Explainable Machine Learning for Early Assessment of COVID-19 Risk Prediction in Emergency Departments

E. Casiraghi, D. Malchiodi, G. Trucco, M. Frasca, L. Cappelletti,
T. Fontana, A. A. Esposito, E. Avola, A. Jachetti, J. Reese, A. Rizzi, P. N. Robinson,
and G. Valentini

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 16, 2020, accepted October 19, 2020, date of publication October 26, 2020, date of current version November 10, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3034032

Explainable Machine Learning for Early Assessment of COVID-19 Risk Prediction in Emergency Departments

ELENA CASIRAGHI[1,2], (Member, IEEE), DARIO MALCHIODI[1,2,3], GABRIELLA TRUCCO[1], MARCO FRASCA[1], LUCA CAPPELLETTI[1], TOMMASO FONTANA[4], ALESSANDRO ANDREA ESPOSITO[5], EMANUELE AVOLA[6], ALESSANDRO JACHETTI[7], JUSTIN REESE[8], ALESSANDRO RIZZI[1], (Member, IEEE), PETER N. ROBINSON[9], AND GIORGIO VALENTINI[1,2,3]

[1]Department of Computer Science "Giovanni degli Antoni," Università degli Studi di Milano, 20133 Milan, Italy
[2]CINI National Laboratory of Artificial Intelligence and Intelligent Systems (AIIS), Università di Roma, 00185 Rome, Italy
[3]Data Science Research Center, Università degli Studi di Milano, 20133 Milan, Italy
[4]Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, 20133 Milan, Italy
[5]Radiology Department, Fondazione IRCCS Ca Granda Ospedale Maggiore Policlinico, 20122 Milan, Italy
[6]Postgraduate School in Radiodiagnostics, Università degli Studi di Milano, 20122 Milan, Italy
[7]Accident and Emergency Department, Fondazione IRCCS Ca Granda Ospedale Maggiore Policlinico, 20122 Milan, Italy
[8]Division of Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[9]The Jackson Laboratory for Genomic Medicine, Farmington, CT 06032, USA

Corresponding authors: Elena Casiraghi (elena.casiraghi@unimi.it) and Dario Malchiodi (dario.malchiodi@unimi.it)

This work was supported in part by the Università degli Studi di Milano through the Piano di Sostegno alla ricerca 2019 Grant.

Objective

Proposing prediction models for early assessment of COVID-19 severity

- Optimizing patients’ flow and waiting times in Emergency Departments (EDs)
- Integrating clinical, laboratory and radiological data
- Specifically designed for easy deployment in EDs (thanks to...)
- Explainable output (feature relevance and prediction rules)

Strengths

- Thorough missing data imputation studies
- Robust feature selection process
- Comparison between different prediction models:
  - Random Forests
  - Generalized Linear Models
  - Associative Trees
  - MLPs
  - Logistic Regression
  - Support Vector Machines

Strengths

- Thorough missing data imputation studies
- Robust feature selection process
- Comparison between different prediction models:

- Random Forests
- Generalized Linear Models
- Associative Trees

- MLPs
- Logistic Regression
- Support Vector Machines

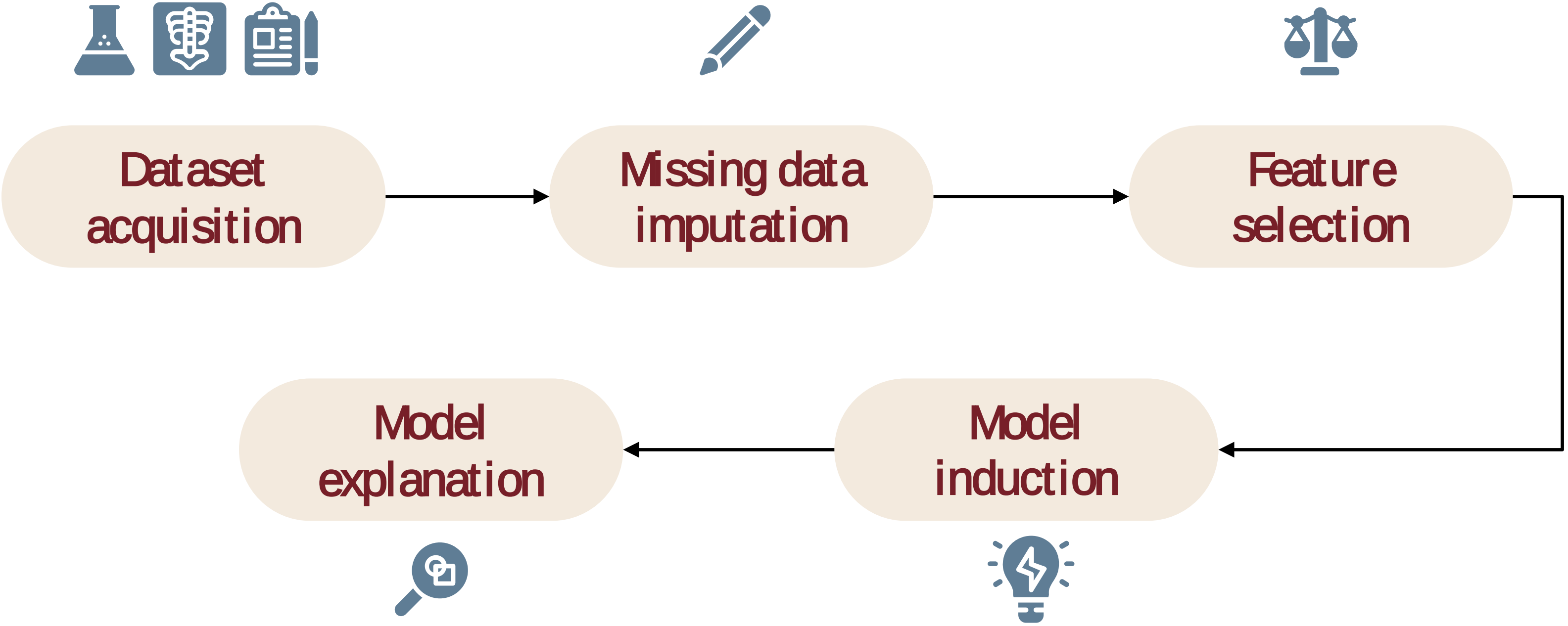
Results not shown due to much lower performance

