Semi-supervised learning, PET/CT segmentation and prediction of the tumoral response to immunotherapy

PhD defense

Hugo Schmutz
Université Côte d'Azur, INRIA

Supervised by Olivier Humbert (IBV, UniCA)
and Pierre-Alexandre Mattei (MAASAI,
INRIA)

Lung cancer is the deadliest cancer worldwide

- The second-most diagnosed form of cancer (2.2 M cases/year, 11% of all cancer cases)

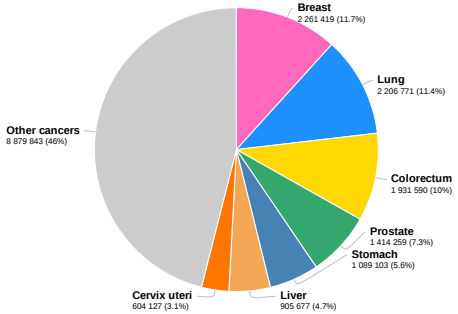


Figure: Incidence per cancer type in 2020, GLOBOCAN [Sung et al., 2021]

Lung cancer is the deadliest cancer worldwide

- The second-most diagnosed form of cancer (2.2 M cases/year, 11% of all cancer cases)
- The leading cause of cancer deaths (1.8M deaths in 2020)

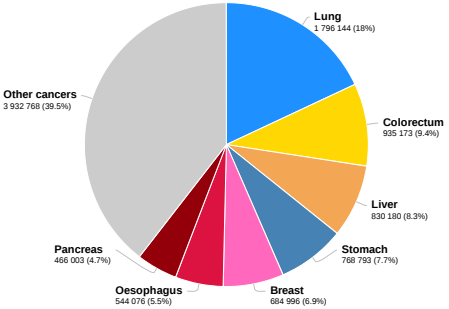


Figure: Death per cancer type in 2020, GLOBOCAN [Sung et al., 2021]

Lung cancer is the deadliest cancer worldwide

- The second-most diagnosed form of cancer (2.2 M cases/year, 11% of all cancer cases)
- The leading cause of cancer deaths (1.8M deaths in 2020)
- Often diagnosed at an advanced stage

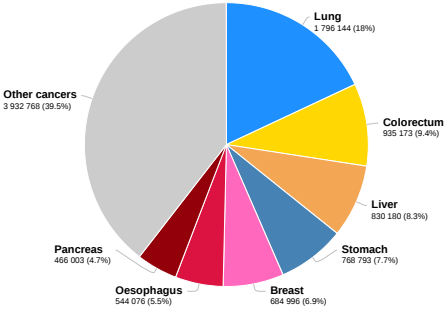


Figure: Death per cancer type in 2020, GLOBOCAN [Sung et al., 2021]

Lung cancer is the deadliest cancer worldwide

- The second-most diagnosed form of cancer (2.2 M cases/year, 11% of all cancer cases)
- The leading cause of cancer deaths (1.8M deaths in 2020)
- Often diagnosed at an advanced stage
- Overall 5-year relative survival rate 28% (Stage I: 70%; Stage IV: 13%, [Woodard et al., 2016])

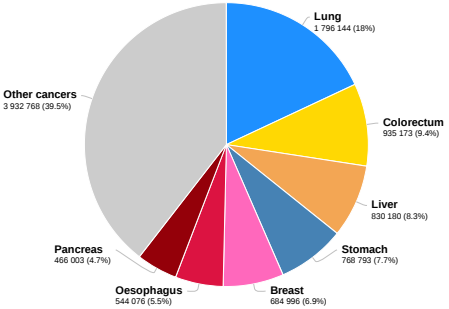


Figure: Death per cancer type in 2020, GLOBOCAN [Sung et al., 2021]

Lung cancer is the deadliest cancer worldwide

- The second-most diagnosed form of cancer (2.2 M cases/year, 11% of all cancer cases)
- The leading cause of cancer deaths (1.8M deaths in 2020)
- Often diagnosed at an advanced stage
- Overall 5-year relative survival rate 28% (Stage I: 70%; Stage IV: 13%, [Woodard et al., 2016])
- Different cancer types
  - 85%: Non-small cell lung cancer (NSCLC)
  - 15%: Small cell lung cancer (SCLC)

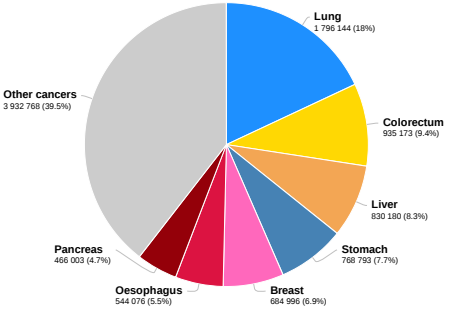


Figure: Death per cancer type in 2020, GLOBOCAN [Sung et al., 2021]

Chosing the most appropriate cure is crucial!

Treatment options:

- Surgery
- Chemotherapy
- Radiotherapy
- Immunotherapy
- ...

Chosing the most appropriate cure is crucial!

Treatment options:

- Surgery
- Chemotherapy
- Radiotherapy
- Immunotherapy
- ...

Depends on the:

- Cancer stage
- Patient's overall health
- Profile of the tumour
- ...

What is anti-tumour immunotherapy?

- Recent treatment revolution against lung cancer (Breakthrough of the Year, Science 2013)

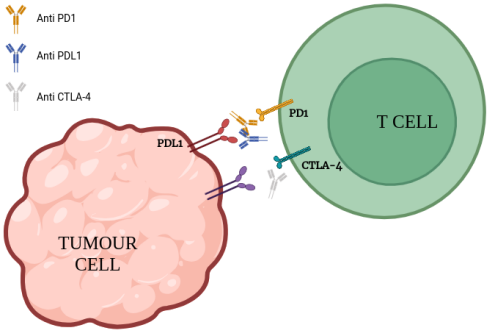


Figure: Created with BioRender.com.

What is anti-tumour immunotherapy?

- Recent treatment revolution against lung cancer (Breakthrough of the Year, Science 2013)
- Restores an effective anti-tumour immunity
- Prevents the deactivation of immune cells through immune checkpoint inhibition (ICPI)

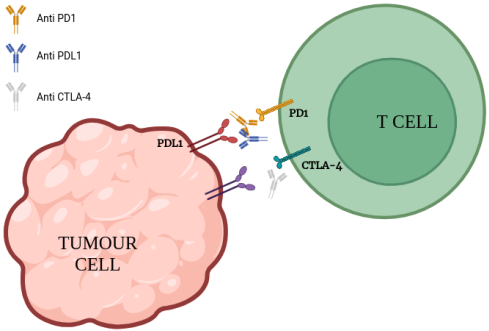


Figure: Created with BioRender.com.

What is anti-tumour immunotherapy?

- Recent treatment revolution against lung cancer (Breakthrough of the Year, Science 2013)
- Restores an effective anti-tumour immunity
- Prevents the deactivation of immune cells through immune checkpoint inhibition (ICPI)
- Objective tumour response in 18% of lung cancer patients [Mazieres et al., 2019]
- Positive impact: between 30% and 40% [Herbst et al., 2016]

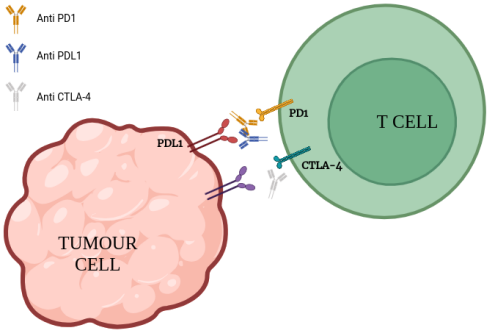


Figure: Created with BioRender.com.

What is anti-tumour immunotherapy?

- Recent treatment revolution against lung cancer (Breakthrough of the Year, Science 2013)
- Restores an effective anti-tumour immunity
- Prevents the deactivation of immune cells through immune checkpoint inhibition (ICPI)
- Objective tumour response in 18% of lung cancer patients [Mazieres et al., 2019]
- Positive impact: between 30% and 40% [Herbst et al., 2016]
- Pseudoprogression leads to inadequate cessation of the treatment [Y. Ma et al., 2019]
- Causes side effects, mostly benign, but sometimes fatal [Kroschinsky et al., 2017]

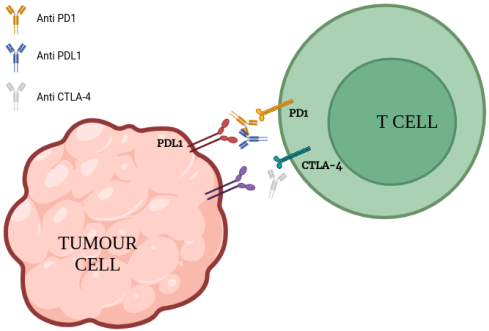


Figure: Created with BioRender.com.

What is anti-tumour immunotherapy?

- Recent treatment revolution against lung cancer (Breakthrough of the Year, Science 2013)
- Restores an effective anti-tumour immunity
- Prevents the deactivation of immune cells through immune checkpoint inhibition (ICPI)
- Objective tumour response in 18% of lung cancer patients [Mazieres et al., 2019]
- Positive impact: between 30% and 40% [Herbst et al., 2016]
- Pseudoprogression leads to inadequate cessation of the treatment [Y. Ma et al., 2019]
- Causes side effects, mostly benign, but sometimes fatal [Kroschinsky et al., 2017]

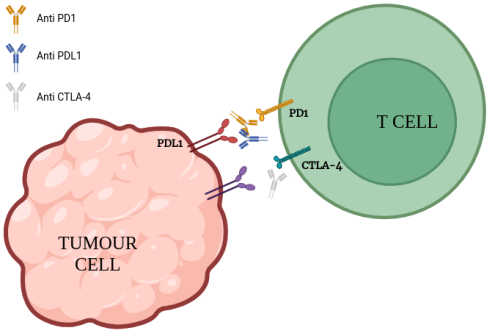


Figure: Created with BioRender.com.

⟶ Anticipating patients' response to immunotherapy is decisive!

What is anti-tumour immunotherapy?

- Recent treatment revolution against lung cancer (Breakthrough of the Year, Science 2013)
- Restores an effective anti-tumour immunity
- Prevents the deactivation of immune cells through immune checkpoint inhibition (ICPI)
- Objective tumour response in 18% of lung cancer patients [Mazieres et al., 2019]
- Positive impact: between 30% and 40% [Herbst et al., 2016]
- Pseudoprogression leads to inadequate cessation of the treatment [Y. Ma et al., 2019]
- Causes side effects, mostly benign, but sometimes fatal [Kroschinsky et al., 2017]

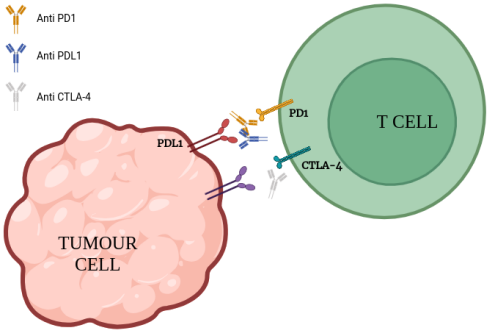


Figure: Created with BioRender.com.

$\longrightarrow$ Anticipating patients' response to immunotherapy is decisive!
Biomarkers can help identify patients that might respond positively

Only two biomarkers have been approved by the US Food and Drug Administration

- Expression level of PD-L1 in tumour cells (PDL1%, in routine, [Herbst, Baas, et al., 2016])
- Average number of mutations present in a tumour cell's DNA (Tumour mutational burden, [Marcus et al., 2021])

High expression levels = higher chance of response

Only two biomarkers have been approved by the US Food and Drug Administration

- Expression level of PD-L1 in tumour cells (PDL1%, in routine, [Herbst, Baas, et al., 2016])
- Average number of mutations present in a tumour cell's DNA (Tumour mutational burden, [Marcus et al., 2021])

High expression levels = higher chance of response

Drawbacks:

- Limited prediction power [Bodor et al., 2020]
- Inter-rater disagreement and intra-tumour and patient variation [Mansfield et al., 2016, Stenzinger et al., 2019]
- Invasive procedures
- TMB requires sequencing analysis (not in used in routine practice and expensive)

Only two biomarkers have been approved by the US Food and Drug Administration

- Expression level of PD-L1 in tumour cells (PDL1%, in routine, [Herbst, Baas, et al., 2016])
- Average number of mutations present in a tumour cell's DNA (Tumour mutational burden, [Marcus et al., 2021])

High expression levels = higher chance of response

Drawbacks:

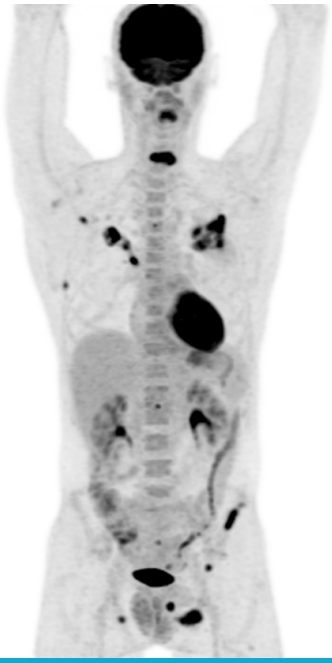
- Limited prediction power [Bodor et al., 2020]
- Inter-rater disagreement and intra-tumour and patient variation [Mansfield et al., 2016, Stenzinger et al., 2019]
- Invasive procedures
- TMB requires sequencing analysis (not in used in routine practice and expensive)

Other known factors and biomarkers: ECOG performance score [Dall'Olio et al., 2020], neutrophil over lymphocyte ratio [Valero et al., 2021], etc.

Medical imaging biomakers do not require invasive procedures

^{18}F-FDG PET/CT scans:

- 3D whole-body images
- CT: morphological information
- ^{18}F-FDG PET: glucose metabolic activity



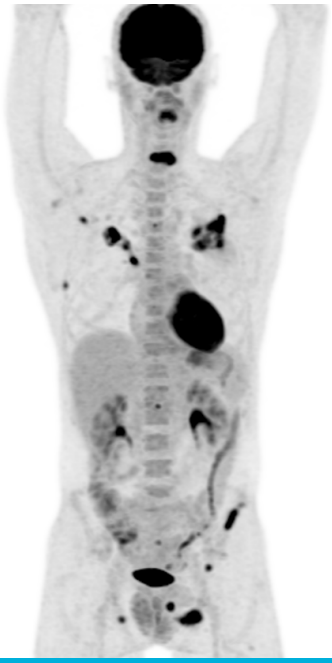
Medical imaging biomakers do not require invasive procedures

^{18}F-FDG PET/CT scans:

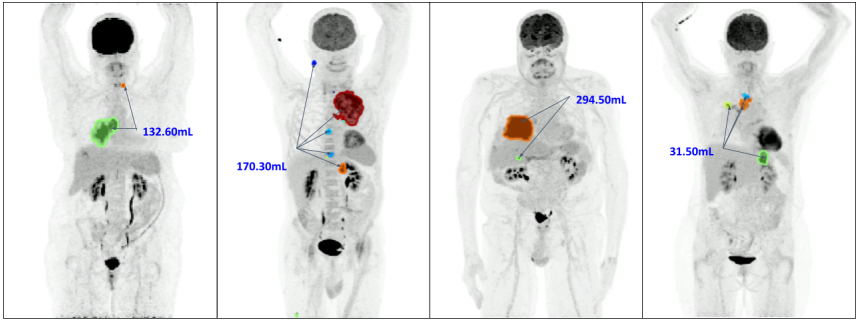
- 3D whole-body images
- CT: morphological information
- ^{18}F-FDG PET: glucose metabolic activity

Advantages:

- Available as a routine exam
- Known biomarkers for other cancer treatments [Alderuccio et al., 2023]
- Non-invasive examinations
- Possibility to visualise all metastatic lesions



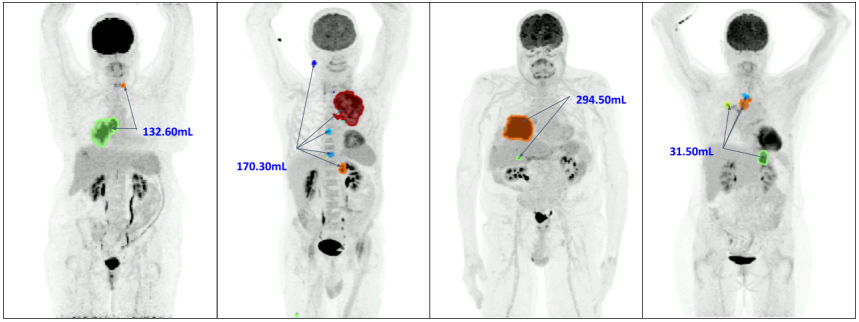
Medical imaging biomakers do not require invasive procedures



[Seban, Mezquita, et al., 2020; Chardin et al., 2020]

- Maximum standard uptake value (SUV_{max})
- Total metabolic tumour volume (tMTV)
- Total lesion glycolysis (TLG)
- Spleen to liver ratio (SLR)
- Bone marrow to liver ratio (BLR)
- Number of lesions
- ...

