Jožef Stefan Institute

SEMI-SUPERVISED AND MULTI-LABEL CLASSIFICATION OF REMOTELY SENSED IMAGES

Sašo Džeroski (w/ Marjan Stoimchev, Sintija Stevanoska, Jurica Levatic, Dragi Kocev, Michelangelo Ceci)

Head of Department of knowledge technologies, Jožef Stefan Institute, Ljubljana, Slovenia

Professor, Jožef Stefan International Postgraduate School, Ljubljana, Slovenia

Visiting professor, European Space Agency, Frascati, Italy

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

MULTI-TARGET PREDICTION & MULTI-LABEL CLASSIFICATION

SINGLE-TARGET vs. MULTI-TARGET PREDICTION (classification, regr.)

	Descriptive space				Target space
Example 1	1	TRUE	0.49	0.69	Yes
Example 2	2	FALSE	0.08	0.07	Yes
Example 3	1	FALSE	0.08	0.07	No
Example 4	2	TRUE	0.49	0.69	Yes
Example 5	3	TRUE	0.49	0.69	No
Example 6	4	FALSE	0.08	0.07	Yes
…	…				…

	Descriptive space				Target space		
Example 1	1	TRUE	0.49	0.69	0.68	0.60	3.91
Example 2	2	FALSE	0.08	0.07	0.56	0.99	7.59
Example 3	1	FALSE	0.08	0.07	0.10	1.69	7.57
Example 4	2	TRUE	0.49	0.69	0.08	0.77	8.86
Example 5	3	TRUE	0.49	0.69	0.11	3.51	2.50
Example 6	4	FALSE	0.08	0.07	0.43	2.10	8.09
…	…				…	…	…

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

THE RATIONALE FOR MULTI-TARGET PREDICTION

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

It makes sense to predict inter-related targets jointly

In weather forecasting, we have multiple tasks

- Predicting the outlook (sunny, overcast, rain): STC
- Predicting the temperature (in degrees Celsius): STR
- Predicting the weather: MTP
  - Outlook
  - Temperature
  - Humidity
  - Quantity of precipitation …

4

MULTI-LABEL CLASSIFICATION: The task and an example

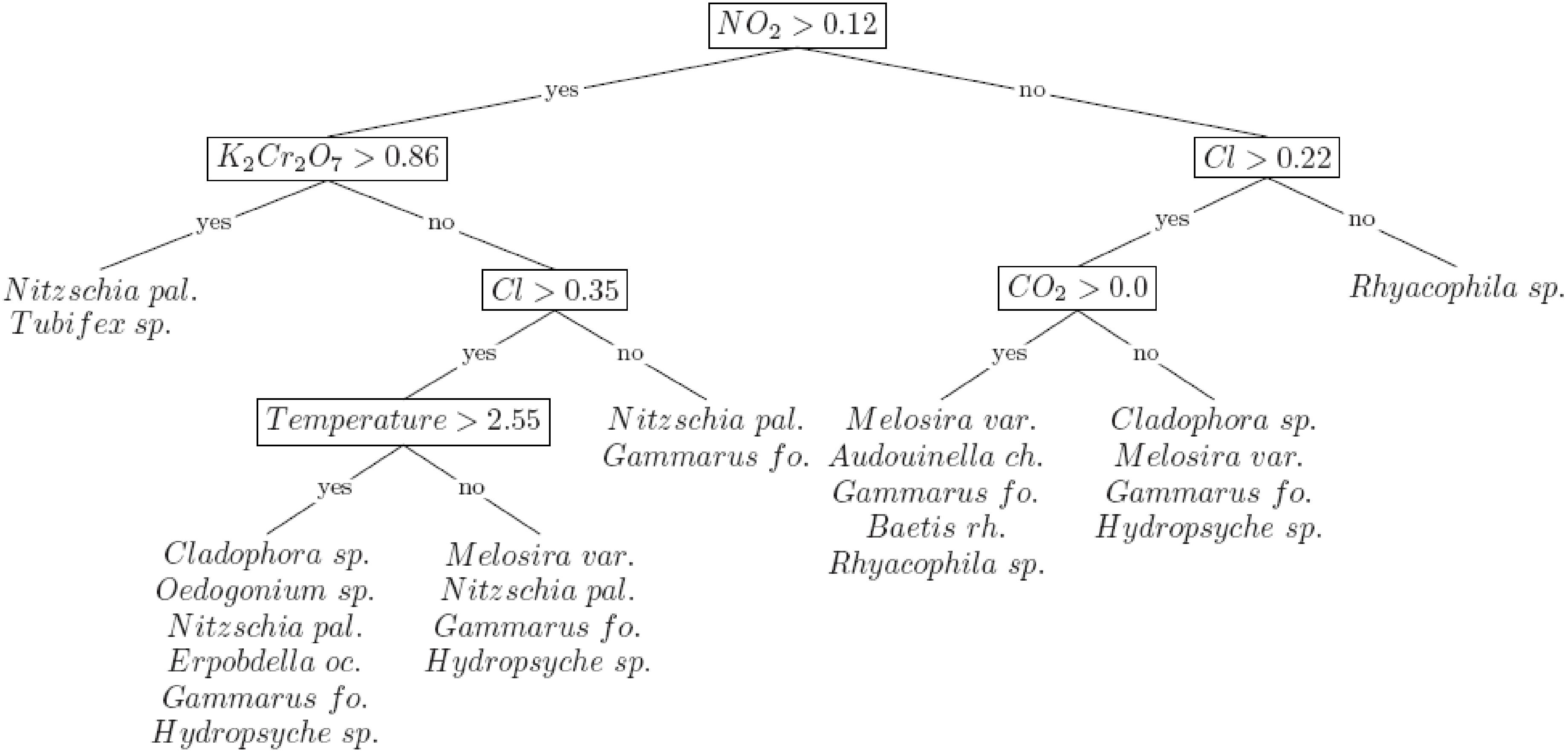
Learning models that simultaneously predict several nominal/binary target variables

Input: *A vector of descriptive variables' values*

Output: *A vector of several binary targets' values*

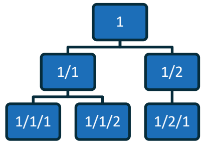
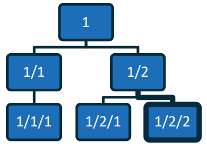
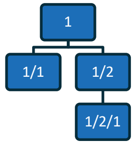
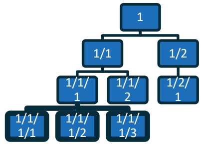
Sample ID	Descriptive variables						Target variables													
	Temperature	$K_2Cr_2O_7$	NO_2	Cl	CO_2	...	*Cladophora sp.*	*Gongrosira incrustans*	*Oedogonium sp.*	*Stigeoclonium tenue*	*Melosira varians*	*Nitzschia palea*	*Audouinella chalybea*	*Erpobdella octoculata*	*Gammarus fossarum*	*Baetis rhodani*	*Hydropsyche sp.*	*Rhyacophila sp.*	*Simulim sp.*	*Tubifex sp.*
ID1	0.66	0.00	0.40	1.46	0.84	...	1	0	0	0	0	1	1	0	1	1	1	1	1	1
ID2	2.03	0.16	0.35	1.74	0.71	...	0	1	0	1	1	1	1	0	1	1	1	1	1	0
ID3	3.25	0.70	0.46	0.78	0.71	...	1	1	0	0	1	0	1	0	1	1	1	0	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

A DECISION TREE FOR MULTI-LABEL CLASSIFICATION



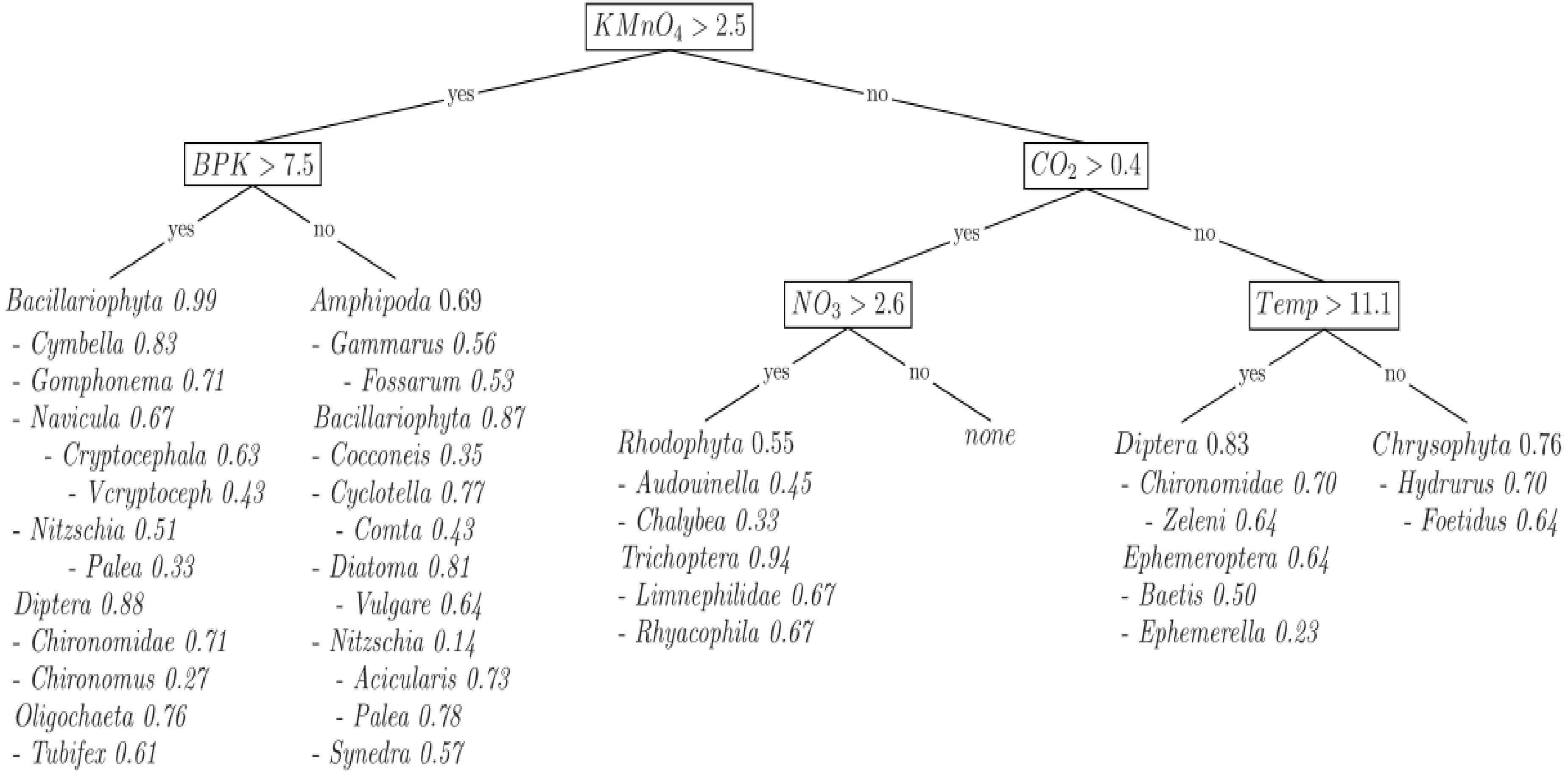
6

HIERARCHICAL MULTI-LABEL CLASSIFICATION

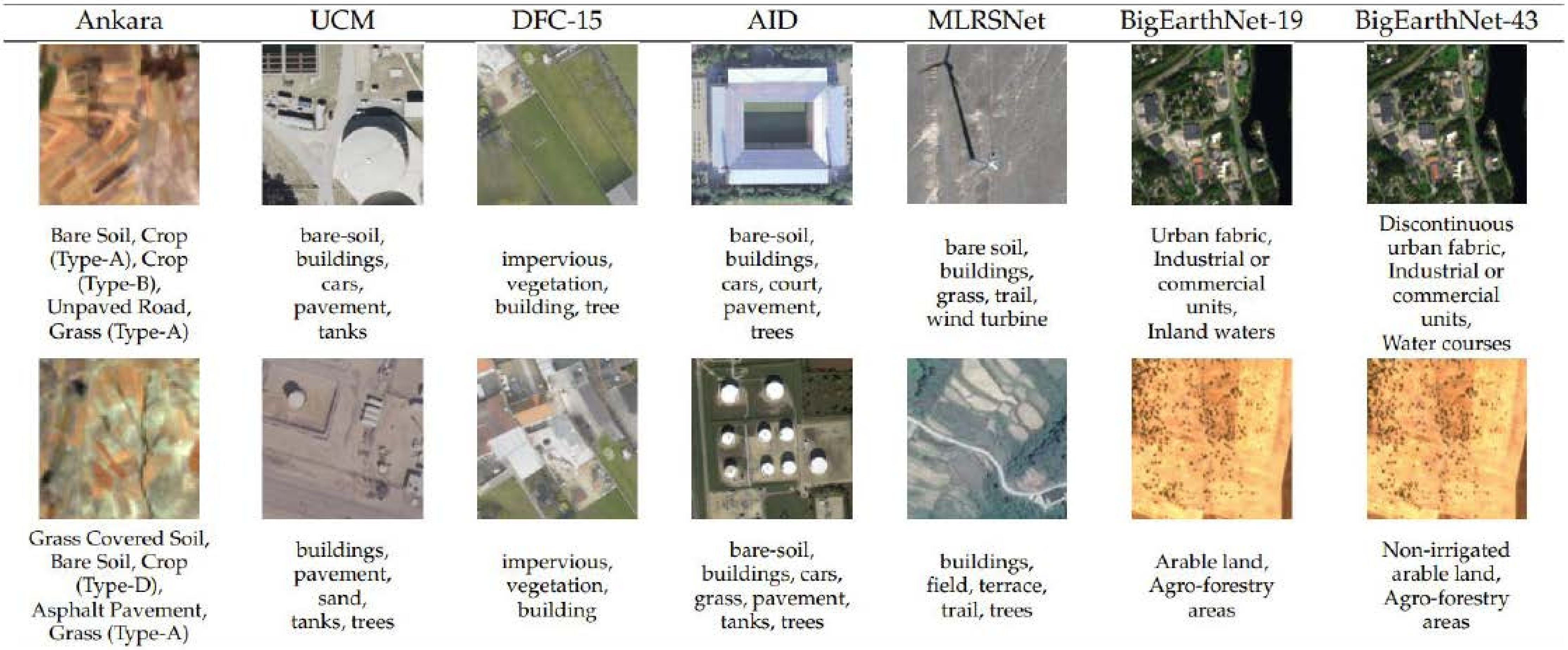
	Descriptive space				Target space
Example 1	1	TRUE	0.49	0.69	
Example 2	2	FALSE	0.08	0.07	
Example 3	1	FALSE	0.08	0.07	
Example 4	2	TRUE	0.49	0.69	
…	…				…

A DECISION TREE FOR HIERARCHICAL MULTI-LABEL CLASSIFICATION

Taking into account the taxonomy of living organisms

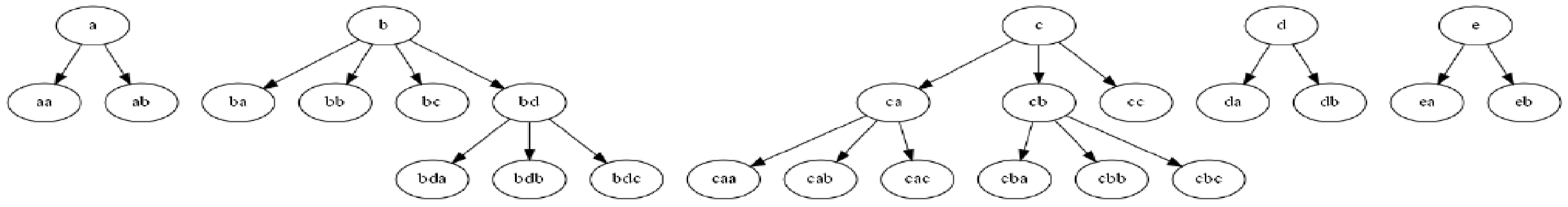


MULTI-LABEL CLASSIFICATION OF REMOTELY SENSED IMAGES



HIERARCHICAL MULTI-LABEL CLASSIFICATION OF REMOTELY SENSED IMAGES (CLC nomenclature)

Rev CLC Level 1	Rev CLC Level 2	Rev CLC Level 3
a Artificial surfaces	aa Urban fabric	
	ab Industrial, commercial and transport units	
b Agricultural areas	ba Arable land	
	bb Permanent crops	
	bc Pastures	
	bd Heterogeneous agricultural areas	bda Complex cultivation patterns bdb Land principally occupied by agriculture, with signific bdc Agro-forestry areas
c Forests and semi-natural areas	ca Forests	caa Broad-leaved forest cab Coniferous forest cac Mixed forest
	cb Shrub and/or herbaceous vegetation association	cba Natural grassland and sparsely vegetated areas cbb Moors, heathland and sclerophyllous vegetation cbc Transitional woodland shrub
	cc Beaches, dunes, sands	
d Wetlands	da Inland wetlands	
	db Coastal wetlands	
e Water bodies	ea Inland waters	
	eb Marine waters	



DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

SEMI-SUPERVISED MULTI-TARGET PREDICTION WITH PREDICTIVE CLUSTERING TREES

SEMI-SUPERVISED MULTI-TARGET PREDICTION

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Different types of structured outputs

- MT/ML Classification, MTR
  Hierarchical MLC/MTR

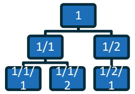
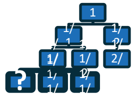
Different supervision levels

- Fully supervised
- Semi-supervised
  - Missing labels
  - Partial labels
- Unsupervised

Two example tasks

- MTR w partial labels
- Semi-supervised HMLC

	Descriptive space				Target space		
Example 1	1	TRUE	0.49	0.69	?	0.60	3.91
Example 2	2	FALSE	0.08	0.07	0.56	0.99	7.59
Example 3	1	FALSE	0.08	0.07	?	?	?
Example 4	2	TRUE	0.49	0.69	0.08	0.77	8.86
Example 5	3	TRUE	0.49	0.69	0.11	?	?
Example 6	4	FALSE	0.08	0.07	0.43	2.10	8.09
…	…				…	…	…

	Descriptive space				Target space
Example 1	1	TRUE	0.49	0.69	
Example 2	2	FALSE	0.08	0.07	?
Example 3	1	FALSE	0.08	0.07	?
Example 4	2	TRUE	0.49	0.69	
…	…				…

LEARNING TREES FOR MULTI-TARGET PREDICTION WITH PREDICTIVE CLUSTERING

To construct a tree T from a training set S:

If **the examples in S have low variance**,

construct a leaf labeled *target(prototype(S))*

Otherwise:

- Select the best attribute A with values v1, …, vn, which **reduces the most the variance** (*measured according to a given distance function d*)
- Partition S into S1, …, Sn according to A
- Recursively construct subtrees T1 to Tn for S1 to Sn
- Result: a tree with root A and subtrees T1, …, Tn

The variance is assessed across the multiple targets

SELECTING THE BEST TEST IN A PCT

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Select the test that maximizes variance reduction

Calculated in line 4

procedure BestTest(E)