Biases to avoid

- Lack of clinical follow-up data (with the risk of inaccurate labels)
- Suboptimal predictor measurements (e.g., the last available ones rather than those acquired in EDs)
- Population not clearly described
- Models described only in part / not suitably tested
- Hyperparameters setting is not robust / not reported

(see L. Wynants et al., ‘‘Prediction models for diagnosis and prognosis of COVID-19: Systematic review and critical appraisal,’’ BMJ, vol. 369, no. 369, 2020)

In a nutshell



Dataset acquisition

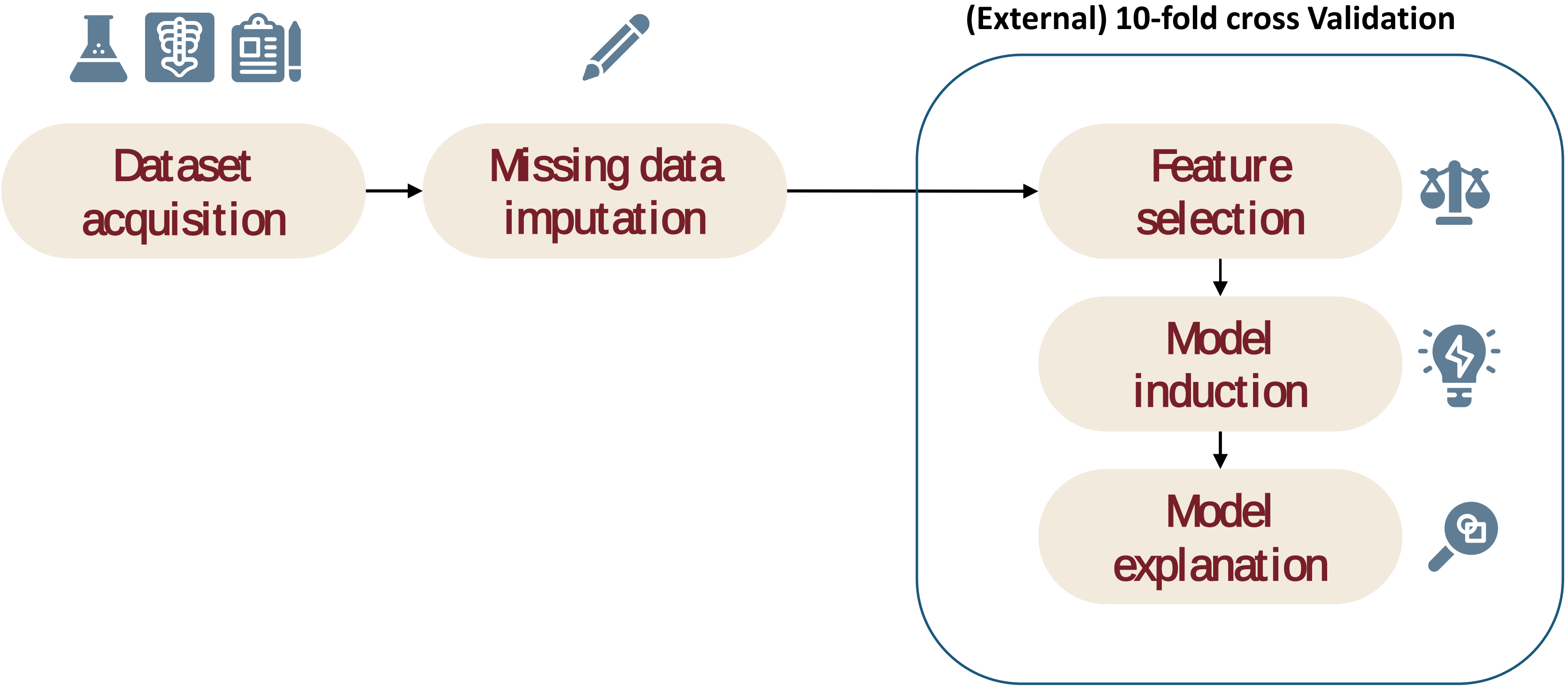
Missing data imputation

Feature selection

Model induction

Model explanation

In a nutshell



(External) 10-fold cross Validation

Dataset acquisition

Missing data imputation

Feature selection

Model induction

Model explanation

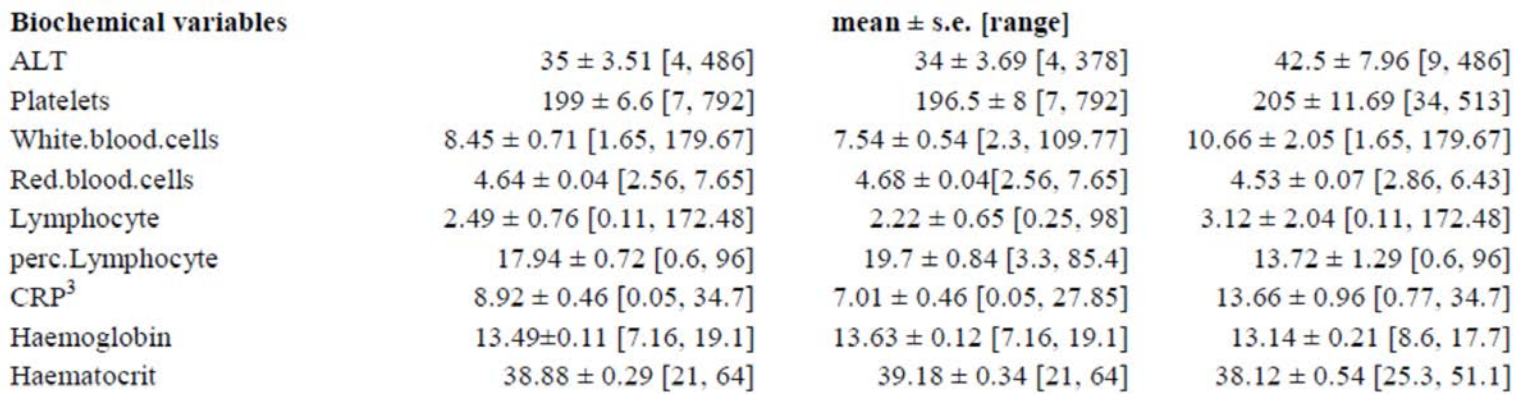
Data Collection and Filtering

- ED admissions in urban multicenter health system
- March, 7th – April 10th 2020
- COVID-19 positive (RT-PCR)
- Five-months clinical follow-up:
  - Low risk (no hospitalization, no serious consequences): 214 patients
  - High risk (intubation, serious consequences, death): 87 patients

- Variables were excluded if > 50% values were missing or they had negligible variance

