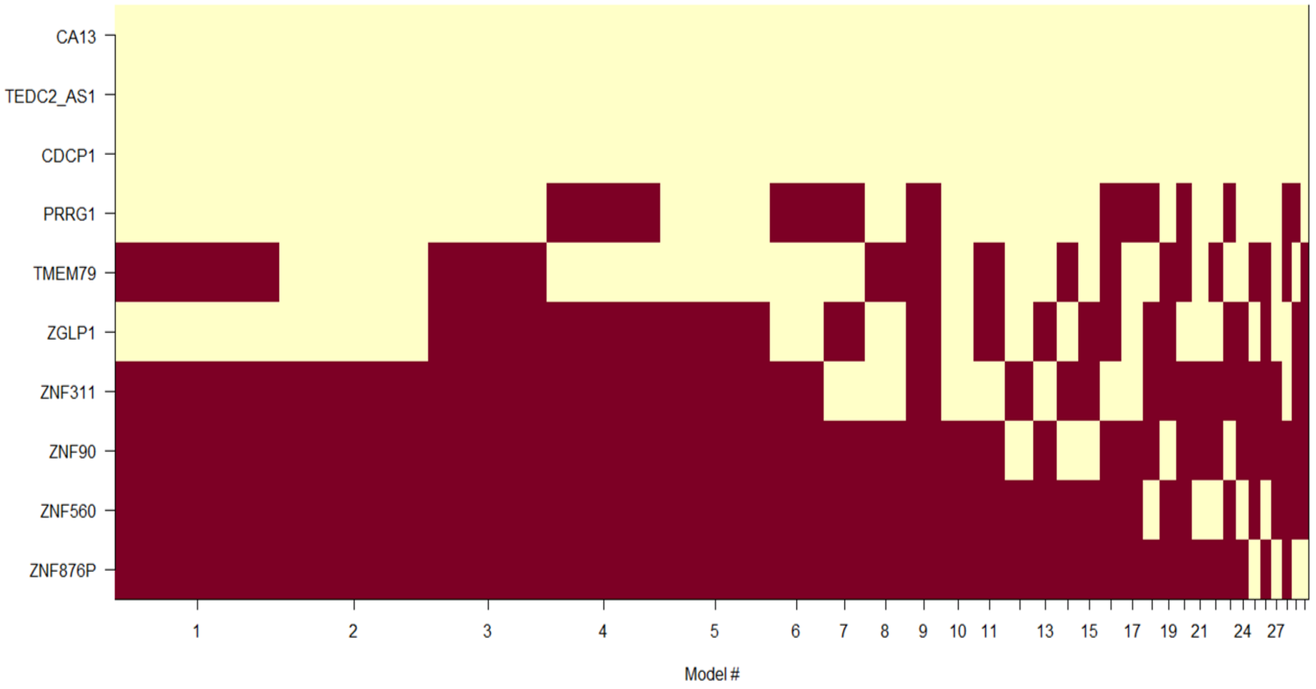
White-box: Supervised	
Bonferroni	adjustment for multiple comparisons
LASSO-COX	L1 regularitation
Elastic-net	L1 regularitation and alpha
BACKWARD selection	It starts with all variables and removes them one at a time
FORWARD selection	It starts with no variables and adds them one at a time
BMA	Bayesian Approach

IV. Results

White-box: Supervised – BMA (Bayesian Model Averaging)

Bayesian Approach

$$P(A|B) = \frac{P(B|A)\ \ P(A)}{P(B)}$$



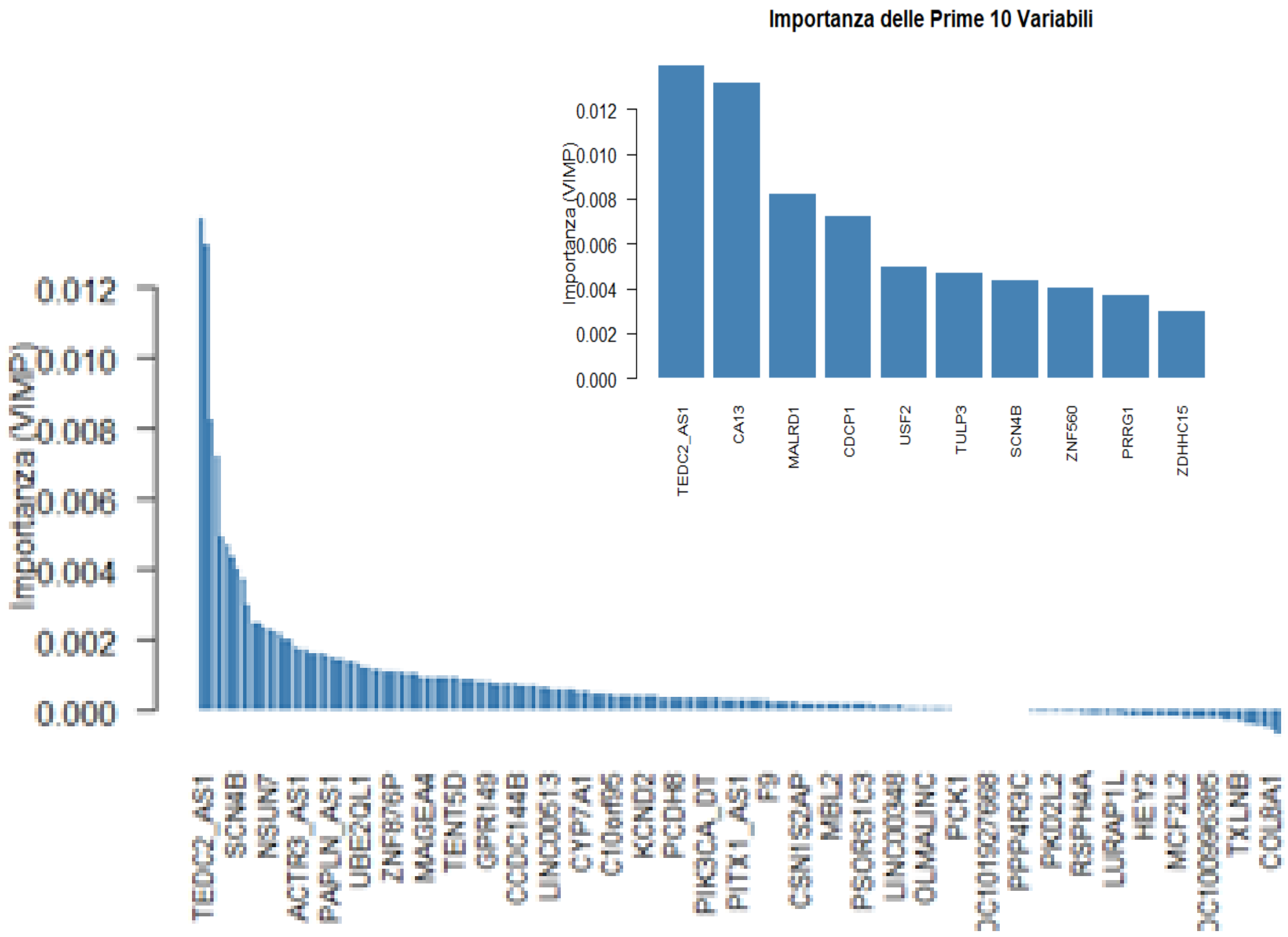
BMA
CA13
TMEM79
ZGLP1
ZNF311
TEDC2_AS1
CDCP1
PRRG1