1: $(t^*, h^*, \mathcal{P}^*) = (none, 0, \emptyset)$
2: **for each** possible test t **do**
3: $\quad \mathcal{P} =$ partition induced by t on E
4: $\quad h = Var(E) - \sum_{E_i \in \mathcal{P}} \frac{|E_i|}{|E|} Var(E_i)$
5: $\quad$ **if** $(h > h^*) \wedge \text{Acceptable}(t, \mathcal{P})$ **then**
6: $\quad\quad (t^*, h^*, \mathcal{P}^*) = (t, h, \mathcal{P})$
7: **return** $(t^*, h^*, \mathcal{P}^*)$

$$Var(E) = \sum_{i=1}^{T} Var(Y_i).$$

SEMI-SUPERVISED LEARNING WITH PCTs

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

New definition of variance that includes both targets and attributes, e.g., for MTR

$$Var(E) = \frac{1}{T+D} \cdot \left(w \cdot \sum_{i=1}^{T} Var(Y_i) + (1-w) \cdot \sum_{j=1}^{D} Var(X_j) \right)$$

T = #target attributes, D = #descriptive attributes

$$E = E_{\text{Labeled}} \cup E_{\text{Unlabeled}}$$

Variances only calculated for non-missing values

SEMI-SUPERVISED LEARNING WITH PCTs

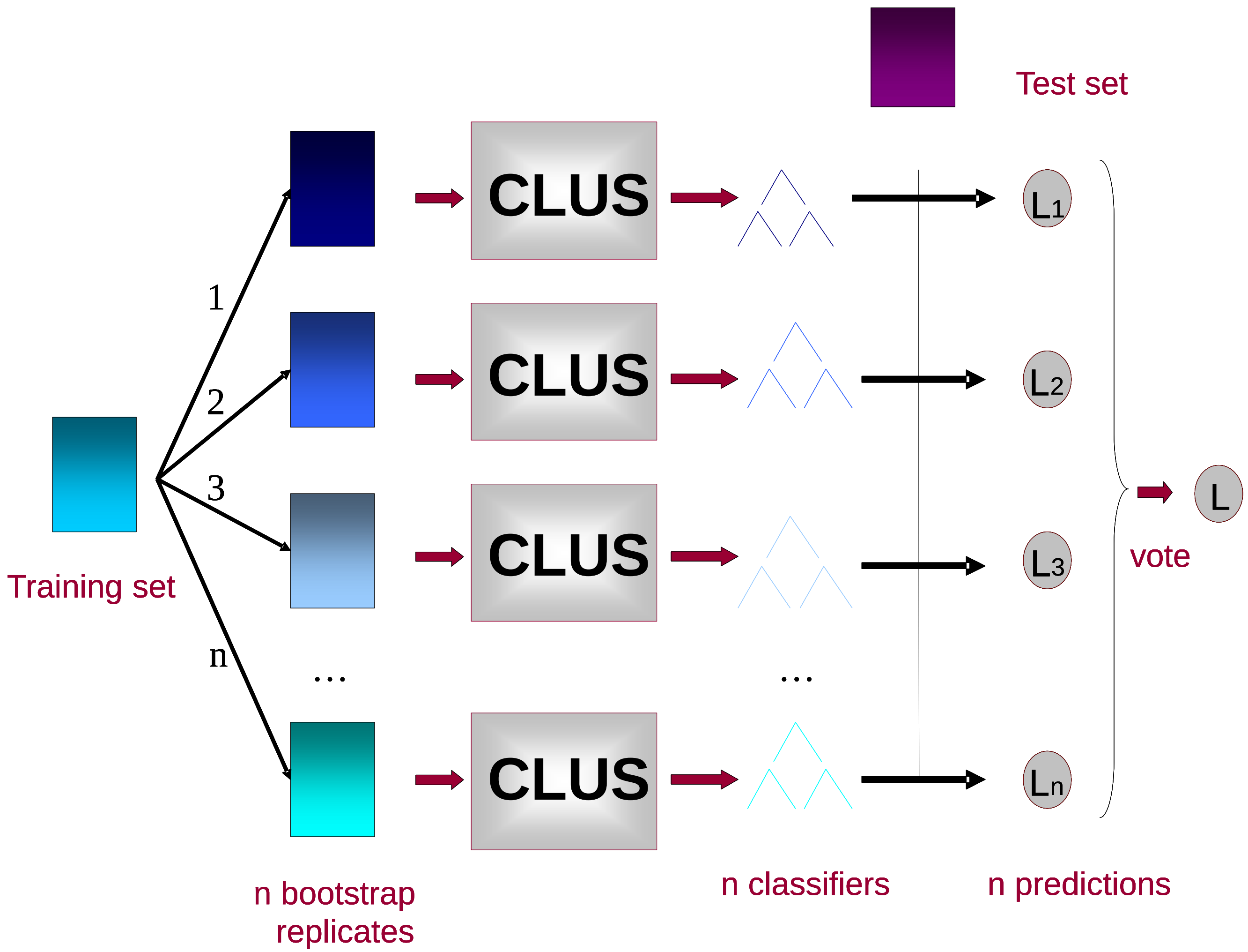
- $Var_f(E,Y) = Var/Gini/\mathit{Weighted\ Euclidean}$ for MTR/MLC/HMLC

- $Var_f(E,X) = \frac{1}{D}\left(\sum_{X_i \text{ numeric}} Var(E,X_i) + \sum_{X_j \text{ nominal}} Gini(E,X_j)\right)$

- w = weight parameter, trades-off focus on
    - Prediction (w=1)
    - Clustering (w=0)
- w tuned by internal cross-validation on labeled part

LEARNING TREE ENSEMBLES

Typical approach: Generate different samples of the data (subsets of rows, subset of columns, or both), then learn a tree on each



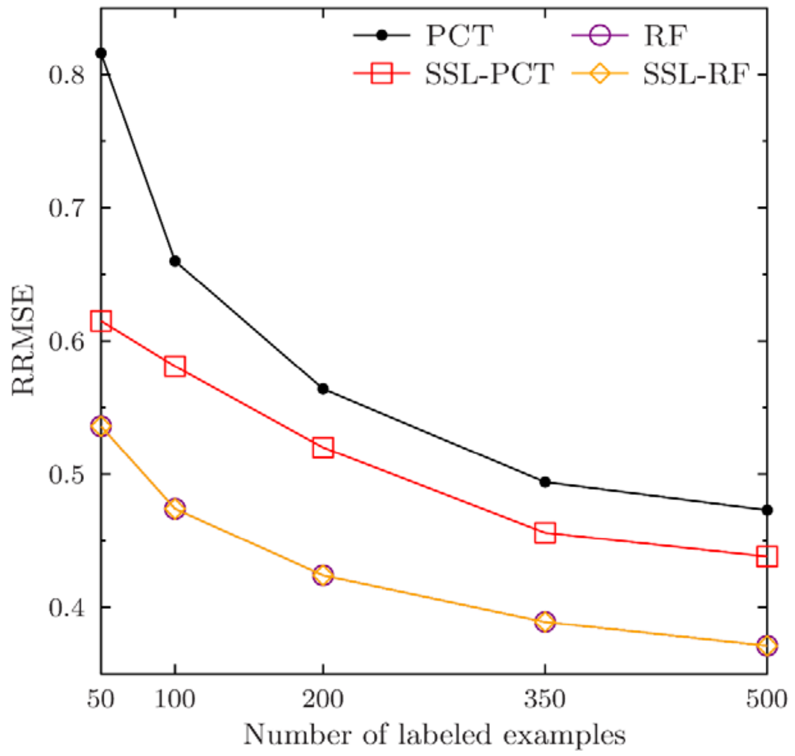
SSL vs. SL: PERFORMANCE

Methods			Number of labeled examples 25	50	100	200	350	500
Binary classification								
PCT	vs.	SSL-PCT	**0.009**	0.388	0.066	**0.005**	**0.019**	**0.019**
RF	vs.	SSL-RF	0.529	0.192	**0.002**	0.099	0.093	**0.012**
Multi-class classification								
PCT	vs.	SSL-PCT	0.248	0.084	**0.014**	**0.007**	0.192	0.081
RF	vs.	SSL-RF	0.563	**0.011**	**0.011**	**0.003**	**0.004**	**0.02**
Regression								
PCT	vs.	SSL-PCT	**0.011**	**0.01**	**0.004**	0.367	0.48	0.583
RF	vs.	SSL-RF	**0.008**	**0.065**	**0.008**	**0.023**	**0.034**	0.126

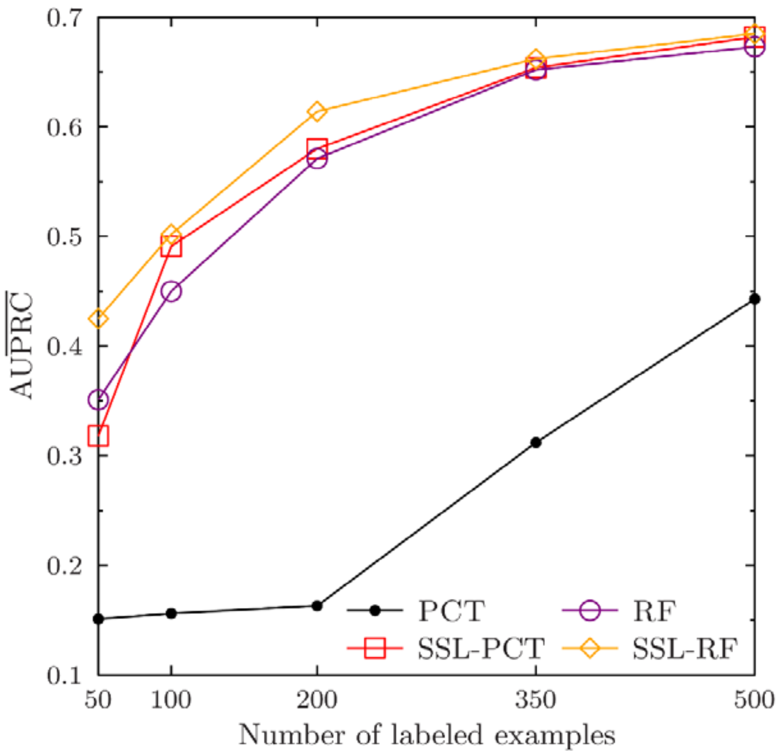
Multi-target regression							
PCT	vs.	SSL-PCT	0.093	**0.022**	**0.028**	**0.022**	**0.009**
RF	vs.	SSL-RF	0.959	0.445	0.445	0.333	0.445
Multi-label classification							
PCT	vs.	SSL-PCT	**0.013**	**0.008**	**0.008**	0.093	0.053
RF	vs.	SSL-RF	0.241	0.415	0.262	0.308	0.575
Hierarchical multi-label classification							
PCT	vs.	SSL-PCT	0.834	0.093	**0.028**	**0.028**	**0.028**
RF	vs.	SSL-RF	0.345	0.345	0.249	0.345	0.345

SSL vs. SL PERFORMANCE

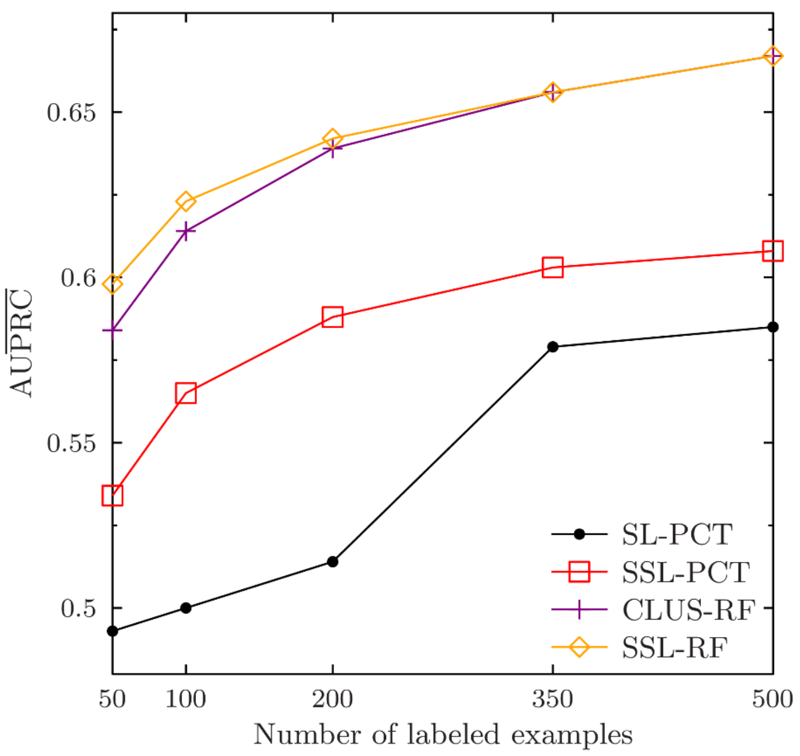
RF1 (MTR)



Medical (MLC)

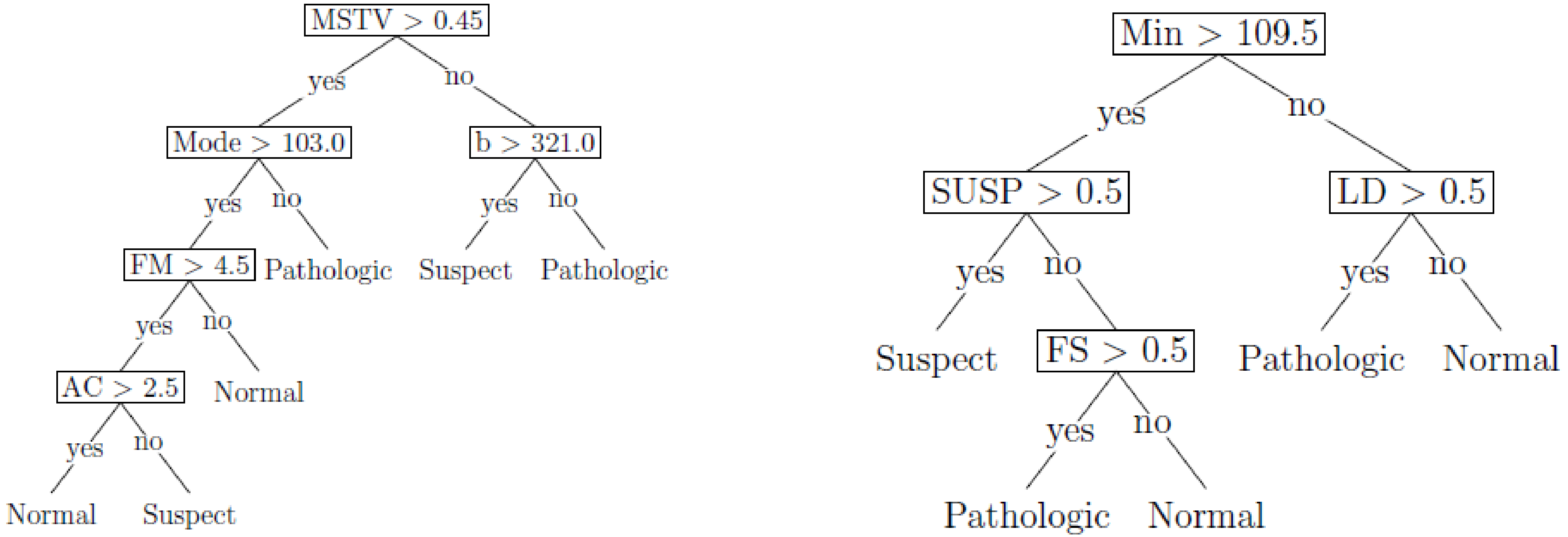


Enron (HMLC)



SSL OF DECISION TREES: ACCURACY & INTERPRETABILITY

Multi-class classification (Cardiotocogramy3 Dataset)



Accuracy=81%, 11 nodes

Accuracy=92%, 9 nodes

(c) SL-PCT, 50 labeled examples

(d) SSL-PCT, 50 labeled and 2076 unlabeled examples

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

remote sensing

MDPI

Article

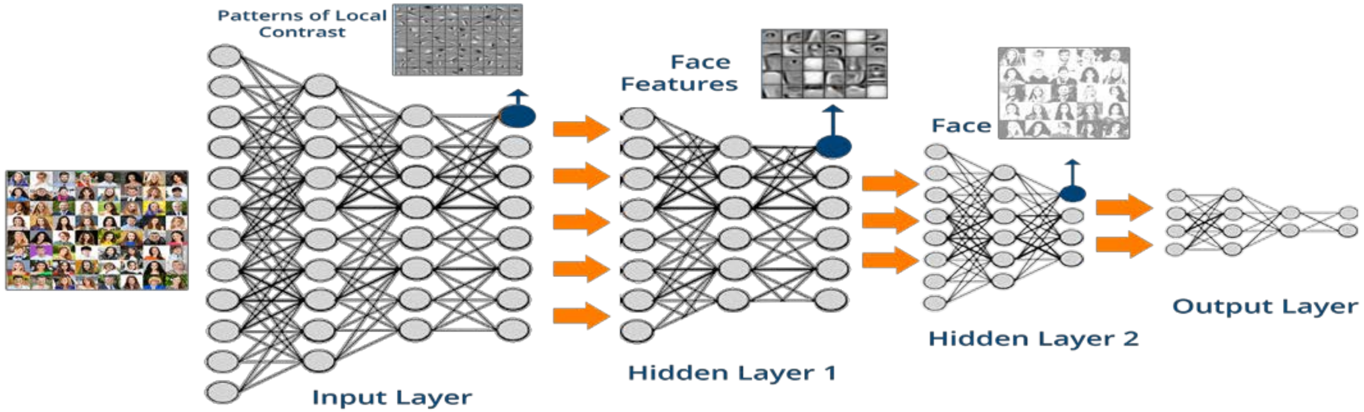
Deep Network Architectures as Feature Extractors for Multi-Label Classification of Remote Sensing Images

Marjan Stoimchev [1,2,*], Dragi Kocev [1,2,3] and Sašo Džeroski [1,2]

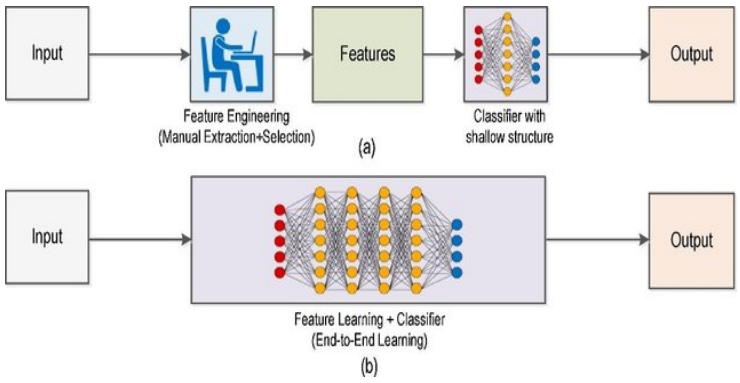
COMBINING DEEP NEURAL NETWORKS AND ENSEMBLES OF PREDICTIVE CLUSTERING TREES FOR MULTI-LABEL CLASSIFCATION OF RSI