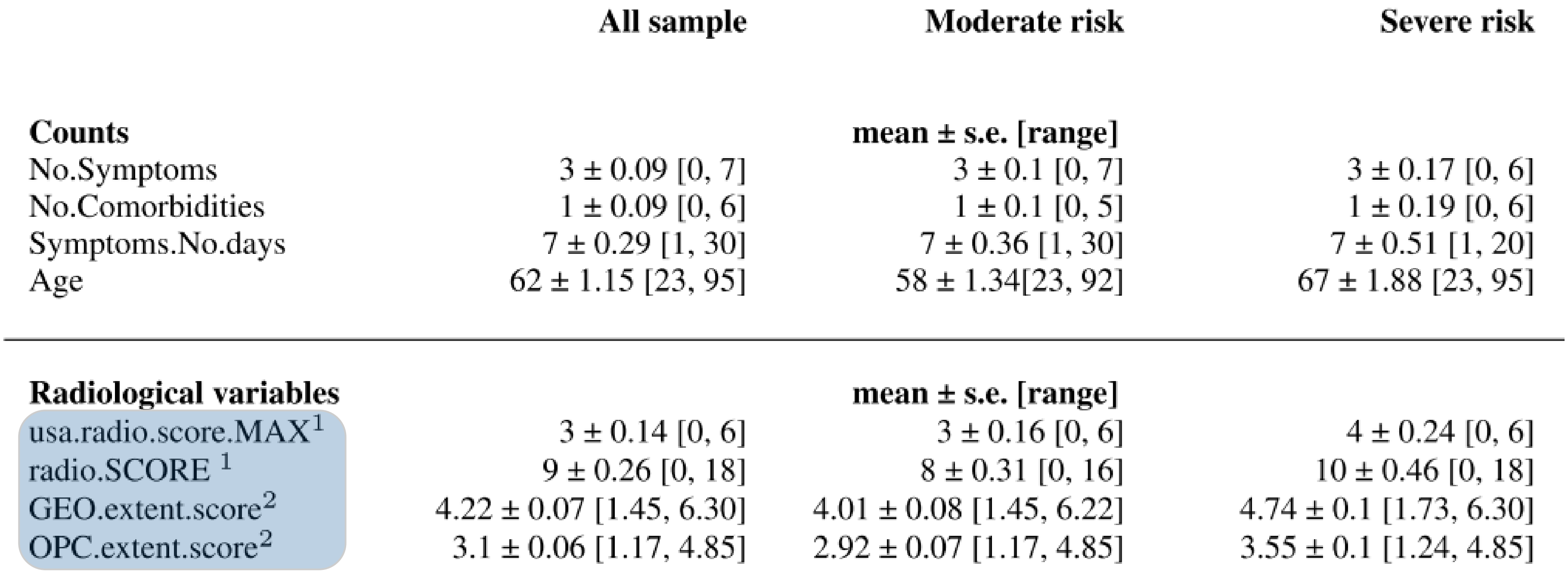
THE Dataset

- 207+94 adult men and women
- Age: 61±1 years [23–95]
- Days with symptoms: 7±0 [1–30]
- Symptoms [fever, cough, etc.]
- Clinical history, comorbidities [cancer, asthma, etc.]
- Laboratory measurement [~~LDH~~, ~~AST~~, white blood cell count, red blood cell count, Lymphocytes, CRP, Haemoglobin, Haematocrit]
- Saturation/oxygen value
- CRX
- Age, gender



Biochemical variables		**mean ± s.e. [range]**	
ALT	35 ± 3.51 [4, 486]	34 ± 3.69 [4, 378]	42.5 ± 7.96 [9, 486]
Platelets	199 ± 6.6 [7, 792]	196.5 ± 8 [7, 792]	205 ± 11.69 [34, 513]
White.blood.cells	8.45 ± 0.71 [1.65, 179.67]	7.54 ± 0.54 [2.3, 109.77]	10.66 ± 2.05 [1.65, 179.67]
Red.blood.cells	4.64 ± 0.04 [2.56, 7.65]	4.68 ± 0.04[2.56, 7.65]	4.53 ± 0.07 [2.86, 6.43]
Lymphocyte	2.49 ± 0.76 [0.11, 172.48]	2.22 ± 0.65 [0.25, 98]	3.12 ± 2.04 [0.11, 172.48]
perc.Lymphocyte	17.94 ± 0.72 [0.6, 96]	19.7 ± 0.84 [3.3, 85.4]	13.72 ± 1.29 [0.6, 96]
CRP[3]	8.92 ± 0.46 [0.05, 34.7]	7.01 ± 0.46 [0.05, 27.85]	13.66 ± 0.96 [0.77, 34.7]
Haemoglobin	13.49±0.11 [7.16, 19.1]	13.63 ± 0.12 [7.16, 19.1]	13.14 ± 0.21 [8.6, 17.7]
Haematocrit	38.88 ± 0.29 [21, 64]	39.18 ± 0.34 [21, 64]	38.12 ± 0.54 [25.3, 51.1]