Medical imaging biomakers do not require invasive procedures



[Seban, Mezquita, et al., 2020; Chardin et al., 2020]

- Maximum standard uptake value (SUV_{max})
- Total metabolic tumour volume (tMTV)
- Total lesion glycolysis (TLG)
- Spleen to liver ratio (SLR)
- Bone marrow to liver ratio (BLR)
- Number of lesions
- ...

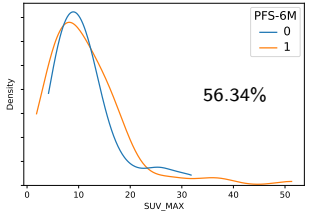
$\longrightarrow$ Requires manual segmentation of PET/CT scans

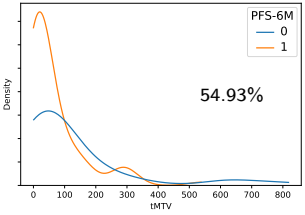
Two main challenges: Performance and collection of the biomarkers

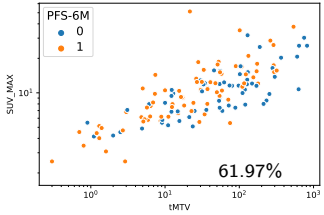
On the performance of biomarkers:

- Inconsistent results through different studies
- Non-correlation of the biomarkers

$\longrightarrow$ Towards a combination of biomarkers







Two main challenges: Performance and collection of the biomarkers

On the performance of biomarkers:

- Inconsistent results through different studies
- Non-correlation of the biomarkers

$\longrightarrow$ Towards a combination of biomarkers

On the collection and standardisation of imaging biomarkers:

- Inter-rater disagreement

Two main challenges: Performance and collection of the biomarkers

On the performance of biomarkers:

- Inconsistent results through different studies
- Non-correlation of the biomarkers

$\longrightarrow$ Towards a combination of biomarkers

On the collection and standardisation of imaging biomarkers:

- Inter-rater disagreement
- Time-consuming and fastidious

Two main challenges: Performance and collection of the biomarkers

On the performance of biomarkers:

- Inconsistent results through different studies
- Non-correlation of the biomarkers

$\longrightarrow$ Towards a combination of biomarkers

On the collection and standardisation of imaging biomarkers:

- Inter-rater disagreement
- Time-consuming and fastidious $\longrightarrow$ Automatic segmentation of PET/CT scans

Two main challenges: Performance and collection of the biomarkers

On the performance of biomarkers:

- Inconsistent results through different studies
- Non-correlation of the biomarkers

$\longrightarrow$ Towards a combination of biomarkers

On the collection and standardisation of imaging biomarkers:

- Inter-rater disagreement
- Time-consuming and fastidious $\longrightarrow$ Automatic segmentation of PET/CT scans
- Lack of segmented (labelled) PET/CT scans

Two main challenges: Performance and collection of the biomarkers

On the performance of biomarkers:

- Inconsistent results through different studies
- Non-correlation of the biomarkers

$\longrightarrow$ Towards a combination of biomarkers

On the collection and standardisation of imaging biomarkers:

- Inter-rater disagreement
- Time-consuming and fastidious $\longrightarrow$ Automatic segmentation of PET/CT scans
- Lack of segmented (labelled) PET/CT scans $\longrightarrow$ Using both labelled and unlabelled data

One objective, two axes

End-to-end pipeline for the collection of biomarkers and prediction of the response to immunotherapy for NSCLC patients

Simple and interpretable, yet powerful combination of heterogeneous biomarkers

Biomarker selection for the prediction of the response to immunotherapy

Automatic collection of PET/CT imaging biomarkers

- Automatic segmentation of PET/CT scans
- Make use of unlabelled data
- Safely: prevent potential degradation

Organisation of the presentation

A. On the performance of biomarkers
   Combination of heterogeneous biomarkers

B. On the collection of biomarkers
   0. Semi-supervised learning generalities
   1. DeSSL: Safe semi-supervised learning via simple debiasing
   2. DeSegSSL: Safe semi-supervised learning for medical image segmentation

Outline

A. On the performance of biomarkers
   Combination of heterogeneous biomarkers

B. On the collection of biomarkers
   0. Semi-supervised learning generalities
   1. DeSSL: Safe semi-supervised learning via simple debiasing
   2. DeSegSSL: Safe semi-supervised learning for medical image segmentation

Population description: a multicentric dataset

142 patients with metastatic NSCLC before initiation of immunotherapy in CAL (Nice) and CHPG (Monaco)

Population description: a multicentric dataset

142 patients with metastatic NSCLC before initiation of immunotherapy in CAL (Nice) and CHPG (Monaco)

Collected biomarkers (from routine practice):

- 12 clinical features (age, weight, height, ECOG, ...)
- 6 biological features (PDL1%, NLR, ...)
- 9 PET/CT parameters (number of lesions, SUV_{max}, tMTV, SLR...)

Population description: a multicentric dataset

142 patients with metastatic NSCLC before initiation of immunotherapy in CAL (Nice) and CHPG (Monaco)

Collected biomarkers (from routine practice):

- 12 clinical features (age, weight, height, ECOG, ...)
- 6 biological features (PDL1%, NLR, ...)
- 9 PET/CT parameters (number of lesions, SUV_{max}, tMTV, SLR...)

Two endpoints:

- 6-month progression-free survival (6M-PFS: 45.07% of reccurence)
- 12-month overall survival (12M-OS: 37.59% of death)

Population description: a multicentric dataset

142 patients with metastatic NSCLC before initiation of immunotherapy in CAL (Nice) and CHPG (Monaco)

Collected biomarkers (from routine practice):

- 12 clinical features (age, weight, height, ECOG, ...)
- 6 biological features (PDL1%, NLR, ...)
- 9 PET/CT parameters (number of lesions, SUV_{max}, tMTV, SLR...)

Two endpoints:

- 6-month progression-free survival (6M-PFS: 45.07% of reccurence)
- 12-month overall survival (12M-OS: 37.59% of death)

22 missing values of PDL1% imputed with MICE [Van Buuren, 2018].

What is Lasso?

Notations:

- (x_i, y_i) denotes the features and the endpoints of the n patients.
- $\theta \in \Theta$ denotes the parameters of the model.
- L a loss function and $\alpha > 0$ a hyperparameter.

Lasso: Leads to sparse parameters' estimator [Tibshirani, 1996]

$$\theta^*_{LASSO}(x, y) = \min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} L(\theta; x_i, y_i) + \alpha ||\theta||_1$$

α is usually chosen using cross-validation.

What is Bolasso?

Bootstrap = random sampling with replacement

Bolasso: bootstrapped lasso for model consistency [Bach, 2008]

Require: data (x, y); number of bootstrap replicates m; α

  for $k = 1$ to m **do**

    Generate bootstrap samples $(\overline{x}_k, \overline{x}_k)$

    Estimate $\theta^*_{LASSO}(\overline{x}_k, \overline{y}_k)$

    Compute the support $J_k = \{j : \theta^*_{LASSO}(\overline{x}_k, \overline{y}_k)_j > 0\}$

  end for

  Compute $J = \bigcap_{k=1}^{m} J_k$

J is the set of features selected by Bolasso. We then, retrain a model using these features.

Relaxation of Bolasso

Bolasso: bootstrapped lasso for model consistency

Require: data (x, y); number of bootstrap replicates m; α

  for $k = 1$ to m **do**

    Generate bootstrap sample $(\overline{x}_k, \overline{y}_k)$

    Estimate $\theta^*_{LASSO}(\overline{x}_k, \overline{y}_k)$

    Compute the support $J_k = \theta^*_{LASSO}(\overline{x}_k, \overline{y}_k) > 0$

  end for

  Compute $J = \frac{1}{m} \sum_{k=1}^{m} J_k$

J is the selection frequency of each features.

Simultaneous feature selection with multi-task lasso

Notations:

- $(x_i, y_i^{PFS}, y_i^{OS})$ denotes the features and the endpoints of the n patients.
- two models : θ^{PFS} and θ^{OS}.

Multi-task lasso: encourages solutions with shared patterns of sparsity for both tasks [Obozinski et al., 2006]

$$\min_{\theta_{PFS},\theta_{OS}\in\Theta^2} \frac{1}{n}\sum_{i=1}^{n} L(\theta_{PFS}; x_i, y_i^{PFS}) + \frac{1}{n}\sum_{i=1}^{n} L(\theta_{OS}; x_i, y_i^{OS}) + \alpha \sum_{k=1}^{d} \sqrt{(\theta_{PFS,k}^2 + \theta_{OS,k}^2)}$$

penalising the ℓ_1-norm of the vector of ℓ_2-norms of the feature-specific coefficient vectors.

Towards the multi-task bolasso

Bolasso: boostrapped lasso for model consistency

Require: data (x, y^{PFS}, y^{OS}); number of bootstrap replicates m; α

  for $k = 1$ to m **do**

    Generate bootstrap samples $(\overline{x}_k, , \overline{y}_k^{PFS}, \overline{y}_k^{OS})$

    Estimate $(\theta_{PFS}, \theta_{OS})^*_{MTLasso}(\overline{x}_k, \overline{y}_{PFS,k}, \overline{y}_{OS,k})$

    Compute the support $J_k = (\theta_{PFS}, \theta_{OS})^*_{MTLasso}(\overline{x}_k, \overline{y}_{PFS,k}, \overline{y}_{OS,k}) > 0$

  end for

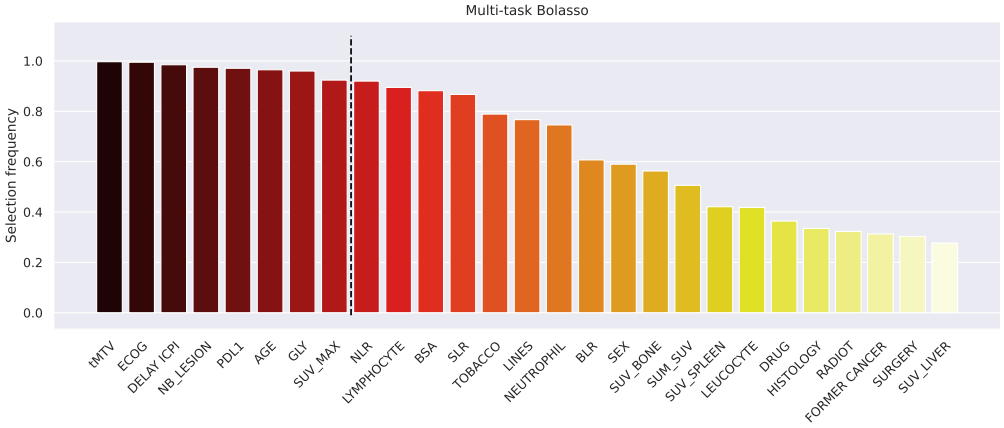
  Compute $J = \frac{1}{m} \sum_{k=1}^{m} J_k$

J is the selection frequency of each feature.

Towards the multi-task bolasso

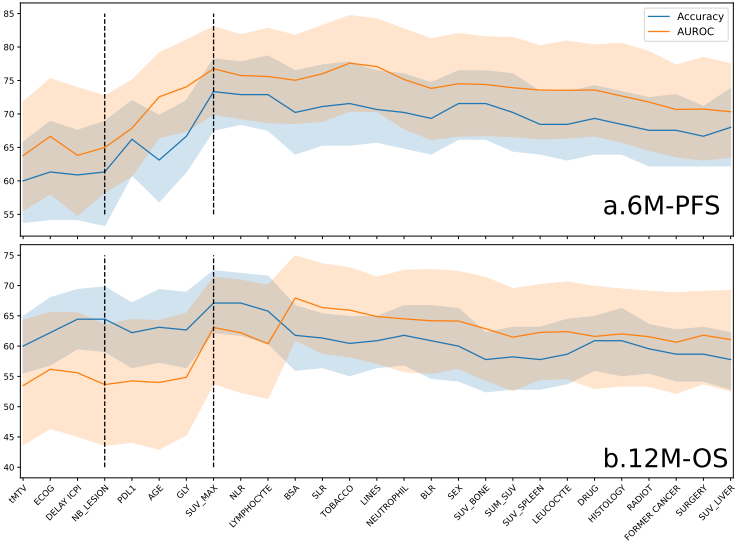
	Feature selection	Robust selection	Multi-task learning
Lasso	✓	✗	✗
Bolasso	✓	✓	✗
Multi-task lasso	✓	✗	✓
Multi-task bolasso	✓	✓	✓

The tMTV and the ECOG were selected more than 99% of the time.



Feature selection frequency given over a 1000 bootstrapping of a multi-task Bolasso.

Dimension reduction does not cause loss of predictive information



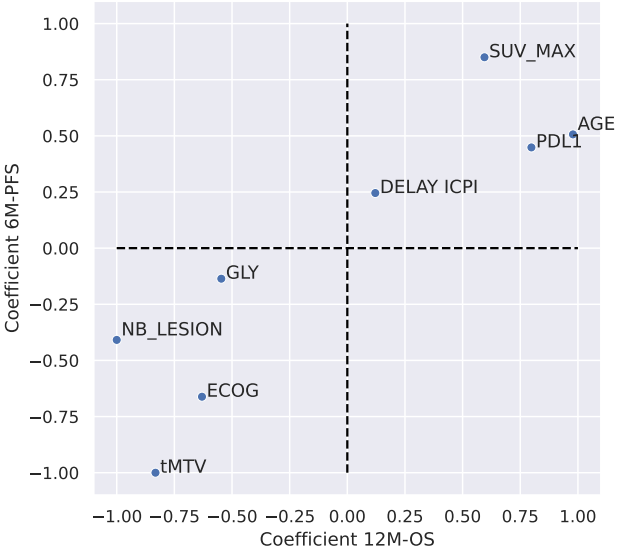
Cross-validation performance on a.6M-PFS and b. 12M-OS of a logistic regression with various numbers of features ranked by their selection frequencies given by the multi-task bolasso.