CONVOLUTIONAL DEEP NEURAL NETWORKS

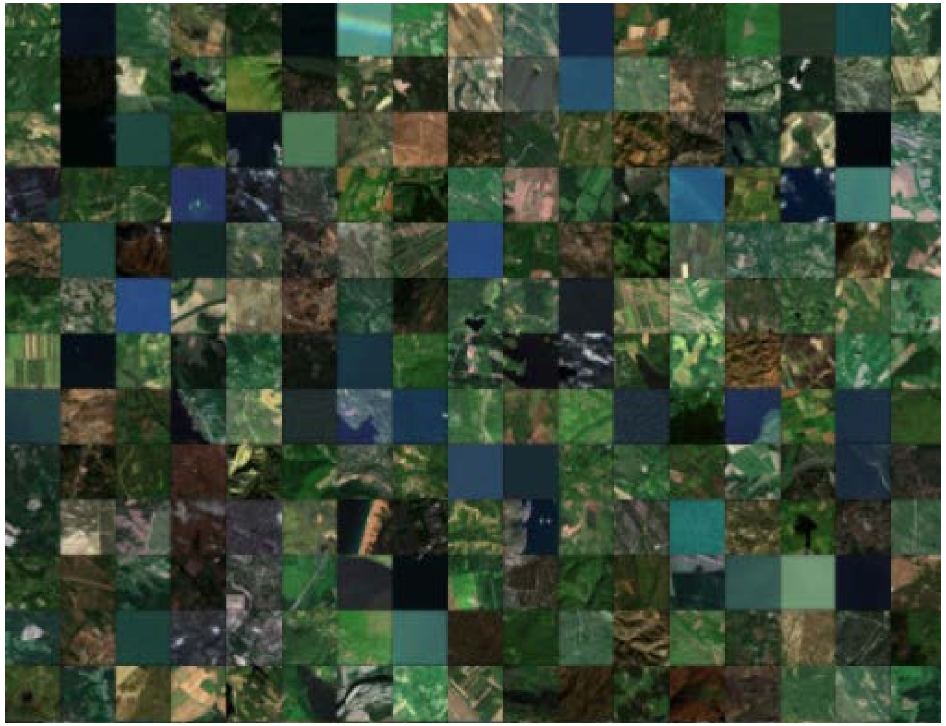
CNNs are DNNs that include computer vision ideas (convolutional filters) and can learn features from images



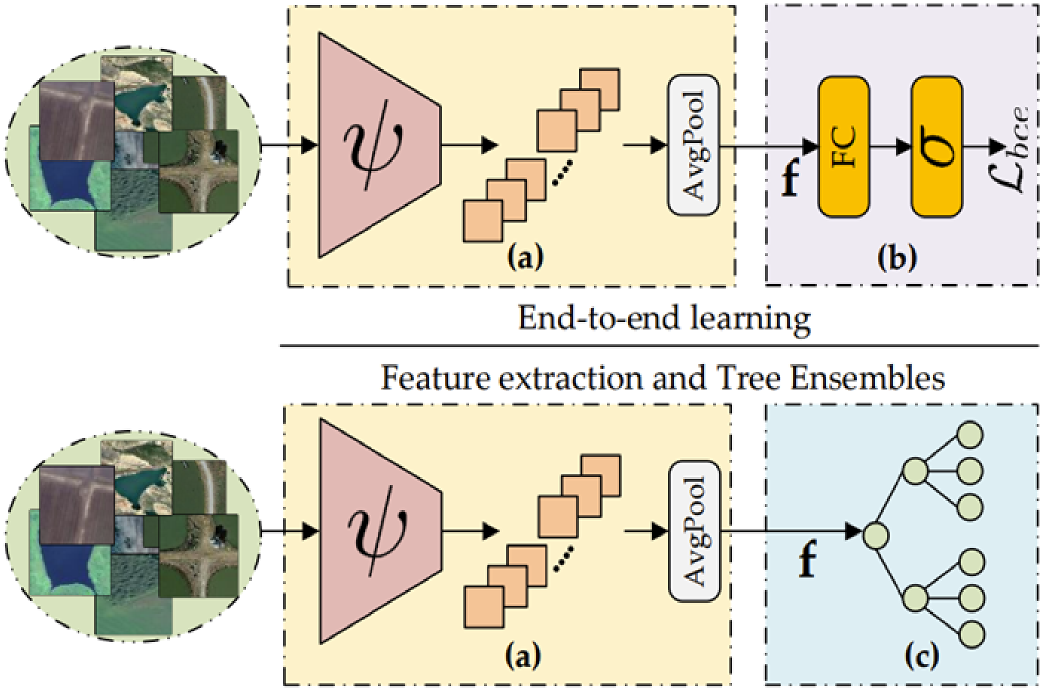
This is the key to the success of NNs: End-to-end learning

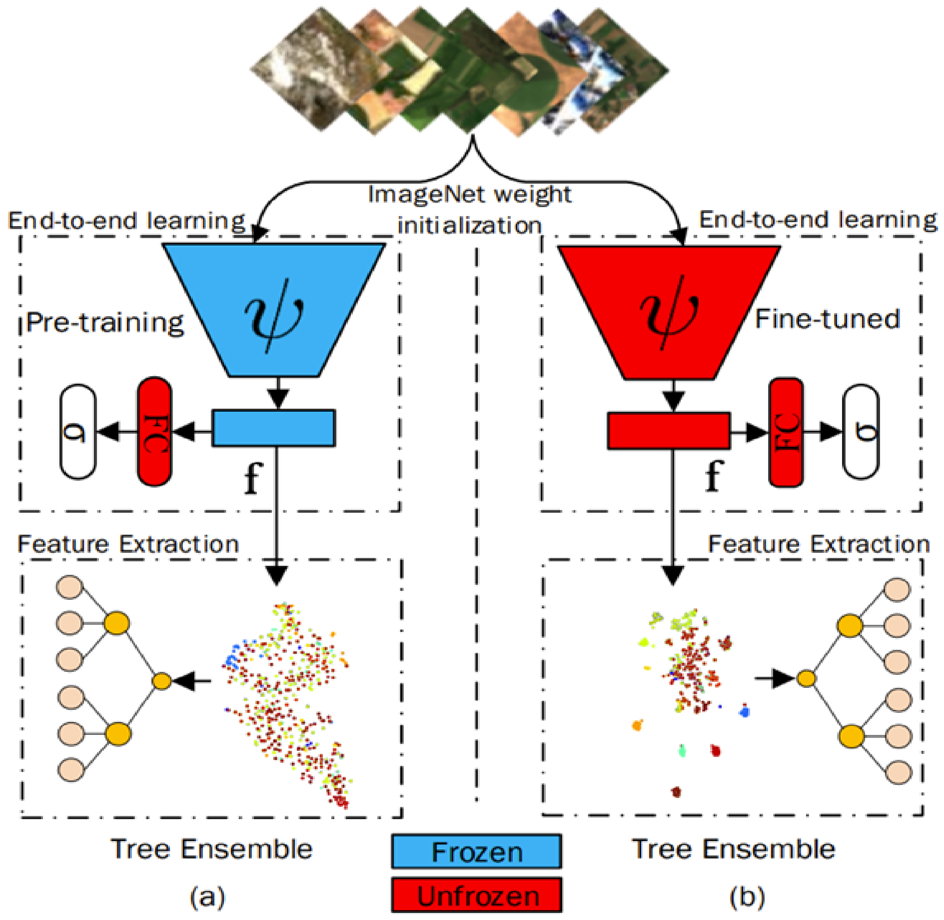


CONVOLUTIONAL NEURAL NETWORKS ARE GREAT FOR ANALYZING IMAGES (incl. remotely sensed/ satellite images)



MLC BY FEATURE EXTRACTION W DNNs + PCTs





MLC BY FEATURE EXTRACTION W DNNs + PCTs vs. End-To-End DNNs

Deep network architectures used:

- *VGGs (16 and 19)*
- *ResNets (34, 50, 152)*
- *EfficientNets (B0, B1, B2)*

Feature extraction:

- *With weights as pre-trained (ImageNet)*
- *With weights fine-tuned*

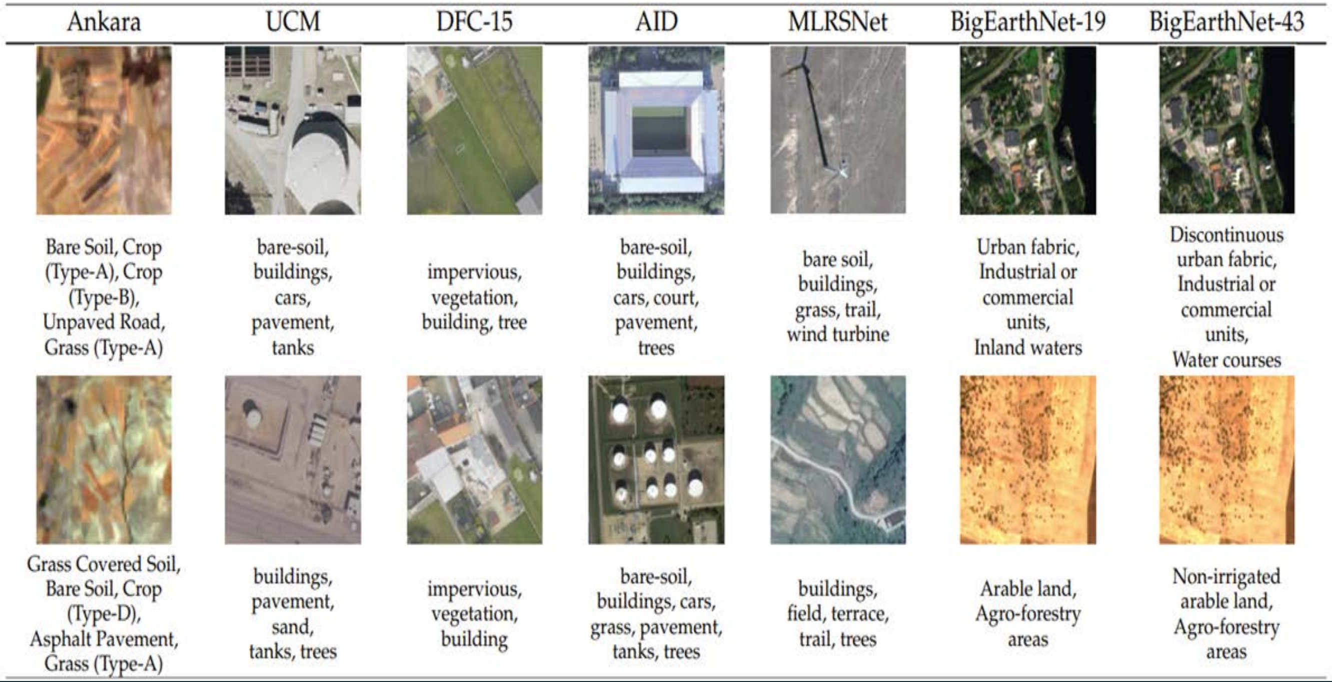
Evaluation measure: *Ranking Loss*

Learning methods:

- *End-to-end*
- *Random Forests / Extra Trees*

Datasets:

Dataset	Image Type	$\lvert\mathcal{L}\rvert$	*Card*	*Dens*	N	N_{train}	N_{test}	$w \times h$
Ankara	Hyperspectral/Aerial RGB	29	9.120	0.536	216	171	45	64 × 64
UC Merced Land Use	Aerial RGB	17	3.334	0.476	2100	1667	433	256 × 256
AID Multilabel	Aerial RGB	17	5.152	0.468	3000	2400	600	600 × 600
DFC-15 Multilabel	Aerial RGB	8	2.795	0.465	3341	2672	669	600 × 600
MLRSNet	Aerial RGB	60	5.770	0.144	109,151	87,325	21,826	256 × 256
BigEarthNet	Hyperspectral/Aerial RGB	19	2.900	0.263	590,326	472,245	118,081	256 × 256
BigEarthNet	Hyperspectral/Aerial RGB	43	2.965	0.247	590,326	472,245	118,081	256 × 256



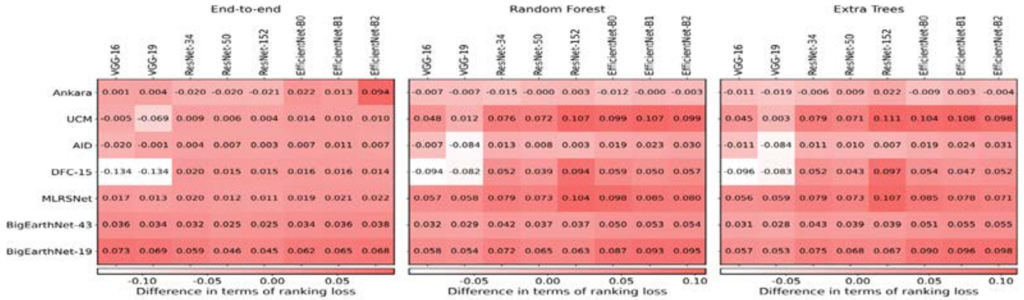
MLC BY FEATURE EXTRACTION W DNNs + PCTs vs. End-To-End DNNs

Conclusions of the experimental comparison

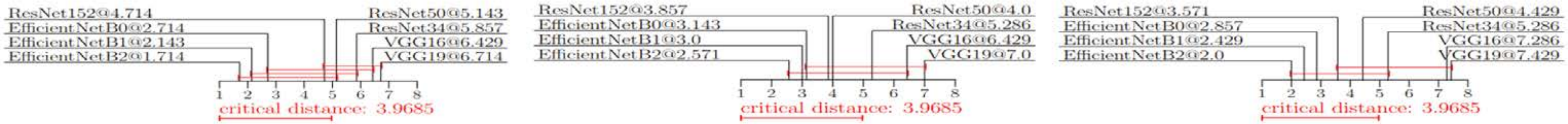
- ***Fine-tuned weights clearly lead to better performance across the board***

$$\mathbf{H} = \begin{bmatrix} \mathrm{rl}_{11} & \mathrm{rl}_{12} & \cdots & \mathrm{rl}_{1n} \\ \mathrm{rl}_{21} & \mathrm{rl}_{22} & \cdots & \mathrm{rl}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{rl}_{m1} & \mathrm{rl}_{m2} & \cdots & \mathrm{rl}_{mn} \end{bmatrix} \in \mathbb{R}^{m \times n}$$

$$\mathrm{rl}_{i,j} = \mathrm{rl}_{i,j}^{\text{fine-tune}} - \mathrm{rl}_{i,j}^{\text{pre-train}}$$



- ***The EfficientNet architecture clearly leads to best performance***



(a) End-to-end (b) Random forest (c) Extra trees

- ***Feature extraction + PCTs is comparable in performance to End-To-End***

(a) Fine-tuning (b) Pre-training

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, VOL. 62, 2024 4706416

Semi-Supervised Multi-Label Classification of Land Use/Land Cover in Remote Sensing Images With Predictive Clustering Trees and Ensembles

Marjan Stoimchev, Jurica Levatić, Dragi Kocev, and Sašo Džeroski

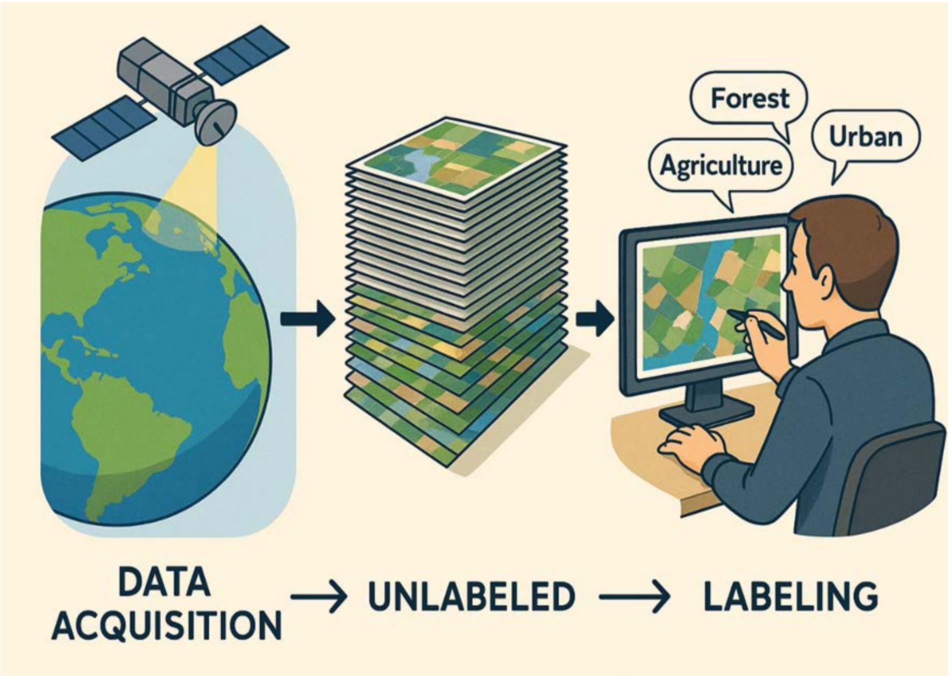
COMBINING DEEP NEURAL NETWORKS AND ENSEMBLES OF PREDICTIVE CLUSTERING TREES FOR SEMI-SUPERVISED LEARNING FROM RSI

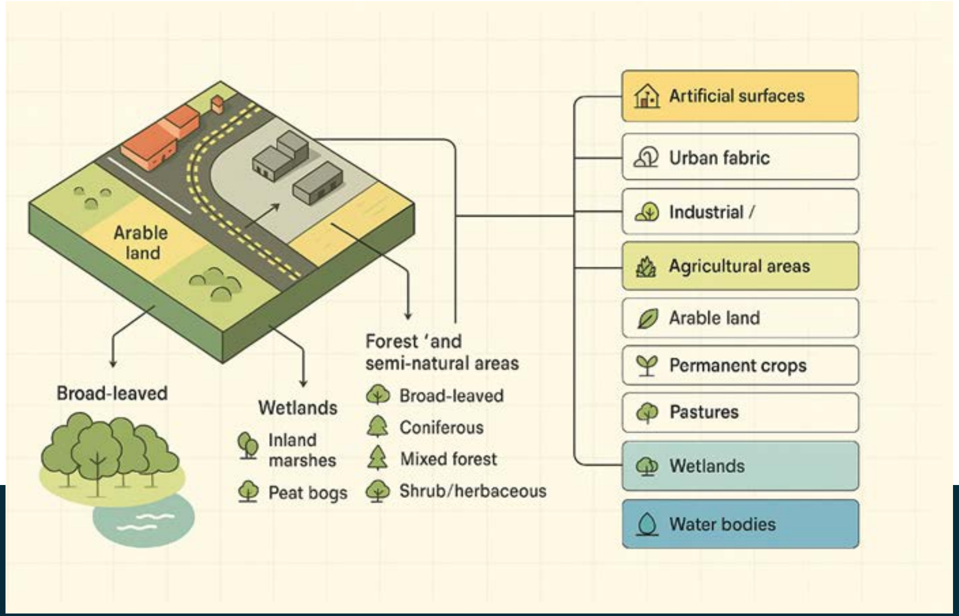
MOTIVATION OF SSL FOR MLC OF RSIs

Current Limitations

- Remote sensing images (RSIs) are **largely unlabeled**
- Manual annotations are **very costly**
- Especially when we have multi-label and hierarchical-multi-label classification