Variable name	All sample	Moderate risk	Severe risk
Symptoms	**% presence (no.)**		
Fever	93 (280)	92.5 (198)	94.3 (82)
Cough	66.8 (201)	68.7 (147)	62.1 (54)
Dyspnea	55.1 (166)	47.7 (102)	73.6 (64)
Respiratory Failure (IR)	13 (39)	8.9 (19)	23 (20)
Myalgias	9.3 (28)	9.3 (20)	9.2 (8)
Other	9.6 (29)	9.8 (21)	9.2 (8)
Syncope	4.3 (13)	5.1 (11)	2.3 (2)
Asthenia	12.3 (37)	11.7 (25)	13.8 (12)
Vomiting.Nausea	5 (15)	4.2 (9)	6.9 (6)
Diarrhea	10.3 (31)	10.7 (23)	9.2 (8)
Headache	3 (9)	3.3 (7)	2.3 (2)
Pharyngeal.pain	3 (9)	3.7 (8)	1.1 (1)
Comorbidities	**% presence (no.)**		
Pneumo.asthma	4.7 (14)	5.1 (11)	3.4 (3)
Pneumo.BPCO	5.3 (16)	4.2 (9)	8 (7)
Neoplasia (last 5 years)	10.6 (32)	7.9 (17)	17.2 (15)
Smoke	5.3 (16)	5.6 (12)	4.6 (4)
Arterial.hypertension	29.9 (90)	26.2 (56)	39.1 (34)
Cardiovascular pathologies	16.6 (50)	11.7 (25)	28.7 (25)
Diabetes	15.9 (48)	12.1 (26)	25.3 (22)
Obesity	6 (18)	5.6 (12)	6.9 (6)
Celebral Stroke	4 (12)	3.7 (8)	4.6 (4)



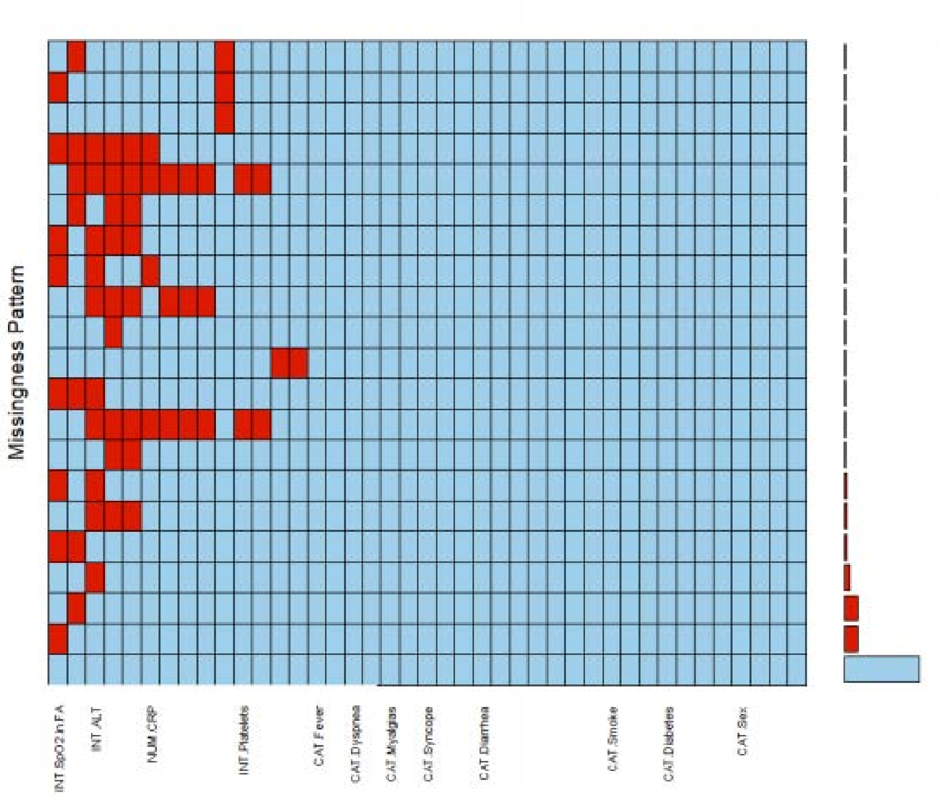
	All sample	Moderate risk	Severe risk
Counts		**mean ± s.e. [range]**	
No.Symptoms	3 ± 0.09 [0, 7]	3 ± 0.1 [0, 7]	3 ± 0.17 [0, 6]
No.Comorbidities	1 ± 0.09 [0, 6]	1 ± 0.1 [0, 5]	1 ± 0.19 [0, 6]
Symptoms.No.days	7 ± 0.29 [1, 30]	7 ± 0.36 [1, 30]	7 ± 0.51 [1, 20]
Age	62 ± 1.15 [23, 95]	58 ± 1.34[23, 92]	67 ± 1.88 [23, 95]
Radiological variables		**mean ± s.e. [range]**	
usa.radio.score.MAX[1]	3 ± 0.14 [0, 6]	3 ± 0.16 [0, 6]	4 ± 0.24 [0, 6]
radio.SCORE [1]	9 ± 0.26 [0, 18]	8 ± 0.31 [0, 16]	10 ± 0.46 [0, 18]
GEO.extent.score[2]	4.22 ± 0.07 [1.45, 6.30]	4.01 ± 0.08 [1.45, 6.22]	4.74 ± 0.1 [1.73, 6.30]
OPC.extent.score[2]	3.1 ± 0.06 [1.17, 4.85]	2.92 ± 0.07 [1.17, 4.85]	3.55 ± 0.1 [1.24, 4.85]

- CXR used as a first-line triage tool
- Less sensitive in early stage
- Less burden

- Indices 1: experienced radiologists
- Indices 2: Covid-Net DNN

Missing data patterns

- No high correlation among patterns
- Independence of missing/observed proportions
- No significant difference between distributions of othe
  attributes
  corresponding to observed and missing values
- Thus, no MNAR



MAR or MCAR?

