

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

Marjan Stoimchev¹, Jurica Levatić², Dragi Kocev³, and Sašo Džeroski⁴

Abstract—The task of remote sensing image (RSI) classification has been studied extensively in the geoscience and remote sensing (RS) community. While deep learning methods have shown great success in solving this task, their reliance on large-scale labeled datasets is a serious limitation when dealing with complex labels and multiple semantic categories. The process of annotating such datasets can be time-consuming and tedious, leading to limited availability of labeled data and reduced performance of supervised learning methods. To address this issue, semi-supervised learning (SSL) methods can be applied, as they use both the limited labeled data and the abundant unlabeled data. In this article, we propose an effective SSL framework for RSI classification, which combines two key concepts. First, we employ a deep convolutional feature extractor to learn feature representations that encode the images into a lower dimensional feature space, capturing the rich semantic context present in RSI. Second, we utilize semi-supervised predictive clustering trees (PCTs) and ensembles thereof to learn from both the labeled and unlabeled data. To evaluate the effectiveness of the proposed framework, we compare it against several state-of-the-art self-supervised and semi-supervised methods from the literature. We conduct extensive experiments on ten publicly available land use/land cover RSI classification datasets: five for multi-class classification (MCC) and five for multi-label classification (MLC). The results demonstrate that the proposed framework has superior predictive performance compared to state-of-the-art methods from the literature, highlighting its effectiveness in semi-supervised RSI classification.

Index Terms—Convolutional autoencoders (CAEs), multiclass classification (MCC), multi-label classification (MLC), remote sensing (RS) images, self-supervised learning, semi-supervised learning (SSL), tree ensembles.

Manuscript received 29 May 2024; accepted 5 July 2024. Date of publication 11 July 2024; date of current version 1 August 2024. This work was supported in part by European Commission through the ASSAS Project under Grant 101059682, the ELIAS Project under Grant 101120237, the INQUIRE Project under Grant 101057499, the PARC Project 101057014, and the TAILOR Project under Grant 952215; and in part by Slovenian Research Agency through the Research Program P2-0103 under Project J1-3033, Project J2-2505, Project J2-4452, Project J2-4660, Project J3-3070, Project J4-3095, Project J5-4575, Project J7-4636, Project J7-4637, and Project N2-0236. (Corresponding author: Marjan Stoimchev.)

Marjan Stoimchev is with the Department of Knowledge Technologies, Jožef Stefan Institute, 1000 Ljubljana, Slovenia, and also with the Jožef Stefan International Postgraduate School, 1000 Ljubljana, Slovenia (e-mail: marjan.stoimchev@ijs.si).

Jurica Levatić, Dragi Kocev, and Sašo Džeroski are with the Department of Knowledge Technologies, Jožef Stefan Institute, 1000 Ljubljana, Slovenia (e-mail: jurica.levatic@ijs.si; dragi.kocev@ijs.si; saso.dzeroski@ijs.si).

Digital Object Identifier 10.1109/TGRS.2024.3426981

I. INTRODUCTION

REMOTE sensing is a valuable tool for studying the Earth's surface and its phenomena from a distance. It allows us to collect data without the need for physical contact with the objects of interest. In the analysis of remote sensing image (RSI) data, machine learning techniques are often employed to identify patterns and extract meaningful information, particularly when dealing with large-scale image datasets [1].

There are two common machine learning tasks in the analysis of RSIs: multiclass classification (MCC) and multi-label classification (MLC). In the older and more prevalent task of MCC, a single label is associated with an image. An example of MCC would be classifying different types of crops in agriculture. The limitation of MCC is that it assumes that an RSI belongs to a single class. However, in real-world situations, multiple classes can be present in a single image. This limitation can be addressed by using MLC, which allows for a more flexible representation of the inherent complexity present in the real world and avoids the oversimplification of the classification process. By considering multiple labels for a single RSI, MLC can lead to more accurate and meaningful results, especially when dealing with complex and heterogeneous landscapes.

The creation of large-scale labeled datasets has allowed deep learning models to surpass the performance of traditional machine learning methods, especially in the fields of image recognition and classification. As a result, researchers from various disciplines, including geoscience and remote sensing (RS), are exploring the potential of deep learning in addressing complex problems in their respective fields [2], [3], [4], [5], [6], [7], [8], [9]. Such approaches can automatically extract knowledge and learn discriminative representations from data in an end-to-end manner and in less-constrained environments. However, obtaining domain knowledge from field experts, in the form of annotation of high volumes of image data, is a tedious and time-consuming task. It can lead to scarcity of labeled data and consequently limited predictive performance of supervised methods.