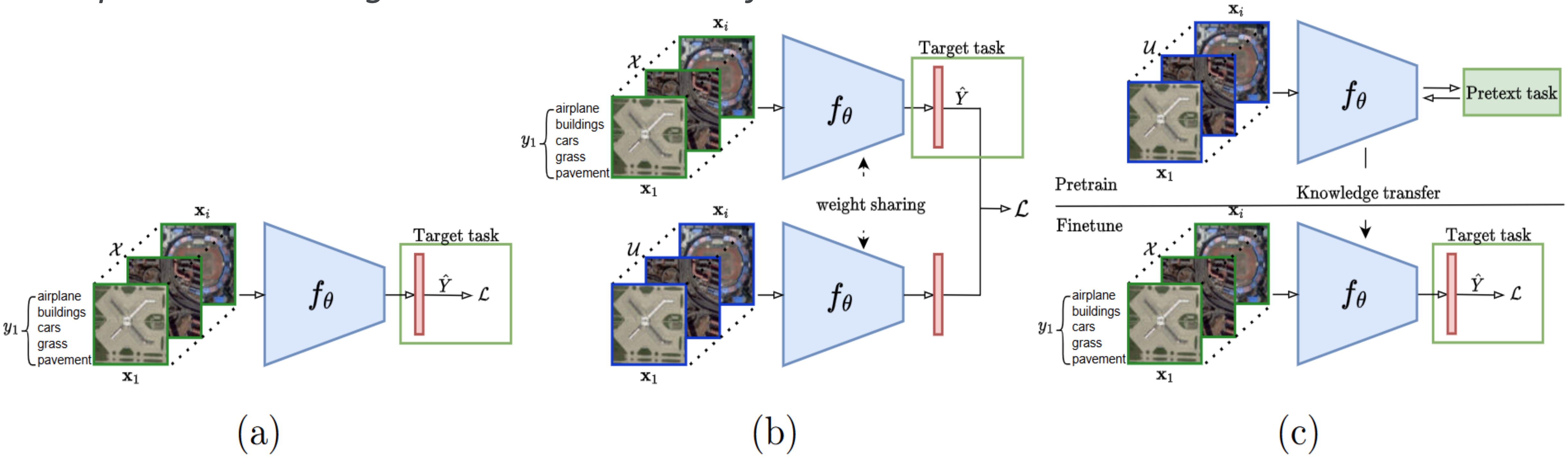
- Semi-supervised learning (SSL) has thrived in **single-label** classification as well as **multi-label** classification of **tabular data**
- Existing deep learning approaches for RSI don't really deal with MLC and especially HMLC
- In particular, they **rarely capture** dependencies within structured labels





DIFFERENT KINDS OF SUPERVISION IN LEARNING FROM RSIs

a) *(Fully) Supervised learning: All images are labeled*
b) *Semi-supervised learning: Labeled and unlabeled images used simultaneously*
c) *SSL via Self-SL: Unlabeled data used first for self-SL (unsupervised), supervised learning on labeled data then follows*



THE BEST OF BOTH WORLDS: CONVOLUTIONAL DNNs & PCT ENSEMBLES

Step 1: Self-supervised pre-training: Image reconstruction

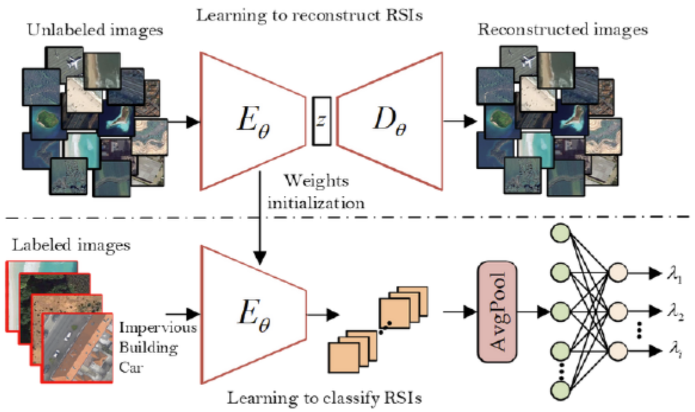
$$\mathcal{L}_{\text{mse}}(\theta, \phi; x) = \left\| x - D_\phi(E_\theta(x)) \right\|^2$$

Step 2: Supervised fine-tuning

$$\mathcal{L}_{\text{ce}} = -\frac{1}{n} \sum_{i=1}^{n} y_i \log(\hat{y}_i)$$

$$\mathcal{L}_{\text{bce}} = -\frac{1}{n} \sum_{i=1}^{n} \left[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right]$$

Convolutional Autoencoder approach



Step 1: Feature extraction

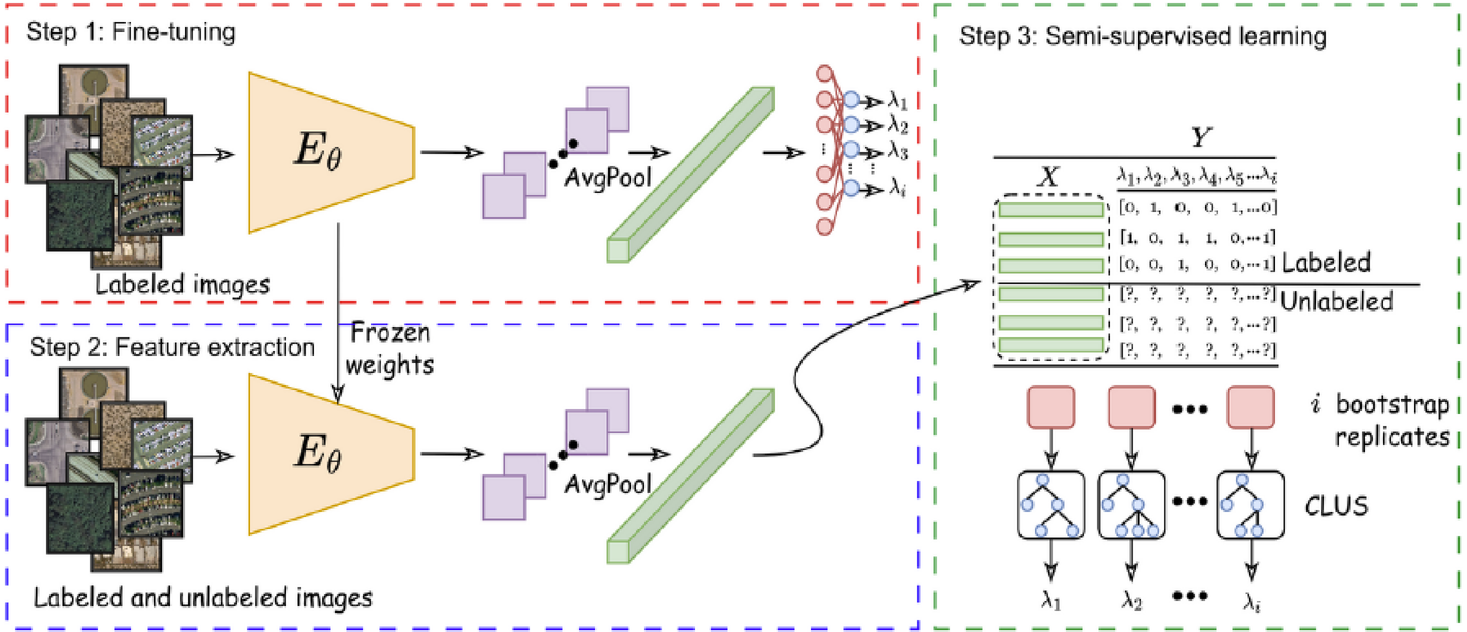
$$x \in \mathbb{R}^{w \times h \times 3}$$

$$\mathbf{f} = E(x; \theta) \in \mathbb{R}^d$$

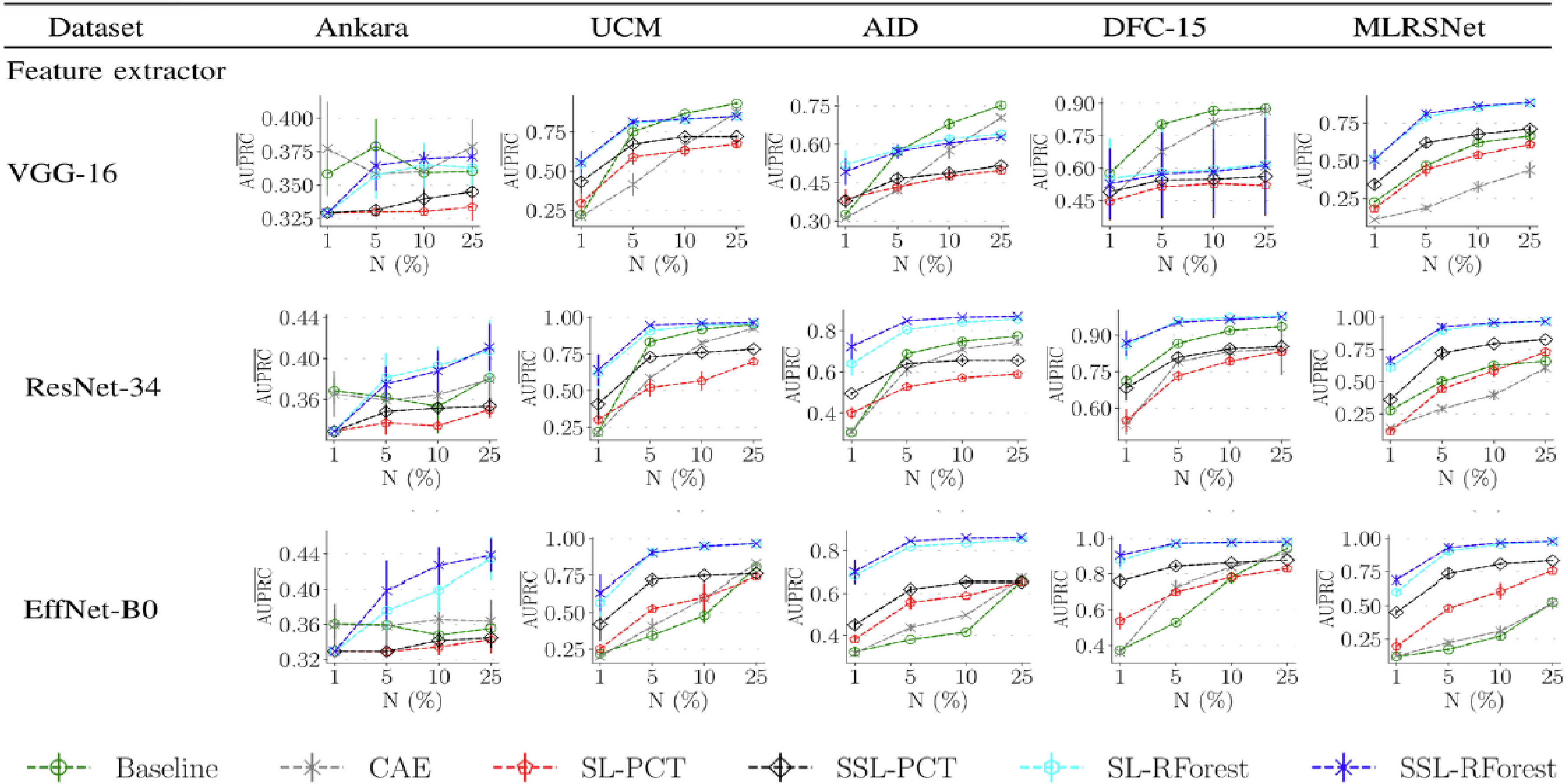
Step 2: Semi-Supervised PCTs and Ensembles

$$Var_f = w\,Var_f(Y) + (1 - w)\,Var_f(X), \quad w \in [0, 1]$$

Semi-Supervised PCTs and Ensembles



SSL vs. SL PERFORMANCE: FEATURE EXTRACTION vs End-to-end L. (MLC)



AGGREGATE RESULTS OF SL/SSL RANKS ACROSS 13 MLC METRICS FOR DIFFERENT DATASETS & DIFF. CNN ARCHITECTURES

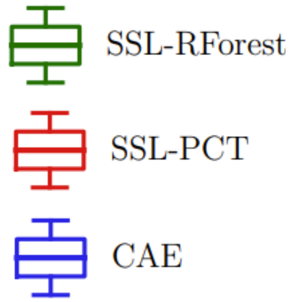
Baseline	CAE	SL-PCT	SSL-PCT	SL-RForest	SSL-RForest
VGG-16	VGG-16	VGG-16	VGG-16	VGG-16	VGG-16
VGG-19	VGG-19	VGG-19	VGG-19	VGG-19	VGG-19
ResNet34	ResNet34	ResNet34	ResNet34	ResNet34	ResNet34
ResNet50	ResNet50	ResNet50	ResNet50	ResNet50	ResNet50
ResNet152	ResNet152	ResNet152	ResNet152	ResNet152	ResNet152
EffNet-B0	EffNet-B0	EffNet-B0	EffNet-B0	EffNet-B0	EffNet-B0
EffNet-B1	EffNet-B1	EffNet-B1	EffNet-B1	EffNet-B1	EffNet-B1
EffNet-B2	EffNet-B2	EffNet-B2	EffNet-B2	EffNet-B2	EffNet-B2

Ranking

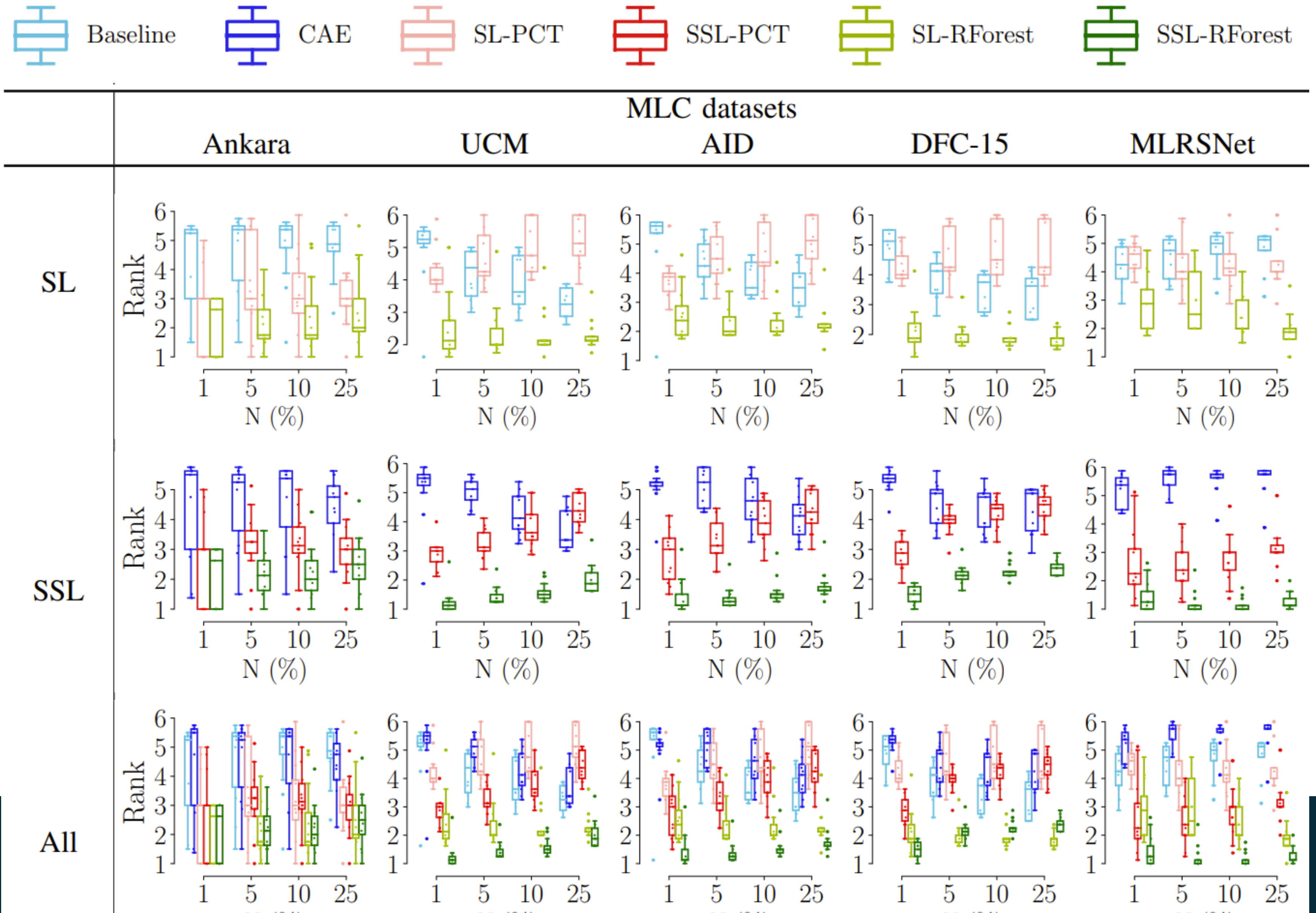
Different metrics

⋮

Repeated for different fractions of labeled data

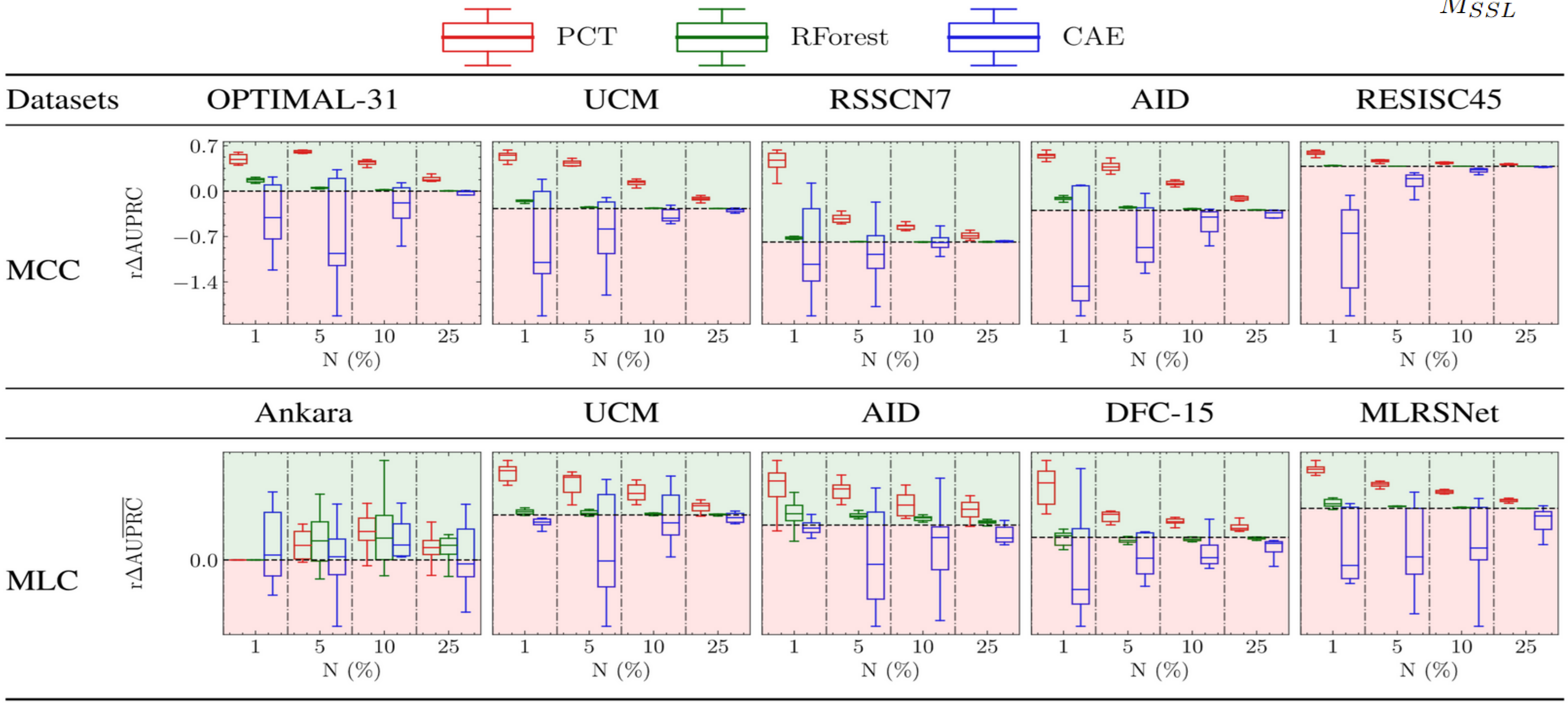


OVERALL RELATIVE PERFORMANCE (MLC)