IV. Results

Black-box: Supervised – SRF: Survival Random Forest

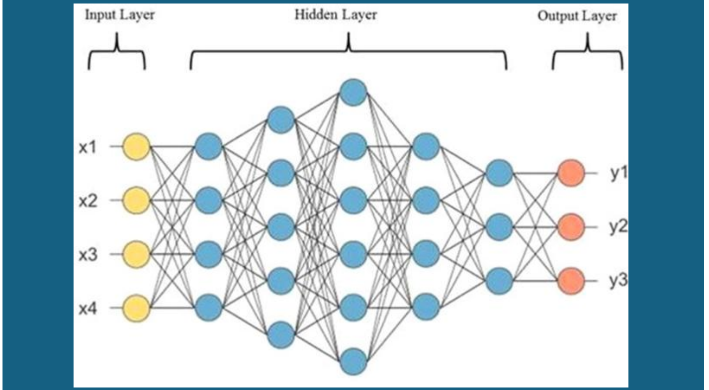
Parameter Tuning

- Mtry
- Nodesize
- Ntree

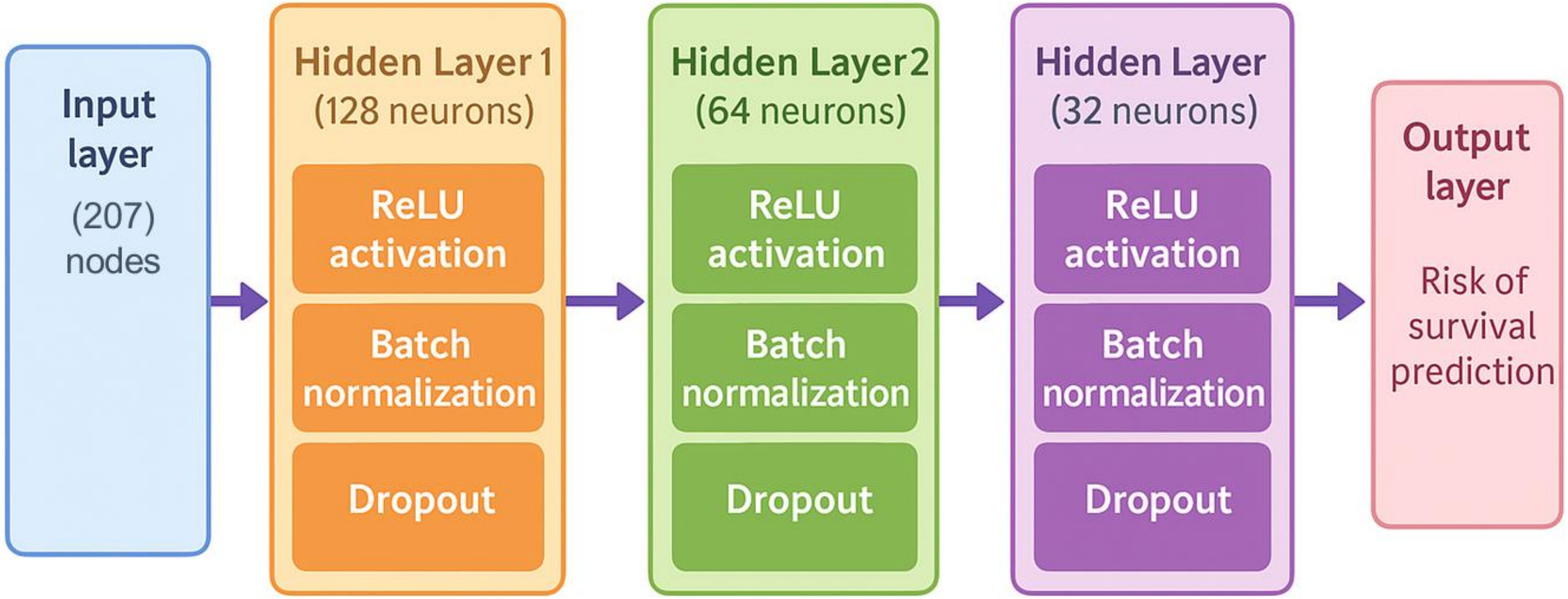


IV. Results

Black-box: Supervised – DeepSurv Cox Neural Network



- Activation function: ReLu - sigmoid
- Batch Normalization: yes
- Dropout: 30%
- Convergence Algorithm: ADAM
- Weight initialization: He method
- Loss function: Accuracy



Input layer (207) nodes

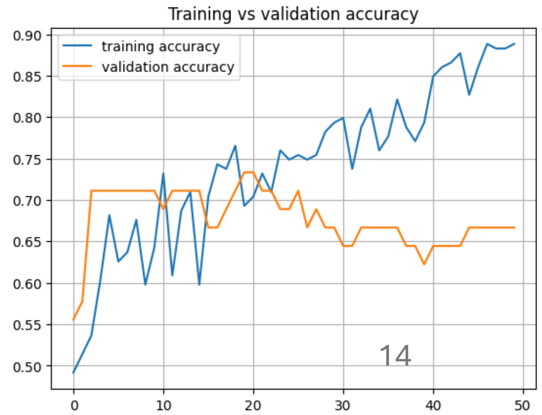
Hidden Layer 1 (128 neurons)
- ReLU activation
- Batch normalization
- Dropout