This is especially true for the domain of RS because of several reasons. First, the complexity of the visual context present in the images and the correlation among the labels are relatively high (e.g., predicting composite concepts in the images). Second, RS imagery might comprise vast geograph-

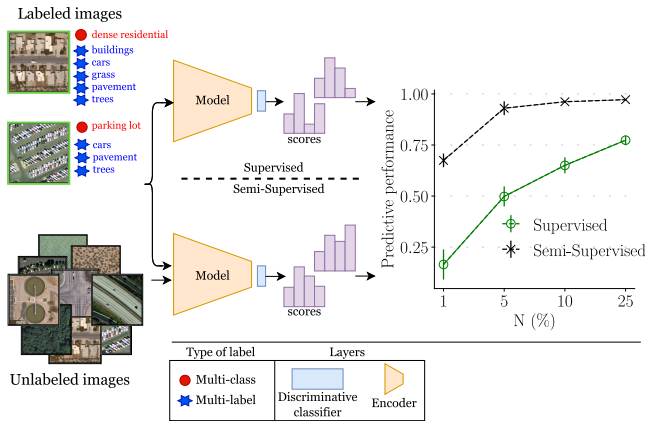


Fig. 1. Schematic illustration of the supervised and SSL settings for multiclass and multi-label RS image classification.

ical areas with some of them remote and difficult to access (e.g., polar and mountainous regions are difficult to access, and conditions can change rapidly; hence, data collection in these environments is both expensive and dangerous). Third, there is a need for specialized knowledge and equipment for accurate labeling (e.g., soil and land degradation assessments require field surveys for soil properties and need specialized equipment and experts). Furthermore, RS imagery is used to observe dynamic phenomena that result in rapid and often unpredictable changes in environmental conditions (e.g., real-time monitoring of natural disasters, such as earthquakes, tsunamis, and floods; monitoring of agricultural landscapes is challenging due to factors such as weather conditions, irrigation and fertilization practices, and pest infestation). Overall, labeled data are indispensable for the success of predictive models using remote sensing imagery, as they enhance the accuracy, reliability, and generalizability of these models, thus improving their application in downstream RS tasks. However, as mentioned earlier, obtaining labeled data is challenging.

There are several approaches for tackling the problem of labeled data availability [10]. The most common one is to exploit proven deep learning models trained on large-scale image datasets, such as ImageNet [11], and to fine-tune them on limited-size datasets for the downstream task. While directly transferring the learned knowledge from ImageNet to RSI can be a successful and easy solution [12], it still requires a significant amount of labeled images of the target domain (typically in the thousands). Another family of methods tries to address the small-sample problem through domain adaptation, where the general idea is to align the distribution of the source and target domains [13]. Significant successes have been achieved with self-supervised learning approaches that focus on various pretext tasks for pretraining, such as image reconstruction by using autoencoders [14], [15], [16] and modeling image similarity or dissimilarity between two or more views through the contrastive learning paradigm [17]. To address the problem of limited labeled data, in this article, we focus on semi-supervised learning (SSL), which leverages both unlabeled and labeled and, hence, can provide greater generalisability [10].

The use of SSL methods reduces the need for tedious manual labeling and can help SSL methods to achieve better performance than fully supervised methods that use only labeled data. The general concept of SSL is illustrated in Fig. 1. Several SSL solutions have been successfully applied in a number of domains, such as image classification [18], object detection [19], and semantic segmentation [20]. However, to the best of our knowledge, the existing SSL approaches for RSI classification primarily focus on the multiclass setting [21], [22], [23], [24], [25], and there is a lack of SSL methods able to tackle both multiclass and multi-label RSI classification tasks. As discussed, RSIs are intrinsically multi-label, and moreover, obtaining multiple labels is even more complex and time-consuming. This exacerbates the problem of labeled data scarcity and emphasizes the need for SSL methods in the MLC of RSIs.

Among the various methods in the realm of SSL, the most dominant ones include consistency regularization-based approaches [18], [26], [27], pseudolabeling [28], [29], self-training [30], and cotraining [31], [32]. Consistency regularization methods involve perturbing input images using random augmentations to establish an invariant output distribution. Pseudolabeling and self-training rely on using a model trained on a small set of labeled examples to generate new predictions based on unlabeled data, creating so-called pseudolabels, typically derived using confidence scores. Similarly, cotraining methods train two or more different models on the same dataset. The models subsequently share their predictions with one another, and only the samples on which they agree are labeled with pseudolabels and utilized for training the next iteration of the models. Another set of SSL methods is based on contrastive learning [33], [34], [35], which involves training a model to determine whether two samples are similar or dissimilar. This is done by maximizing the agreement between two representations of the same sample, subjected to different perturbations.