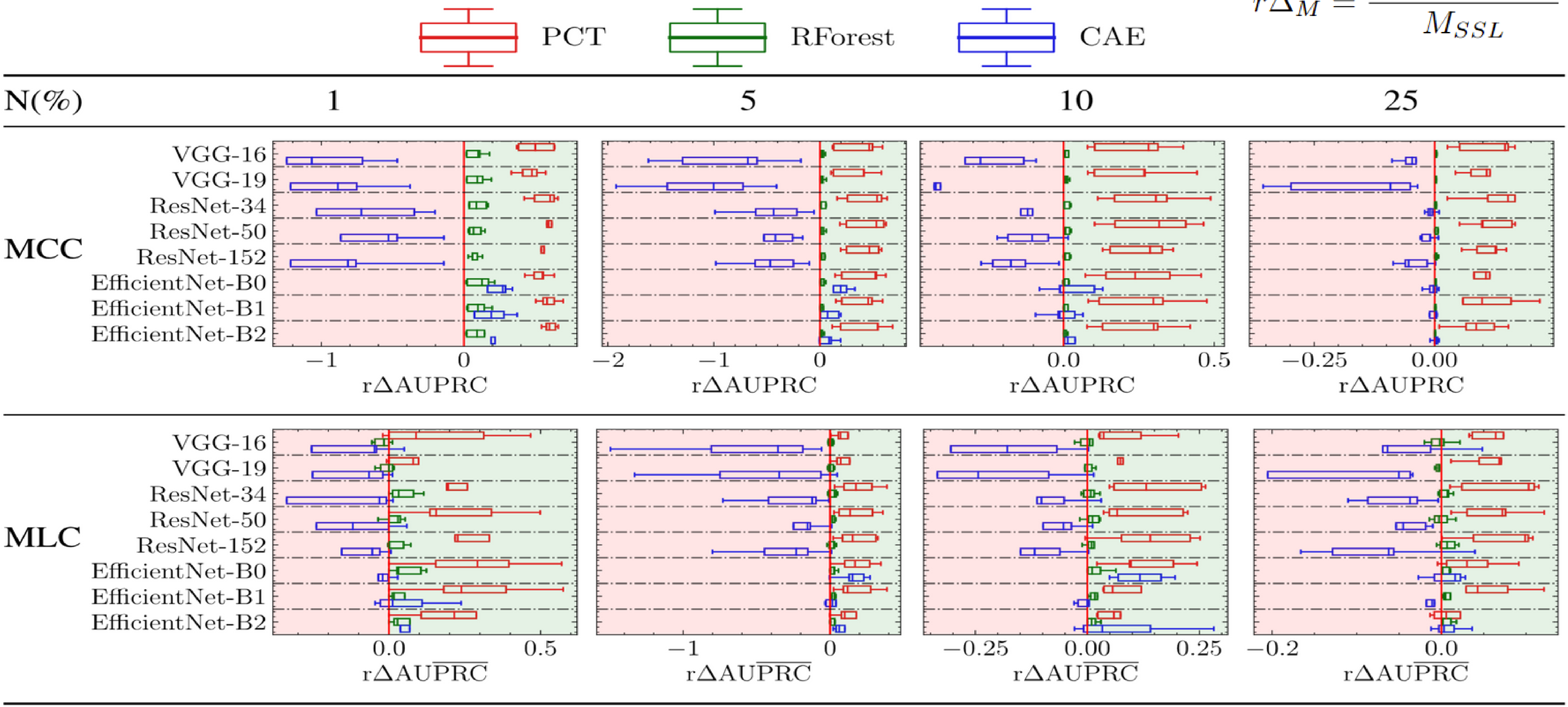
SSL vs. SL PERFORMANCE DIFFERENCE (per dataset)

$$r\Delta_M = \frac{M_{SSL} - M_{SL}}{M_{SSL}}$$



SSL vs. SL PERFORMANCE DIFFERENCE (per architecture)

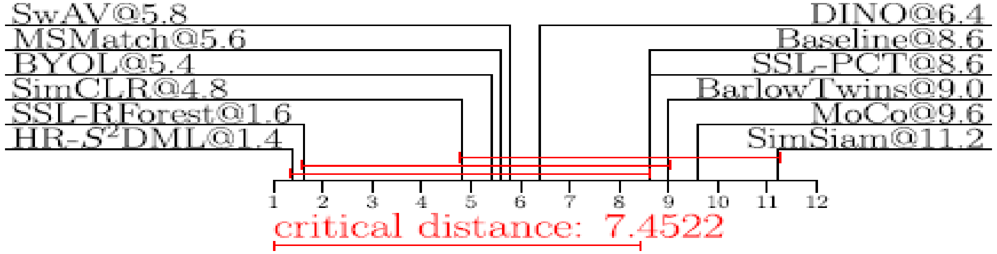
$$r\Delta_M = \frac{M_{SSL} - M_{SL}}{M_{SSL}}$$

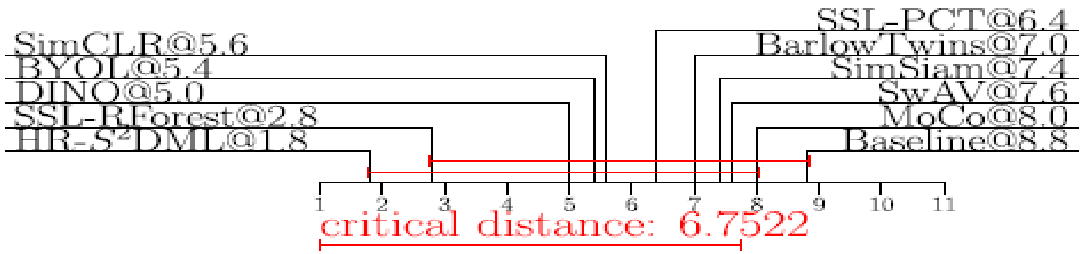


COMPARISON OF OUR SSL APPROACHES TO COMPETING ONES (MCC&MLC)

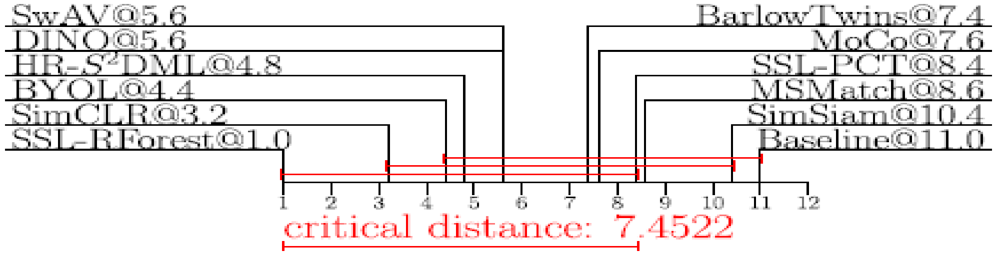
N(%) **MCC** **MLC**

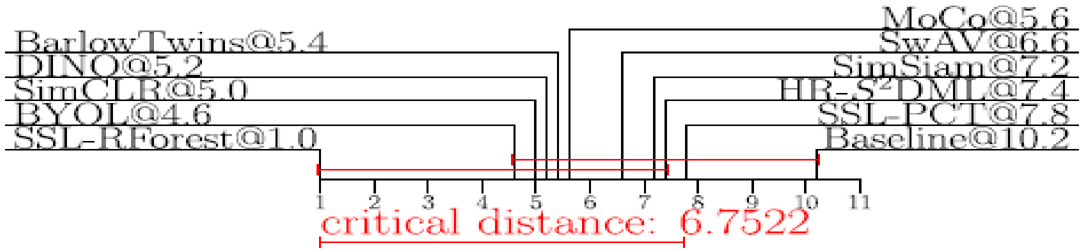
1



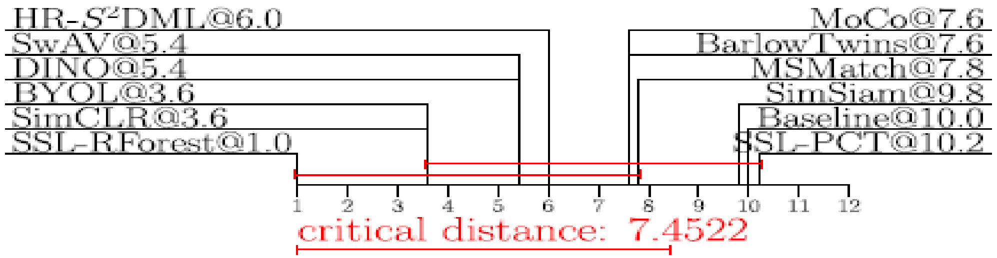


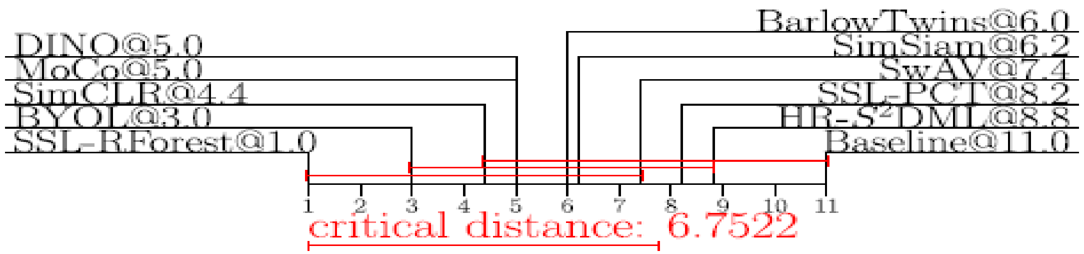
5



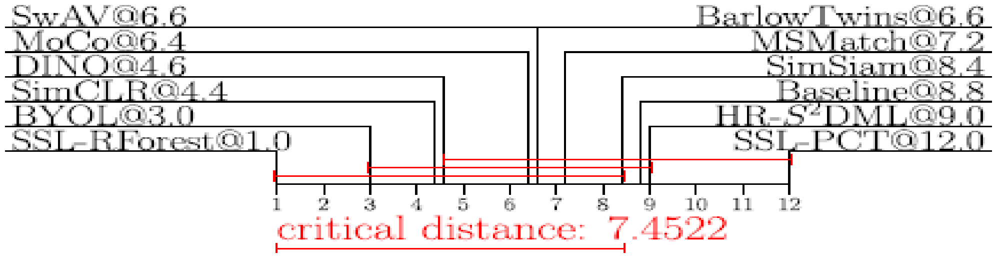


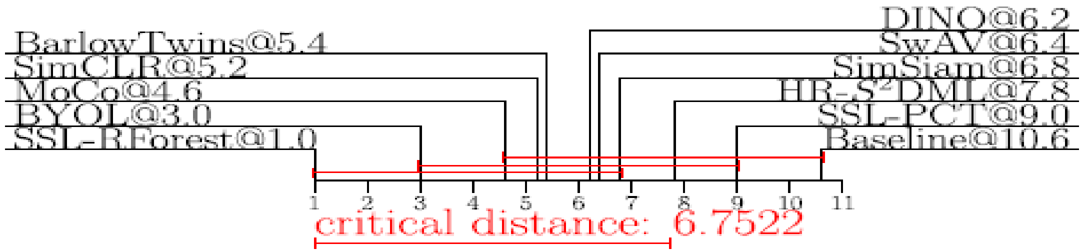
10





25

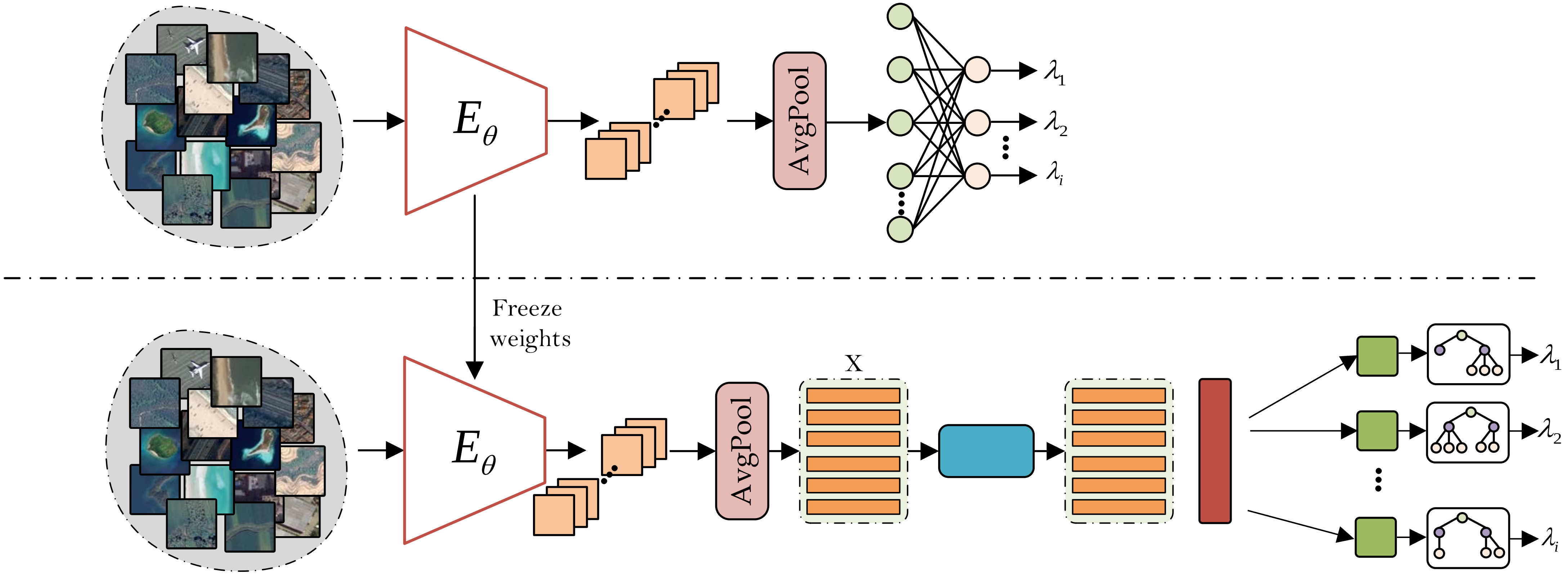




DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

IMPROVING THE EFFICIENCY OF SSL WITH ENSEMBLES OF PREDICTIVE CLUSTERING TREES WHEN LEARNING FROM RSI

INSTEAD OF USING EXTRACTED FEATURES WITH PCTs FOR SSL, USE PCA FIRST AND REDUCE THE NUMBER OF FEATURES (FOR EFFICIENCY)



DIMENSIONALITY REDUCTION FOR EFFICIENT SSL FROM RSI

- Datasets (MCC)
  - OPTIMAL-31
  - UCM
  - RSSCN7
  - AID
  - RESISC45

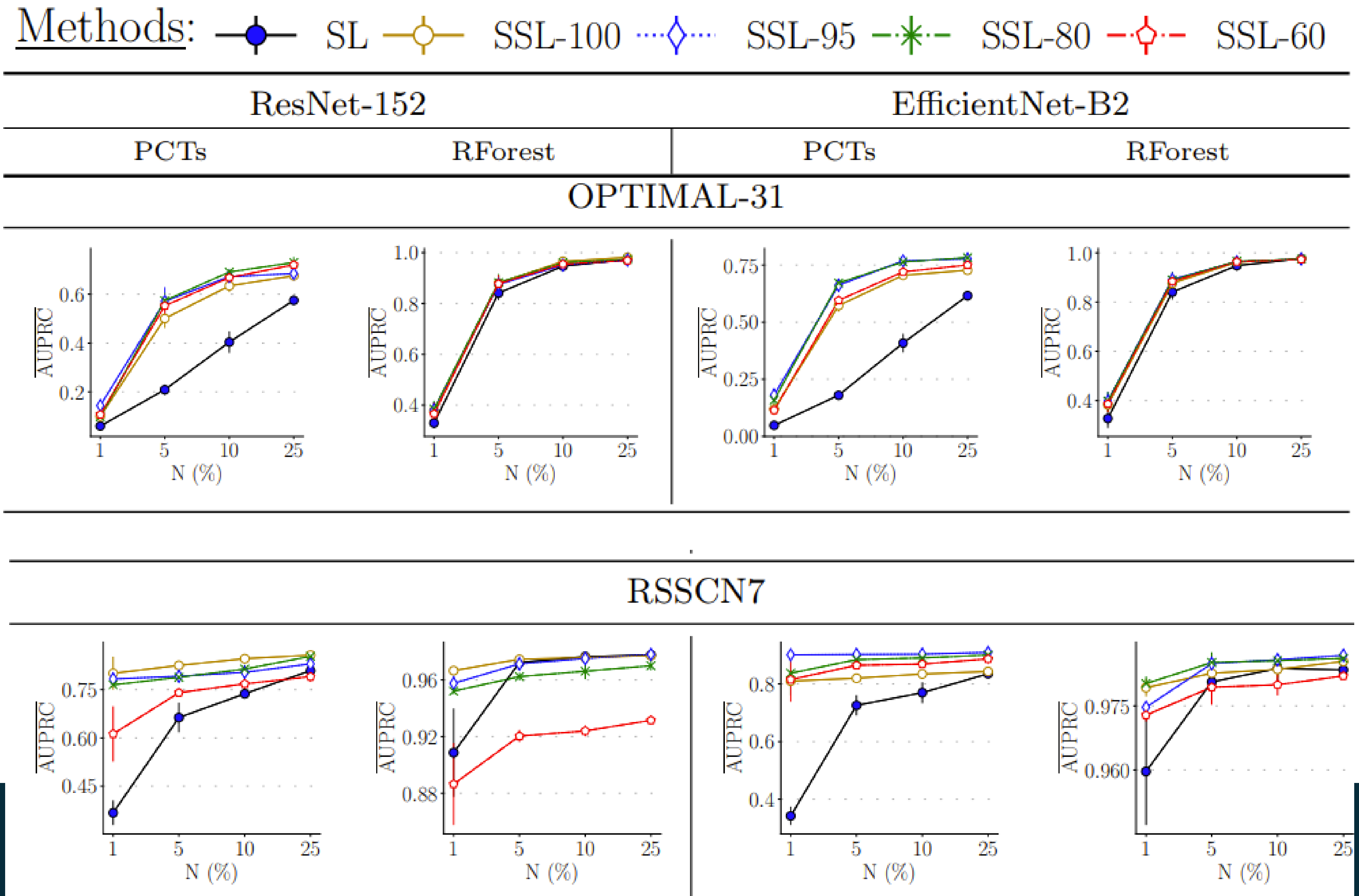
Dataset splits:
- 70% train, 10% validation, 20% test
- The splits are stratified
- Different percentages of (randomly sub-sampled) labeled data from the training sets, with the following amounts: 1%, 5%, 10%, and 25%