Hidden Layer2 (64 neurons)
- ReLU activation
- Batch normalization
- Dropout

Hidden Layer (32 neurons)
- ReLU activation
- Batch normalization
- Dropout

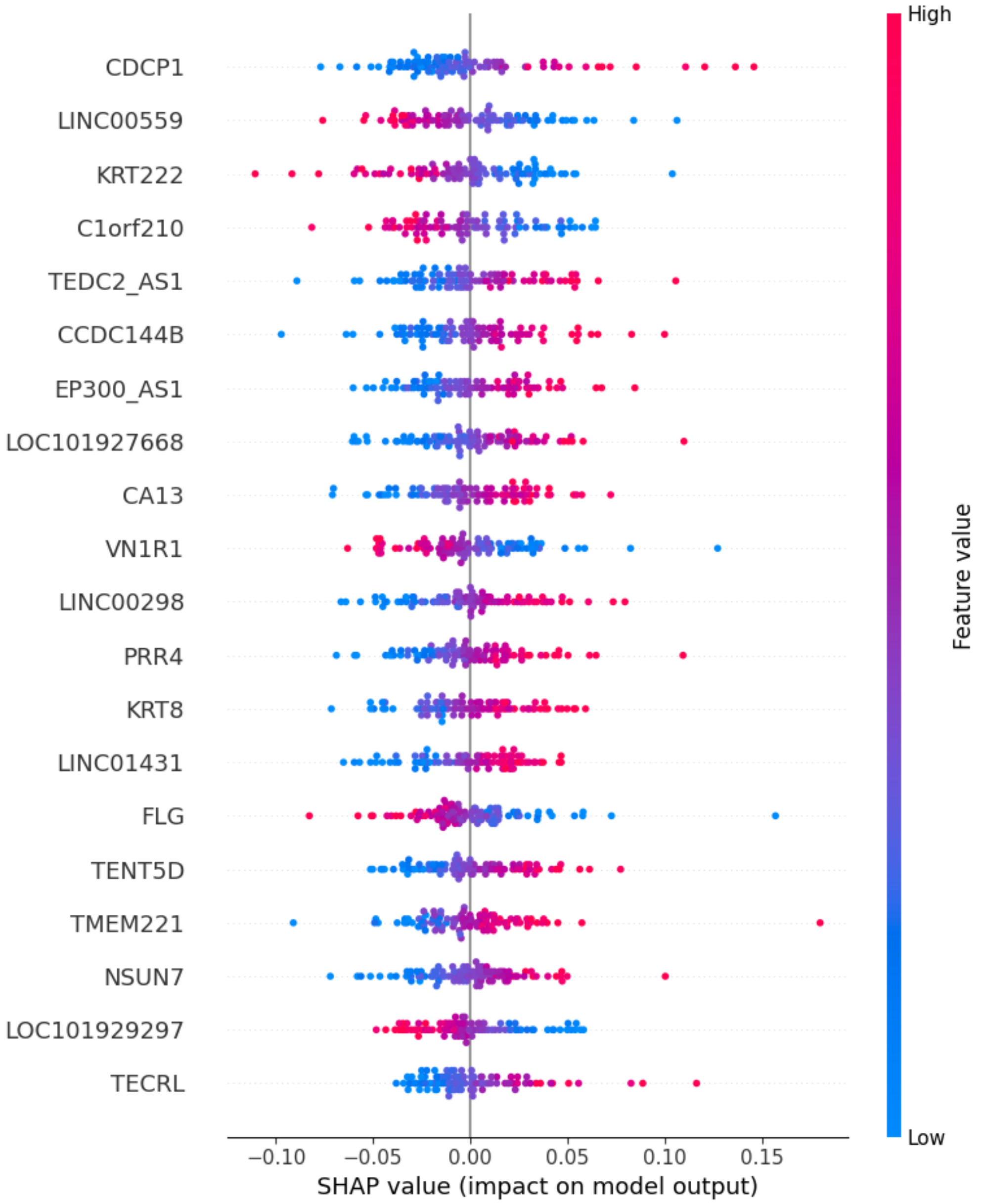
Output layer

Risk of survival prediction


Training vs validation accuracy

IV. Results

XAI: Explainable Artificial Intelligence



In a SHAP summary plot, each point represents an individual instance's SHAP value for a specific feature.

The color of the point indicates the original value of that feature: **red** denotes high feature values **blue** signifies low feature values.

IV. Results

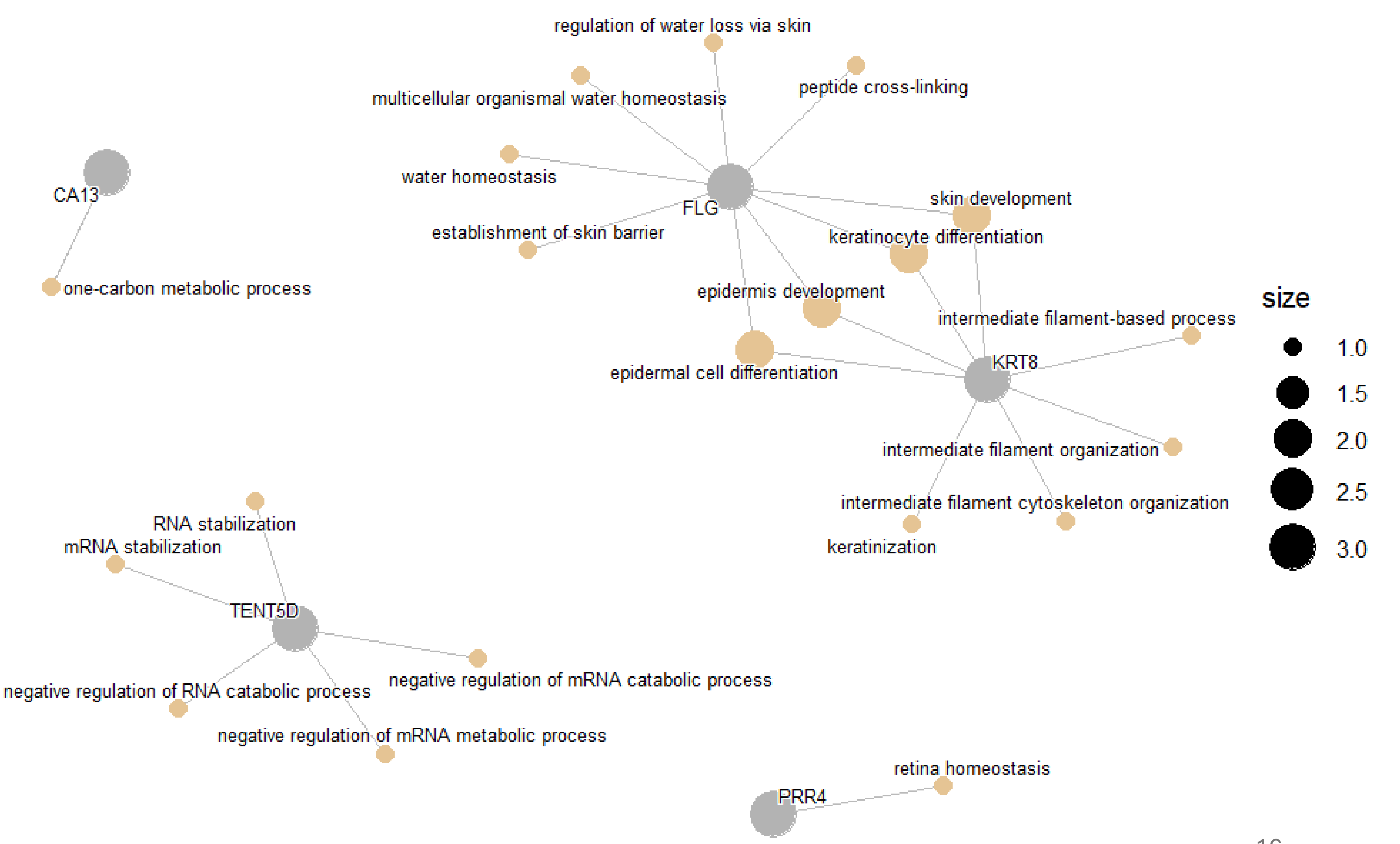
Gene Ontology analysis of genes selected by the Cox Neural Network

Biological Process (BP)

Molecular Function (MF)

Cellular Component (CC)

