Applied Statistics A.A. 2019/2020

Automated Chest X-Ray Diagnostics

Francesco Giammaria

Stefano Gusmeroli

Edoardo Peretti

Matteo Savino

Alessio Tranchida

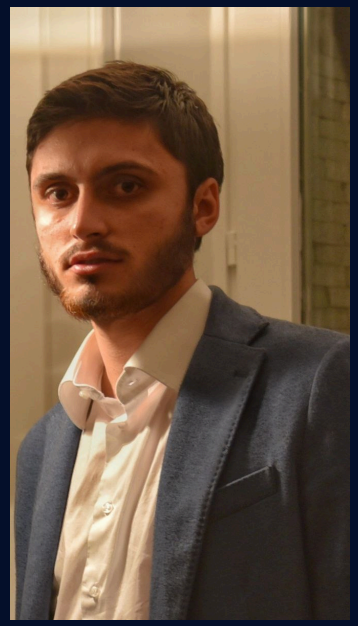
POLITECNICO
MILANO 1863

Tutor: Dr. Daniele Loiacono

This is us

Stefano Gusmeroli

Edoardo Peretti



Matteo Savino

Alessio Tranchida

Francesco Giammaria

DATA PRESENTATION

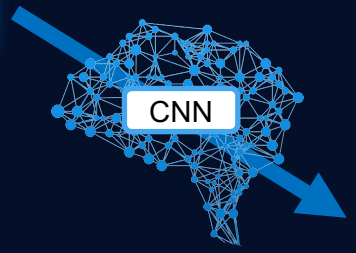
DATA PRESENTATION

More than 60,000 patients

More than 180,000 images

14 pathologies

input → Deep Convolutional Neural Network (CNN) → output → Embeddings

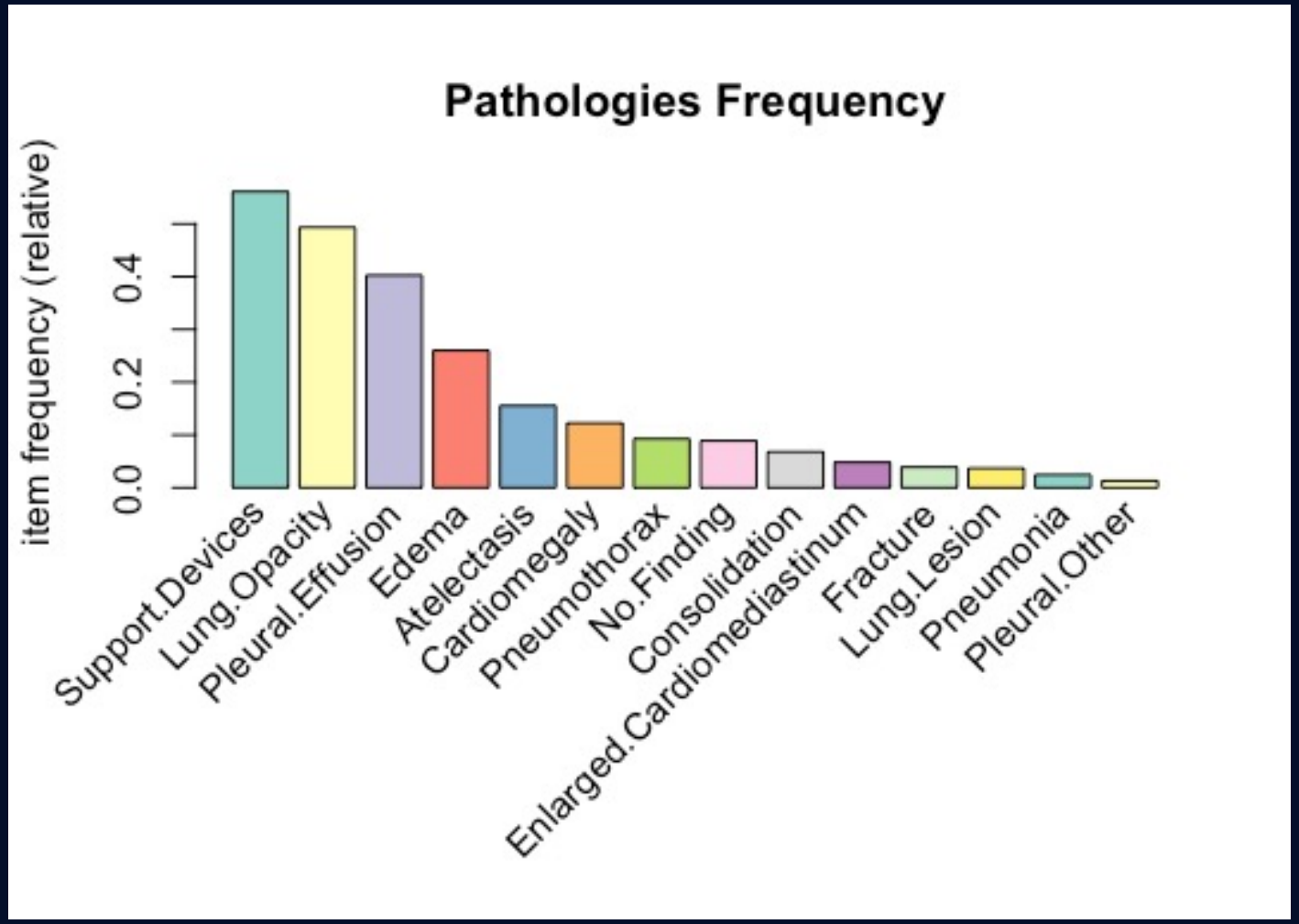


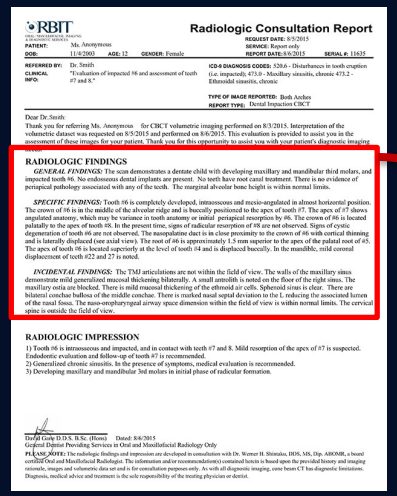
Embeddings

(189116 x 931)