- Jamshidian and Jalal test (homoscedasticity test for data having identical missingness patterns is a proxy for MCAR)
- Homoscedasticity tested using Hawkins test for complete datasets
- Completeness obtained after distFree imputation (only assuming independence)
- Result: MCAR

Missing data imputation

- Multiple Imputation by Chained Equations
  - Each variable cyclically predicted from remaining ones (+randomness source)
  - Several “donor” predictions relying on Predictive Mean Matching (micePMM) or Random Forests (**miceRF**), imputation selected among donors
  - Prediction cycles repeated several times

- **missForest**
  - Repeatedly train RFs to predict one variable from remaining ones (only one hyperparameter, no distributional assumptions)

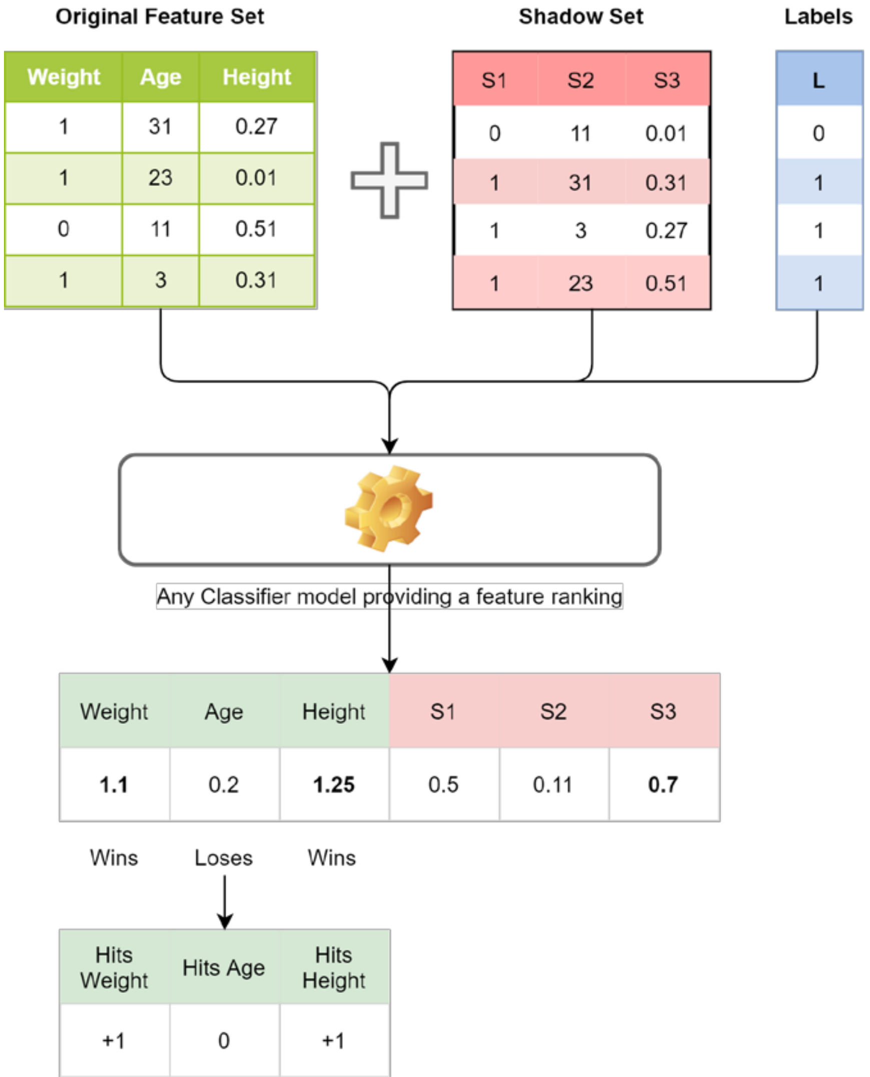
- distFree
  - Imputation via ML linear predictors + random errors

Boruta Feature selection

- Add “fake” features (random permutation of each original feature)

- Train RF, retain/discard original features which resulted as important* for a statistically significant number of times, set remaining features as tentative