BIOLOGICAL PROCESS



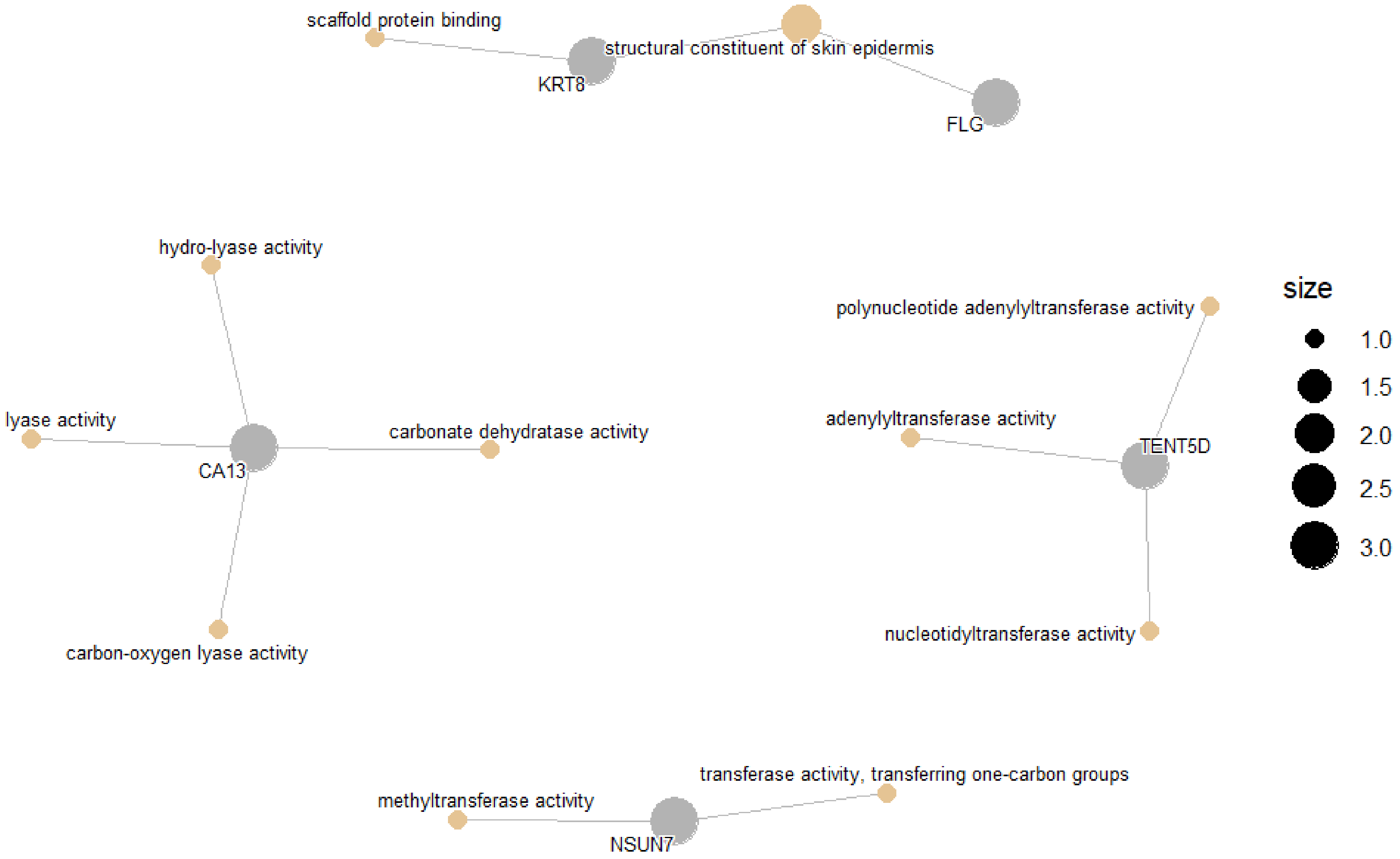
regulation of water loss via skin
multicellular organismal water homeostasis
peptide cross-linking
water homeostasis
CA13
FLG
skin development
establishment of skin barrier
keratinocyte differentiation
one-carbon metabolic process
epidermis development
size
intermediate filament-based process
1.0
KRT8
epidermal cell differentiation
1.5
2.0
intermediate filament organization
2.5
intermediate filament cytoskeleton organization
3.0
keratinization
RNA stabilization
mRNA stabilization
TENT5D
negative regulation of mRNA catabolic process
negative regulation of RNA catabolic process
negative regulation of mRNA metabolic process
retina homeostasis
PRR4

IV. Results

Gene Ontology analysis of genes selected by the Cox Neural Network

Biological Process (BP)
Molecular Function (MF)
Cellular Component (CC)

MOLECULAR FUNCTION


scaffold protein binding
structural constituent of skin epidermis
KRT8
FLG
hydro-lyase activity
polynucleotide adenylyltransferase activity
size
1.0
1.5
2.0
2.5
3.0
lyase activity
carbonate dehydratase activity
adenylyltransferase activity
TENT5D
CA13
carbon-oxygen lyase activity
nucleotidyltransferase activity
transferase activity, transferring one-carbon groups
methyltransferase activity
NSUN7