	input_data.Path	X2	X3	X4	X5	X6	X8
1	CheXpert-v1.0-small/tra...	5.044617e-08	0.009756592	0.005829324	0.13361263	0.6152759	0
2	CheXpert-v1.0-small/tra...	0.000000e+00	0.008479902	0.005943561	0.10519153	0.6668760	0
3	CheXpert-v1.0-small/tra...	0.000000e+00	0.009295503	0.007226741	0.10922829	0.5146037	0
4	CheXpert-v1.0-small/tra...	0.000000e+00	0.008557481	0.006199519	0.09658385	0.8638877	0
5	CheXpert-v1.0-small/tra...	6.659148e-06	0.004773823	0.005670027	0.14989142	0.5566312	0
6	CheXpert-v1.0-small/tra...	0.000000e+00	0.009678223	0.006153073	0.13729700	0.5397011	0
7	CheXpert-v1.0-small/tra...	0.000000e+00	0.008135295	0.005715157	0.05424061	0.3945267	0
8	CheXpert-v1.0-small/tra...	0.000000e+00	0.009958112	0.006867255	0.07581601	0.9344040	0
9	CheXpert-v1.0-small/tra...	5.897515e-06	0.007355212	0.006465926	0.17092736	0.5021766	0

14 pathologies





Radiologic Consultation Report

RADIOLOGIC FINDINGS

RADIOLOGIC IMPRESSION

Labels

Uncertainties

	Cardiomegaly	Lung.Opacity	Edema
1	0	1	1
2	0	1	0
3	0	1	0
4	0	U	0
5	0	1	U
6	1	1	0
7	0	1	0
8	1	1	1
9	0	1	0
10	U	1	0