- For the feature extractors we use ResNet-152 and EfficientNet-B2
- Trained for 25 epochs
- Adam optimizer
- Batch size: 128
- Learning rate of 1e-4

PCA
- We 60%, 80%, and 95% of the variance explained by the principal components
- Baseline: 100% explained variance (whole feature space)

SSL-PCTs and ensembles for MCC
- M5 pruning
- 100 unpruned trees for supervised and semi-supervised tree ensembles
- We use 3-fold internal cross-validation on the training part of the dataset to optimize the w parameter

DIMENSIONALITY REDUCTION FOR EFFICIENT SSL FROM RSI



DIMENSIONALITY REDUCTION FOR EFFICIENT SSL FROM RSI

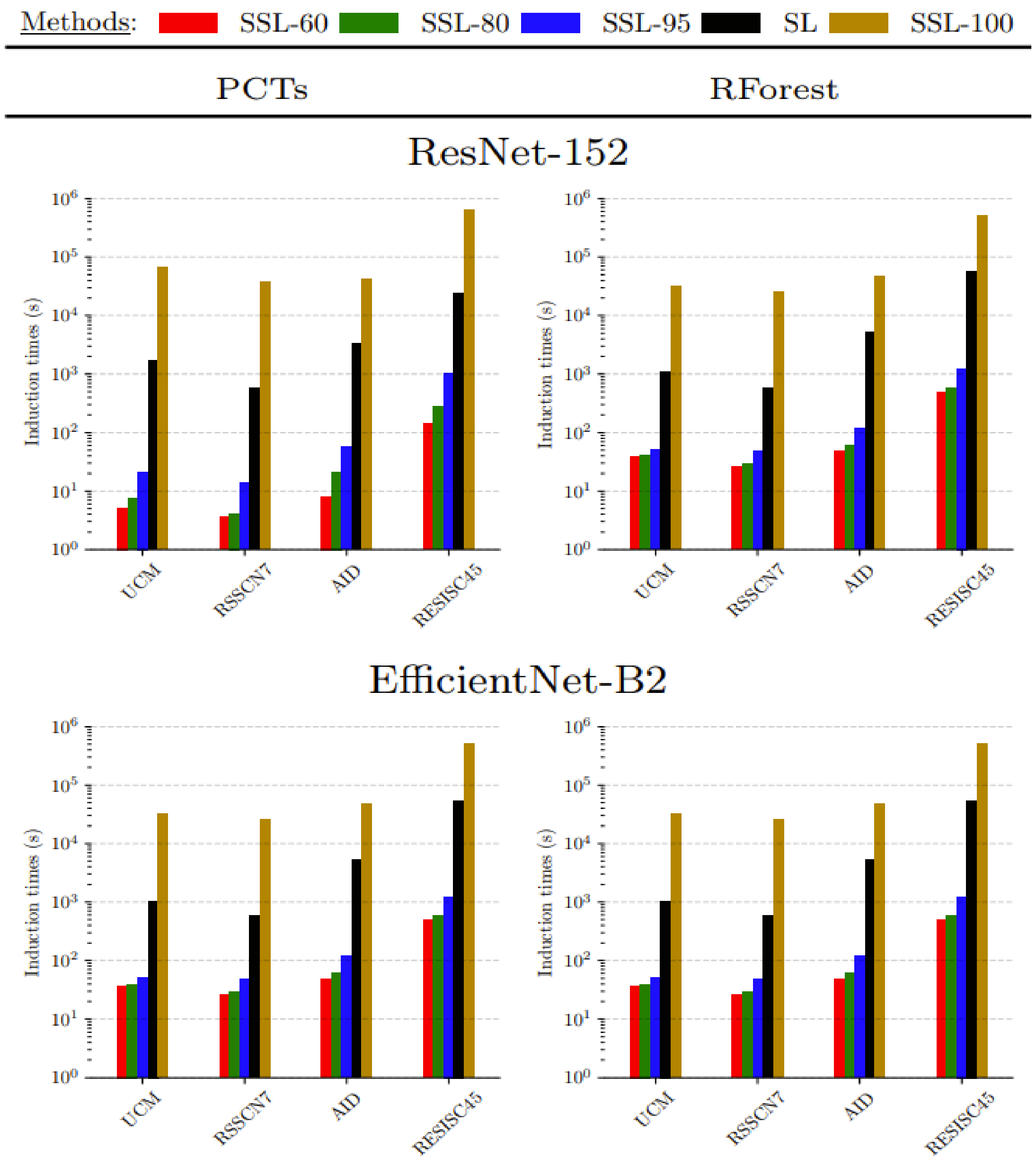


Table 1: The dimensions of the feature spaces before and after the application of PCA for each dataset and network architecture. EV is the cumulative explained variance in(%).

			EV(%)		
Dataset	Features extracted	Original	95	80	60
OPTIMAL-31	ResNet-152	2048	48	21	12
UCM		2048	159	19	12
RSSCN7		2048	163	12	5
AID		2048	54	20	12
RESISC45		2048	61	26	15
OPTIMAL-31	EfficientNet-B2	1408	192	27	16
UCM		1408	36	15	10
RSSCN7		1408	30	6	4
AID		1408	214	27	16
RESISC45		1408	416	47	24

Induction times in seconds for SL and SSL algorithms for 25% labeled data and different levels of dimensionality reduction

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

14882 IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, VOL. 18, 2025

SSL-MAE: Adaptive Semisupervised Learning Framework for Multilabel Classification of Remote Sensing Images Using Masked Autoencoders

Marjan Stoimchev, Jurica Levatić, Dragi Kocev, and Sašo Džeroski

ADAPTING THE LOSS FUNCTION WITHIN END-TO-END APPROACHES FOR SSL FROM RSI

ADAPT THE LOSS FUNCTION IN SS End-to-end LEARNING with MAEs

1. Image masking

$$x_p^l, x_p^u \in \mathbb{R}^{N_p \times (C \times P^2)}$$

2. Encoder

$$\tilde{z} = E(x, \theta) : \mathbb{R}^{2N_p \times (C \times P^2)} \longrightarrow \mathbb{R}^{2N_p \times d}$$

3. Classification Head

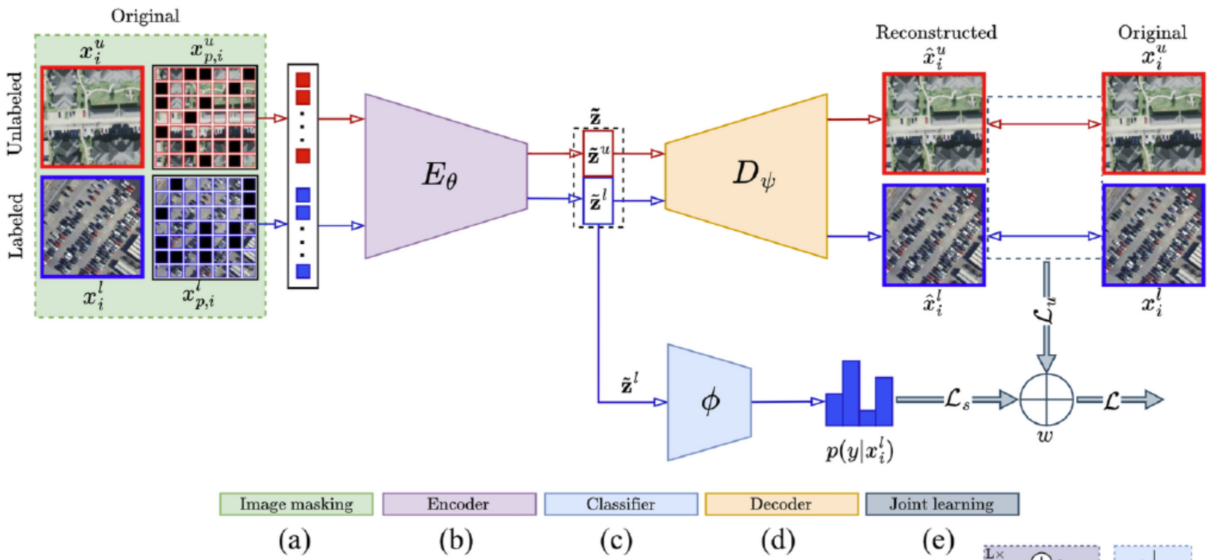
$$\mathbf{f}_v = \frac{1}{N_p^l} \sum_{i=1}^{N_p^l} \tilde{z}_v^{l(i)} \in \mathbb{R}^d$$

$$\mathcal{L}_s = \frac{1}{B_l} \sum_{i=1}^{B_l} \mathcal{H}\left(y_i, p(y \mid x_i^l)\right)$$

4. Decoder

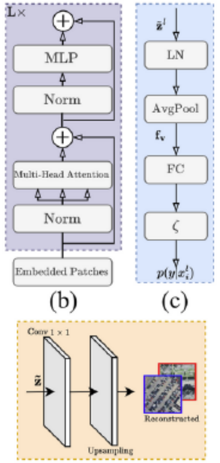
$$\hat{x} = D(\tilde{z}, \psi) \in \mathbb{R}^{C \times W \times H}$$

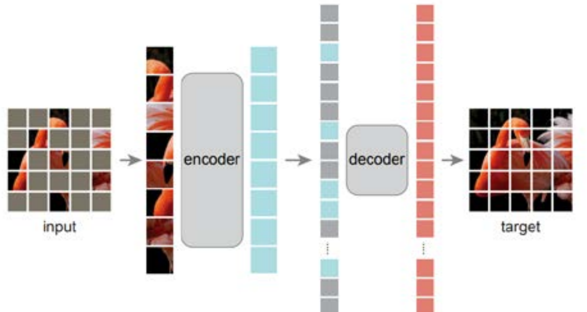
$$\mathcal{L}_u = \frac{1}{B} \sum_{i=1}^{B} \left(\frac{1}{\Omega(x_i^M)} \sum_{m \in M_i} \|\hat{x}_{i,m} - x_{i,m}\|_1 \right)$$



5. Adaptive Joint Learning

$$\mathcal{L} = w\mathcal{L}_s + (1 - w)\mathcal{L}_u$$





Pretext task: Masked image modeling (with masked auto-encoders)

EXPERIMENTAL COMPARISON: DESIGN

Experimental Setup

- Inductive learning setting
- Different percentages of (randomly sub-sampled) labeled data fractions: (1%, 5%, 10%, and 25%)
- Five repetitions (seeds)
- Learning curves

Datasets

MCC datasets:

- OPTIMAL-31, UCM, AID, RSSCN7, RESISC45

MLC datasets:

- UCM, AID, DFC-15, MLRSNet, BigEarthNet-43

Optimization of w :

- Grid search (GS)
- Learnable weight:

$$\mathcal{L}(\theta, \theta) = \sigma(\theta)\mathcal{L}_s(\theta) + (1 - \sigma(\theta))\mathcal{L}_u(\theta)$$

$$\theta \leftarrow \theta - \eta \frac{\partial \mathcal{L}}{\partial \theta}, \quad \theta \leftarrow \theta - \eta \frac{\partial \mathcal{L}}{\partial \theta}$$

$$\frac{\partial \mathcal{L}}{\partial \theta} = \sigma(\theta)(1 - \sigma(\theta))[\mathcal{L}_s(\theta) - \mathcal{L}_u(\theta)]$$

Comparison against state-of-the-art methods

SSL methods:

- SoftMatch, FixMatch, FreeMatch, SimMatch, FlexMatch, CoMatch

Self-SL methods:

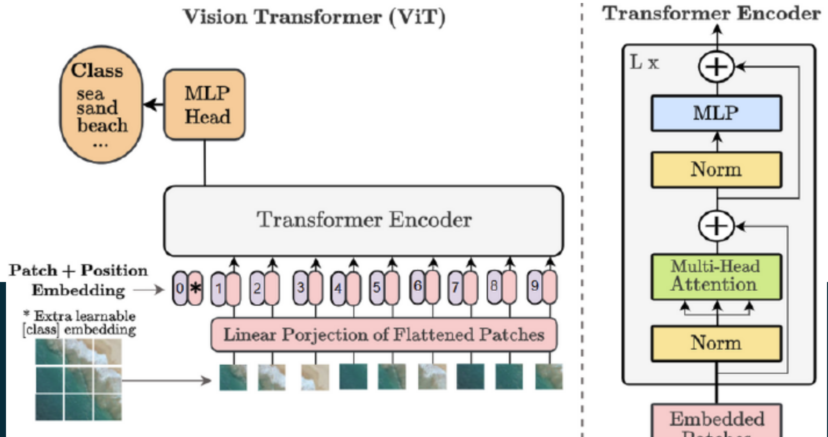
- SimMIM, SimCLR, DINO, VICReg, MoCo, BYOL, Tri-ReD

Supervised Baseline method:

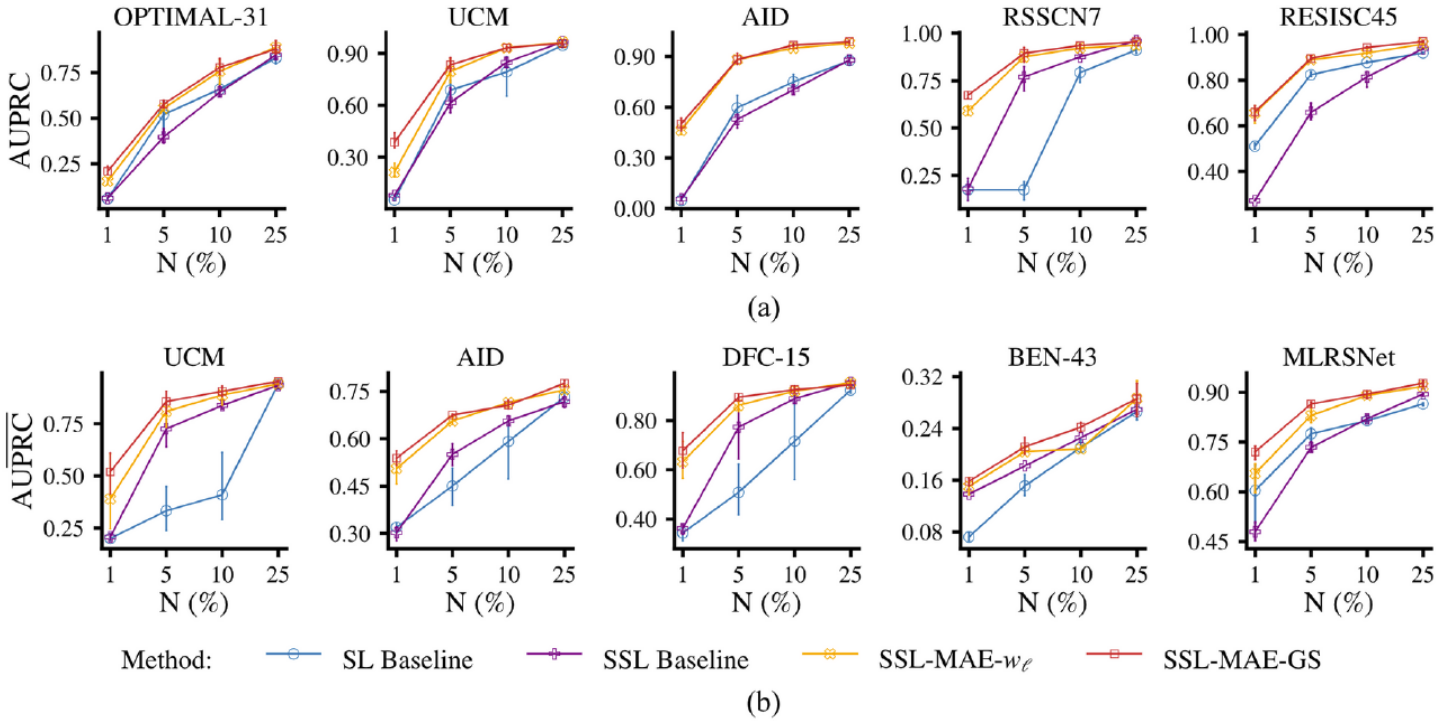
$$\mathcal{L} = w\mathcal{L}_s + (\cancel{1 - w})\mathcal{L}_u$$

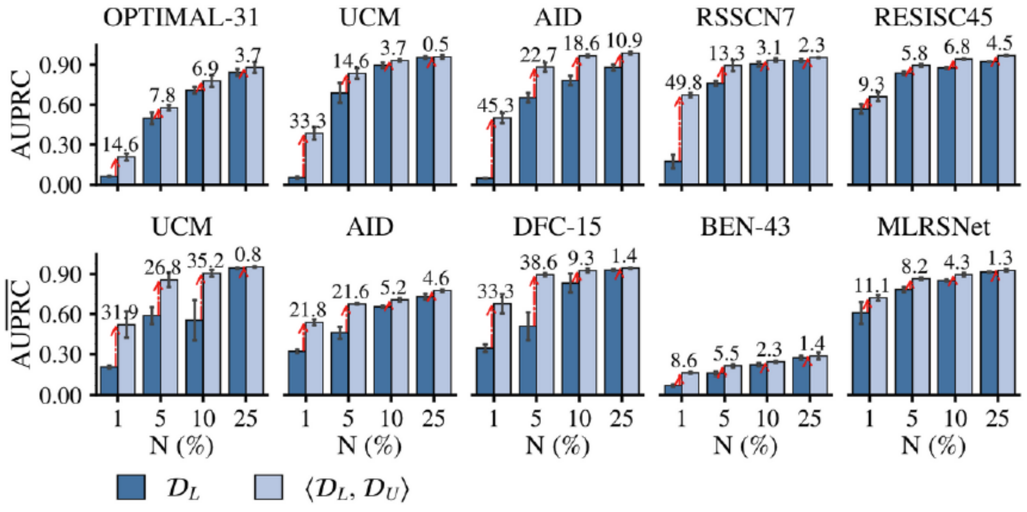
Classification tasks

- Multi-Class Classification (MCC)
- Multi-Label Classification (MLC)