Dimension reduction does not cause loss of predictive information

	Accuracy		AUROC	
	6M-PFS	12M-OS	6M-PFS	12M-OS
M-4	61.33 (14.65)	64.44 (10.52)	64.99 (14.68)	53.65 (20.33)
M-8	**73.33** (10.89)	**67.11** (18.10)	**76.75** (12.74)	**63.07** (18.10)
M-27	68.00 (11.47)	61.08 (16.42)	70.33 (14.07)	61.08 (16.42)
MTV	60.00 (12.17)	60.00 (9.11)	63.77 (15.94)	53.48 (19.49)
PDL1	52.89 (11.54)	65.78 (9.99)	56.41 (12.11)	49.40 (11.45)
SLR	57.78 (14.94)	64.88 (11.54)	59.76 (13.63)	51.78 (18.05)

Cross-validation performances of a logistic regression on different subsets of features (top 4, top 8 and all) compared to standard biomarkers (PDL1 expression, tMTV and SLR).

Coefficient signs correspond to literature



Normalised logistic regression coefficients for both outcomes.

- Clinical
  - AGE [Fulop et al., 2011; Kugel et al., 2018; Morinaga et al., 2023]
  - ECOG [Dall'Olio et al. 2020]
  - DELAY ICPI
- Biological
  - PDL1
  - GLY [Monzavi-Karbassi et al., 2016; Kakehi et al., 2018]
- PET/CT
  - SUV_{max} [Yang Wang et al., 2020]
  - tMTV [S. I. Kim et al. 2020; K. Zhu et al. 2022,]
  - NB_LESION

Biomarker selection improves over univariate biomarkers on an external validation set

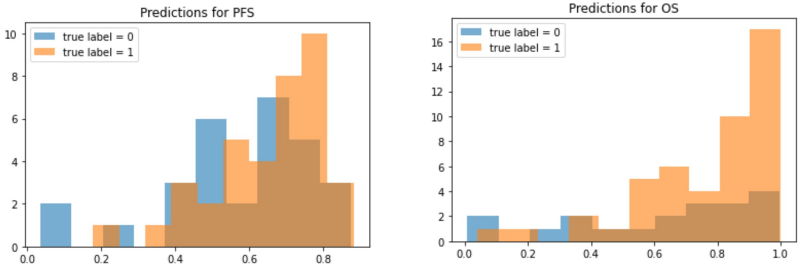
30 patients from CHB, Rouen and CAL, Nice. 46.66% of reccurence and 60.0% of death.

	6M-PFS		12M-OS	
Model	Accuracy	AUROC	Accuracy	AUROC
M-4 LR	80.00	86.61	70.00	81.48
M-8 LR	**83.33**	88.39	**73.33**	**90.74**
M-8 DT	73.33	73.66	**73.33**	73.61
M-8 RF	**83.33**	85.49	60.00	85.65
M-8 XGBoost	76.67	83.48	**73.33**	77.31
M-8 SVM	**83.33**	84.38	70.00	87.04
M-8 MLP	80.00	86.16	**73.33**	90.74
tMTV	**83.33**	**92.86**	66.66	79.17
PDL1	56.67	55.13	40.00	54.63

Blinded validation on Curie dataset

67 patients from Institut Curie, Paris. 43.93% of recurrence and 28.78% of death.

- 6M-PFS: accuracy of 63.63%, AUROC of 61.79%
- 12M-OS: accuracy of 74.24%, AUROC of 65.06%



$\longrightarrow$ Data and label shift.

Conclusion

Conclusion:

- Multi-task bolasso is an efficient feature selection method
- 8 biomarkers: simple and interpretable models
- Obtained from current routine care.

Conclusion

Conclusion:

- Multi-task bolasso is an efficient feature selection method
- 8 biomarkers: simple and interpretable models
- Obtained from current routine care.

Publications:

- EANM 2022 highlights
- Abstract in ASCO 2023
- Paper under review at the EJNMMI

Conclusion

Conclusion:

- Multi-task bolasso is an efficient feature selection method
- 8 biomarkers: simple and interpretable models
- Obtained from current routine care.

Publications:

- EANM 2022 highlights
- Abstract in ASCO 2023
- Paper under review at the EJNMMI

Limitations and future directions:

- **Limited size** of the exploratory and validation cohorts
- Further work is needed for **test-time adapation** [Lipton et al., 2018; Garg et al., 2020, Goyal et al., 2022]
  $\longrightarrow$ FederatedPET project (1200 patients on 10 centers)

Conclusion

Conclusion:

- Multi-task bolasso is an efficient feature selection method
- 8 biomarkers: simple and interpretable models
- Obtained from current routine care.

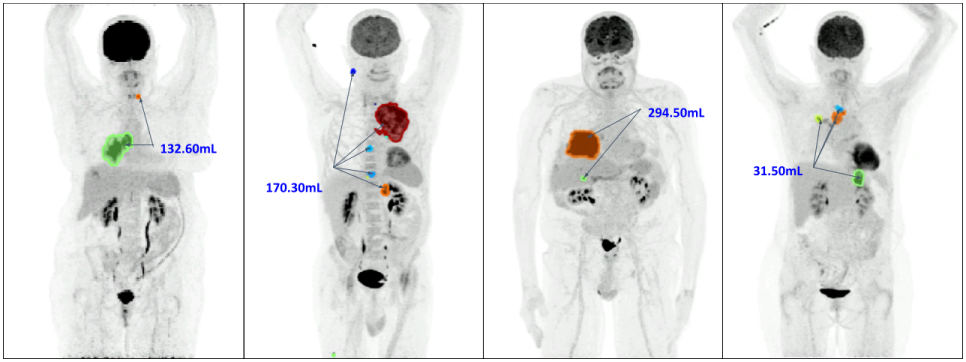
Publications:

- EANM 2022 highlights
- Abstract in ASCO 2023
- Paper under review at the EJNMMI

Limitations and future directions:

- **Limited size** of the exploratory and validation cohorts
- Further work is needed for **test-time adapation** [Lipton et al., 2018; Garg et al., 2020, Goyal et al., 2022]
  $\longrightarrow$ FederatedPET project (1200 patients on 10 centers)
- The bolasso relaxation can cause selection of **correlated** features
- Fastidious and time consuming collection of the three imaging biomakers: tMTV, SUV_{max} and the number of lesions

What is next? Towards the automatic collection of the three PET/CT biomarkers



Collection of the tMTV, SUV_{max} and the number of lesions.
Semi-supervised learning methods for segmentation are adaptations of SSL methods for classification

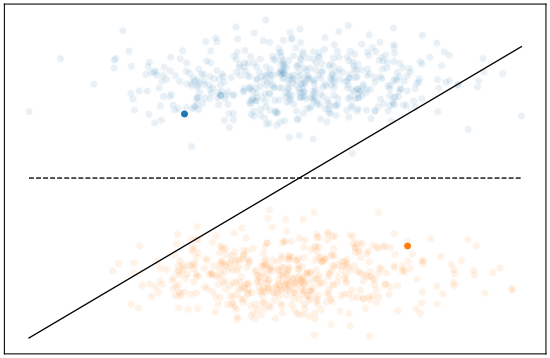
Outline

A. On the performance of biomarkers
   Combination of heterogeneous biomarkers

B. On the collection of biomarkers
   0. Semi-supervised learning generalities
   1. DeSSL: Safe semi-supervised learning via simple debiasing
   2. DeSegSSL: Safe semi-supervised learning for medical image segmentation

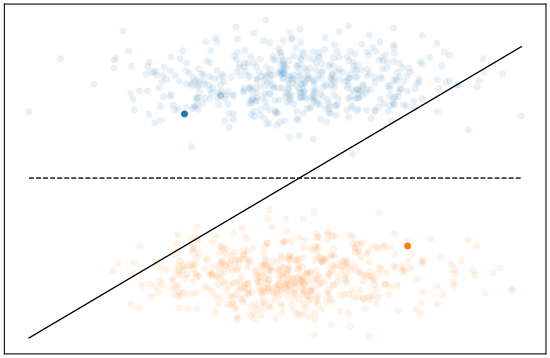
Deep semi-supervised learning ? What for ?

Goal: Using both labelled and unlabelled data to build better learners



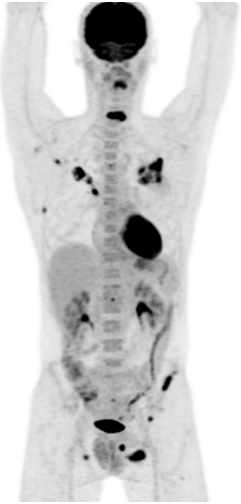
Deep semi-supervised learning ? What for ?

Goal: Using both labelled and unlabelled data to build better learners



Why bother ?

- unlabelled data are cheap
- labelled data can be hard to get



Learning theory relies on the unbiased estimator of the risk

- x the features and y the labels (categorical).
- θ the model's parameters
- L a loss function

The ultimate goal of a learning algorithm is to find θ^* among a fixed class of functions:

$$\theta^* = \underset{\theta \in \Theta}{\arg\min}\, \mathcal{R}(\theta) = \underset{\theta \in \Theta}{\arg\min}\, \mathbb{E}[L(\theta; x, y)]$$

Learning theory relies on the unbiased estimator of the risk

- x the features and y the labels (categorical).
- θ the model's parameters
- L a loss function

The ultimate goal of a learning algorithm is to find θ^* among a fixed class of functions:

$$\theta^* = \arg\min_{\theta \in \Theta} \mathcal{R}(\theta) = \arg\min_{\theta \in \Theta} \mathbb{E}[L(\theta; x, y)]$$

In practice: $P(x, y)$ is unknown. We have access to a training set: $\{(x_1, y_1), ..., (x_n, y_n)\}$

Learning theory relies on the unbiased estimator of the risk

- x the features and y the labels (categorical).
- θ the model's parameters
- L a loss function

The ultimate goal of a learning algorithm is to find θ^* among a fixed class of functions:

$$\theta^* = \arg\min_{\theta \in \Theta} \mathcal{R}(\theta) = \arg\min_{\theta \in \Theta} \mathbb{E}[L(\theta; x, y)]$$

In practice: $P(x, y)$ is unknown. We have access to a training set: $\{(x_1, y_1), ..., (x_n, y_n)\}$
Empirical risk:

$$\hat{\mathcal{R}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} L(\theta; x_i, y_i)$$

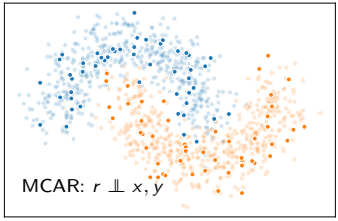
$\longrightarrow$ Unbiased estimator $\Rightarrow$ basic property that is needed for the development of traditional learning theory and asymptotic statistics. [van der Vaart, 1998; Shalev-Shwartz et al., 2014]

Semi-supervised learning is a missing data problem.

- labelled data: $\{(x_i, y_i) : i \in \mathcal{L}\}$
- unlabelled data: $\{x_i : i \in \mathcal{U}\}$
- $|\mathcal{L}| + |\mathcal{U}| = n_l + n_u = n$
- Missing data mechanism $r \in \{0, 1\}$.

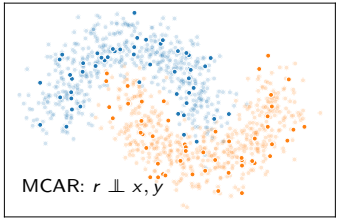
Semi-supervised learning is a missing data problem.

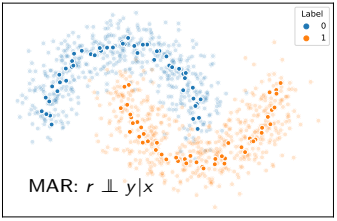
- labelled data: $\{(x_i, y_i) : i \in \mathcal{L}\}$
- unlabelled data: $\{x_i : i \in \mathcal{U}\}$
- $|\mathcal{L}| + |\mathcal{U}| = n_l + n_u = n$
- Missing data mechanism $r \in \{0, 1\}$.



Semi-supervised learning is a missing data problem.

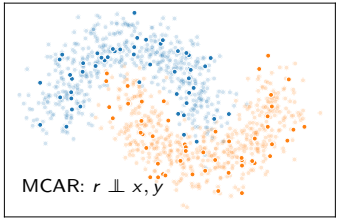
- labelled data: $\{(x_i, y_i) : i \in \mathcal{L}\}$
- unlabelled data: $\{x_i : i \in \mathcal{U}\}$
- $|\mathcal{L}| + |\mathcal{U}| = n_l + n_u = n$
- Missing data mechanism $r \in \{0, 1\}$.

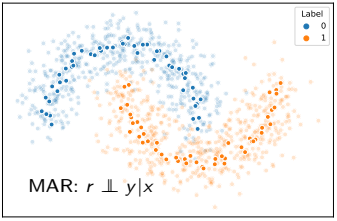


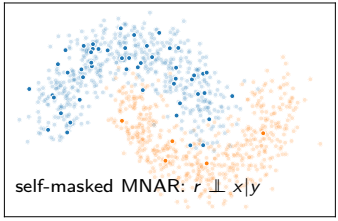


Semi-supervised learning is a missing data problem.

- labelled data: $\{(x_i, y_i) : i \in \mathcal{L}\}$
- unlabelled data: $\{x_i : i \in \mathcal{U}\}$
- $|\mathcal{L}| + |\mathcal{U}| = n_l + n_u = n$
- Missing data mechanism $r \in \{0, 1\}$.

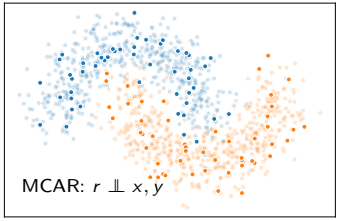


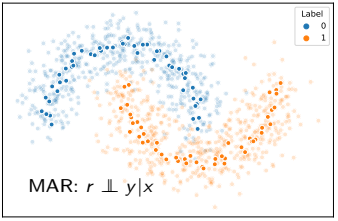


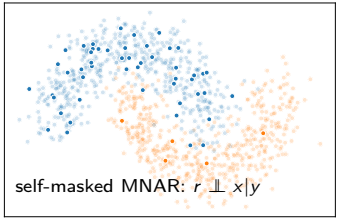


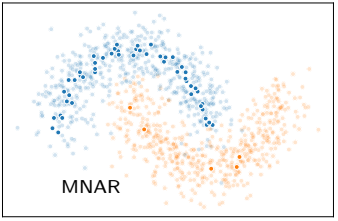
Semi-supervised learning is a missing data problem.

- labelled data: $\{(x_i, y_i) : i \in \mathcal{L}\}$
- unlabelled data: $\{x_i : i \in \mathcal{U}\}$
- $|\mathcal{L}| + |\mathcal{U}| = n_l + n_u = n$
- Missing data mechanism $r \in \{0, 1\}$.









Under MCAR, the complete case is unbiased

- labelled data: $\{(x_i, y_i) : i \in \mathcal{L}\}$
- unlabelled data: $\{x_i : i \in \mathcal{U}\}$
- $|\mathcal{L}| + |\mathcal{U}| = n_l + n_u = n$
- Missing data mechanism $r \in \{0, 1\}$.