Guidelines

Embeddings Analysis

Treatment of Uncertainties

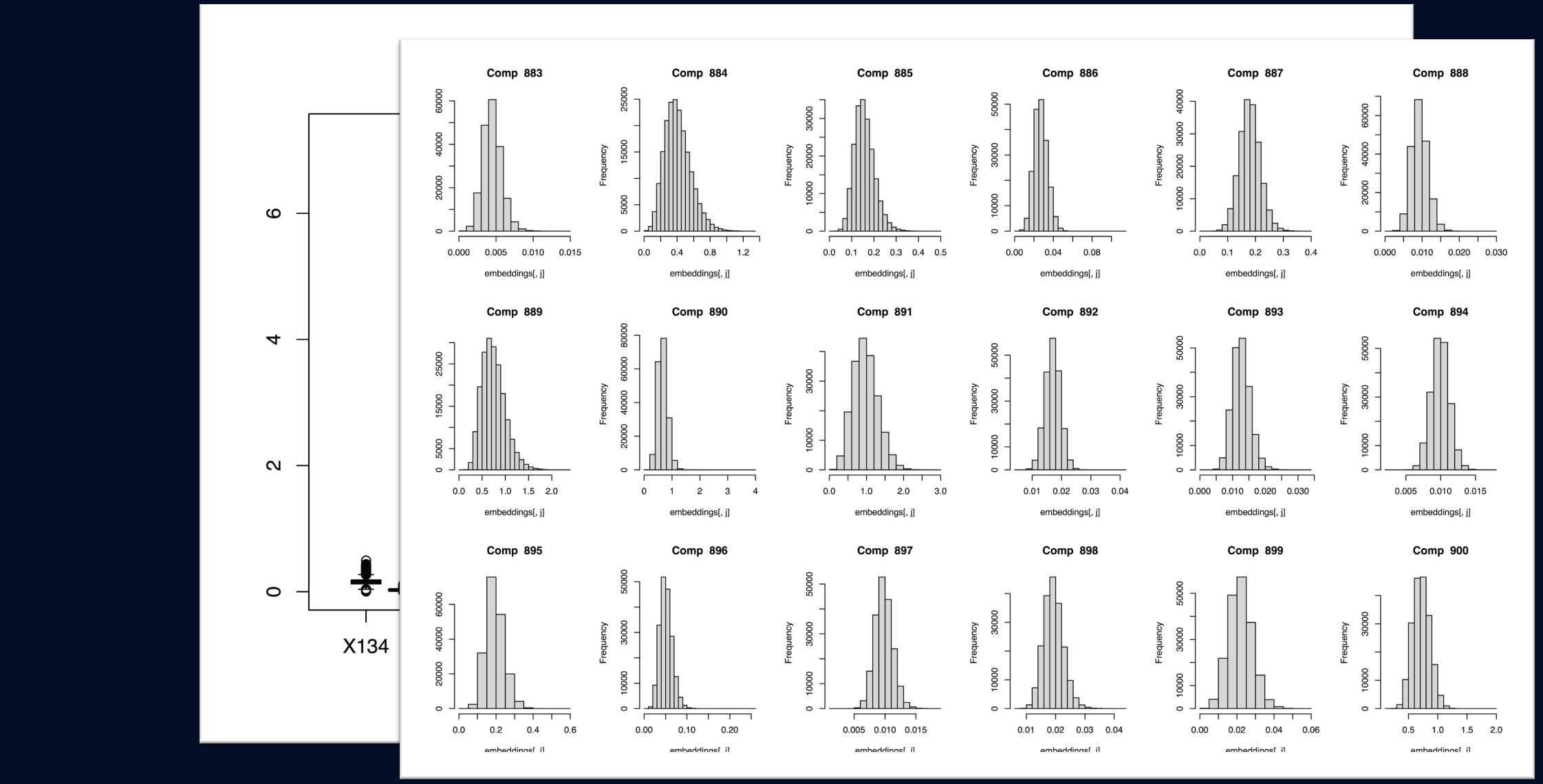
Pathologies Relationships

GRAPHICAL REPRESENTATION

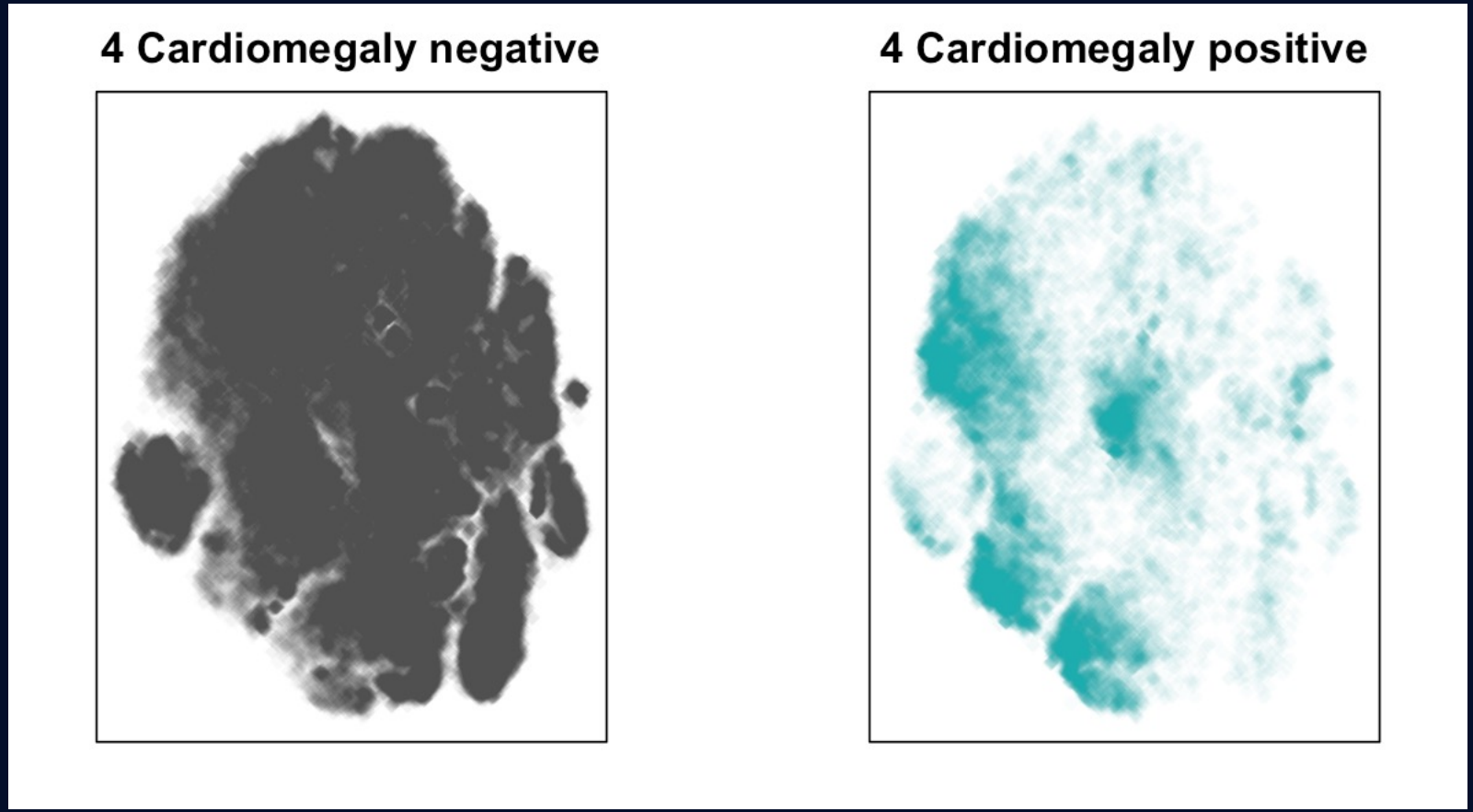