IV. Results

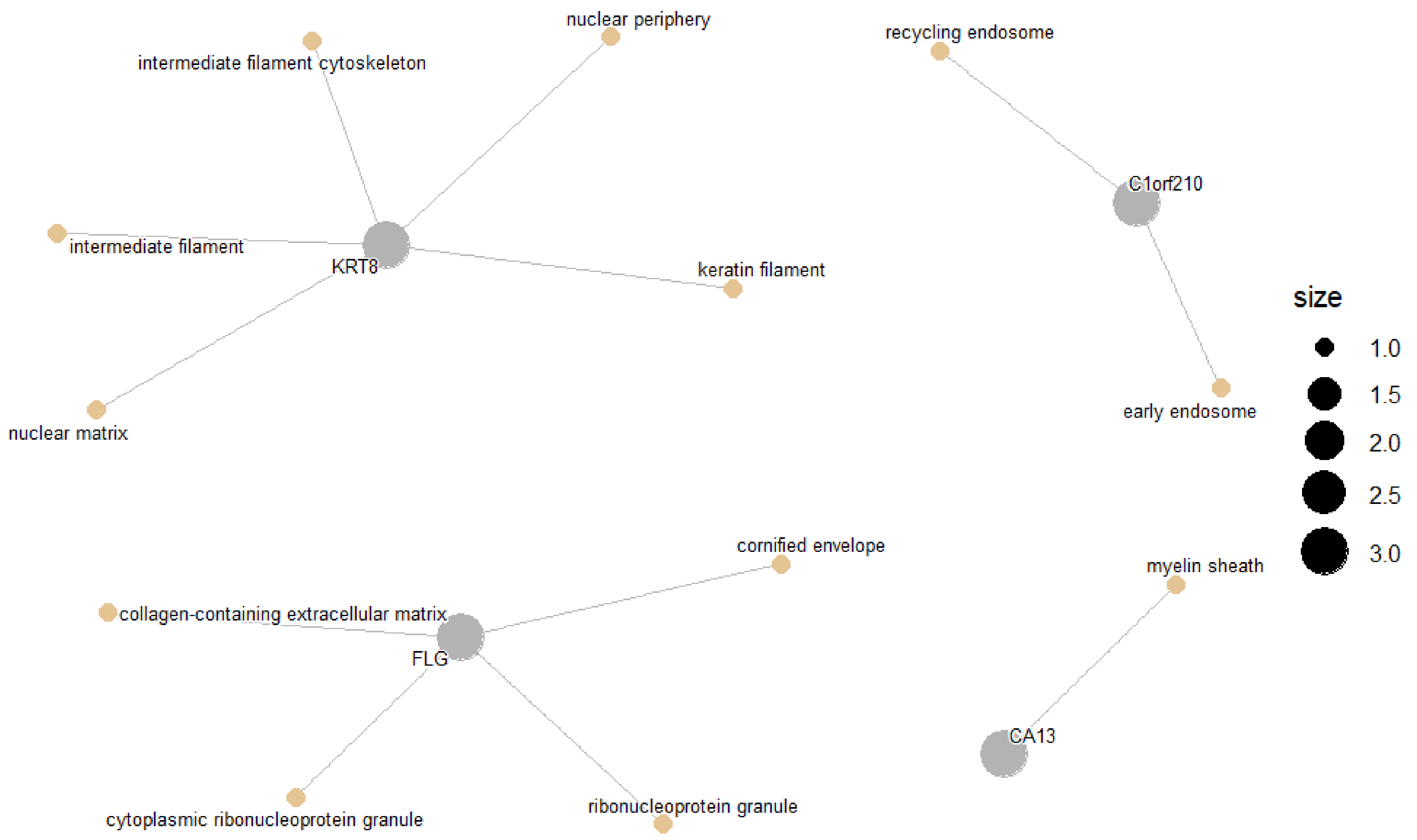
Gene Ontology analysis of genes selected by the Cox Neural Network

Biological Process (BP)

Molecular Function (MF)

Cellular Component (CC)