Missing completely at random (MCAR): $p(r = 1|x, y) = p(r = 1) = \pi$

Complete case. Get rid of unlabelled data:

$$\hat{\mathcal{R}}_{CC}(\theta) = \frac{1}{n_l} \sum_{i \in \mathcal{L}} L(\theta; x_i, y_i)$$

$\longrightarrow$ Under MCAR, also unbiased $\Rightarrow$ basic property that is needed for the development of traditional learning theory and asymptotic statistics.

Using unlabelled data leads to a biased estimator of the risk

Including unlabelled data in the risk estimator:

$$\hat{\mathcal{R}}_{SSL}(\theta) = \frac{1}{n_l} \sum_{i \in \mathcal{L}} L(\theta; x_i, y_i) + \frac{\lambda}{n_u} \sum_{i \in \mathcal{U}} H(\theta; x_i), \qquad \lambda > 0$$

Using unlabelled data leads to a biased estimator of the risk

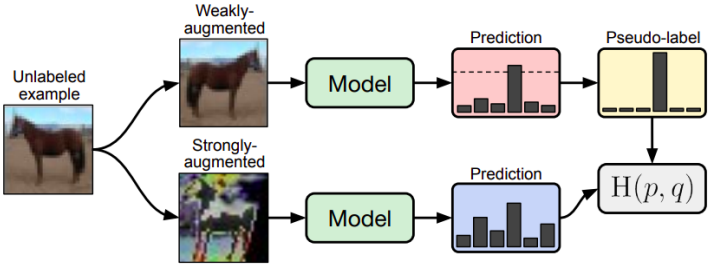
Including unlabelled data in the risk estimator:

$$\hat{\mathcal{R}}_{SSL}(\theta) = \frac{1}{n_l} \sum_{i \in \mathcal{L}} L(\theta; x_i, y_i) + \frac{\lambda}{n_u} \sum_{i \in \mathcal{U}} H(\theta; x_i), \qquad \lambda > 0$$

Examples: $p_\theta(.|x)$ the outputs of the model

- Entropy minimization: $H(\theta; x) = -\sum p_\theta(y|x) \log(p_\theta(y|x))$
  [Grandvalet and Bengio 2005]
- Consistency based: $H(\theta; x) = Div(p_\theta(.|x), p_\theta(.|pert(x))$
  [Tarvainen et al. 2017, Laine and Aila 2017, Miyato et al. 2018, ...]
- Pseudo-label (PL): $H(\theta; x) = -\log(\max_y p_\theta(y|x)) \mathbb{1}[\max_y p_\theta(y|x) > \tau]$
  [Scudder 1965, Lee 2013, Sohn et al. 2020, Xie et al. 2020, ...]

Fixmatch [Sohn et al., 2020]



$$\hat{y} = \arg\max_{y} p_\theta(y|x_{weak})$$

$$H(\theta; x) = -\mathbb{1}[p_\theta(\hat{y}|x_{weak}) > \tau] \log(p_\theta(\hat{y}|x_{strong}))$$

Using unlabelled data leads to a biased estimator of the risk

Including unlabelled data in the risk estimator:

$$\hat{\mathcal{R}}_{SSL}(\theta) = \frac{1}{n_l} \sum_{i \in \mathcal{L}} L(\theta; x_i, y_i) + \frac{\lambda}{n_u} \sum_{i \in \mathcal{U}} H(\theta; x_i)$$

Good performance on a variety of (deep) learning tasks, **but:**

Using unlabelled data leads to a biased estimator of the risk

Including unlabelled data in the risk estimator:

$$\hat{\mathcal{R}}_{SSL}(\theta) = \frac{1}{n_l} \sum_{i \in \mathcal{L}} L(\theta; x_i, y_i) + \frac{\lambda}{n_u} \sum_{i \in \mathcal{U}} H(\theta; x_i)$$

Good performance on a variety of (deep) learning tasks, **but:**
Problems:

- **Potential degradation** reported in previous works [Schölkopf et al. 2012, V.Engelen & Hoos 2020, Zhu et al. 2022]

Using unlabelled data leads to a biased estimator of the risk

Including unlabelled data in the risk estimator:

$$\hat{\mathcal{R}}_{SSL}(\theta) = \frac{1}{n_l} \sum_{i \in \mathcal{L}} L(\theta; x_i, y_i) + \frac{\lambda}{n_u} \sum_{i \in \mathcal{U}} H(\theta; x_i)$$

Good performance on a variety of (deep) learning tasks, **but:**
Problems:

- **Potential degradation** reported in previous works [Schölkopf et al. 2012, V.Engelen & Hoos 2020, Zhu et al. 2022]
- **Few theoretical guarantees** using strong distributional assumptions [Mey & Loog 2019]

Using unlabelled data leads to a biased estimator of the risk

Including unlabelled data in the risk estimator:

$$\hat{\mathcal{R}}_{SSL}(\theta) = \frac{1}{n_l} \sum_{i \in \mathcal{L}} L(\theta; x_i, y_i) + \frac{\lambda}{n_u} \sum_{i \in \mathcal{U}} H(\theta; x_i)$$

Good performance on a variety of (deep) learning tasks, **but:**
Problems:

- **Potential degradation** reported in previous works [Schölkopf et al. 2012, V.Engelen & Hoos 2020, Zhu et al. 2022]
- **Few theoretical guarantees** using strong distributional assumptions [Mey & Loog 2019]
- **No asymptotic consistency**: may fail even with an infinite number of labelled datapoints

Using unlabelled data leads to a biased estimator of the risk

Including unlabelled data in the risk estimator:

$$\hat{\mathcal{R}}_{SSL}(\theta) = \frac{1}{n_l} \sum_{i \in \mathcal{L}} L(\theta; x_i, y_i) + \frac{\lambda}{n_u} \sum_{i \in \mathcal{U}} H(\theta; x_i)$$

Good performance on a variety of (deep) learning tasks, **but:**
Problems:

- **Potential degradation** reported in previous works [Schölkopf et al. 2012, V.Engelen & Hoos 2020, Zhu et al. 2022]
- **Few theoretical guarantees** using strong distributional assumptions [Mey & Loog 2019]
- **No asymptotic consistency**: may fail even with an infinite number of labelled datapoints
- Choice of H can be confusing [Corduneanu & Jaakkola 2003, Krause et al. 2010]

Using unlabelled data leads to a biased estimator of the risk

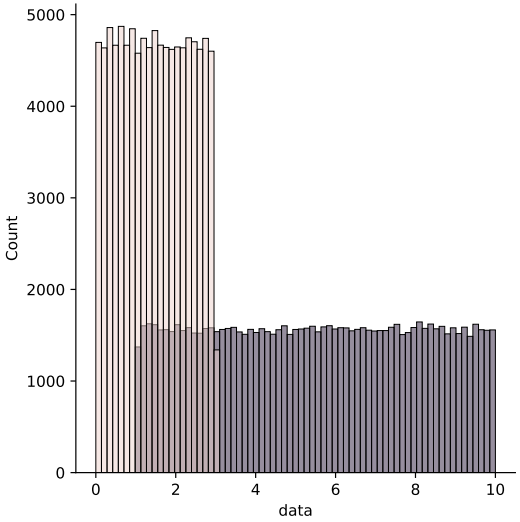
Including unlabelled data in the risk estimator:

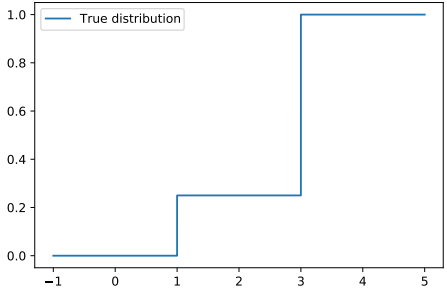
$$\hat{\mathcal{R}}_{SSL}(\theta) = \frac{1}{n_l} \sum_{i \in \mathcal{L}} L(\theta; x_i, y_i) + \frac{\lambda}{n_u} \sum_{i \in \mathcal{U}} H(\theta; x_i)$$

Good performance on a variety of (deep) learning tasks, **but:**
Problems:

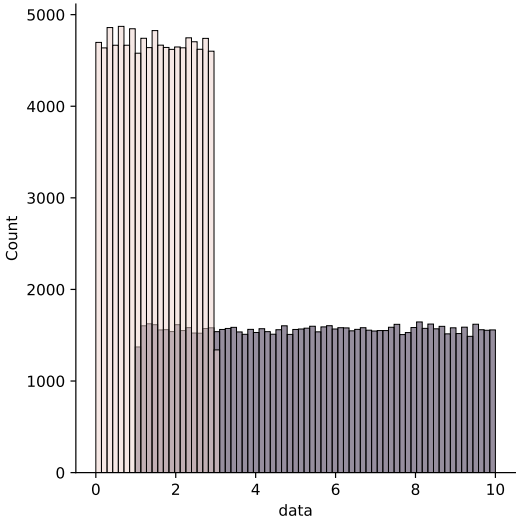
- **Potential degradation** reported in previous works [Schölkopf et al. 2012, V.Engelen & Hoos 2020, Zhu et al. 2022]
- **Few theoretical guarantees** using strong distributional assumptions [Mey & Loog 2019]
- **No asymptotic consistency**: may fail even with an infinite number of labelled datapoints
- Choice of H can be confusing [Corduneanu & Jaakkola 2003, Krause et al. 2010]
- An additional hyperparameter λ and **no realistic validation** [Oliver et al. 2018]

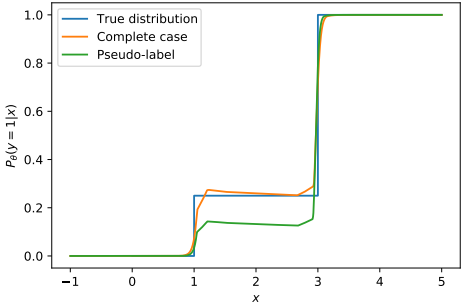
FEAR THE UNLABELLED !



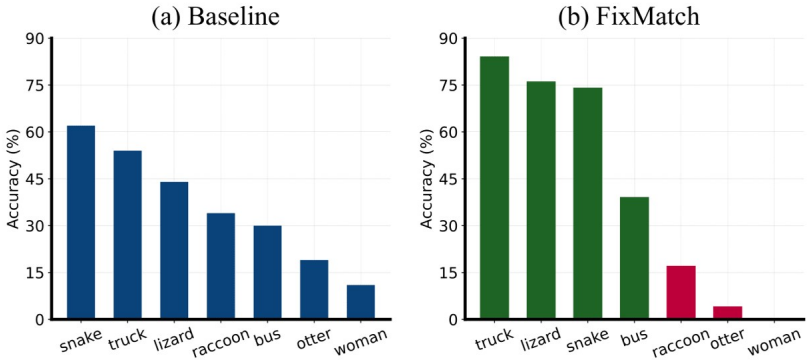


FEAR THE UNLABELLED !



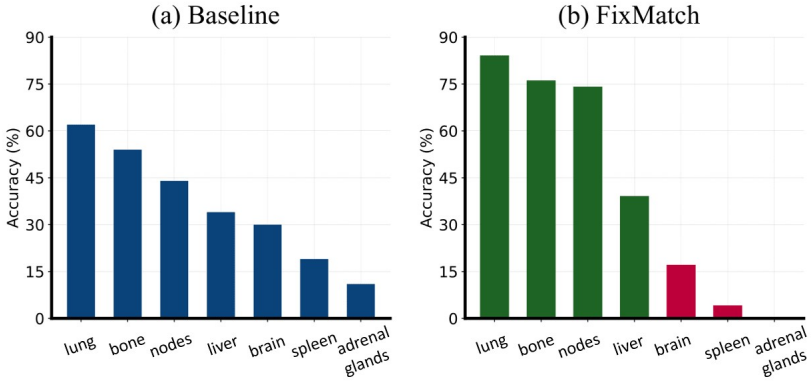


The rich get richer ! [Zhu et al. 2022]



Top-1 accuracy of 7 randomly selected categories with different training methods on CIFAR-100. (Left) Complete case. (Right) FixMatch [Chen et al., 2022]

Analogy to PET/CT segmentation



Not a real experiment !

Towards safe SSL for a reliable use in critical settings

Safe semi-supervised learning: an SSL algorithm is safe if it has theoretical guarantees that are similar or stronger to the complete case baseline.

Outline

A. On the performance of biomarkers
   Combination of heterogeneous biomarkers

B. On the collection of biomarkers
   0. Semi-supervised learning generalities
   1. DeSSL: Safe semi-supervised learning via simple debiasing
   2. DeSegSSL: Safe semi-supervised learning for medical image segmentation

DeSSL: Debiased version of SSL

We propose to remove the bias:

$$\hat{\mathcal{R}}_{DeSSL}(\theta) = \frac{1}{n_l}\sum_{i\in\mathcal{L}} L(\theta; x_i, y_i) + \frac{\lambda}{n_u}\sum_{i\in\mathcal{U}} H(\theta; x_i) - \frac{\lambda}{n_l}\sum_{i\in\mathcal{L}} H(\theta; x_i)$$

DeSSL: Debiased version of SSL

We propose to remove the bias:

$$\hat{\mathcal{R}}_{DeSSL}(\theta) = \frac{1}{n_l}\sum_{i\in\mathcal{L}} L(\theta; x_i, y_i) + \frac{\lambda}{n_u}\sum_{i\in\mathcal{U}} H(\theta; x_i) - \frac{\lambda}{n_l}\sum_{i\in\mathcal{L}} H(\theta; x_i)$$

- Under the MCAR assumption, estimator is unbiased estimator of the true risk.

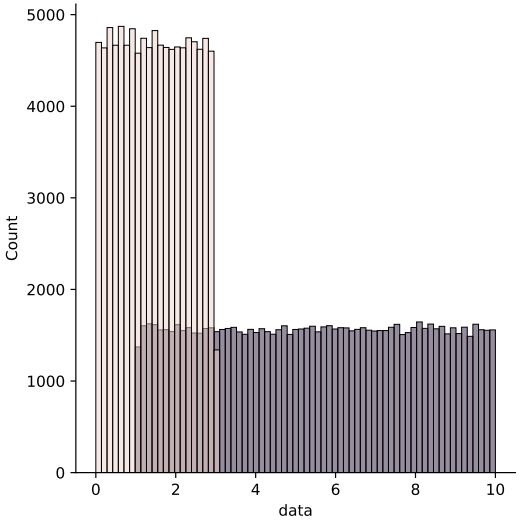
DeSSL: Debiased version of SSL

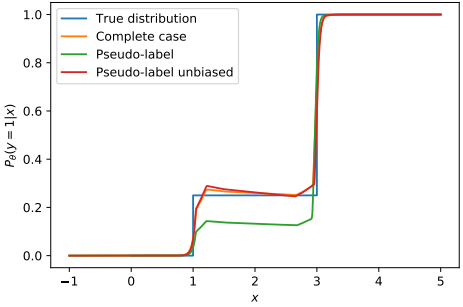
We propose to remove the bias:

$$\hat{\mathcal{R}}_{DeSSL}(\theta) = \frac{1}{n_l}\sum_{i\in\mathcal{L}} L(\theta; x_i, y_i) + \frac{\lambda}{n_u}\sum_{i\in\mathcal{U}} H(\theta; x_i) - \frac{\lambda}{n_l}\sum_{i\in\mathcal{L}} H(\theta; x_i)$$

- Under the MCAR assumption, estimator is unbiased estimator of the true risk.
- Close relationship with control variates [Owen 2013].
- Other motivations:
  - Penalising the confidence of a model on labelled data [Pereyra et al., 2017]
  - Maximising the plausibility [Barndorff-Nielsen 1976].
  - The risk estimate is a Lagrangian!
- $\longrightarrow$ Optimality of debiasing with the labelled dataset

Pseudo-label unbiased success under the cluster assumption





$\hat{\mathcal{R}}_{DeSSL}(\theta)$ is unbiased but is it an accurate risk estimator ?