Boruta Feature selection



Original Feature Set

Weight	Age	Height
1	31	0.27
1	23	0.01
0	11	0.51
1	3	0.31

Shadow Set

S1	S2	S3
0	11	0.01
1	31	0.31
1	3	0.27
1	23	0.51

Labels

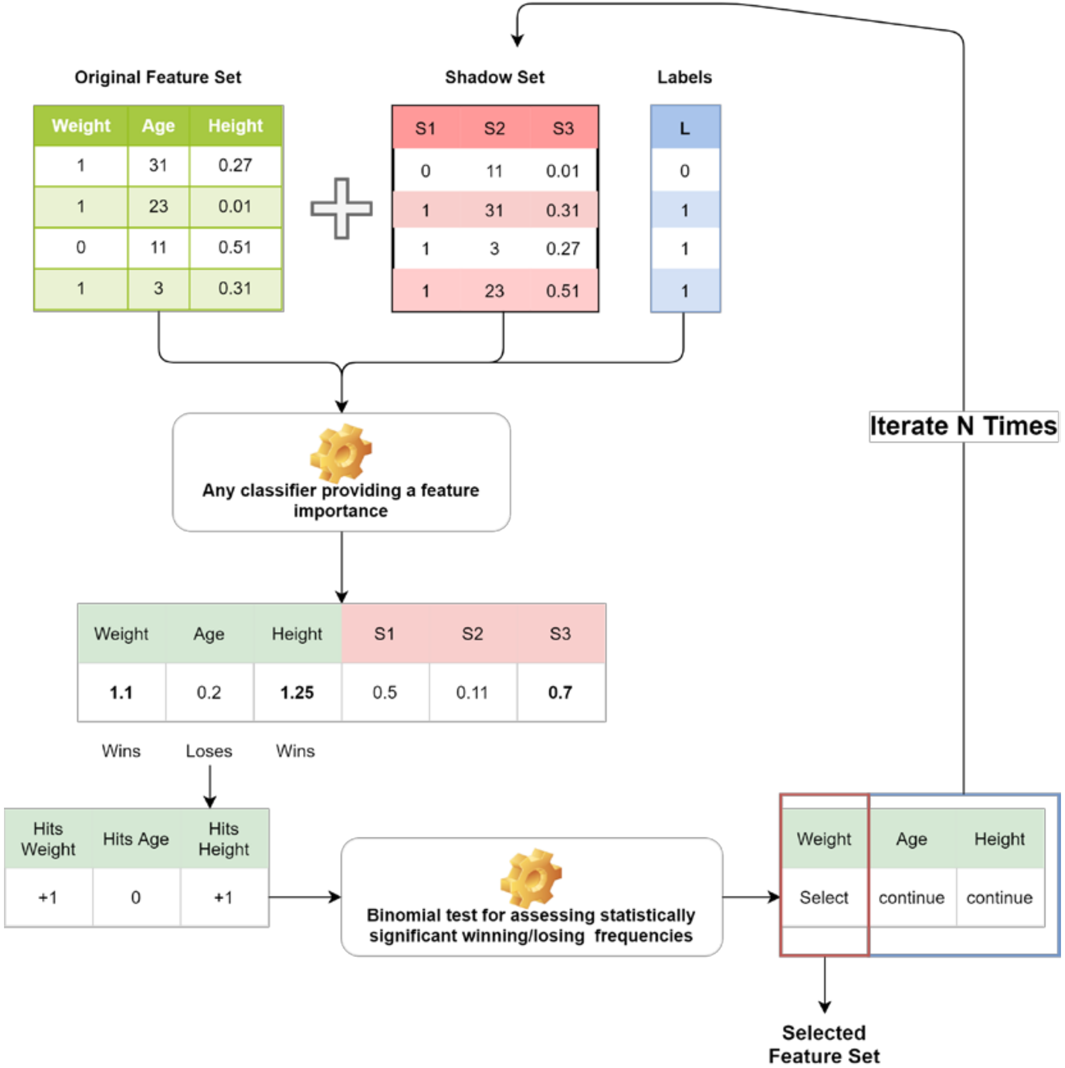
L
0
1
1
1

Any Classifier model providing a feature ranking

Weight	Age	Height	S1	S2	S3
1.1	0.2	**1.25**	0.5	0.11	**0.7**
Wins	Loses	Wins			

Hits Weight	Hits Age	Hits Height
+1	0	+1

Boruta Feature selection



Original Feature Set

Weight	Age	Height
1	31	0.27
1	23	0.01
0	11	0.51
1	3	0.31

Shadow Set

S1	S2	S3
0	11	0.01
1	31	0.31
1	3	0.27
1	23	0.51

Labels

L
0
1
1
1

Any classifier providing a feature importance

Weight	Age	Height	S1	S2	S3
1.1	0.2	**1.25**	0.5	0.11	**0.7**
Wins	Loses	Wins			

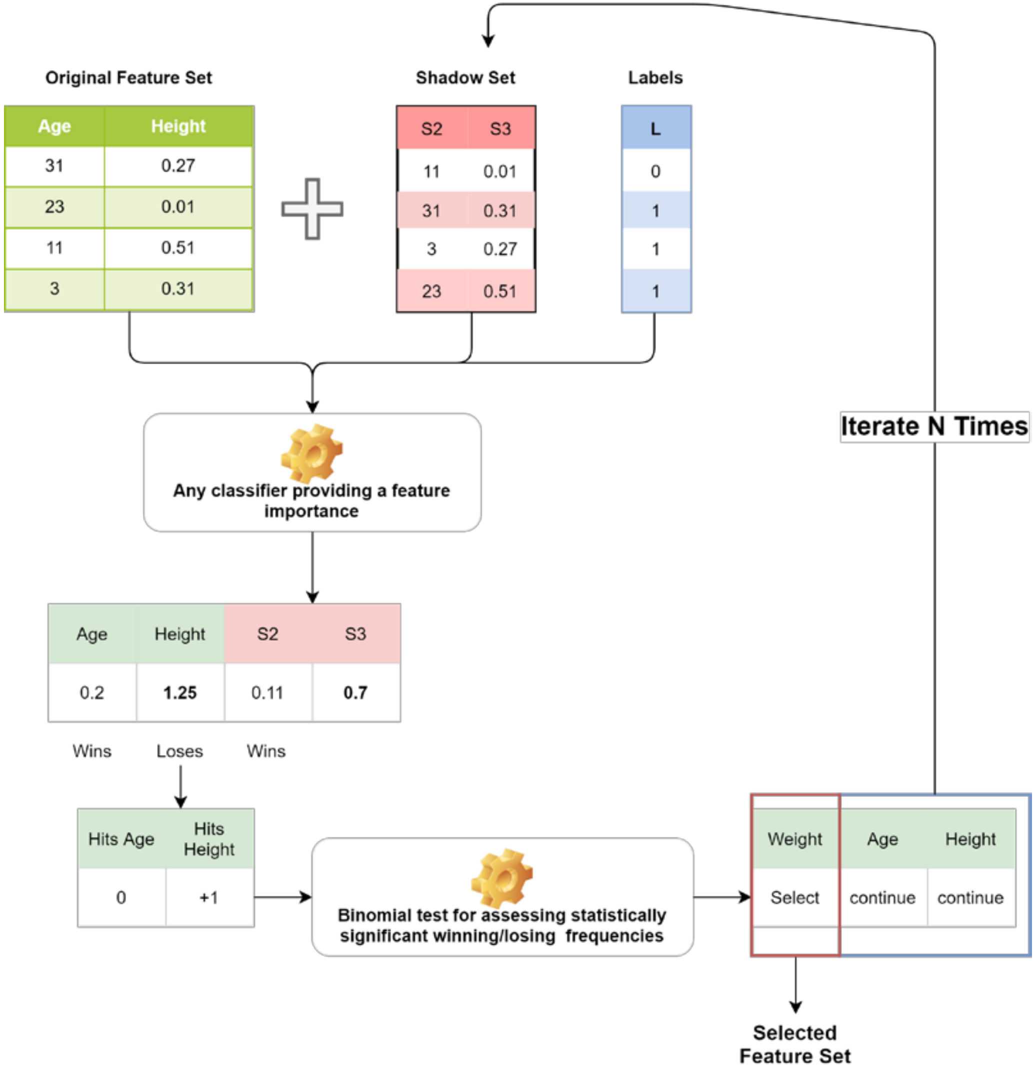
Hits Weight	Hits Age	Hits Height
+1	0	+1