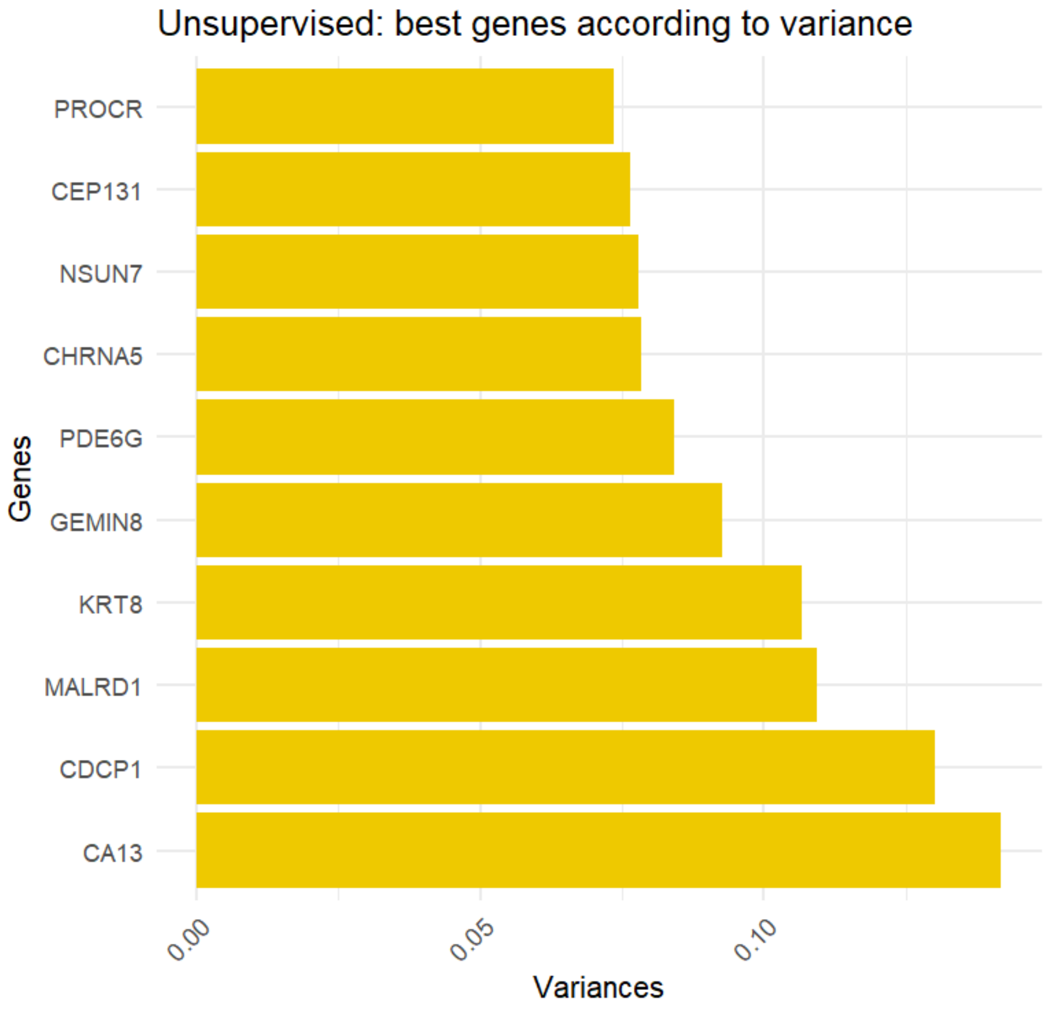
COMPONENT CELLULAR



IV. Results

● Unsupervised Learning: t-SNE

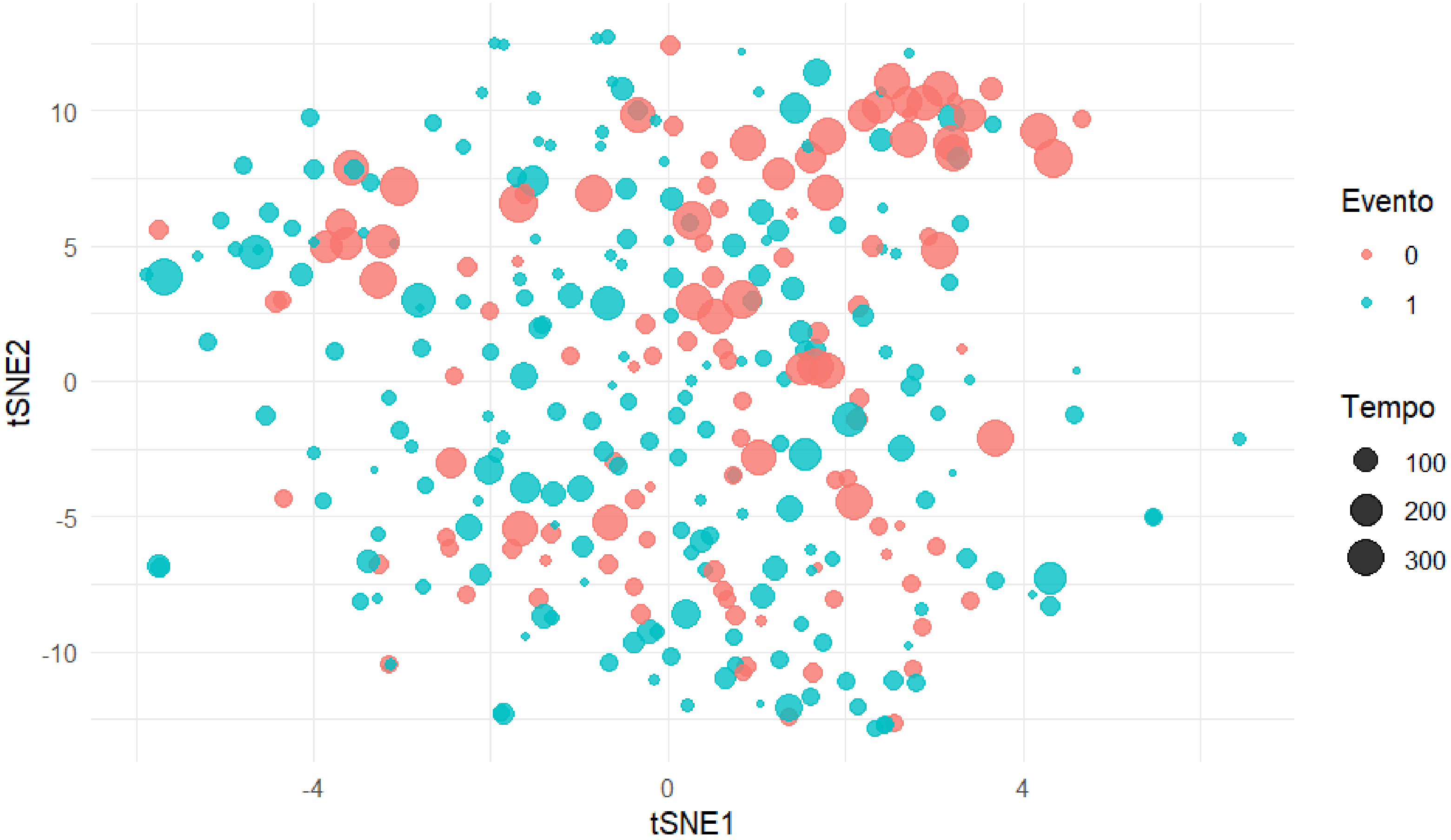
t-SNE (t-Distributed Stochastic Neighbor Embedding) is a method based on variance for each feature



IV. Results

● Unsupervised Learning: t-SNE

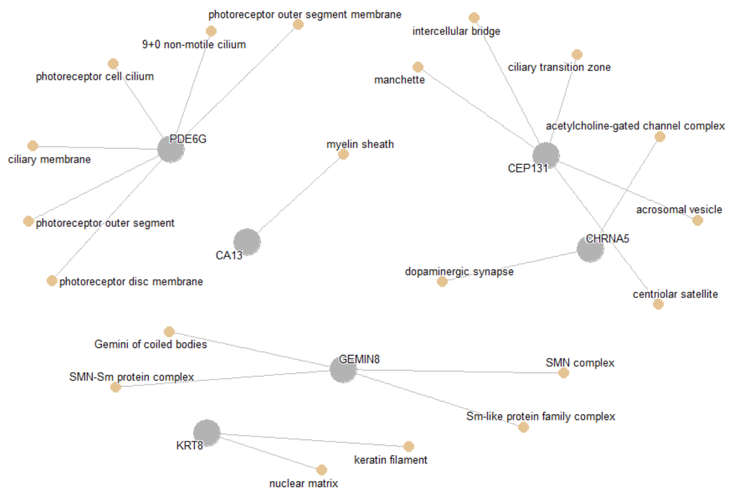
t-SNE (t-Distributed Stochastic Neighbor Embedding)
is a powerful tool for visualizing high-dimensional data.
It is widely used in data science and machine learning for its ability to reduce dimensions while preserving the local structure of the data.



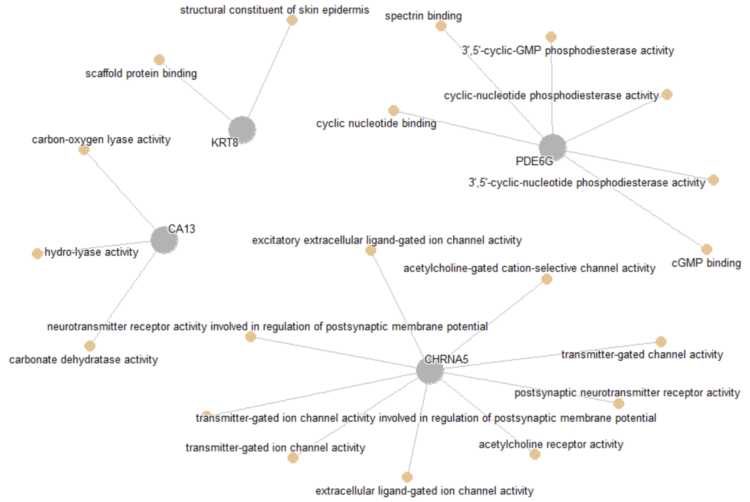
IV. Results

Gene Ontology analysis of genes selected by the t-SNE

Biological Process

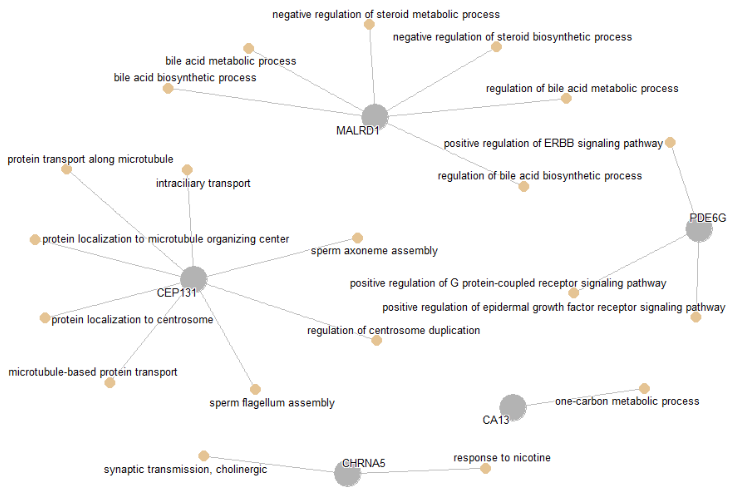


Molecular Function



structural constituent of skin epidermis
spectrin binding
3',5'-cyclic-GMP phosphodiesterase activity
scaffold protein binding
cyclic-nucleotide phosphodiesterase activity
cyclic nucleotide binding
KRT8
carbon-oxygen lyase activity
PDE6G
3',5'-cyclic-nucleotide phosphodiesterase activity
CA13
excitatory extracellular ligand-gated ion channel activity
hydro-lyase activity
cGMP binding
acetylcholine-gated cation-selective channel activity
neurotransmitter receptor activity involved in regulation of postsynaptic membrane potential
carbonate dehydratase activity
CHRNA5
transmitter-gated channel activity
postsynaptic neurotransmitter receptor activity
transmitter-gated ion channel activity involved in regulation of postsynaptic membrane potential
transmitter-gated ion channel activity
acetylcholine receptor activity
extracellular ligand-gated ion channel activity