Theorem: It exists λ_{opt} such that:

$$\begin{aligned}\mathbb{V}(\hat{\mathcal{R}}_{DeSSL}(\theta)|r)|_{\lambda_{opt}} &= (1 - \frac{n_u}{n}{\rho_{L,H}}^2)\mathbb{V}(\hat{\mathcal{R}}_{CC}(\theta)|r) \\ &\leq \mathbb{V}(\hat{\mathcal{R}}_{CC}(\theta)|r)\end{aligned}$$

where $\rho_{L,H} = \mathrm{Corr}(L(\theta; x, y), H(\theta; x))$.

- Justification on the heuristic idea that H should be a surrogate of L.
- Formula for λ_{opt}

$\hat{\mathcal{R}}_{DeSSL}(\theta)$ is an accurate risk estimator but $\nabla\hat{\mathcal{R}}_{DeSSL}(\theta)$ is even better

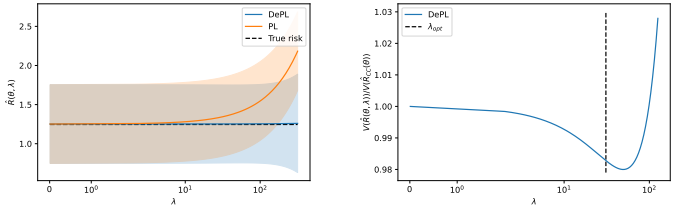


Figure: (Left) Risk estimate value for PseudoLabel (PL) and DePseudoLabel (DePL) compared to the true value of the risk. (Right) The influence of λ on the ratio $\mathbb{V}(\hat{\mathcal{R}}_{DePL}(\theta)|r)/\mathbb{V}(\hat{\mathcal{R}}_{CC}(\theta)|r)$. (Down) The influence of λ on the ratio $\mathbb{V}(\nabla\hat{\mathcal{R}}_{DePL}(\theta)|r)/\mathbb{V}(\nabla\hat{\mathcal{R}}_{CC}(\theta)|r)$.

$\hat{\mathcal{R}}_{DeSSL}(\theta)$ is an accurate risk estimator but $\nabla\hat{\mathcal{R}}_{DeSSL}(\theta)$ is even better

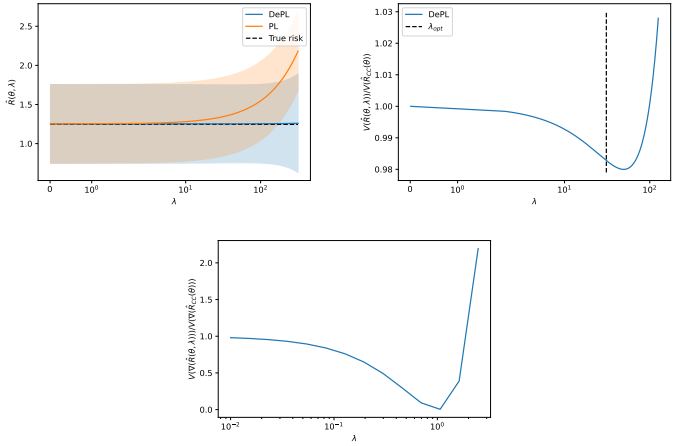


Figure: (Left) Risk estimate value for PseudoLabel (PL) and DePseudoLabel (DePL) compared to the true value of the risk. (Right) The influence of λ on the ratio $\mathbb{V}(\hat{\mathcal{R}}_{DePL}(\theta)|r)/\mathbb{V}(\hat{\mathcal{R}}_{CC}(\theta)|r)$. (Down) The influence of λ on the ratio $\mathbb{V}(\nabla\hat{\mathcal{R}}_{DePL}(\theta)|r)/\mathbb{V}(\nabla\hat{\mathcal{R}}_{CC}(\theta)|r)$.

DeSSL is calibrated and benefits from generalisation error bounds

Calibration:

Theorem: If L is a proper scoring rule, then DeSSL is also a proper scoring rule.

- We can expect DeSSL to be as **well-calibrated** as the complete case
- SSL generally overfits

DeSSL is calibrated and benefits from generalisation error bounds

Calibration:

Theorem: If L is a proper scoring rule, then DeSSL is also a proper scoring rule.

- We can expect DeSSL to be as **well-calibrated** as the complete case
- SSL generally overfits

Generalisation error bound:

Theorem: Under classical assumptions on L and H, DeSSL benefits of generalisation error bounds derived from the **Rademacher complexity**,

$$\mathcal{R}(\theta) \leq \hat{\mathcal{R}}_{DeSSL}(\theta) + 2R_n + \kappa\sqrt{\frac{\log(4/\delta)}{n}}.$$

DeSSL is consistent and improves the supervised baseline

Consistency:

Theorem: $\hat{\theta}_{DeSSL} = \arg\min \hat{\mathcal{R}}_{DeSSL}$ is asymptotically consistent with respect to n.

- DeSSL provides **asymptotically consistent** models.
- SSL may fail with an infinite number of labelled data when DeSSL will not

DeSSL is consistent and improves the supervised baseline

Consistency:

Theorem: $\hat{\theta}_{DeSSL} = \arg\min \hat{\mathcal{R}}_{DeSSL}$ is asymptotically consistent with respect to n.

- DeSSL provides **asymptotically consistent** models.
- SSL may fail with an infinite number of labelled data when DeSSL will not

Asymptotic normality:

Theorem: $\sqrt{n}(\hat{\theta}_{DeSSL} - \theta^*) \longrightarrow \mathcal{N}(0, \Sigma_{DeSSL})$ and it exists λ_{opt} such that:

$$\mathrm{Tr}(\Sigma_{DeSSL}) \leq \mathrm{Tr}(\Sigma_{CC})$$

DeSSL outperforms the complete case baseline in term of parameters estimation

DeFixmatch improves Fixmatch accuracy and calibration

Table: Test accuracy, worst class accuracy and cross-entropy of Complete Case, Fixmatch and DeFixmatch on 5 folds of CIFAR-10.

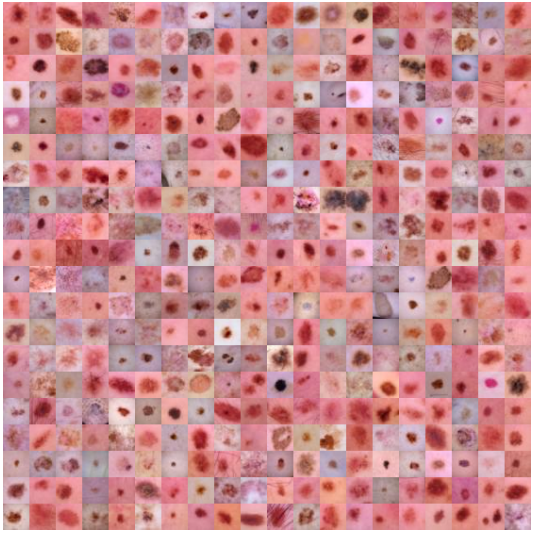
	CIFAR-10 ($n_l = 4000$)		
	Complete Case	Fixmatch	DeFixmatch
Accuracy	87.27 $\pm$ 0.25	93.87 $\pm$ 0.13	**95.44 $\pm$ 0.10**
Worst class accuracy	70.08 $\pm$ 0.93	82.25 $\pm$ 2.27	**87.16 $\pm$ 0.46**
Cross entropy	0.60 $\pm$ 0.01	0.27 $\pm$ 0.01	**0.20 $\pm$ 0.01**
Brier score	0.214 $\pm$ 0.005	0.101 $\pm$0.003	**0.076 $\pm$ 0.001**

DeSSL mitigates the disparate effect of SSL

Table: Mean accuracy per class and mean benefit ratio ($\mathcal{BR}$, [Zhu et al. 2022]) on 5 splits.

	Complete Case	Fixmatch	DeFixmatch
	Accuracy	$\mathcal{BR}$	$\mathcal{BR}$
airplane	86.94	0.88	0.94
automobile	95.26	0.68	0.89
bird	**80.46**	**0.68**	**0.80**
cat	**70.08**	**0.56**	**0.78**
deer	88.88	0.78	0.94
dog	**79.66**	**0.53**	**0.81**
frog	93.12	0.80	0.94
horse	90.96	0.83	0.92
ship	94.12	0.67	0.84
truck	93.18	0.84	0.93

DeSSL mitigates the disparate effect of SSL



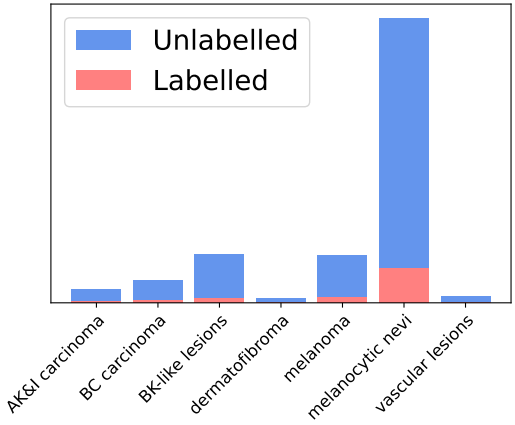
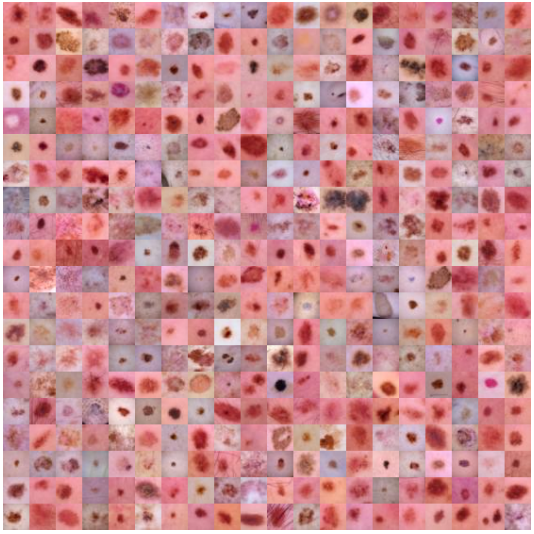


Figure: Class distribution DermaMNIST [J. Yang, R. Shi, and Ni, 2021].

DeSSL mitigates the disparate effect of SSL



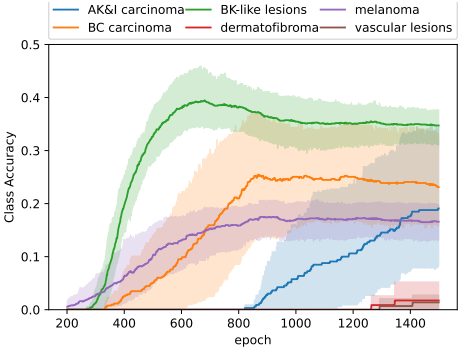
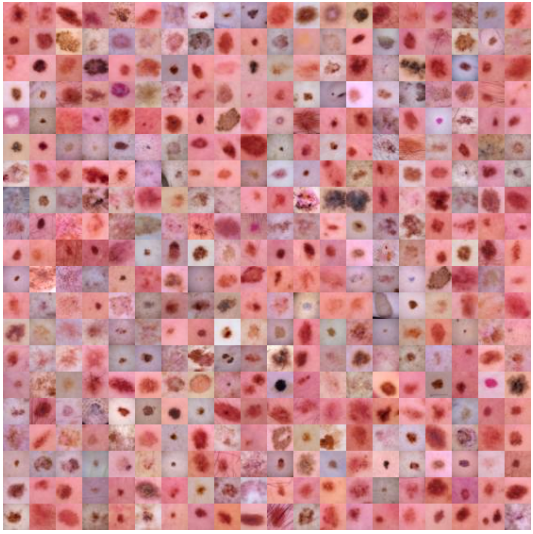


Figure: Class accuracies (without the majority class) on DermaMNIST trained with $n_l = 1000$ labelled data on five folds for Complete Case.

DeSSL mitigates the disparate effect of SSL



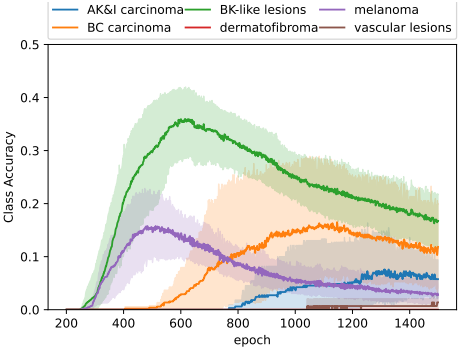
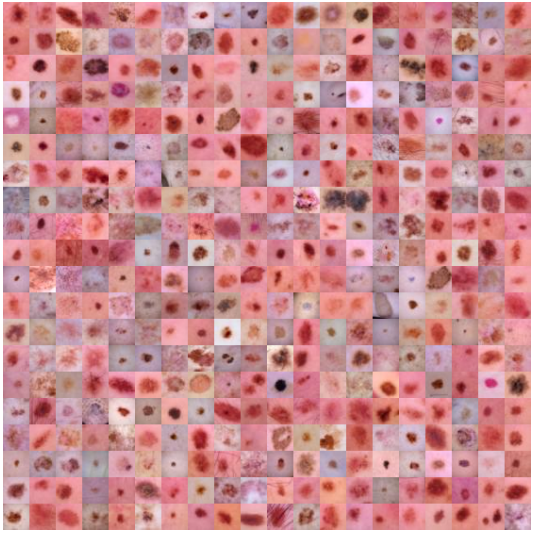


Figure: Class accuracies (without the majority class) on DermaMNIST trained with $n_l = 1000$ labelled data on five folds for PseudoLabel.

DeSSL mitigates the disparate effect of SSL



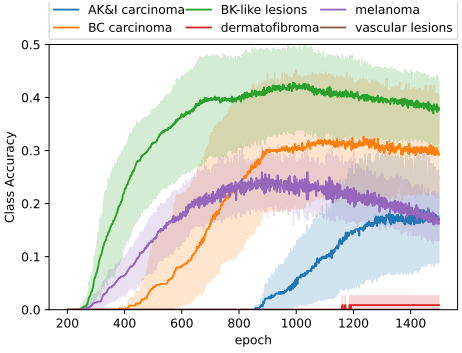


Figure: Class accuracies (without the majority class) on DermaMNIST trained with $n_l = 1000$ labelled data on five folds for DePseudoLabel.

Integration to USB benchmarck [Wang et al., 2022]

- Unified implementation of **14 SSL methods**
- Classical computer vision benchmark (Cifar10, Cifar100, SVHN, STL10)
- + More challenging tasks: computer vision, audio and NLP

Integration to USB benchmarck [Wang et al., 2022]

- Unified implementation of **14 SSL methods**
- Classical computer vision benchmark (Cifar10, Cifar100, SVHN, STL10)
- + More challenging tasks: computer vision, audio and NLP

- DeFixmatch is competitive with the other SOTA methods.