Binomial test for assessing statistically significant winning/losing frequencies

Weight	Age	Height
Select	continue	continue

Iterate N Times

Selected Feature Set



Original Feature Set

Age	Height
31	0.27
23	0.01
11	0.51
3	0.31

Shadow Set

S2	S3
11	0.01
31	0.31
3	0.27
23	0.51

Labels

L
0
1
1
1

Age	Height	S2	S3
0.2	**1.25**	0.11	**0.7**

Hits Age	Hits Height
0	+1

Weight	Age	Height
Select	continue	continue

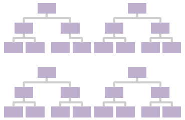
Any classifier providing a feature importance
Wins
Loses
Wins
Binomial test for assessing statistically significant winning/losing frequencies
Iterate N Times
Selected Feature Set

How to solve the “tentative” drama?

- Use internal $\mathbf{k}_{\mathbf{RF\_ext}}$-fold cross-validation
- For each fold j in 1 … $\mathbf{k}_{\mathbf{RFext}}$
    - Run $\mathbf{k}_{\mathbf{B\_int}}$ times* Borutaand keep features judged as selected or tentative for > $\mathbf{k}_{\mathbf{B\text{-}int}}$/2 times
    - Train a rebalanced* RF on selected features and measure feature importance
- keep “most important” features are kept

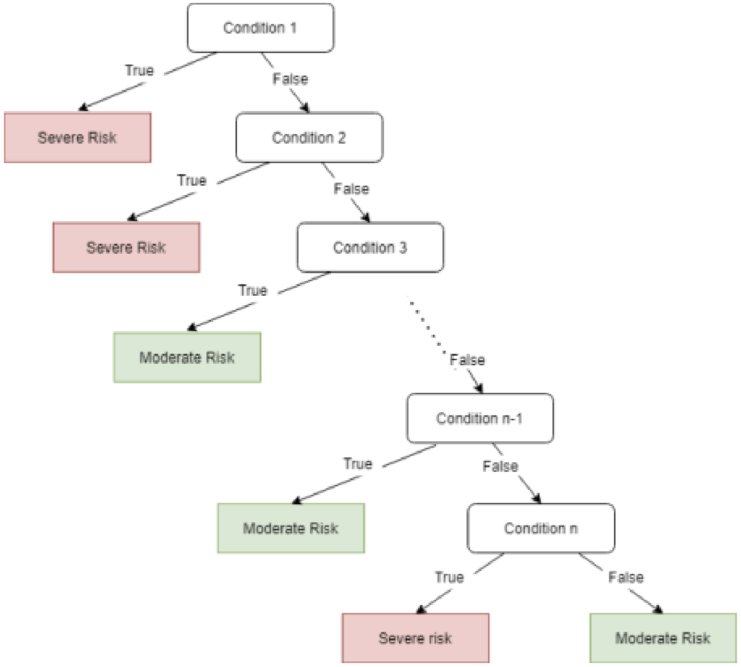
Risk prediction: Random Forest

- Hyper-parameters: n. of trees (fixed after preliminary grid search), n. of variables per split (greedy search), min. size of leaves (fixed following clinical experts advice)
- Class imbalance dealt via tailored bootstrap sampling
- Model assessment in a 10-fold CV
- Importance of predictors normalized in each fold, followed by averaging

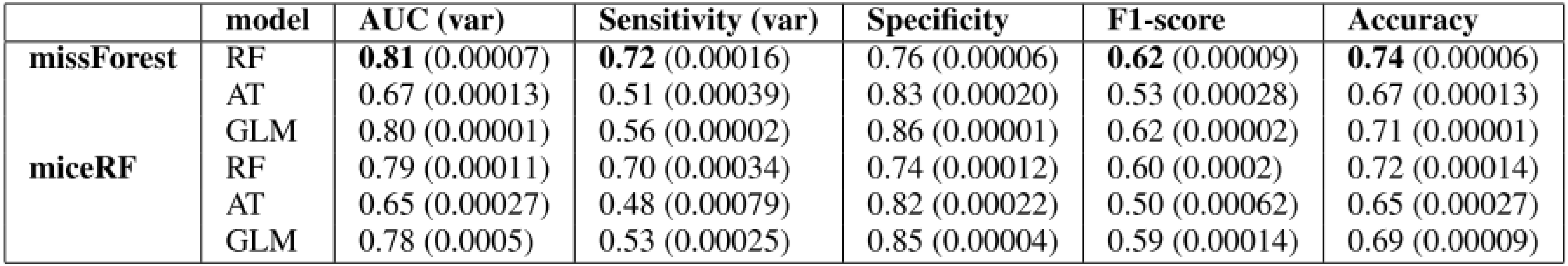


Model explainability

- RF are explainable, but tend to produce a big number of complex rules
- Solution: translate trees in the forest into Associative Trees