Component Cellular



negative regulation of steroid metabolic process
negative regulation of steroid biosynthetic process
bile acid metabolic process
bile acid biosynthetic process
regulation of bile acid metabolic process
MALRD1
positive regulation of ERBB signaling pathway
protein transport along microtubule
intraciliary transport
regulation of bile acid biosynthetic process
PDE6G
protein localization to microtubule organizing center
sperm axoneme assembly
CEP131
positive regulation of G protein-coupled receptor signaling pathway
positive regulation of epidermal growth factor receptor signaling pathway
protein localization to centrosome
regulation of centrosome duplication
microtubule-based protein transport
sperm flagellum assembly
one-carbon metabolic process
CA13
response to nicotine
synaptic transmission, cholinergic
CHRNA5

IV. Results

Bologna
Data
Policlinico
Sant’Orsola

-13938

External dataset

# 74 patients
# 36619 raw RNA-seq samples

↓

RNA-seq data normalization performed with **edgeR**

↓

# 74 patients
# 22681 raw RNA-seq samples

IV. Results

Bologna Dati Policlinico Sant'Orsola

# Event: 40
# Median fu: 17[6-29] months

27 relapse

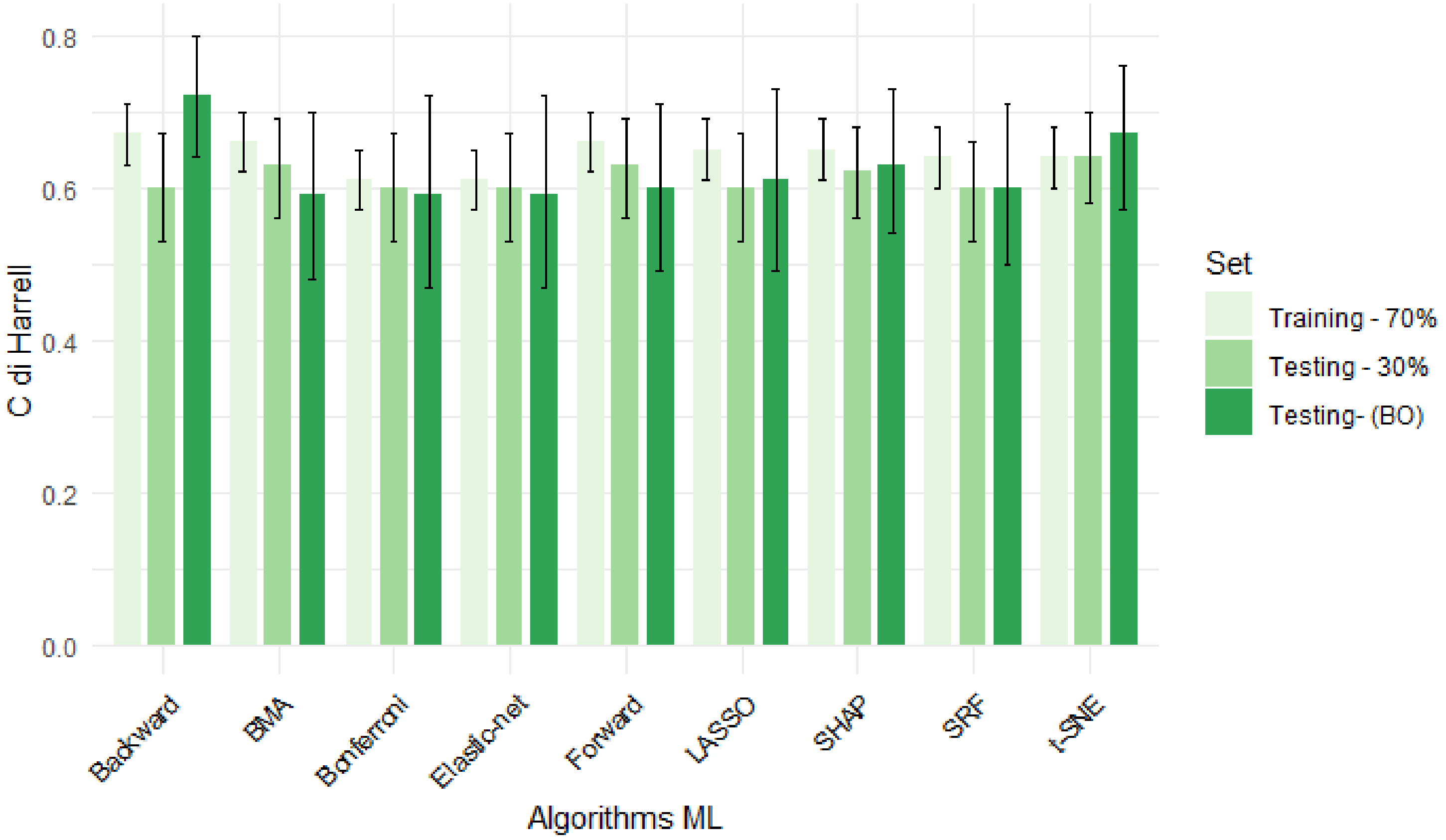
External dataset

Clinic variables	N	%
Gender		
F	28	37.90%
M	46	62.10%
Age,median[Q1-Q3	65[23-92]	
Risk		
HIGH	22	
INTERMEDIATE	24	
LOW	23	

Mutation			
npm1			
	neg	38	58%
	pos	28	42%
	8 NA		
evi1			
	neg	32	57%
	pos		
	42 NA		
n_ras			
	neg	10	63%
	pos	6	38%
	58 NA		
Flt3_tkd			
	neg	29	91%
	pos	3	9%
	42 NA		
Cebpa			
	neg	30	40%
	pos	-	-
	44 NA		
Flt3_itd			
	neg	26	81%
	pos	6	19%
	42 NA		
K_ras			
	neg	14	88%
	pos	2	12%
	58 NA		
Idh1			
	neg	28	93%
	pos	2	7%
	44 NA		
Idh2			
	neg	28	93%
	pos	2	7%
	44 NA		
TP53			
	neg	21	81%
	pos	5	19%
	48 NA		