	GTZAN		UrbanSound8k		FSDnoisy	ESC-50	
# Label	100	400	100	400	1773	250	500
Complete Case	52.73±2.86	32.04±0.51	42.65±1.63	27.6±1.19	34.74±1.58	49.83±1.71	38.75±1.47
Fixmatch	41.47±1.62	**21.89±1.01**	40.02±6.62	20.83±2.31	31.05±1.27	43.58±2.79	32.0±1.08
DeFixmatch	**47.15±3.66**	22.31±2.06	**38.43±5.42**	**20.47±1.55**	**29.53±0.79**	**41.83±0.11**	**31.74±0.20**

Conclusion

- DeSSL comes with theoretical guarantees using only the MCAR assumption
- Estimator unbiased, reduction of variance, asymptotically consistent, well calibrated
- Formula for the hyperparameter
- Mitigates the disparate effect of SSL
- Performs better than the biased estimator on various datasets

Future directions:

- Computation of λ_{opt}
- How to build batches? $\longrightarrow$ Stratified sampling

Publications:

- Published at ICLR 2023
- Github repo: `https://github.com/hugoschmutz/defixmatch`
- DeFixmatch is in USB:
  `https://github.com/microsoft/Semi-supervised-learning`

Conclusion

- DeSSL comes with theoretical guarantees using only the MCAR assumption
- Estimator unbiased, reduction of variance, asymptotically consistent, well calibrated
- Formula for the hyperparameter
- Mitigates the disparate effect of SSL
- Performs better than the biased estimator on various datasets

Future directions:

- Computation of λ_{opt}
- How to build batches? $\longrightarrow$ Stratified sampling

Collaborations:

- Extension to self-masked MNAR [Sportisse et al., ICML 2023]
- SSL python library: `https://semipy.github.io/` [Boiteau et al., 2023]

Outline

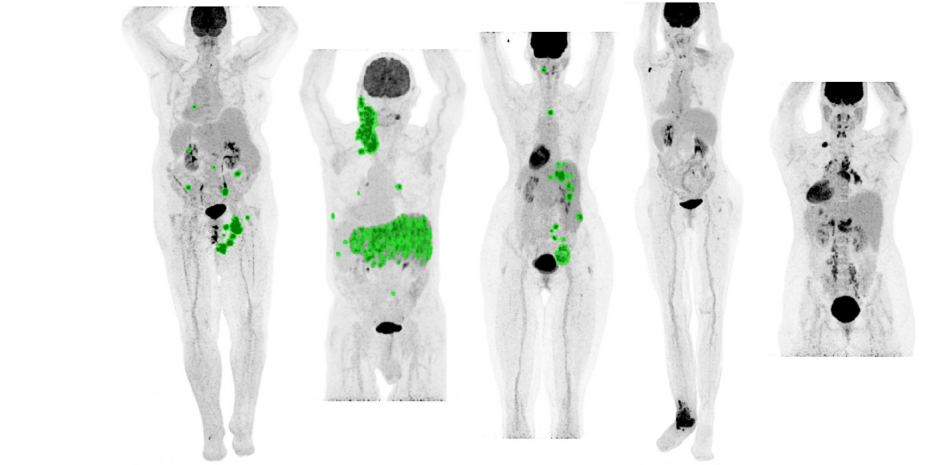
A. On the performance of biomarkers
   Combination of heterogeneous biomarkers

B. On the collection of biomarkers
   0. Semi-supervised learning generalities
   1. DeSSL: Safe semi-supervised learning via simple debiasing
   2. DeSegSSL: Safe semi-supervised learning for medical image segmentation

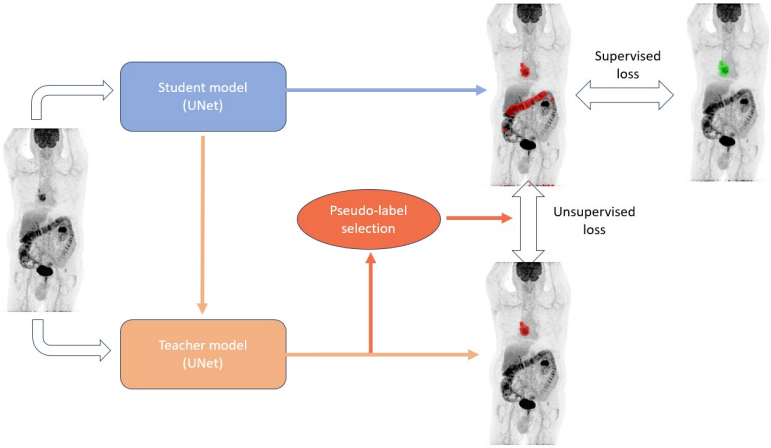
The MCAR assumption for segmentation

- $x \in \mathbb{R}^{2*H_x \times W_x \times D_x}$ and $y \in \{0,1\}^{H_x \times W_x \times D_x}$
- n_l labelled and n_u unlabelled samples.

Missing completely at random (MCAR) i.e. y being missing is independent x.



No safe SSL method has been proposed for SSL segmentation...



$$H(\theta; x) = \frac{1}{|V_x|} \sum_{v \in V_x} h(\theta, x_v, \hat{y_v}) * mask(\hat{y})_v$$

,

V_x the set of pixels in image x.

- Pseudo-label generation
  - Pseudo-label [Bai et al. 2017]
  - Mean-teacher [Yu et al. 2019]
- Pseudo-label selection
  - No selection [Xu et al. 2022]
  - Softmax output
  - Monte Carlo dropout [Yu et al. 2019
- h choices [Ma et al. 2021]
  - Dice
  - Cross-entropy
  - Compound loss

DeSegSSL: DeSSL is directly scalable to segmentation

$$\hat{\mathcal{R}}_{DeSegSSL}(\theta) = \frac{1}{n_l}\sum_{i\in\mathcal{L}} L(\theta; x_i, y_i) + \frac{\lambda}{n_u}\sum_{i\in\mathcal{U}} H(\theta; x_i) - \frac{\lambda}{n_l}\sum_{i\in\mathcal{L}} H(\theta; x_i)$$

DeSegSSL: DeSSL is directly scalable to segmentation

$$\hat{\mathcal{R}}_{DeSegSSL}(\theta) = \frac{1}{n_l}\sum_{i\in\mathcal{L}} L(\theta; x_i, y_i) + \frac{\lambda}{n_u}\sum_{i\in\mathcal{U}} H(\theta; x_i) - \frac{\lambda}{n_l}\sum_{i\in\mathcal{L}} H(\theta; x_i)$$

- $\hat{\mathcal{R}}_{DeSegSSL}(\theta)$ is unbiased
- Variance reduction theorem holds with the same λ_{opt}
- Generalisation error bounds holds

An open access dataset [Gatidis et al. 2022]

The MICCAI 2022 autoPET challenge dataset:

- 501 positive studies from 489 patients
- lymphoma (154), melanoma (188), NSCLC (168)
- voxels size: (2.04mm, 2.04mm, 3mm)
- Two scenarios: $n_l = 50$ and $n_l = 200$

Training details

- model: 3D-UNet
- Patch-based training
- L = dice cross entropy
- H = dice cross entropy

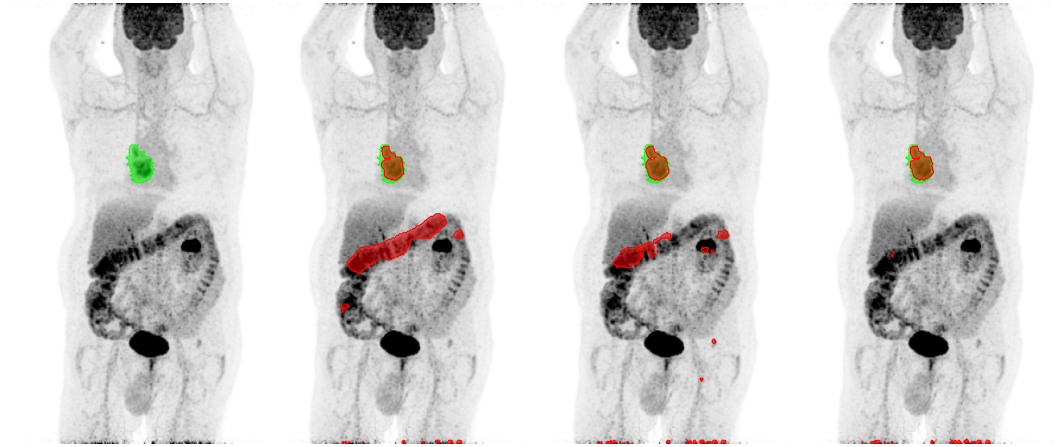
Metrics

- Dice score
- False positive and negative

The debiasing method performs well on classic metrics

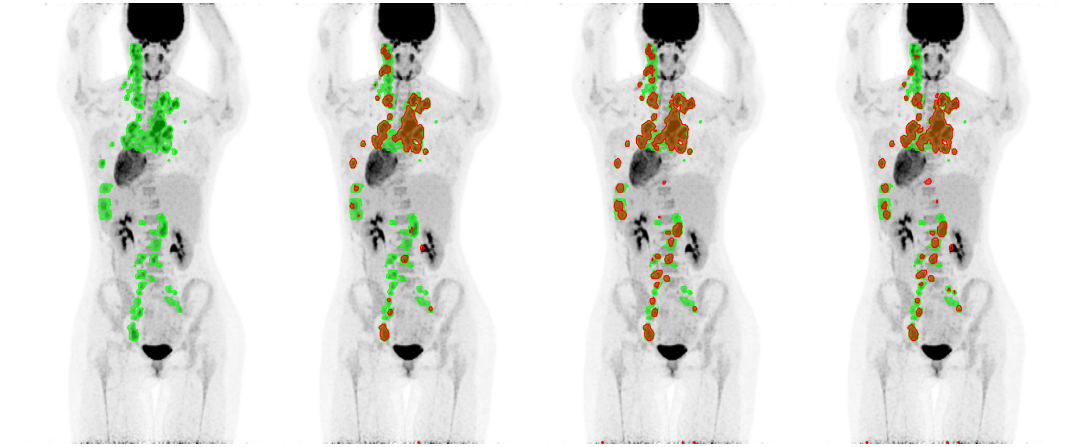
			200			50		
	Uncertainty	Method	Dice	FP	FN	Dice	FP	FN
Pretraining	None	CC	0.616	988.8	440.4	0.502	662.1	1082.4
Finetuning	None	CC	0.627	2039.9	242.8	0.594	2177.5	269.0
	None	SegMT	0.622	1451.5	**186.1**	**0.623**	**1197.1**	448.3
		DeSegMT	**0.631**	**1188.0**	203.3	0.607	1487.2	**352.0**
	Softmax	SegMT	**0.631**	**1330.5**	305.4	0.611	1207.3	**341.3**
		DeSegMT	0.622	1695.7	**191.7**	**0.621**	**1032.9**	407.2

Example of segmentation



(Left) Ground truth (Middle left) Complete case (Middle right) SegMT (Right) DeSegMT

Example of segmentation



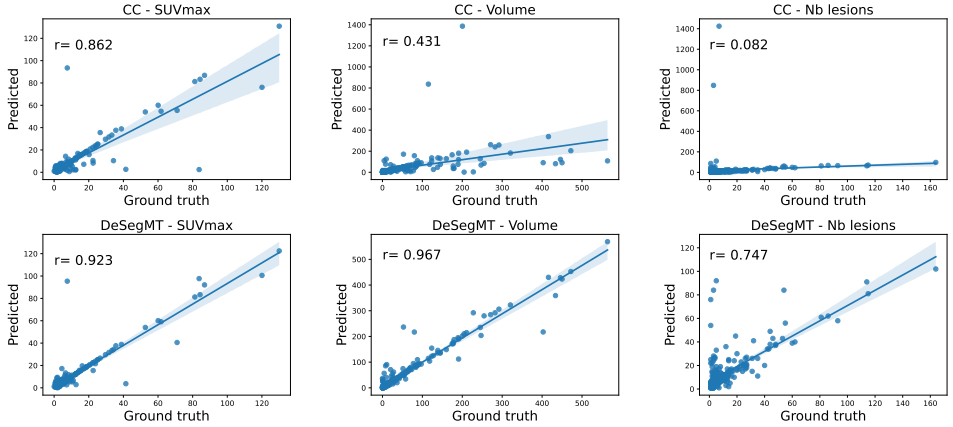
(Left) Ground truth (Middle left) Complete case (Middle right) SegMT (Right) DeSegMT

AI4PET: a 199 patients dataset containing our 142 initial patients

- 199 NSCLC patients treated with immunotherapy in CAL (Nice) and CHPG (Monaco)
- 818 PET/CT scans (before initiation of ICPIs + follow-up)
- 84.53% positive
- 71 patients in test set (325 scans)
- 140 labelled scans (including validation) + 353 unlabelled
- voxels size: (2.62mm, 2.62mm, 2.62mm)

	Uncertainty	Method	Dice	FP	FN
Pretraining		CC	0.591	2557.25	737.70
Finetuning		CC	0.604	2775.79	606.67
	None	SegMT	0.672	1007.28	226.89
		DeSegMT	0.675	1227.66	193.68
	Softmax	SegMT	0.678	1129.84	198.83
		DeSegMT	0.667	1372.36	191.33

SSL improves consequently the prediction of biomarkers



(UP) Complete case (Bottom) DeSegMT with softmax pseudo-label selection

Does the automatic collection of biomarkers work?

	Accuracy		AUROC	
	6M-PFS	12M-OS	6M-PFS	12M-OS
M-8 (manual)	**73.33** (10.89)	67.11 (18.10)	**76.75** (12.74)	63.07 (18.10)
M-8 (automatic on Val)	66.67 (13.77)	**70.67** (8.71)	75.01 (14.91)	**67.89** (17.78)
M-8 (automatic on Train&Val)	70.22 (10.85)	68.89 (9.64)	74.93 (15.63)	65.42 (18.35)

Cross-validation performances of a logistic regression with automatic extraction of tMTV, SUV_{max} and the number of lesions versus manual collection.

Conclusion

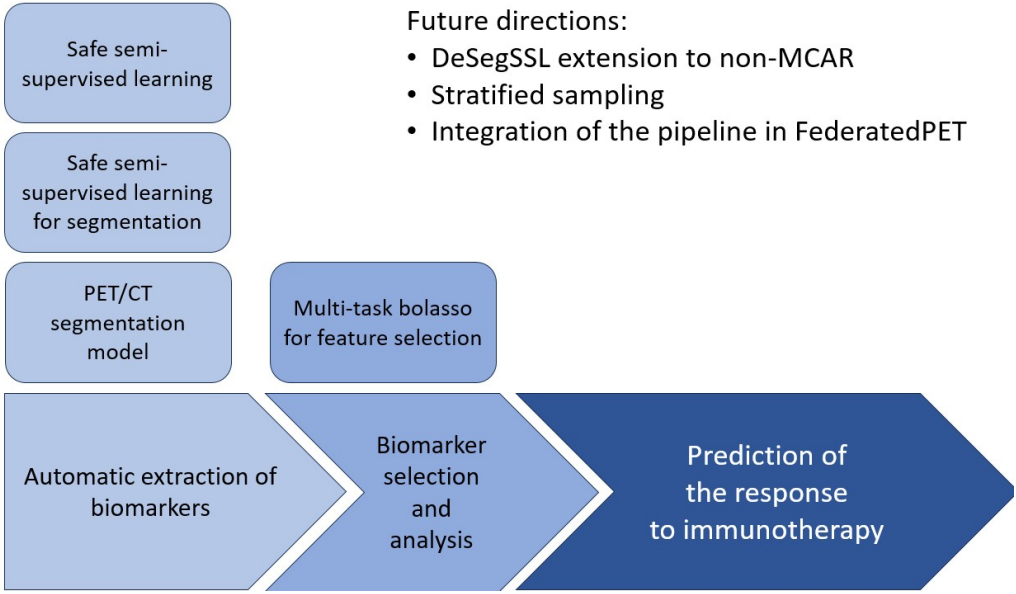
- DeSegSSL is the first safe segmentation SSL method
- Competes with the biased estimator on the autoPET and AI4PET datasets

Future directions:

- Extending the benchmark of models
- Extend to non-MCAR scenario

Publications: DeSegSSL will be released on Arxiv soon and submitted to a journal.