POLITECNICO
MILANO 1863

Automated Chest X-Ray Diagnostics

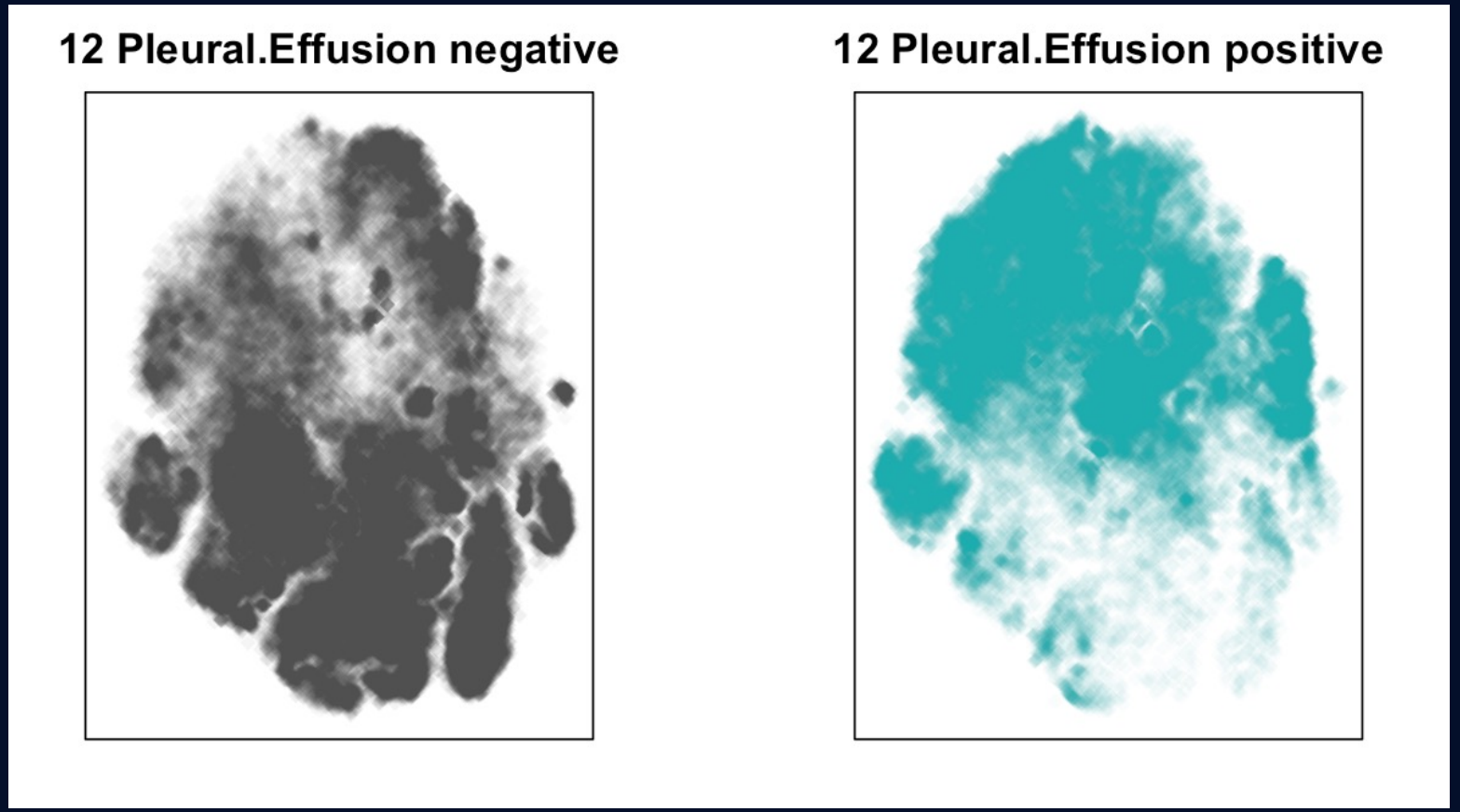
Applied Statistics
A.A. 2019/2020



t-SNE (t-distibuted stochastic neighbor embedding)