EXPERIMENTAL COMPARISON: RESULTS





Performance of SL & SSL methods on MCC (top) and MLC (bottom) dataset

AUPRC and AU average PRC as metrics for MCC and MLC

Absolute performance (left) and performance improvement for SSL vs SL for MAE-GS

Comparison to competing SSL and SelfSL approaches

MCC Datasets	OPTIMAL-31 N (%)				UCM N (%)				AID N (%)				RSSCN7 N (%)				RESISC45 N (%)				Avg. ranks N (%)			
Methods	1	5	10	25	1	5	10	25	1	5	10	25	1	5	10	25	1	5	10	25	1	5	10	25
SL Baseline	0.062	0.450	0.610	0.889	0.053	0.688	0.793	0.966	0.048	0.814	0.903	0.988	0.175	0.743	0.881	0.907	0.027	0.892	0.944	0.969	9.0	5.0	5.1	3.0
SoftMatch	0.092	0.274	0.419	0.512	0.147	0.549	0.806	0.944	0.091	0.309	0.625	0.720	0.340	0.699	0.848	0.920	0.217	0.514	0.632	0.868	7.2	8.1	7.7	7.6
FixMatch	0.097	0.331	0.628	0.846	0.142	0.620	0.851	0.955	0.081	0.347	0.678	0.916	0.356	0.717	0.874	0.942	0.300	0.821	0.895	0.948	6.6	6.4	5.6	4.7
FreeMatch	0.099	0.251	0.372	0.529	0.166	0.535	0.755	0.946	0.105	0.307	0.630	0.844	0.346	0.748	0.833	0.909	0.212	0.522	0.634	0.842	6.2	7.4	8.2	7.4
SimMatch	0.093	0.417	0.630	0.636	0.207	0.664	0.871	0.954	0.125	0.503	0.787	0.921	0.359	0.746	0.879	0.932	0.492	0.767	0.907	0.551	5.0	5.2	4.6	6.3
FlexMatch	0.102	0.339	0.419	0.498	0.228	0.572	0.812	0.955	0.122	0.392	0.652	0.816	0.395	0.747	0.859	0.932	0.192	0.514	0.629	0.853	4.8	6.3	7.1	6.8
CoMatch	0.220	0.505	0.739	0.868	0.285	0.696	0.904	0.941	0.193	0.639	0.826	0.934	0.521	0.788	0.888	0.952	0.587	0.885	0.918	0.954	2.4	3.4	3.4	4.6
SSL-MAE-w_ℓ	0.153	0.555	0.757	0.889	0.212	0.797	0.933	0.966	0.463	0.881	0.950	0.979	0.590	0.876	0.922	0.937	0.656	0.889	0.919	0.958	2.6	2.2	2.1	2.6
SSL-MAE-GS	0.208	0.577	0.778	0.879	0.386	0.835	0.933	0.960	0.501	0.882	0.968	0.988	0.673	0.894	0.936	0.955	0.661	0.895	0.944	0.969	1.2	1.0	1.2	2.0

MCC Datasets	OPTIMAL-31 N (%)				UCM N (%)				AID N (%)				RSSCN7 N (%)				RESISC45 N (%)				Avg. ranks N (%)			
Methods	1	5	10	25	1	5	10	25	1	5	10	25	1	5	10	25	1	5	10	25	1	5	10	25
SL Baseline	0.062	0.450	0.610	0.889	0.053	0.688	0.793	0.966	0.048	0.814	0.903	0.988	0.175	0.743	0.881	0.907	0.027	0.892	0.944	0.969	10.0	5.4	6.0	3.8
SimMIM	0.061	0.397	0.642	0.847	0.078	0.616	0.849	0.967	0.056	0.527	0.703	0.880	0.179	0.769	0.876	0.960	0.272	0.659	0.814	0.938	10.0	9.9	9.0	8.4
SimCLR	0.063	0.060	0.806	0.893	0.066	0.075	0.936	0.976	0.063	0.800	0.897	0.943	0.231	0.882	0.939	0.961	0.781	0.918	0.948	0.968	9.8	8.8	7.4	6.2
DINO	0.068	0.073	0.798	0.911	0.084	0.070	0.062	0.687	0.055	0.745	0.870	0.926	0.210	0.898	0.929	0.969	0.792	0.922	0.946	0.968	4.4	4.8	6.2	5.6
VICReg	0.056	0.060	0.799	0.895	0.078	0.072	0.918	0.971	0.064	0.710	0.822	0.918	0.171	0.878	0.924	0.958	0.773	0.920	0.947	0.971	5.7	6.3	5.6	5.7
MoCo	0.049	0.055	0.840	0.928	0.061	0.063	0.944	0.985	0.056	0.847	0.919	0.965	0.166	0.911	0.939	0.969	0.860	0.944	0.961	0.974	6.4	4.7	2.2	2.4
BYOL	0.061	0.071	0.853	0.929	0.056	0.069	0.952	0.987	0.056	0.829	0.915	0.954	0.231	0.905	0.957	0.974	0.828	0.944	0.961	0.975	5.2	4.1	1.7	1.9
Tri-ReD	0.197	0.551	0.769	0.863	0.369	0.843	0.918	0.962	0.374	0.817	0.912	0.968	0.693	0.920	0.940	0.947	0.239	0.655	0.834	0.930	3.2	7.0	8.3	9.0
SSL-MAE-w_ℓ	0.153	0.555	0.757	0.889	0.212	0.797	0.933	0.965	0.463	0.881	0.950	0.979	0.590	0.876	0.922	0.937	0.656	0.889	0.919	0.958	3.0	4.2	5.7	6.5
SSL-MAE-GS	0.208	0.577	0.778	0.879	0.386	0.835	0.933	0.960	0.501	0.882	0.968	0.988	0.673	0.894	0.936	0.955	0.661	0.895	0.944	0.969	2.0	2.6	4.4	5.8

MLC Datasets	UCM N (%)				AID N (%)				DFC-15 N (%)				BEN-43 N (%)				MLRSNet N (%)				Avg. ranks N (%)			
Methods	1	5	10	25	1	5	10	25	1	5	10	25	1	5	10	25	1	5	10	25	1	5	10	25
SL Baseline	0.201	0.333	0.410	0.944	0.320	0.450	0.592	0.747	0.344	0.760	0.873	0.919	0.072	0.151	0.211	0.266	0.587	0.833	0.888	0.923	7.8	7.6	6.6	8.0
SimMIM	0.208	0.726	0.838	0.934	0.303	0.550	0.657	0.717	0.362	0.773	0.889	0.954	0.138	0.182	0.226	0.270	0.479	0.734	0.820	0.895	7.3	7.2	6.0	8.2
SimCLR	0.220	0.231	0.926	0.951	0.330	0.680	0.719	0.767	0.376	0.887	0.924	0.958	0.078	0.208	0.244	0.297	0.775	0.894	0.920	0.941	4.6	4.2	2.9	4.1
DINO	0.258	0.233	0.665	0.959	0.307	0.666	0.738	0.774	0.385	0.880	0.923	0.955	0.081	0.212	0.238	0.298	0.733	0.892	0.914	0.942	4.4	4.4	5.3	3.4
VICReg	0.228	0.221	0.898	0.948	0.303	0.635	0.704	0.758	0.401	0.878	0.923	0.957	0.077	0.206	0.241	0.303	0.757	0.887	0.918	0.944	4.9	6.4	7.0	3.8
MoCo	0.256	0.261	0.926	0.955	0.304	0.701	0.757	0.786	0.382	0.883	0.931	0.955	0.075	0.209	0.241	0.294	0.746	0.898	0.921	0.946	5.6	3.0	2.0	3.1
BYOL	0.222	0.239	0.927	0.960	0.319	0.692	0.751	0.787	0.391	0.898	0.930	0.958	0.076	0.224	0.251	0.317	0.692	0.878	0.906	0.942	5.6	3.0	2.2	1.6
Tri-ReD	0.516	0.851	0.903	0.948	0.526	0.662	0.703	0.757	0.673	0.789	0.855	0.880	0.148	0.201	0.205	0.275	0.416	0.748	0.843	0.918	3.0	2.1	3.6	2.4
SSL-MAE-w_ℓ	0.390	0.809	0.888	0.941	0.506	0.656	0.714	0.755	0.633	0.861	0.919	0.953	0.150	0.204	0.208	0.286	0.655	0.830	0.891	0.918	3.0	6.0	6.8	7.3
SSL-MAE-GS	0.520	0.857	0.905	0.953	0.538	0.675	0.706	0.775	0.677	0.895	0.925	0.944	0.158	0.211	0.242	0.286	0.720	0.865	0.895	0.927	1.8	3.2	4.4	5.5

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Fully- and Semi-Supervised Hierarchical Multi-Label Image Classification with Graph Learning

Marjan Stoimchev, Boshko Koloski, Jurica Levatić, Dragi Kocev, and Sašo Džeroski

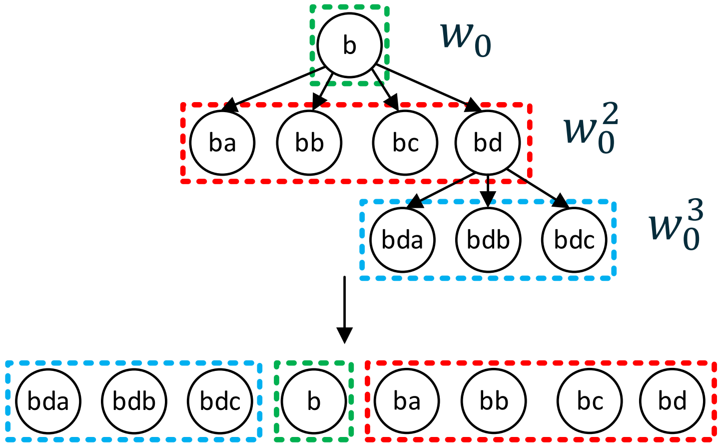
HMLC OF RSI BY COMBINING VISION TRANSFORMERS AND GRAPH NEURAL NETWORKS

INITIAL APPROACH: HIERARCHICAL MULTI-LABEL CLASSIFICATION OF RSI BY ADAPTING THE LOSS FUNCTION IN End-to-end LEARNING

- The classifier head contains number of neurons equal to the number of leaves plus the intermediate nodes in the hierarchy

$$\mathcal{L}_{bce} = -\frac{1}{N}\sum_{i=1}^{N}[y_i log(\hat{y_i}) + (1-y_i)log(1-\hat{y_i})]$$

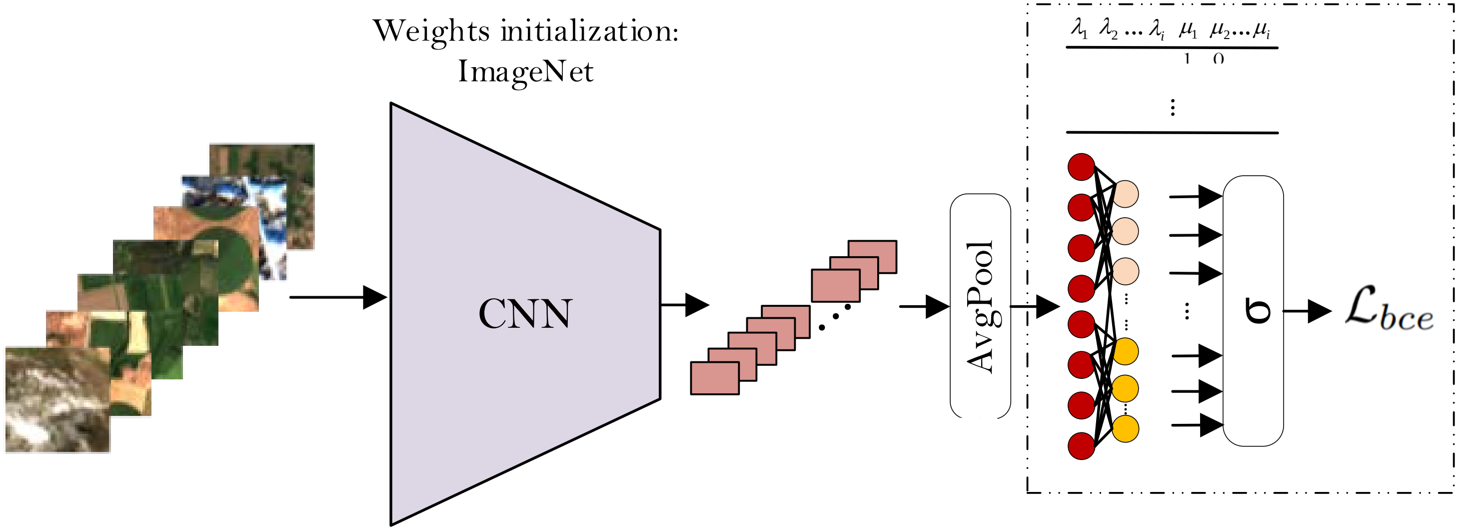
$$\mathcal{L}_{wbce} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{c} w_j \cdot [y_i log(\hat{y_i}) + (1-y_i)log(1-\hat{y_i})]$$



$w_0 \in (0,1]$

$w(node) = w_0 w(node \cdot parent)$

$w_0 = 1$: MLC



HMLC of RSIs: The BigEarthNet dataset

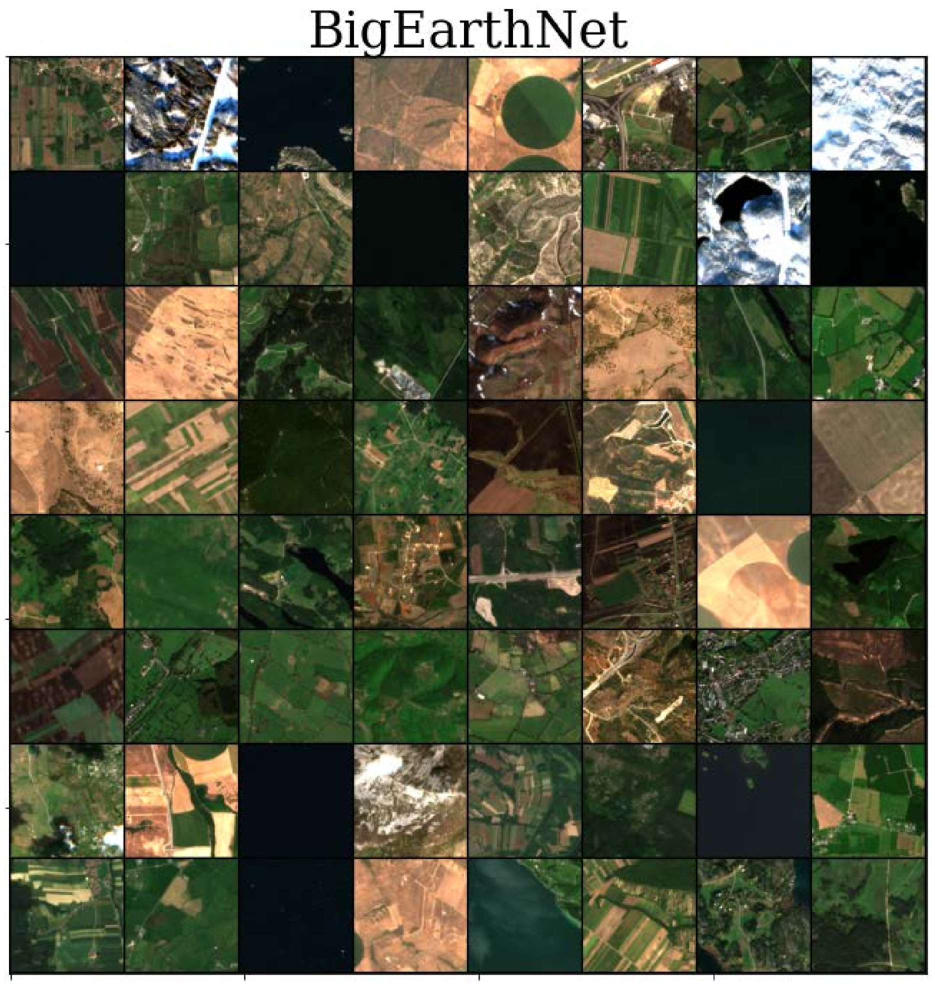
- BigEarthNet

  -590,326 non-overlapping image patches

- We used two Corine Land Cover (CLC) nomenclatures for the hierarchy:

  - Original CLC with 43 labels -> all hierarchical labels: 63
  - Reduced CLC with 19 labels -> all hierarchical labels: 27

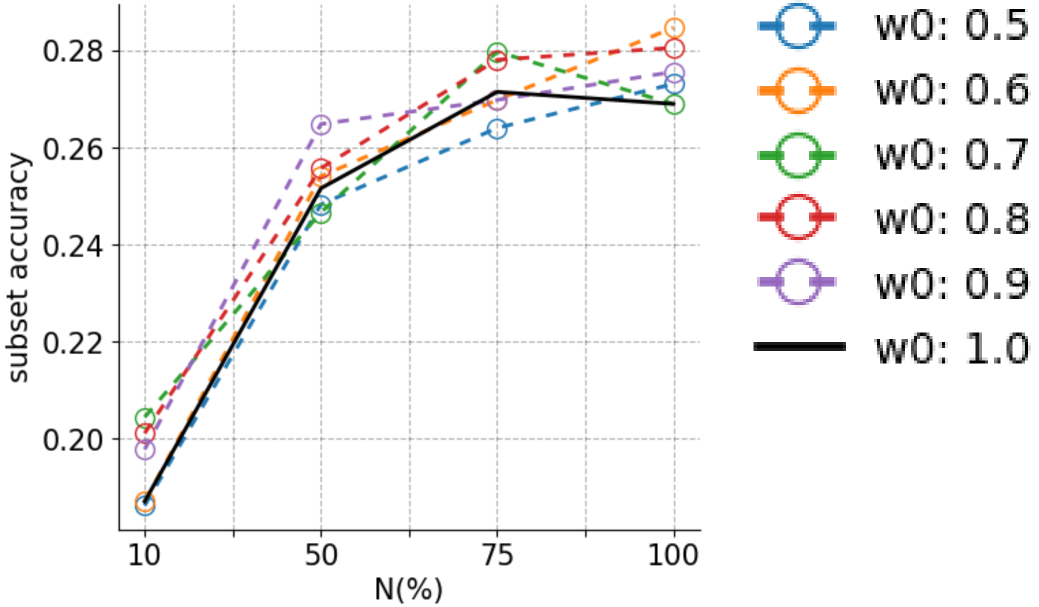
- We subsample around 1% of the dataset to provide the initial results



HMLC of RSIs: The BigEarthNet dataset, Initial experiments

Dataset splits:
- 70% train, 10% validation, 20% test
- The splits are stratified

Training details:
- For the feature extractors we use ConvNext-V2
- Trained for 50 epochs
- Adam optimizer
- Batch size: 128
- Learning rate of 1e-4
- The weight parameter for producing the weights in the hierarchy is set to: 0.6, 0.7, 0.8, 0.9 and 1



Fully- and Semi-Supervised Hierarchical Multi-Label Image Classification with Graph Learning

Motivation

Current Limitations:

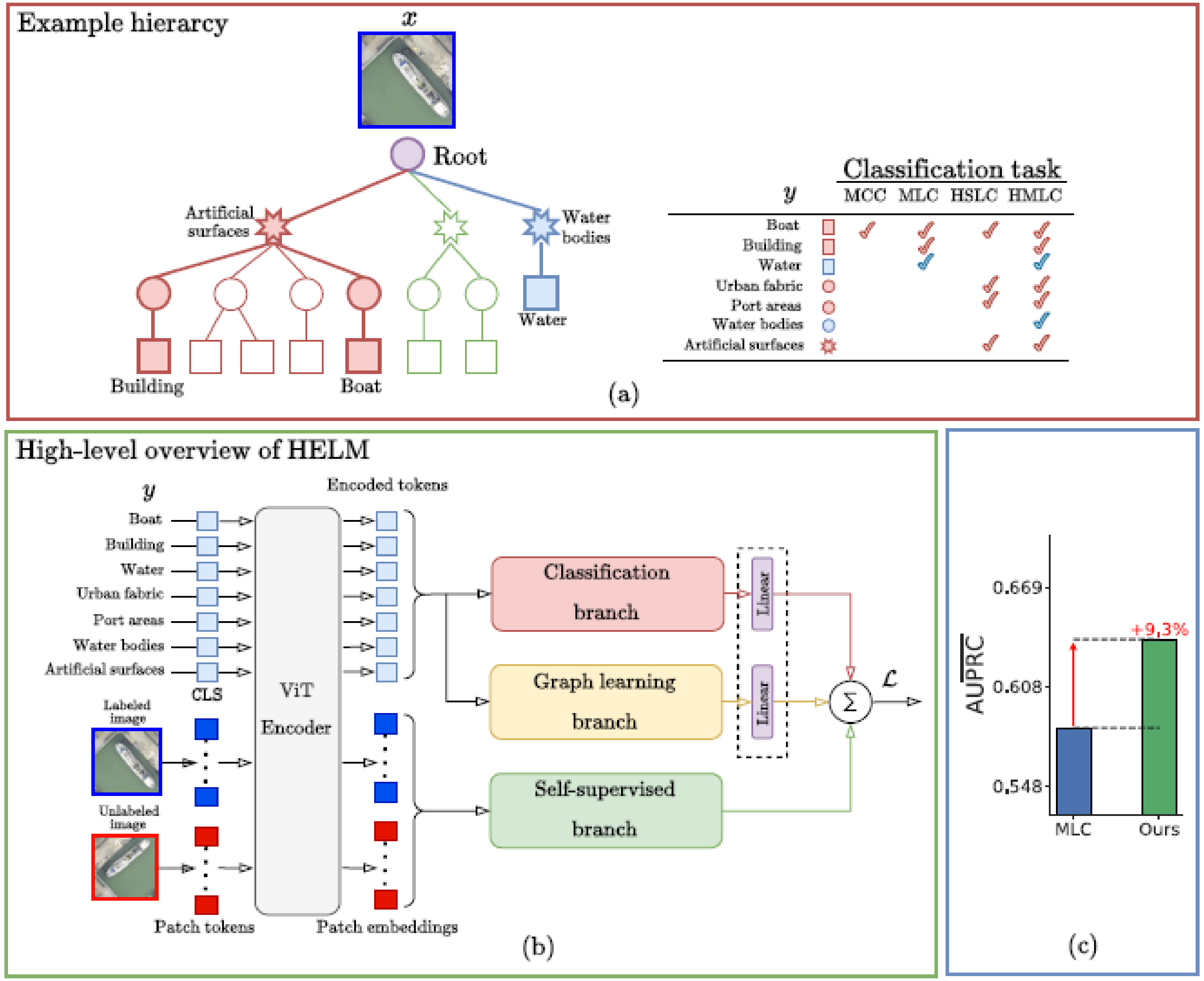
- **Single-path assumption** in hierarchies
- **Underutilization** of hierarchical structure
- **Supervised-only** approaches
- **Scarcity** of HMLC image datasets

Annotation Challenge:

- Hierarchical labeling is expensive and time-consuming
- Expert knowledge required for complex taxonomies
- Limited labeled data vs. abundant unlabeled data

Our Contributions → HELM Framework:

- Hierarchy-specific class tokens for label interactions
- Graph Convolutional Networks for hierarchical structure encoding
- Semi-supervised framework leveraging unlabeled data
- New HMLC benchmark datasets from remote sensing



HELM Methodology

Hierarchy-Specifiic Class Tokens

- Extend Vision Transformer (ViT) with M hierarchy-specific CLS tokens
- $M = M_l + M_h$ (leaf labels + intermediate hierarchy nodes)
- Input sequence: $T = [T_{\text{CLS}} \| T_p] \in \mathbb{R}^{(M+N_p)\times d}$

$$\tilde{\mathbf{z}} = E(T, \theta) \in \mathbb{R}^{(M+N_p)\times d}$$

where $E(\cdot, \theta)$ is the ViT encoder with parameters θ

Hierarchical Structure Encoding

Construct directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ from label hierarchy:

$$\tilde{\mathbf{z}}_g = \phi(\tilde{\mathbf{z}}_{\text{CLS}}; \mathcal{G}) \in \mathbb{R}^{M\times d}$$

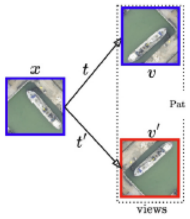
$$\mathbf{f}_g = \frac{1}{M}\sum_{m=1}^{M} \tilde{\mathbf{z}}_g^{(m)}$$

Classification loss (operates only on labels samples):

$$\mathcal{L}_g = \frac{1}{B_l}\sum_{i=1}^{B_l} \mathcal{H}(y_i, p_g(y|x_i))$$

Bootstrap Your Own Latent (BYOL)

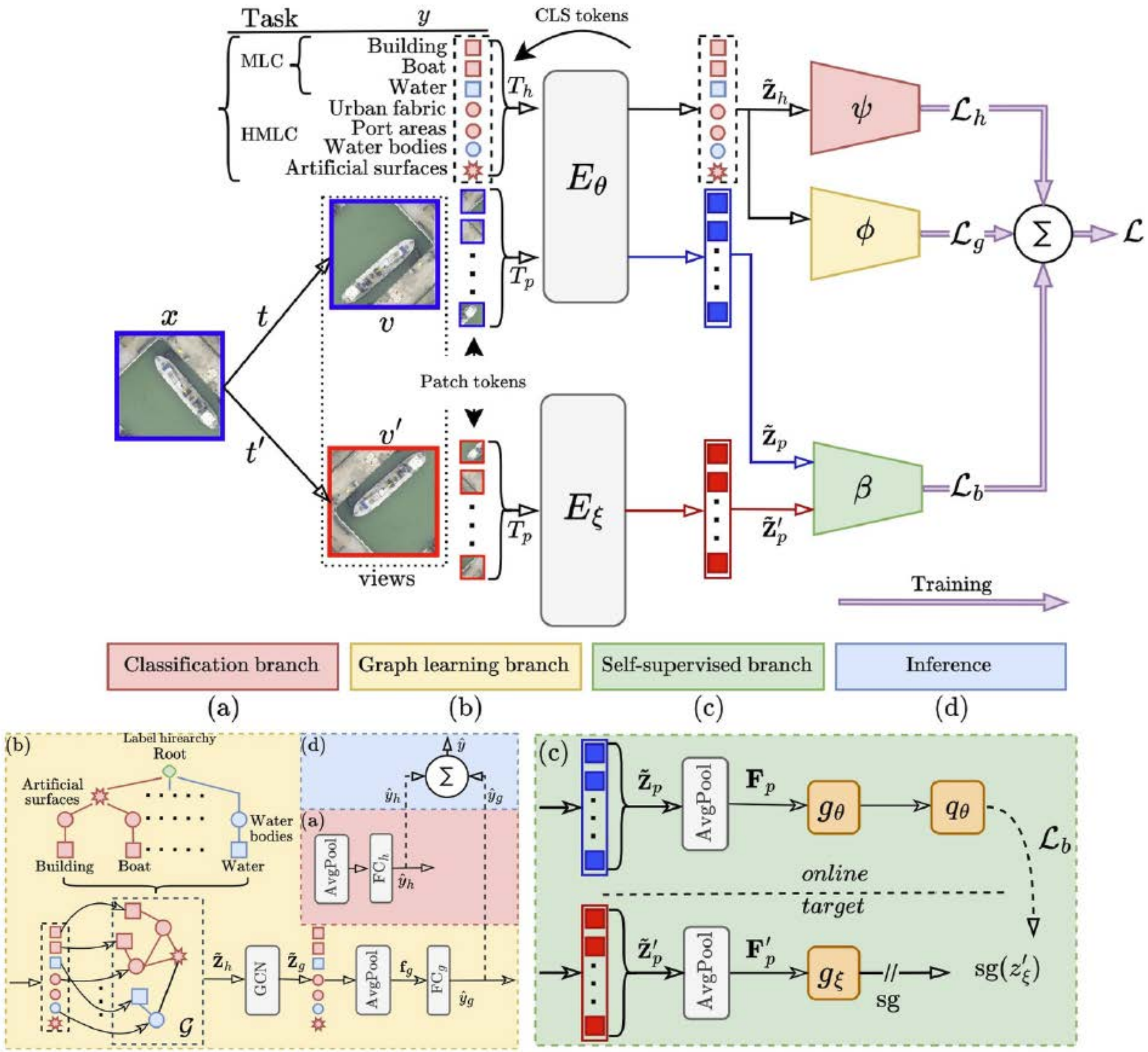
Augmentation views:



$$\mathcal{L}_b = 2 - 2 \cdot \frac{\langle q_\theta(g_\theta(\mathbf{F}_p)), g_\xi(\mathbf{F}'_p)\rangle}{\|q_\theta(g_\theta(\mathbf{F}_p))\| \cdot \|g_\xi(\mathbf{F}'_p)\|}$$

Target network update: $\xi \leftarrow \tau\xi + (1-\tau)\theta$

Momentum coefficient: $\tau \in [0, 1)$



HELM Evaluation: Experimental Design

Evaluations:

- Supervised
- Semi-supervised: Labeled data proportions {1%, 5%, 10%, 25%} - inductive learning setting
- Repeat the experiments three times

Method	$\mathcal{L}_s$	$\mathcal{L}_g$	$\mathcal{L}_b$	Setting
MLC Baseline	✓			Supervised
HMLC Baseline	✓			Supervised
HELM_g	✓	✓		Supervised/SSL
HELM_b	✓		✓	Supervised/SSL
HELM (Full)	✓	✓	✓	Supervised/SSL

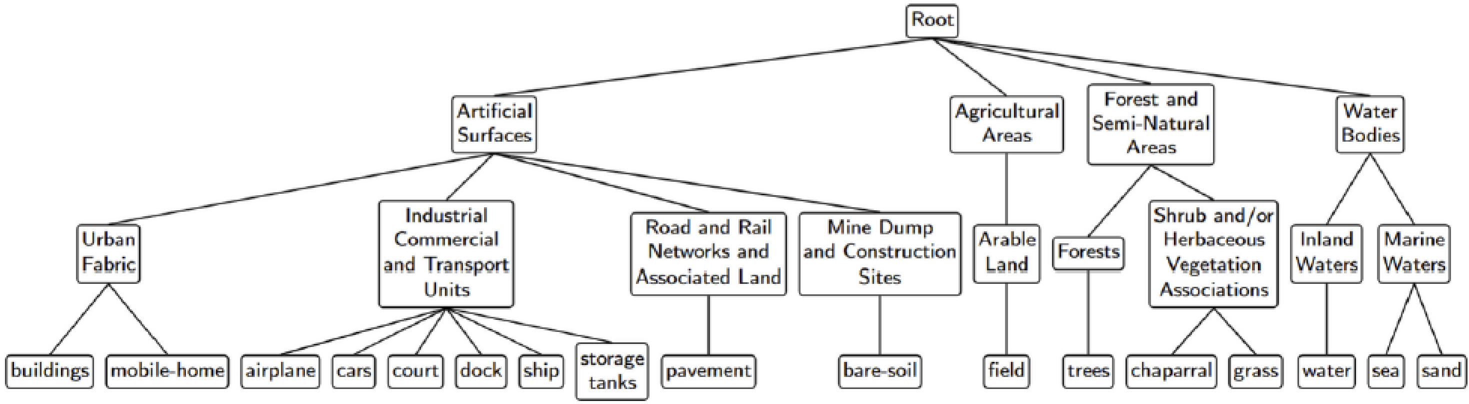
State-of-the-Art Comparison:

1. Coherent Hierarchical Multi-Label Classification Networks
2. HiMulConE
3. HMI

Datasets

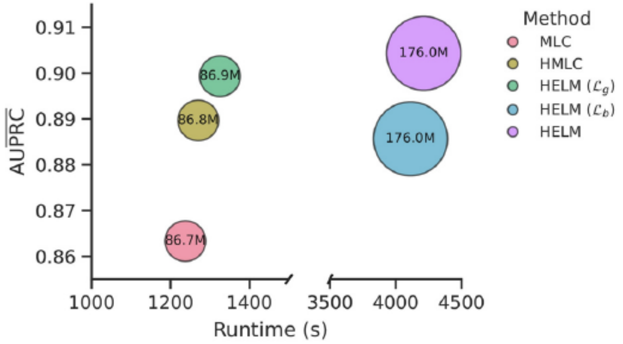
Dataset	N	N_{train}	N_{test}	$\lvert\mathcal{L}\rvert$							
				1	2	3	4	5	6	ℓ	h
UCM	2,100	1,667	433	4	9	17	-	-	-	17	30
AID	3,000	2,400	600	4	9	17	-	-	-	17	30
DFC-15	3,341	2,672	669	3	7	7	-	-	-	8	17
MLRSNet	109,151	87,336	21,815	7	15	22	60	-	-	60	104

Hierarchies



HELM Evaluation: SL Results

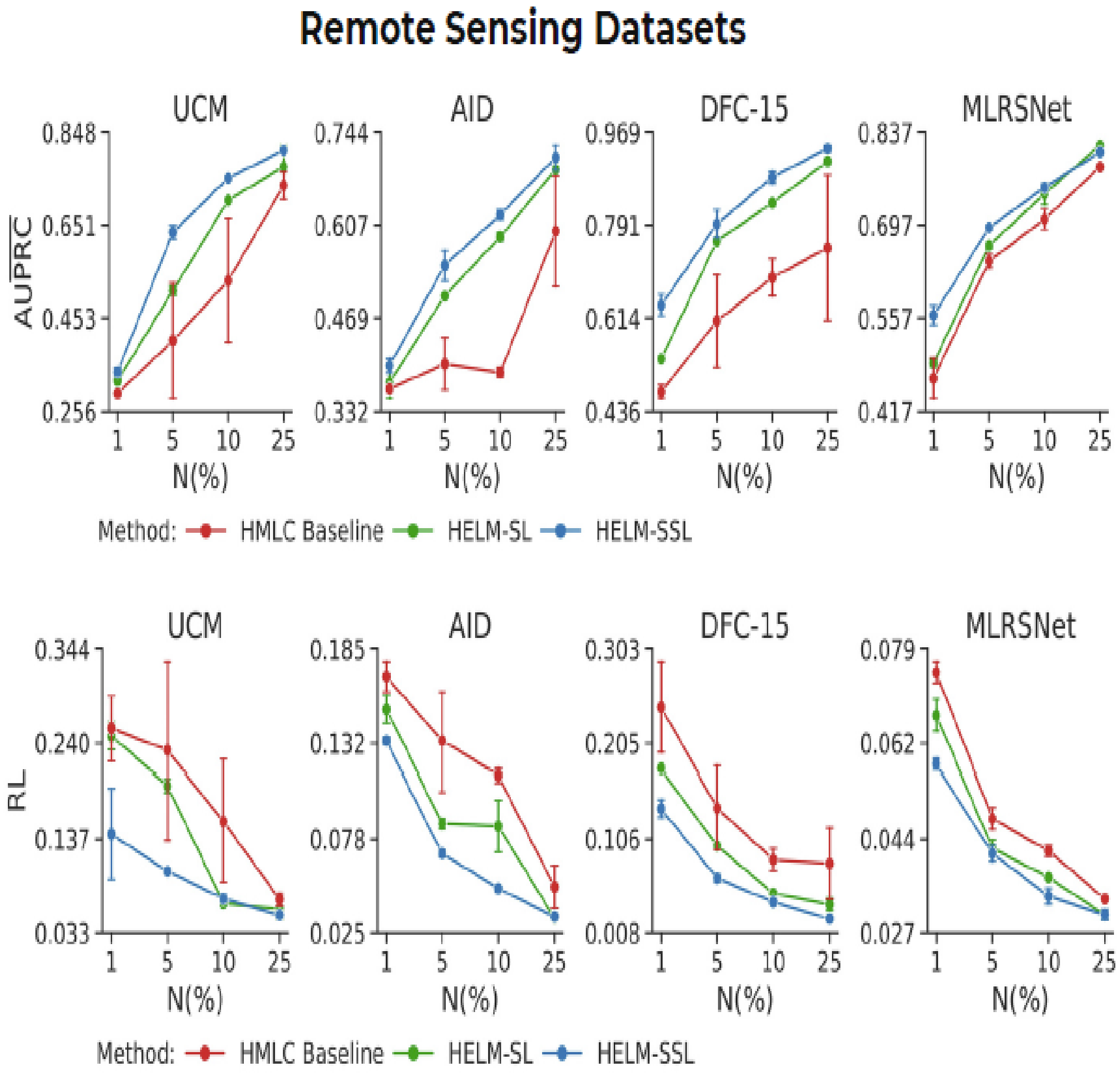
Method	$\mathcal{L}_s$	$\mathcal{L}_g$	$\mathcal{L}_b$	Setting
MLC Baseline	✓			Supervised
HMLC Baseline	✓			Supervised
HELM$_g$	✓	✓		Supervised/SSL
HELM$_b$	✓		✓	Supervised/SSL
HELM (Full)	✓	✓	✓	Supervised/SSL



Method	Remote Sensing Datasets				Rank
	UCM	AID	DFC-15	MLRSNet	
$\overline{\text{AUPRC}}$ (↑)					
MLC	0.863	0.767	0.967	0.838	5.00
HMLC	0.890	0.827	0.971	0.863	3.00
HELM$_g$	0.899 (+4.2, +1.0)	0.842 (+9.8, +1.8)	**0.979** (+1.2, +0.8)	0.869 (+3.7, +0.7)	2.50
HELM$_b$	0.885 (+2.5, -0.6)	**0.852** (+11.1, +3.0)	0.969 (+0.2, -0.2)	**0.873** (+4.2, +1.2)	3.13
HELM	**0.904** (+4.8, +1.6)	0.849 (+10.7, +2.7)	0.977 (+1.0, +0.6)	0.871 (+3.9, +0.9)	**1.38**
Ranking Loss (↓)					
MLC	0.031	0.025	0.010	0.039	4.75
HMLC	0.031	0.021	0.008	0.027	2.88
HELM$_g$	0.024 (+22.6, +22.6)	0.019 (+24.0, +9.5)	0.007 (+30.0, +12.5)	0.025 (+35.9, +7.4)	2.31
HELM$_b$	0.029 (+6.5, +6.5)	0.019 (+24.0, +9.5)	0.012 (-20.0, -50.0)	0.025 (+35.9, +7.4)	3.75
HELM	**0.022** (+29.0, +29.0)	**0.017** (+32.0, +19.0)	**0.006** (+40.0, +25.0)	**0.024** (+38.5, +11.1)	**1.31**

HELM Evaluation: SSL Results

Method	AUPRC (↑) UCM	AUPRC (↑) AID	AUPRC (↑) DFC-15	AUPRC (↑) MLRSNet	Ranking Loss (↓) UCM	Ranking Loss (↓) AID	Ranking Loss (↓) DFC-15	Ranking Loss (↓) MLRSNet
	Remote Sensing Datasets							
C-HMCNN [36]	0.834	0.764	0.962	0.792	0.038	0.024	0.012	0.041
HiMulConE [15]	0.843	0.770	0.970	0.865	0.031	0.020	**0.006**	0.035
HMI [76]	0.661	0.647	0.923	0.437	0.080	0.073	0.043	0.138
HELM (Ours)	**0.904**	**0.849**	**0.977**	**0.871**	**0.022**	**0.017**	**0.006**	**0.025**



DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

MANY THANKS!

TO YOU, FOR YOUR ATTENTION.
TO PARTNERS IN COLLABORATIVE PROJECTS.
TO COLLEAGUES@JSI, FOR THEIR COLLABORATION