Conclusion



Future directions:

- DeSegSSL extension to non-MCAR
- Stratified sampling
- Integration of the pipeline in FederatedPET

Contribution summary

- Comte, V., **Schmutz, H.**, Chardin, D., Orlhac, F., Darcourt, J., & Humbert, O. Development and validation of a radiomic model for the diagnosis of dopaminergic denervation on [18F]FDOPA PET/CT. *European Journal of Nuclear Medicine and Molecular Imaging (2022)*
- Bergamin, F., Mattei, P.A., Havtorn, J. D., Senetaire, H, **Schmutz, H.**, Maaløe, L., Hauberg, S., & Frellsen, J. Model-agnostic out-of-distribution detection using combined statistical tests. *International Conference on Artificial Intelligence and Statistics (2022)*
- **Schmutz, H.**, Humbert, O., & Mattei, P.-A. Don't fear the unlabelled: safe semi-supervised learning via simple debiasing *Proceedings of the Eleventh International Conference on Learning Representations (2023)*
- Sportisse, A., **Schmutz, H.**, Humbert, O., Bouveyron, C., & Mattei, P.A. Are labels informative in semi-supervised learning? Estimating and leveraging the missing-data mechanism. *Proceedings of the 40th International Conference on Machine Learning (2023)*
- **Schmutz, H.**, Mattei, P.-A., Sara Contu, S., Florent, F., Decazes, P., Guisier, F., Schiappa, R. Comte, V., Tricarico, P., Martin, N., Chardin, D., Humbert, O.. Multi-task learning to combine heterogeneous biomarkers for the prediction of outcomes in non-small cell lung cancer patients treated with immunotherapy. *Submitted to a journal*.

$\longrightarrow$ DeSegSSL will be released on Arxiv soon.

Thank you for your attention!

Exploratory and validation cohort description

Characteristic	Exploratory cohort n=142: 132 CAL + 10 CHPG	External validation cohort n=30: 18 HBC + 12 CAL
Age (Years), median (range)	65 (39 - 91)	73 (59 – 90)
Male sex, n (%)	92 (64.79)	NK
ECOG performance status, n (%)		
0	20 (14.08)	11 (36.67)
1	43 (30.28)	13 (43.33)
2	79 (55.63)	6 (20.00)
Tumour histology, n (%)		
NON-SQUAMOUS CELL CARCINOMA	110 (77.46)	NK
SQUAMOUS CELL CARCINOMA	32 (22.54)	NK
PD-L1 tumour expression, n (%)		
<1%	11 (7.75)	2 (16.67)
1 - 49%	45 (31.69)	2 (6.67)
$\geq$ 50%	64 (45.07)	23 (76.67)
Unknown	22 (15.50)	0 (0.0)
Treatment, n (%)		
Pembrolizumab	87 (61.26)	24 (80.00)
Nivolumab	52 (36.62)	6 (20.00)
Atezolizumab	3 (2.11)	0 (0.00)
Current or former smoker, n (%)		
Current	53 (37.32)	11 (36.67)
Former	65 (45.77)	17 (56.67)
No	24 (16.90)	2 (6.67)
Nb of previous chemotherapy lines, n (%)		
None	27 (19.01)	15 (50.00)
1	65 (45.77)	9 (30.00)
2	24 (16.90)	5 (16.67)
3 or more	26 (18.31)	1 (3.33)
BSA median (range)	0.182 (0.11 – 0.28)	NA

Exploratory and validation cohort description

Characteristic	Exploratory cohort n=142: 132 CAL + 10 CHPG	External validation cohort n=30: 18 HBC + 12 CAL
SURGERY (yes), n (%)	30 (21.13)	NK
RADIOT (yes), n (%)	56 (39.44)	NK
FORMER CANCER (yes), n (%)	27 (19.01)	NK
DELAY ICPI (days), median (range)	286 (3 – 7785)	66 (13 – 2317)
GLY ($mmol.L^{-1}$), median (range)	5.6 (4.2 -12.4)	5.9 (4.6 – 9.0)
NEUTROPHIL ($10^9/L$), median (range)	4.89 (0.97- 25.13)	NK
LYMPHOCITE ($10^9/L$), median (range)	1.44 (0.30 - 3.5)	NK
LEUCOCYTE ($10^9/L$), median (range)	7.28 (2.70 – 27.80)	NK
NLR, median (range)	2.01 (0.50 -13.49)	NK
Outcome, n (%)		
6M-FPS (recurrence)	64 (45.07)	14 (46.66)
12M-OS (death)	53 (37.59)	18 (60.00)

Exploratory and validation cohort description

Characteristic	Exploratory cohort n=142: 132 CAL + 10 CHPG	External validation cohort n=30: 18 HBC + 12 CAL
NB_LESION, n (%)		
1	7 (4.93)	3 (10.00)
2	12 (8.45)	3 (10.00)
3	19 (13.38)	4 (13.33)
4	16 (11.27)	2 (6.67)
≥ 5	88 (61.97)	18 (60.00)
SUM_SUV, median (range)	30.18 (1.84 – 133.97)	NK
SUV_MAX, median (range)	10.04 (1.84 – 51.28)	11.85 (5.0 – 29.2)
tMTV (mL), median (range)	34.35 (0 – 818.10)	128.65 (3.0 – 949.7)
SUV_LIVER, median (range)	2.29 (1.28 – 3.98)	NK
SUV_SPLEEN, median (range)	1.89 (0.92 – 3.38)	NK
SUV_BONE, median (range)	1.64 (0.67 – 3.45)	NK
SLR, median (range)	0.83 (0.47 – 1.51)	NK
BLR, median (range)	0.70 (0.38 – 1.60)	NK

Multi-task stability selection

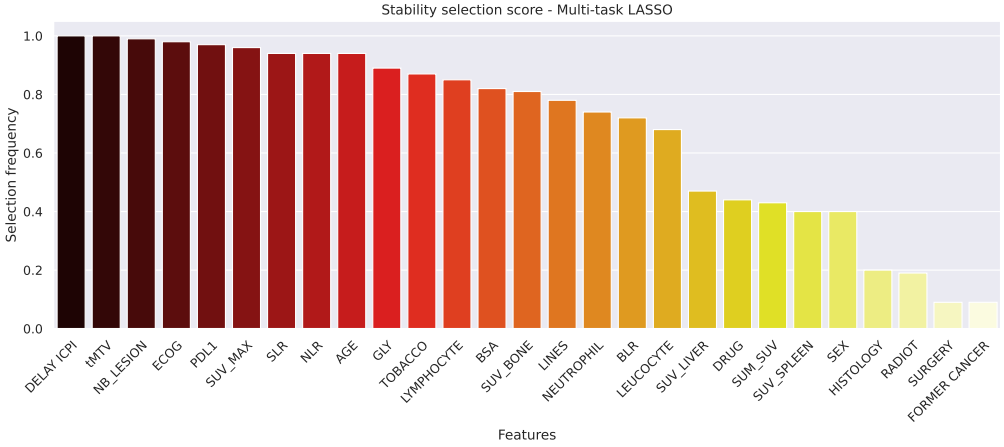
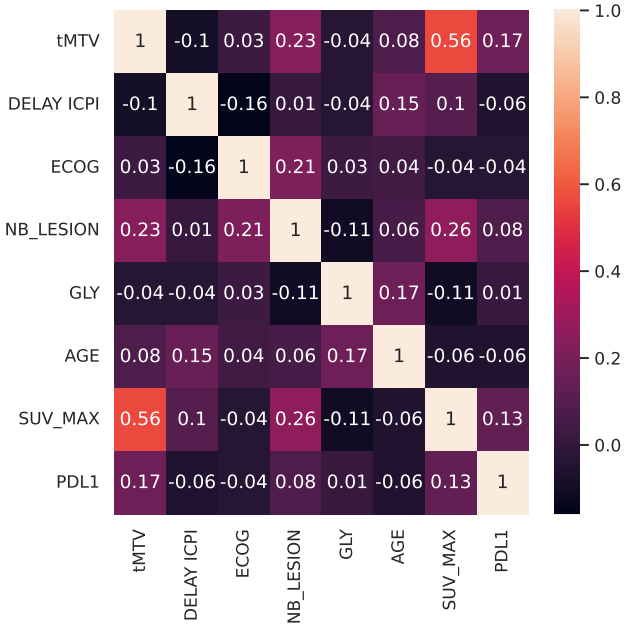


Figure: Features stability selection score using a multi-task LASSO.

Correlation of the features



Is $\hat{\mathcal{R}}_{DeSSL}(\theta)$ an accurate risk estimator ?

Theorem: The function $\lambda \mapsto \mathbb{V}(\hat{\mathcal{R}}_{DeSSL}(\theta))$ reaches its minimum for:

$$\lambda_{opt} = \frac{n_u}{n} \frac{\mathrm{Cov}(L(\theta; x, y), H(\theta; x))}{\mathbb{V}(H(\theta; x))}$$

and

$$\begin{aligned} \mathbb{V}(\hat{\mathcal{R}}_{DeSSL}(\theta))|_{\lambda_{opt}} &= (1 - \frac{n_u}{n}\rho^2_{L,H})\mathbb{V}(\hat{\mathcal{R}}_{CC}(\theta)) \\ &\leq \mathbb{V}(\hat{\mathcal{R}}_{CC}(\theta)) \end{aligned}$$

where $\rho_{L,H} = \mathrm{Corr}(L(\theta; x, y), H(\theta; x))$.

Justification on the heuristic idea that H should be a surrogate of L.

On the optimality of debiasing with the unlabelled set

Theorem: We consider a subset $\mathcal{A}$ of the training set and the following unbiased estimator under the MCAR assumption:

$$\hat{\mathcal{R}}_{DeSSL,\mathcal{A}}(\theta) = \frac{1}{n_l}\sum_{i\in\mathcal{L}} L(\theta; x_i, y_i) + \frac{\lambda}{n_u}\sum_{i\in\mathcal{U}} H(\theta; x_i) - \lambda\sum_{i\in\mathcal{A}} H(\theta; x_i).$$

The function $\mathbb{V}(\hat{\mathcal{R}}_{DeSSL,\mathcal{A}}(\theta)|r)$ reaches its minimum in $\mathcal{A}$ when the subset is the full dataset $\mathcal{A} = \mathcal{L}\cup\mathcal{U}$ or the labelled dataset $\mathcal{A} = \mathcal{L}$ and both are equivalent.

DeSSL models are calibrated

Theorem: If the original loss is a proper scoring rule, then DeSSL is also a proper scoring rule.

- We can expect DeSSL to be as well-calibrated as the complete case.
- What if the model is non probabilistic ? Fisher consistency

Asymptotic consistency

$\hat{\theta}$ is asymptotically consistent with respect to n if $d(\hat{\theta}, \theta^*) \xrightarrow{p} 0$.

Theorem: Under usual regularity conditions of M-estimators for both L and H, $\hat{\theta} = \arg\min \hat{\mathcal{R}}_{DeSSL}$ is asymptotically consistent with respect to n.

SSL may fail with an infinite number of labelled data when DeSSL will not.

Asymptotic normality

Theorem:
Suppose L and H are smooth functions in $\mathcal{C}^2(\Theta, \mathbb{R})$. Assume $\mathcal{R}(\theta)$ admit a second-order Taylor expansion at θ^* with a non-singular second order derivative V_{θ^*}. Under the MCAR assumption, we have that $\hat{\theta}_{DeSSL}$ is asymptotically normal and we can minimise the trace of the asymptotic covariance. Indeed, $\mathrm{Tr}(\Sigma_{DeSSL})$ reaches its minimum at

$$\lambda_{opt} = (1-\pi)\frac{\mathrm{Tr}(V_{\theta^*}^{-1} K_{\theta^*} V_{\theta^*}^{-1})}{\mathrm{Tr}(V_{\theta^*}^{-1}\mathbb{E}\left[\nabla H(\theta^*; x)\nabla H(\theta^*; x)^T\right] V_{\theta^*}^{-1})},$$

and at λ_{opt} :

$$\mathrm{Tr}(\Sigma_{DeSSL}) - \mathrm{Tr}(\Sigma_{CC}) = -\frac{1-\pi}{\pi}\frac{\mathrm{Tr}(V_{\theta^*}^{-1} K_{\theta^*} V_{\theta^*}^{-1})^2}{\mathrm{Tr}(V_{\theta^*}^{-1}\mathbb{E}\left[\nabla H(\theta^*; x)\nabla H(\theta^*; x)^T\right] V_{\theta^*}^{-1})} \leq 0.$$

DeSSL does benefit of generalisation error bounds

$$R_n = \mathbb{E}_{(\varepsilon_i)_{i\leq n}}\left[\sup_{\theta\in\Theta}\left(\frac{1}{n\pi}\sum_{i\in\mathcal{L}}\varepsilon_i L(\theta;x_i,y_i) - \frac{\lambda}{n\pi}\sum_{i\in\mathcal{L}}\varepsilon_i H(\theta;x_i) + \frac{\lambda}{n(1-\pi)}\sum_{i\in\mathcal{U}}\varepsilon_i H(\theta;x_i)\right)\right],$$

Theorem: We assume that labels are MCAR and that both L and H are bounded. Then, there exists a constant $\kappa > 0$, that depends on λ, L, H, and the ratio of observed labels, such that, with probability at least $1-\delta$, for all $\theta \in \Theta$,

$$\mathcal{R}(\theta) \leq \hat{\mathcal{R}}_{DeSSL}(\theta) + 2R_n + \kappa\sqrt{\frac{\log(4/\delta)}{n}}.$$

USB dataset description: CV

- **CIFAR-100:** natural image (32×32) recognition dataset consisting 100 classes. 500 training samples and 100 test samples per class.
- **STL-10:** natural color image (96×96) recognition dataset consisting 10 classes. 500 training samples and 800 test samples per class. 100,000 unlabelled samples (open-world).
- **EuroSat:** Sentinel-2 satellite images covering 13 spectral bands and consisting of 10 classes with 27,000 labeled and geo-referenced samples.
- **TissueMNIST:** medical dataset of human kidney cortex cells, segmented from 3 reference tissue specimens and organized into 8 categories. 236,386 training samples. gray-scale image (28×28).
- **Semi-Aves:** Aves (birds) classification,5,959 images of 200 bird species are labelled and 26,640 images are unlabelled. Naturally imbalanced.

USB dataset description: audio and NLP

Audio

- **GTZAN:** Music genre classification of 10 classes and 100 audio recordings for each class.
- **UrbanSound8k:** 8,732 labelled sound events of urban sounds of 10 classes, with the maximum length of 4 seconds.
- **FSDNoisy18k:** sound event classification dataset across 20 classes. Small amount of labelled data and noisy unlabelled data.
- **ESC-50:** 2,000 environmental audio recordings for 50 sound classes.