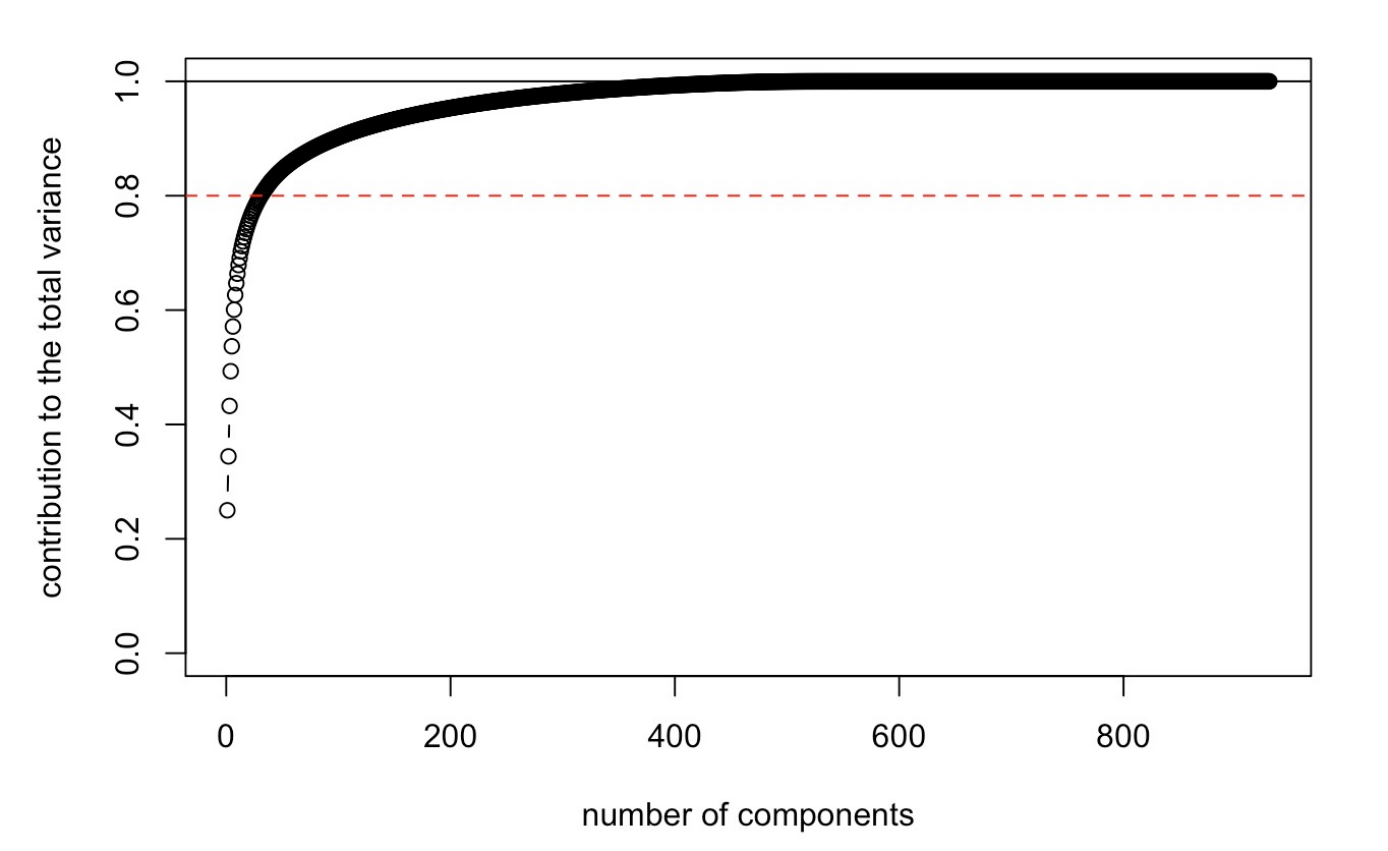


t-SNE


12 Pleural.Effusion negative
12 Pleural.Effusion positive

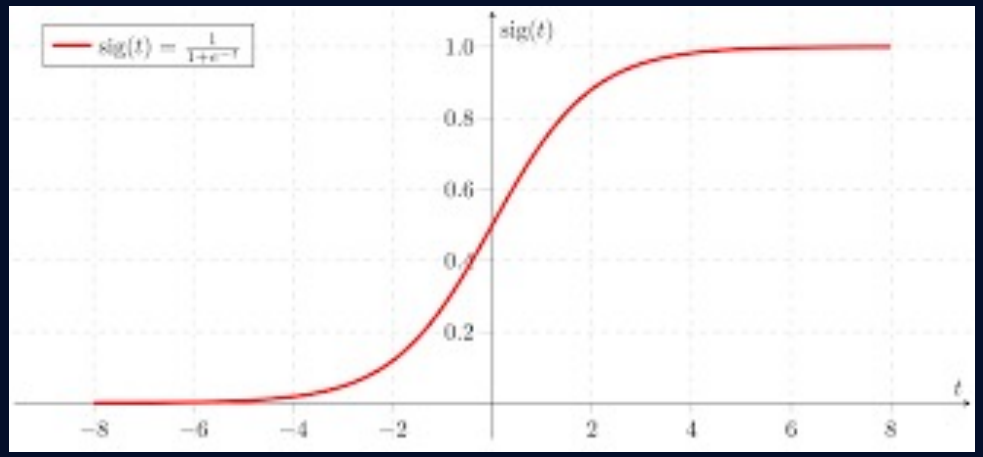
CLASSIFICATION MODELS

DIMENSIONALITY REDUCTION ➡ PCA



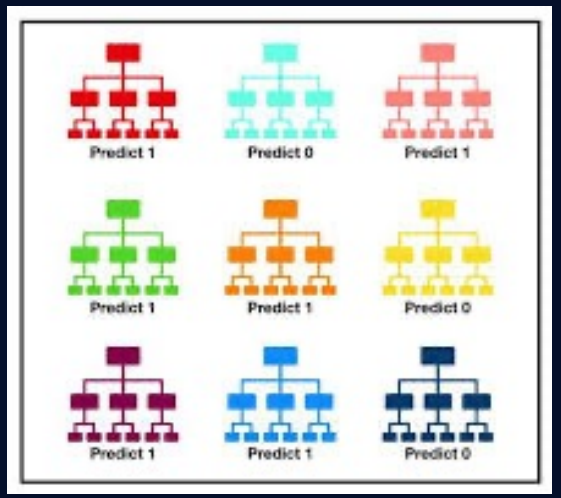
LOGISTIC REGRESSION

- Bagging Model



RANDOM FOREST

- 100 trees
- features random selection
- Quality of split: Gini index

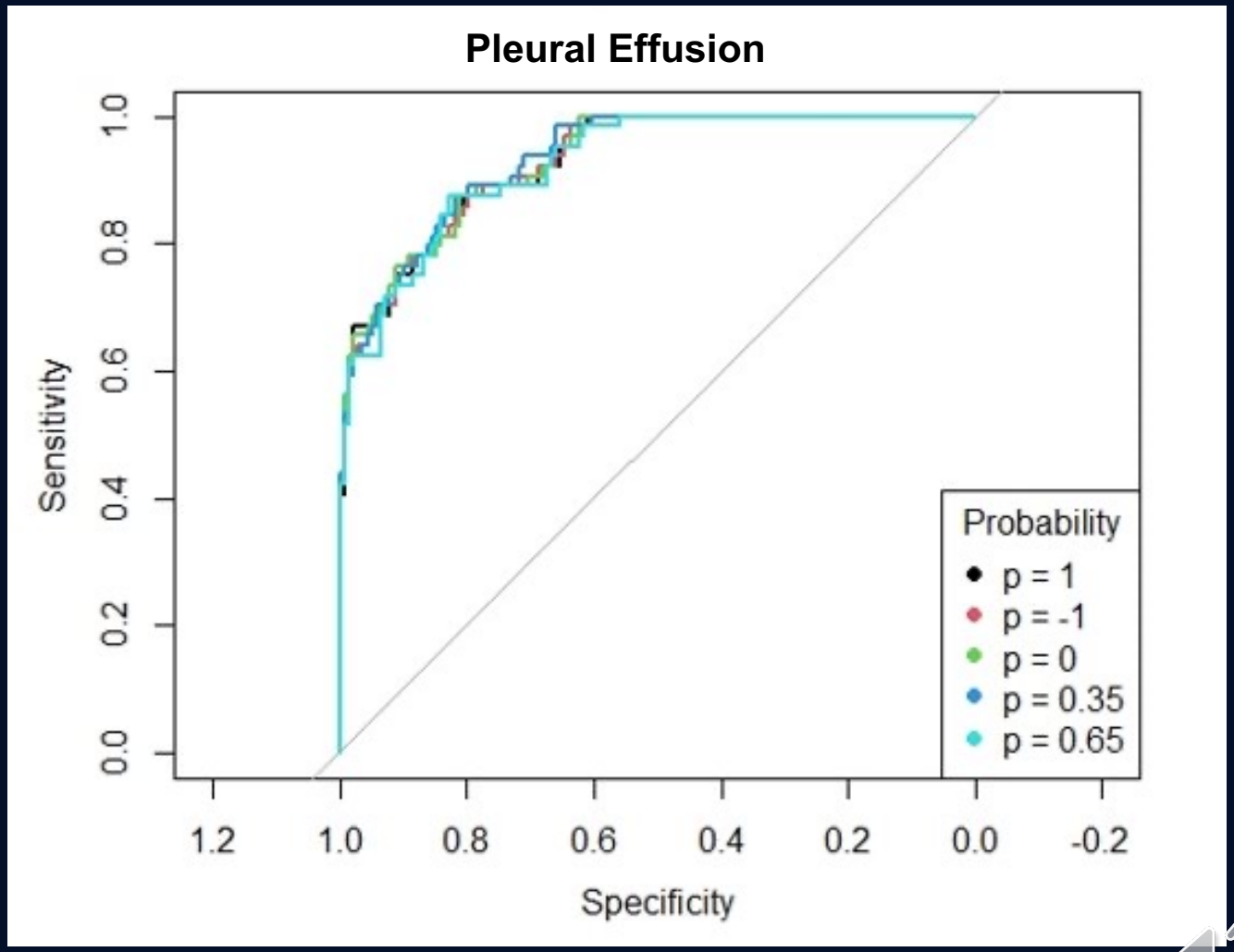


Treatment of Uncertanties

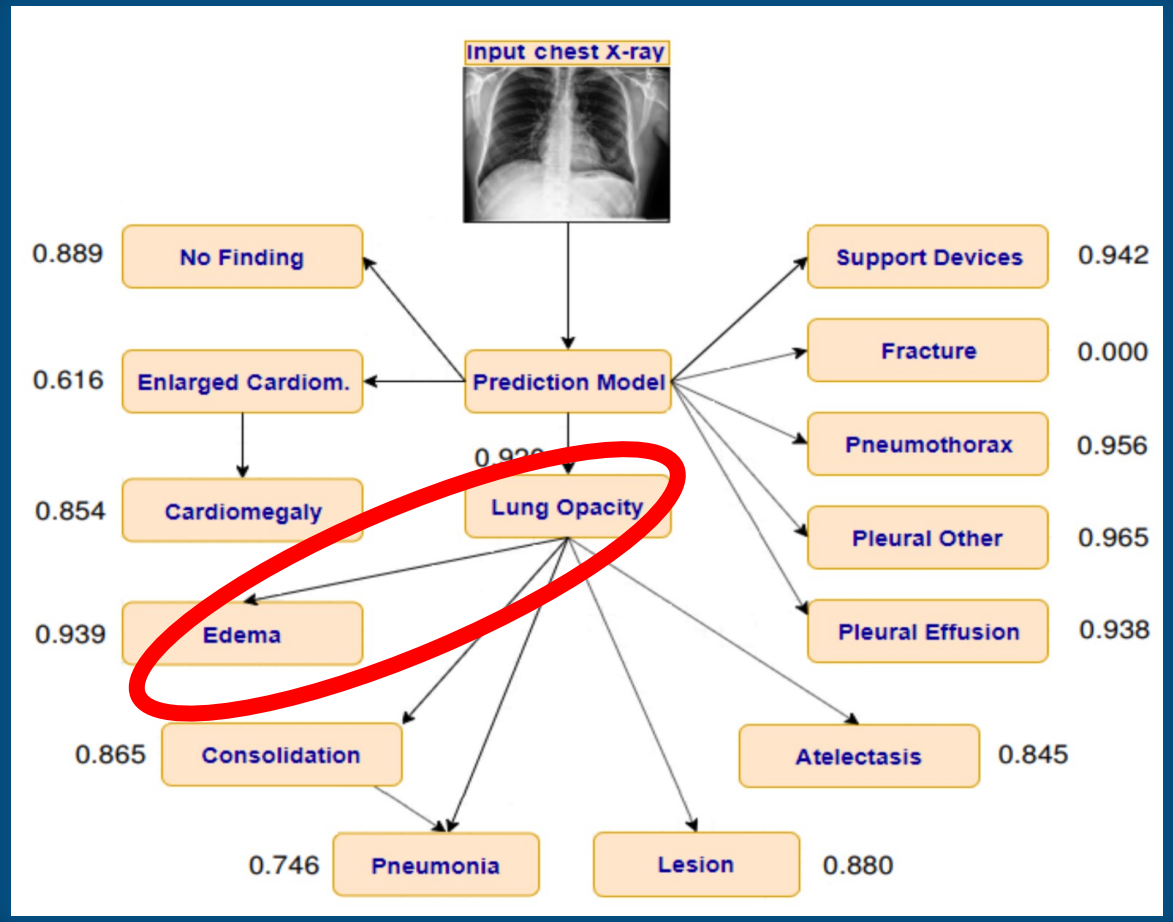
Uncertain values can be:

- Discarded (p = -1)
- All assigned to 0 (p = 0)
- All assigned to 1 (p = 1)
- Assigned to 1 with probability p

(Logit Models for different uncertainties)



ASSOCIATIONS



[Pham et al. 2019]

Conditional Models

LUNG.OPACITY ➡ EDEMA

EDEMA									
X345	X346	X347	X348	X349	X350	X351	X352	X353	Lung.Opacity
0.005	0.034	0.94	0.56	0.23	0.66	0.26	0.001	0.569	1
0.567	0.734	0.005	0.745	0.234	0.657	0.678	0.392	0.004	1
0.178	0.167	0.001	0.070	0.788	0.167	0.281	0.192	0.182	0
0.121	0.431	0.379	0.781	0.112	0.011	0.087	0.965	0.289	0
0.078	0.256	0.389	0.719	0.001	0.781	0.228	0.118	0.801	1