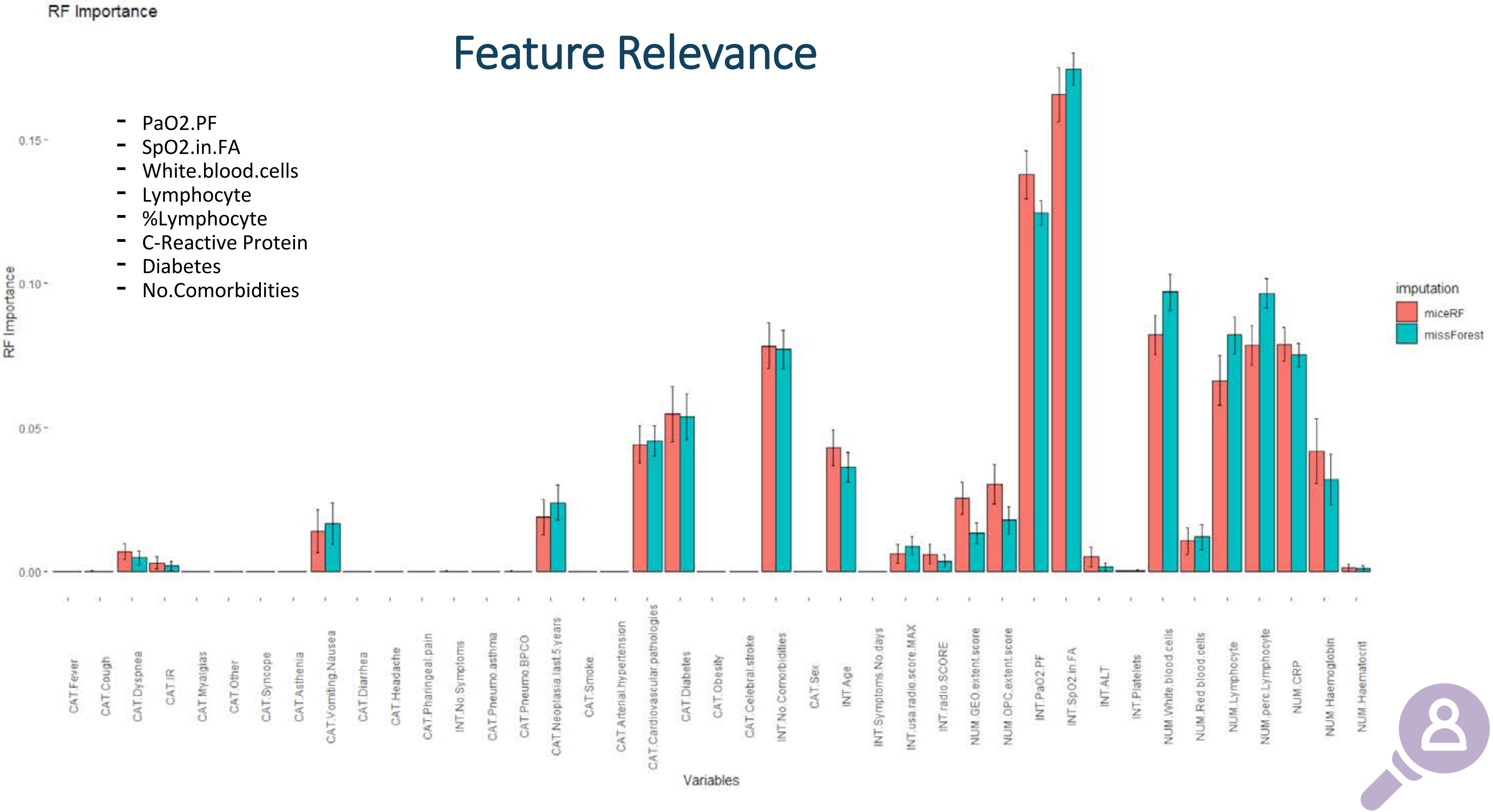
Results



	model	AUC (var)	Sensitivity (var)	Specificity	F1-score	Accuracy
missForest	RF	**0.81** (0.00007)	**0.72** (0.00016)	0.76 (0.00006)	**0.62** (0.00009)	**0.74** (0.00006)
	AT	0.67 (0.00013)	0.51 (0.00039)	0.83 (0.00020)	0.53 (0.00028)	0.67 (0.00013)
	GLM	0.80 (0.00001)	0.56 (0.00002)	0.86 (0.00001)	0.62 (0.00002)	0.71 (0.00001)
miceRF	RF	0.79 (0.00011)	0.70 (0.00034)	0.74 (0.00012)	0.60 (0.0002)	0.72 (0.00014)
	AT	0.65 (0.00027)	0.48 (0.00079)	0.82 (0.00022)	0.50 (0.00062)	0.65 (0.00027)
	GLM	0.78 (0.0005)	0.53 (0.00025)	0.85 (0.00004)	0.59 (0.00014)	0.69 (0.00009)

- AUC drove selection of imputation + learning algorithm
- Other metrics drove selection among learnt models
- Mean & Variance within 10-CV computed via Rubin's rule
- Statistical analyses confirm missForest+RF as best performing combination

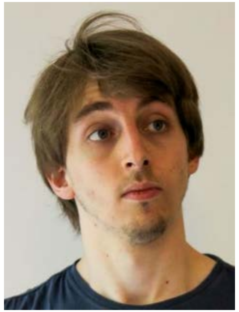
Feature Relevance



- PaO2.PF
- SpO2.in.FA
- White.blood.cells
- Lymphocyte
- %Lymphocyte
- C-Reactive Protein
- Diabetes
- No.Comorbidities

Implementation

- R code publicly available at https://github.com/AnacletoLAB/DataAnalysisR/tree/main/Rcode_clinicalDataNew
- A bit tailored on the experiments (use AYOR)
- Documentation not available
- Might however be a good starting point for a modular implementation on N3C Enclave



Thanks!

Questions ?

What’s next?

- Imputation as an “internal” processing step

- Graph based feature imputation

- try network-medicine approaches

Test the model on:

- Cremona Dataset (private by now but it may - sooner or later - be in N3C?)

- HUST-19 dataset http://ictcf.biocuckoo.cn/HUST-19.php