NLP:

- **Amazon Review:** Sentiment classification dataset. There are 5 classes (scores). Each class (score) contains 600,000 training samples and 130,000 test samples.
- **AG News:** News topic classification dataset containing 4 classes. itemize

Integration to USB benchmarck [Wang et al., 2022]

	CIFAR-100		STL10		Euro-SAT		TissueMNIST		Semi-Aves
# Label	200	400	40	100	20	40	80	400	3959
Complete Case	35.88±0.36	26.76±0.83	19.0±2.9	10.87±0.49	26.49±1.6	16.12±1.35	60.36±3.83	54.08±1.55	41.2±0.17
Fixmatch	**29.6±0.9**	**19.56±0.52**	**16.15±1.89**	8.11±0.68	**13.44±3.53**	5.91±2.02	55.37±4.5	51.24±1.56	**31.9±0.06**
DeFixmatch	31.52±1.85	21.12±1.74	17.68±7.94	**7.94±1.31**	14.71±6.52	**3.72±0.79**	**54.07±6.19**	**48.95±1.14**	32.01±0.26

Integration to USB benchmarck [Wang et al., 2022]

	CIFAR-100		STL10		Euro-SAT		TissueMNIST		Semi-Aves
# Label	200	400	40	100	20	40	80	400	3959
Complete Case	35.88±0.36	26.76±0.83	19.0±2.9	10.87±0.49	26.49±1.6	16.12±1.35	60.36±3.83	54.08±1.55	41.2±0.17
Fixmatch	**29.6±0.9**	**19.56±0.52**	**16.15±1.89**	8.11±0.68	**13.44±3.53**	5.91±2.02	55.37±4.5	51.24±1.56	**31.9±0.06**
DeFixmatch	31.52±1.85	21.12±1.74	17.68±7.94	**7.94±1.31**	14.71±6.52	**3.72±0.79**	**54.07±6.19**	**48.95±1.14**	32.01±0.26

	GTZAN		UrbanSound8k		FSDnoisy	ESC-50	
# Label	100	400	100	400	1773	250	500
Complete Case	52.73±2.86	32.04±0.51	42.65±1.63	27.6±1.19	34.74±1.58	49.83±1.71	38.75±1.47
Fixmatch	41.47±1.62	**21.89±1.01**	40.02±6.62	20.83±2.31	31.05±1.27	43.58±2.79	32.0±1.08
DeFixmatch	**47.15±3.66**	22.31±2.06	**38.43±5.42**	**20.47±1.55**	**29.53±0.79**	**41.83±0.11**	**31.74±0.20**

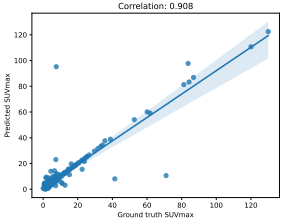
Integration to USB benchmarck [Wang et al., 2022]

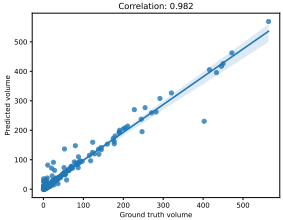
	CIFAR-100		STL10		Euro-SAT		TissueMNIST		Semi-Aves
# Label	200	400	40	100	20	40	80	400	3959
Complete Case	35.88±0.36	26.76±0.83	19.0±2.9	10.87±0.49	26.49±1.6	16.12±1.35	60.36±3.83	54.08±1.55	41.2±0.17
Fixmatch	**29.6±0.9**	**19.56±0.52**	**16.15±1.89**	8.11±0.68	**13.44±3.53**	5.91±2.02	55.37±4.5	51.24±1.56	**31.9±0.06**
DeFixmatch	31.52±1.85	21.12±1.74	17.68±7.94	**7.94±1.31**	14.71±6.52	**3.72±0.79**	**54.07±6.19**	**48.95±1.14**	32.01±0.26

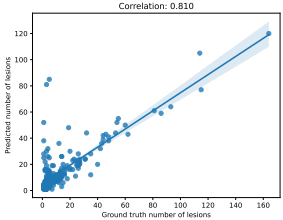
	GTZAN		UrbanSound8k		FSDnoisy	ESC-50	
# Label	100	400	100	400	1773	250	500
Complete Case	52.73±2.86	32.04±0.51	42.65±1.63	27.6±1.19	34.74±1.58	49.83±1.71	38.75±1.47
Fixmatch	41.47±1.62	**21.89±1.01**	40.02±6.62	20.83±2.31	31.05±1.27	43.58±2.79	32.0±1.08
DeFixmatch	**47.15±3.66**	22.31±2.06	**38.43±5.42**	**20.47±1.55**	**29.53±0.79**	**41.83±0.11**	**31.74±0.20**

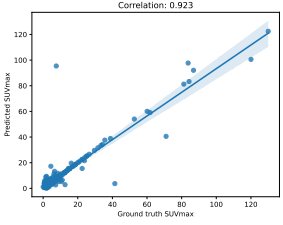
	AG News		Amazon Review	
# Label	40	200	250	1000
Complete Case	**15.06±1.08**	14.25±0.97	52.31±1.28	47.53±0.69
Fixmatch	30.17±1.87	11.71±1.95	**47.61±0.83**	**43.05±0.54**
Defixmatch	31.17+-2.55	**11.68+-1.16**	49.92+60.94	44.62+-1.67

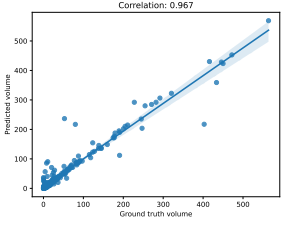
Biomarkers prediction: SegMT versus DeSegMT

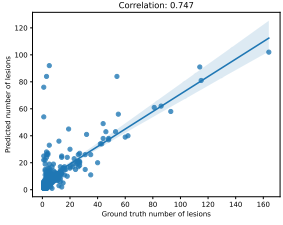












The debiasing method performs well on classic metrics...

			200			50		
	Uncertainty	Method	Dice	FP	FN	Dice	FP	FN
Pretraining	None	CC	0.616	988.8	440.4	0.502	662.1	1082.4
Finetuning	None	CC	0.627	2039.9	242.8	0.594	2177.5	269.0
	None	SegPL	**0.638**	**1173.6**	248.0	**0.629**	**1025.7**	441.7
		DeSegPL	0.625	1590.5	**204.9**	0.587	1176.2	**368.6**
		SegMT	0.622	1451.5	**186.1**	**0.623**	**1197.1**	448.3
		DeSegMT	**0.631**	**1188.0**	203.3	0.607	1487.2	**352.0**
	Softmax	SegPL	0.612	1652.8	**169.8**	0.602	1289.0	**333.7**
		DeSegPL	**0.641**	**1318.5**	272.0	**0.614**	**1198.6**	437.0
		SegMT	**0.631**	**1330.5**	305.4	0.611	1207.3	**341.3**
		DeSegMT	0.622	1695.7	**191.7**	**0.621**	**1032.9**	407.2

... and also on biomarker prediction.

			200			50		
	Uncertainty	Method	NMSE SUV_{max}	NMSE volume	MAE N_lesions	NMSE SUV_{max}	NMSE volume	MAE N_lesions
Pretraining	None	CC	6.51	10.03	1.17	2.27	152.01	1.07
Finetuning	None	CC	17.02	83.38	1.2	17.14	210.10	1.17
	None	SegPL	**8.56**	**37.25**	1.17	13.30	62.70	1.2
		DeSegPL	11.42	83.82	**1.13**	**7.10**	**37.37**	**1.17**
		SegMT	15.21	70.58	1.15	**12.71**	128.97	**1.17**
		DeSegMT	**12.54**	**46.52**	**1.1**	13.37	**59.13**	**1.17**
	Softmax	SegPL	15.32	45.05	1.2	**5.79**	407.9	1.2
		DeSegPL	**12.02**	**23.24**	**1.15**	12.28	**254.75**	**1.03**
		SegMT	13.93	31.385	1.2	**4.18**	232.58	1.17
		DeSegMT	**9.21**	**26.67**	**1.17**	10.13	**178.10**	**1.08**

The debiasing method performs well on classic metrics...

H = Cross-entropy

			200			50		
	Uncertainty	Method	Dice	FP	FN	Dice	FP	FN
Pretraining	None	CC	0.583	393.6	160.4	0.551	108.4	210.7
Finetuning	None	CC	0.621	359.3	210.1	0.576	608.5	88.5
	None	SegPL	**0.632**	514.3	**102.8**	0.560	826.3	**119.7**
		DeSegPL	0.630	**432.9**	302.0	**0.595**	**188.7**	161.4
		SegMT	**0.631**	676.4	**63.6**	0.591	411.1	155.6
		DeSegMT	0.616	**394.8**	147.6	**0.616**	**276.2**	**132.2**
	Softmax	SegPL	**0.631**	**444.3**	135.4	**0.581**	**459.4**	**93.9**
		DeSegPL	0.606	481.2	**102.4**	0.579	481.2	102.4
		SegMT	0.632	323.9	**97.23**	0.561	633.5	**96.7**
		DeSegMT	**0.661**	**163.33**	165.9	**0.602**	**511.1**	135.3

... and also on biomarkers prediction.

H = Cross-entropy

			200			50		
	Uncertainty	Method	NMSE SUV_{max}	NMSE volume	MAE N_lesions	NMSE SUV_{max}	NMSE volume	MAE N_lesions
Pretraining	None	CC	w	23.2	1.17		0.323	0.80
Finetuning	None	CC		13.4	1.17		8.85	1.23
	None	SegPL		**102.8**	**1.20**	**119.7**	**15.87**	**1.33**
		DeSegPL		302.0	**1.20**	161.4	**1.27**	**1.00**
		SegMT		**63.6**	**1.20**	155.6	3.58	**1.07**
		DeSegMT		147.6	1.23	**132.2**	**1.89**	1.23
	Softmax	SegPL		**135.4**	**1.27**	**93.9**	**9.06**	**1.33**
		DeSegPL		**102.4**	**1.23**	102.4	33.55	**1.23**
		SegMT		**97.23**	1.17	**96.7**	13.1	**1.33**
		DeSegMT		165.9	**1.13**	135.3	**2.65**	**1.13**

Dice, false positive and negative distribution

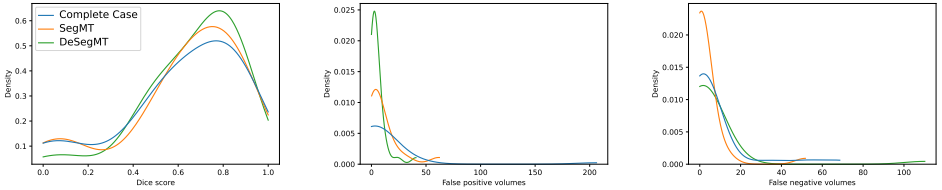
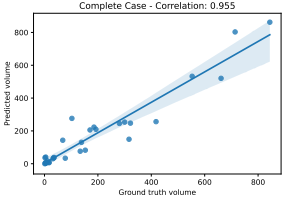
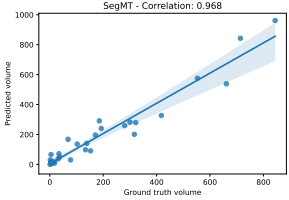
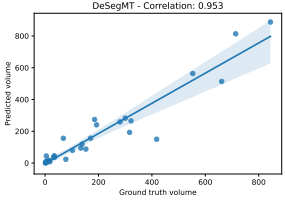


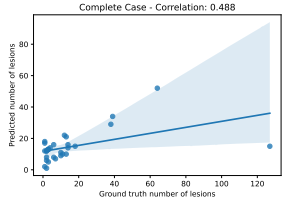
Figure: Kernel density estimation on the test set of (Left) Dice score (Middle) False positive volumes (Right) False negative volumes.

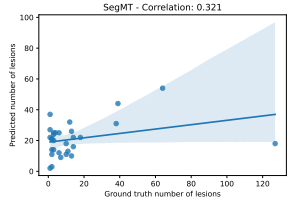
Biomarkers prediction

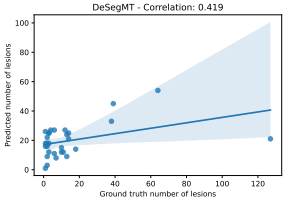




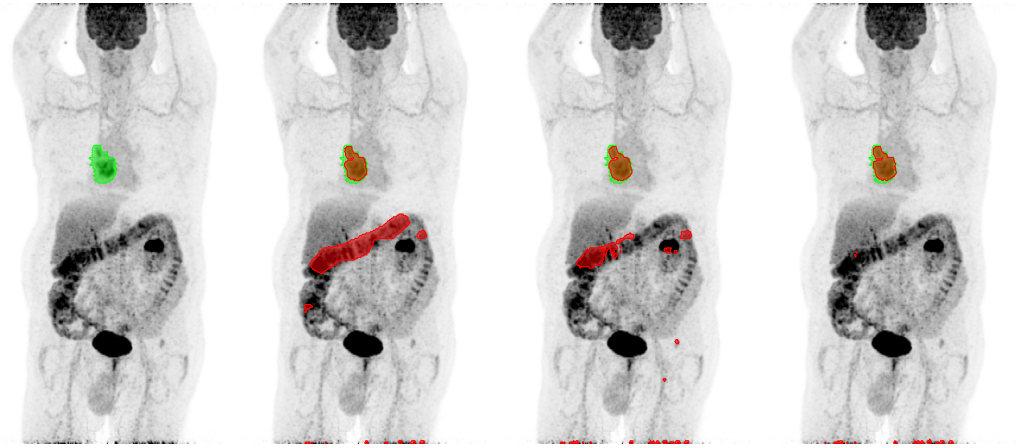




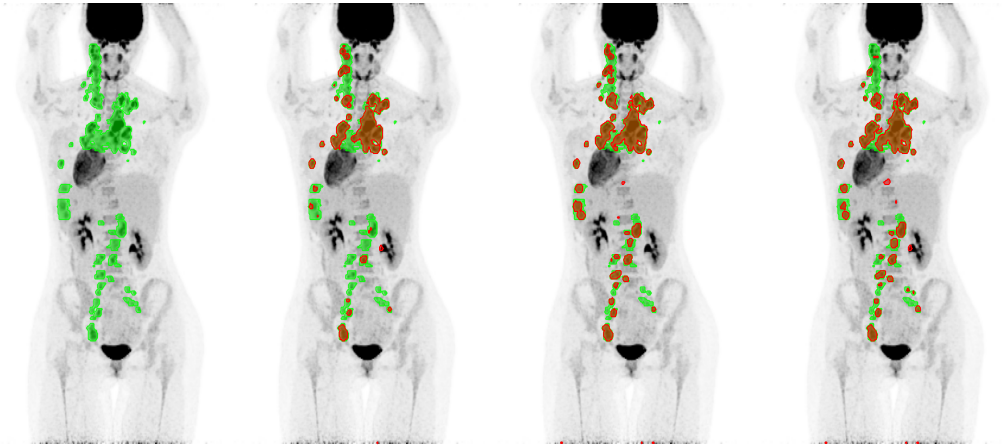




Example of segmentation



Example of segmentation



Dice, false positive and negative distribution for AI4PET

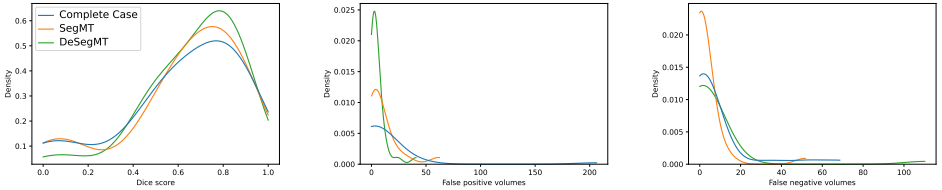


Figure: Kernel density estimation on the test set of (Left) Dice score (Middle) False positive volumes (Right) False negative volumes.

Biomarkers prediction for AI4PET