Corrected Labels

Edema	Lung. Opacity	Cardiomegaly
1	u	1
0	u	1
0	0	1
1	0	0
1	1	1

Edema	Lung. Opacity	Cardiomegaly
1	1	1
0	u	1
0	0	1
1	1	0
1	1	1

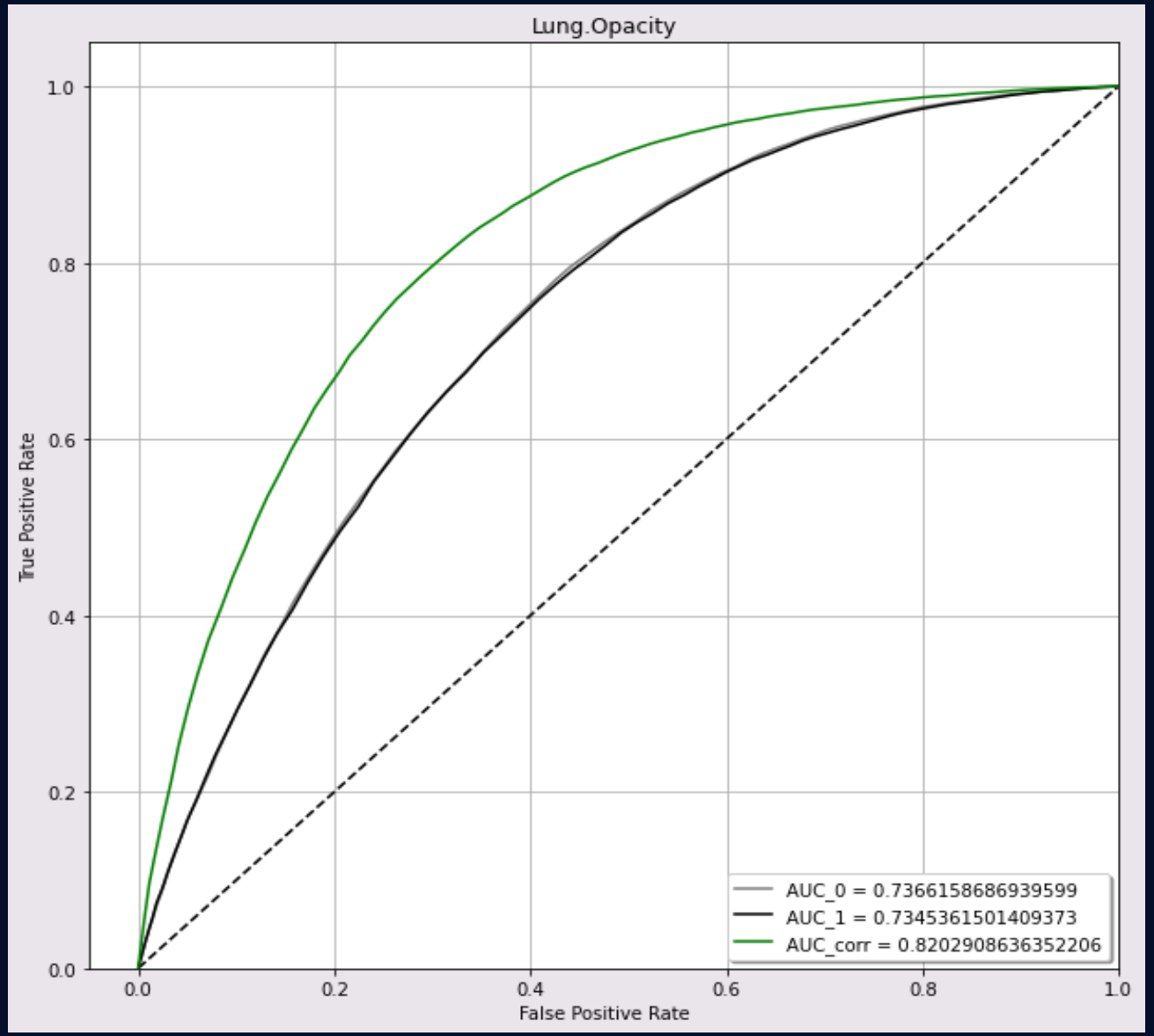
Conditional models:
increase in AUC up to 10÷20%, e.g. Cardiomegaly and Edema

Corrected labels: further increase of 1.5%

RANDOM FOREST

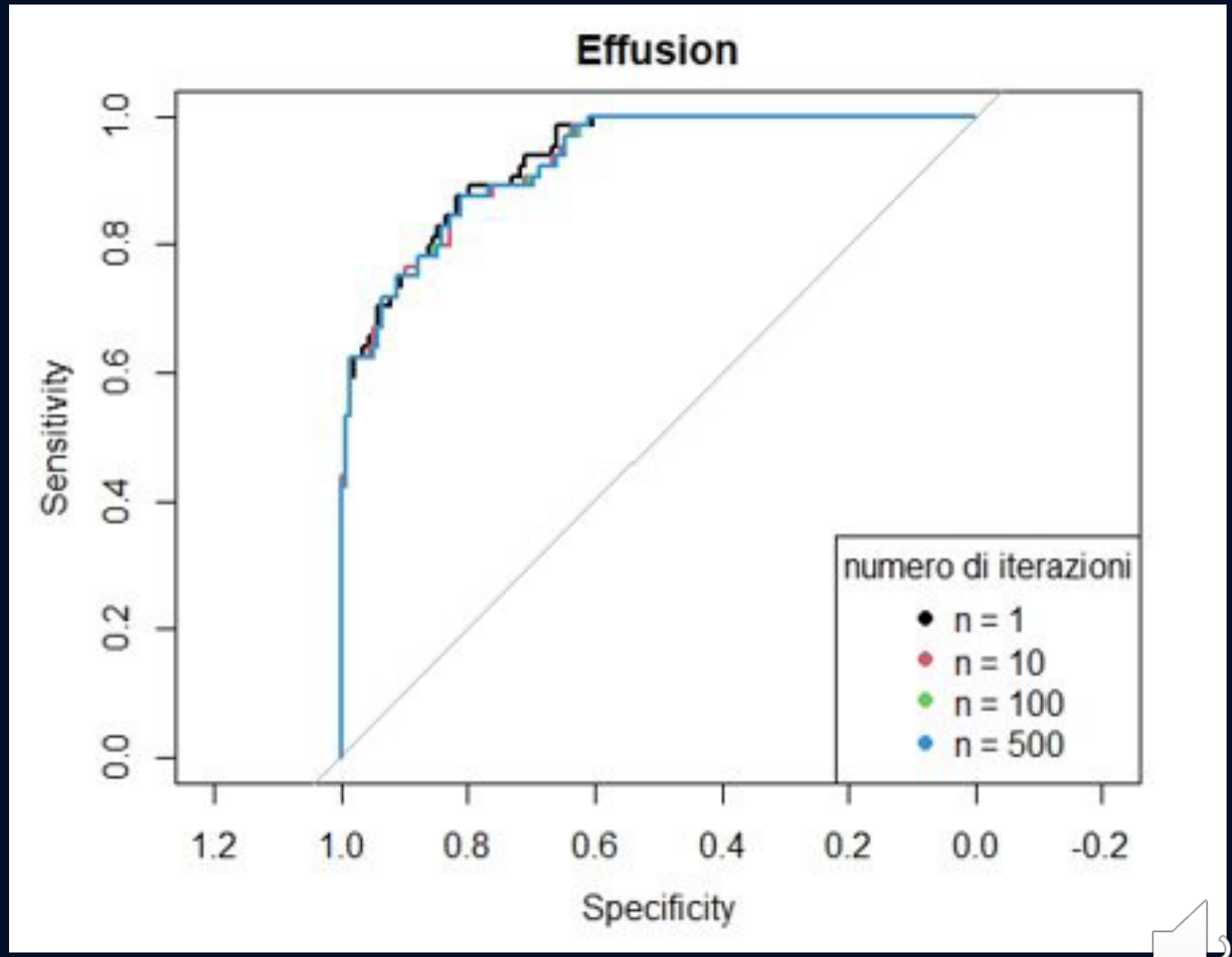
Lung.Opacity:

U = 0
U = 1
U = corrected



Bagging

- **Bagging**: sample with replacement (bootstrap) + aggregation
- Ensemble model with bootstrapped training sets
- At each iteration **uncertainties are resampled**



CONCLUSION

CONCLUSION

- ✓ Conditional models are very useful
- ✓ Corrected labels are helpful
- ✗ Treatment of uncertainties (because uncertainties << observations)
- ✗ Bagging for logistic regression

THANK YOU

We look forward to answer your question!!!

If you want, join us at the following link:

https://github.com/edpere/CheXpert-analysis

CheXpert competition: https://stanfordmlgroup.github.io/competitions/chexpert/

Bibliography:

- *Interpreting chest X-rays via CNNs that exploit disease dependencies and uncertainty labels,* Hieu H. Pham, Tung T. Le, Dat Q. Tran, Dat T. Ngo, Ha Q. Nguyen 2019
- *CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison*, Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, et al.

POLITECNICO
MILANO 1863

Automated Chest X-Ray Diagnostics

Applied Statistics
A.A. 2019/2020

DATA PRESENTATION

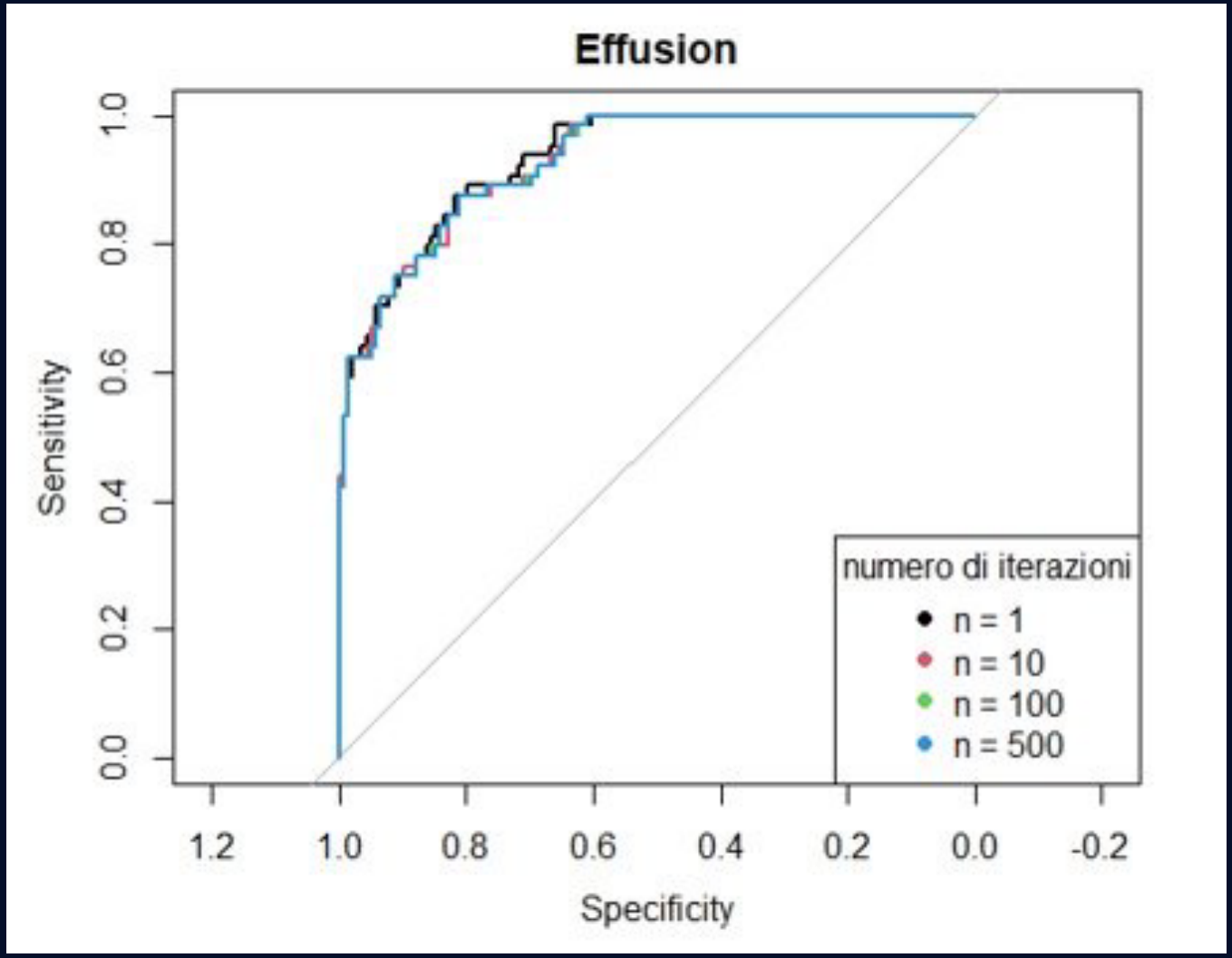
GRAPHICAL REPRESENTATION

CLASSIFICATION MODELS

CONCLUSION

Lasso

- We looked for the optimal hyper-parameter α using the **validation set**
- In the end we do not see great improvements



Conclusions

- ✓ Conditional models are very useful
- ✓ Corrected labels can be helpful
- ✗ Treatment of uncertainties (because of big dataset)
- ✗ Bagging for logistic regression
- ✗ Regularization (e.g. lasso) (no overfitting)