IV. Results

METRICS: C di HARRELL



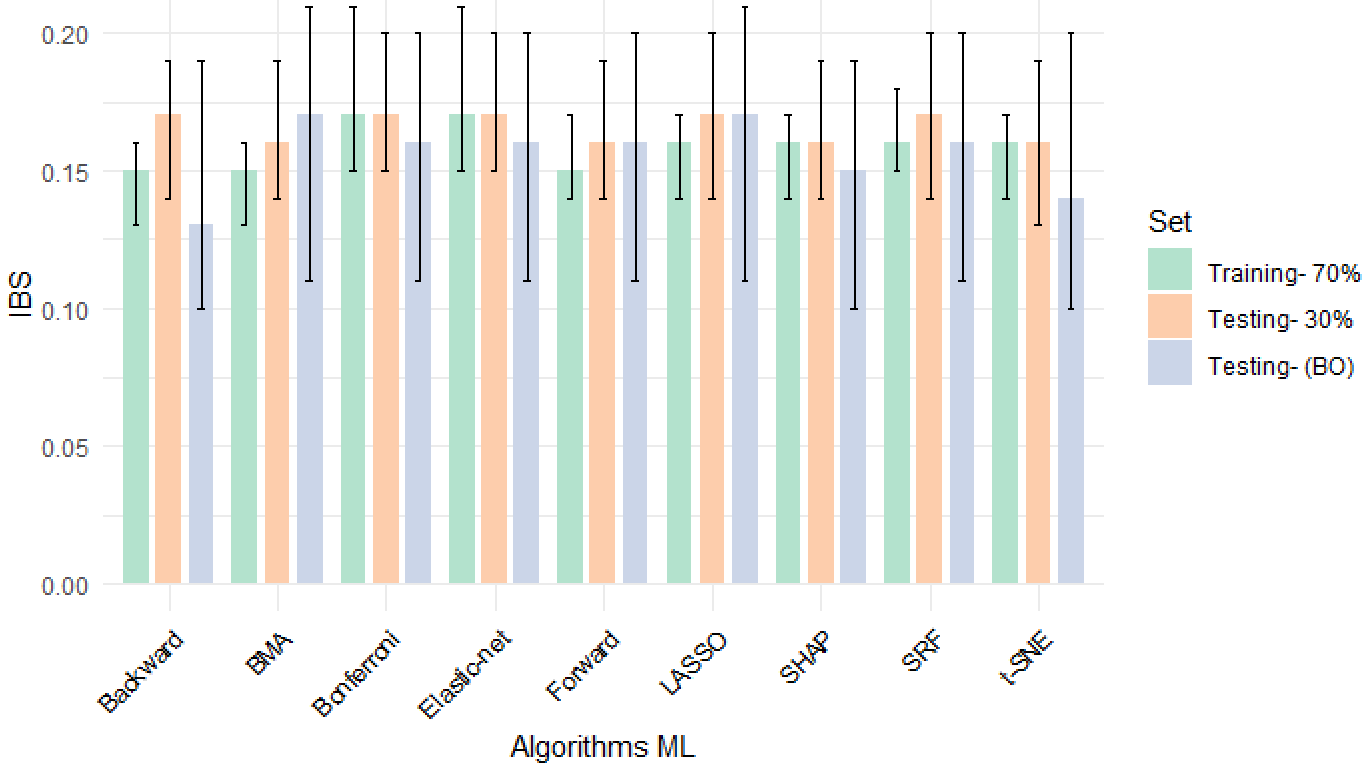
Higher is better

C di Harrell											
SET	N	# eventi	Bonferroni	LASSO	Elastic-net	Forward	Backward	SRF	BMA	SHAP	t-SNE
Training 70%	321	203	0.61[0.57-0.65]	0.65[0.61-0.69]	0.61[0.57-0.65]	0.66[0.62-0.70]	*0.67[0.63-0.71]*	0.64[0.60-0.68]	0.66[0.62- 0.70]	0.65[0.61-0.69]	0.64[0.60-0.68]
Validation 30%	136	87	0.60[0.53-0.67]	0.60[0.53-0.67]	0.60[0.53-0.67]	0.63[0.56-0.69]	0.60[0.53-0.67]	0.60[0.53-0.66]	0.63[0.56-0.69]	0.62[0.56-0.68]	*0.64[0.58-0.70]*
Testing - Bologna	68	36	0.59[0.47-0.72]	0.61[0.49-0.73]	0.59[0.47-0.72]	0.60[0.49-0.71]	*0.72[0.64-0.80]*	0.60[0.50-0.71]	0.59[0.48-0.70]	0.63[0.54-0.73]	0.67[0.57-0.76]

IV. Results

METRICS: IBS

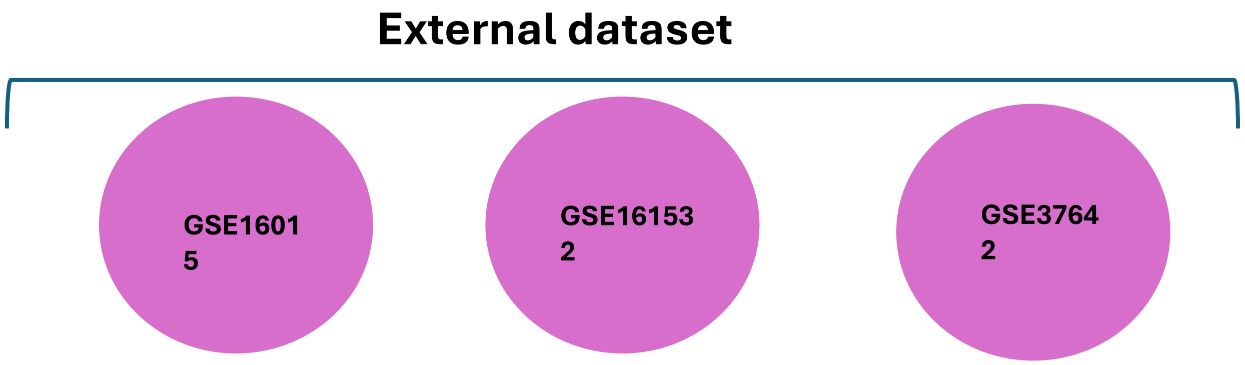
Lower is better



IBS											
SET	**N**	**# eventi**	**Bonferroni**	**LASSO**	**Elastic-net**	**Forward**	**Backward**	**SRF**	**BMA**	**SHAP**	**t-SNE**
Training 70%	321	203	0.17[0.15-0.21]	0.16[0.14-0.17]	0.17[0.15-0.21]	0.15[0.14-0.17]	0.15[0.13-0.16]	0.16[0.15-0.18]	0.15[0.13-0.16]	0.16[0.14-0.17]	0.16[0.14-0.17]
Validation 30%	136	87	0.17[0.15-0.20]	0.17[0.14-0.20]	0.17[0.15-0.20]	0.16[0.14-0.19]	0.17[0.14-0.19]	0.17[0.14-0.20]	0.16[0.14-0.19]	0.16[0.14-0.19]	0.16[0.13-0.19]
Testing - Bologna	68	36	0.16[0.11-0.20]	0.17[0.11-0.21]	0.16[0.11-0.20]	0.16[0.11-0.20]	0.13[0.1-0.19]	0.16[0.11-0.20]	0.17[0.11-0.21]	0.15[0.10-0.19]	0.14[0.10-0.20]

NEXTS STEP

1. Validating the identified signature(s) on external datasets


External dataset

2. Repeat all the previous analyses for the logistic outcome related to risk stratification