In addition, there are SSL methods based on generative modeling, where models are trained to generate new data that are similar to a given dataset. Such approaches have a range of potential applications, such as generating new images or text or creating synthetic data to expand a small labeled dataset. One well-known method of generative modeling is the approach of generative adversarial networks (GANs) [36], [37]. In this approach, two neural networks are trained in an adversarial fashion: a generator network tries to deceive the discriminator network that aims to correctly differentiate between real and generated images. Another approach to generative modeling is the use of variational autoencoders (VAEs) [38]. VAEs are a type of neural network that can learn to encode data into a lower dimensional representation and then decode the representation back into the original data. VAEs can be used to generate new data by sampling from the learned representation and then decoding the samples into the original data space. Despite the apparent need for SSL, to the best of our knowledge, there is very limited research on SSL for multi-label RSI classification, with only a few such methods available [39], [40] to solve the task at hand.

To address the lack of semi-supervised methods for multi-label RSI classification, in this article, we propose an effective framework that combines ideas from deep learning on the one hand and SSL based on predictive clustering trees (PCTs) and random forest ensembles thereof [41], [42] on the other hand. This framework integrates the advantages of both approaches: 1) learning feature representations using deep convolutional-based feature extractors that capture the rich semantic context of RSIs and 2) leveraging a small amount of labeled data with abundant unlabeled data in an SSL fashion. The latter (PCTs and their ensembles) can address both MCC and MLC tasks, considering the label dependencies inherent in MLC tasks. Note that the proposed framework is general and can be readily applied to other domains facing a scarcity of labeled images.

The main contributions of our work can be summarized as follows.

- 1) We propose an SSL framework for multi-label and multiclass RSI classification, which combines decision trees for SSL with the discriminative feature representations learned by a backbone CNN feature extractor.
- 2) We design and execute a comprehensive experimental analysis of the performance of the proposed framework using multiple RSI datasets for MCC and MLC, various network architectures, and different evaluation measures. We analyze the results quantitatively (through visual inspection of the classification of example images) and statistically (through the Friedman test and Nemenyi posthoc analysis).
- 3) We show that semi-supervised methods, particularly SSL PCTs and SSL random forests, robustly and consistently outperform their supervised counterparts across the various experimental setups. This indicates the effectiveness of incorporating unlabeled data in improving classification accuracy and its potential for versatile use in RSI classification.
- 4) We demonstrate that the SSL random forest method is the top performer among the considered state-of-the-art self-supervised and SSL methods, particularly for datasets with more than 1% labeled data. The significant improvement in the predictive performance of SSL over supervised learning was particularly evident in single PCTs. These findings were confirmed through statistical tests and average rank diagrams.
- 5) We finally show that the semi-supervised convolutional autoencoder (CAE) method shows improvements over the supervised baseline only for certain network architectures, such as EfficientNet. This suggests that it is very important to use the appropriate network architecture for SSL in end-to-end learning scenarios.

II. METHODOLOGY

In this section, we present the proposed approach for semi-supervised MCC and MLC of RSI (see Fig. 2). We describe our two-stage learning method that is composed of a feature extraction and a classification part. The feature extraction uses a standard CNN-based feature extractor to represent images in

a lower dimensional space. The classification part employs a learning method for SSL based on PCTs and tree-ensembles for MLC (and MCC).

A. Feature Extraction

The feature extraction component is a crucial part of our proposed method. Its task is to capture and represent the information present in the images. In this context, a simple CNN backbone feature extractor $E(\mathbf{x}, \theta)$ is used. It takes an RSI $\mathbf{x} \in \mathbb{R}^{w \times h \times 3}$ as input, where w and h represent the image dimensions, 3 represents the three channels of RGB images, and θ signifies the parameters of the CNN that need to be learned during training. As a result, a d -dimensional feature representation $\mathbf{f} \in \mathbb{R}^d$ is generated from the last pooling layer. The CNN backbone is fully convolutional to ensure versatility and applicability to images of varying sizes, as needed in many computer vision tasks, including RSI classification.

B. Semi-Supervised PCTs and Ensembles

To achieve accurate and efficient classification of RSI with limited labeled data, we utilize PCTs and ensembles of PCTs in an SSL setting for both MCC and MLC tasks [41], [42] (see Fig. 2).

PCTs provide a sophisticated extension of traditional decision trees by allowing predictions of complex, structured output data. PCTs view decision trees as a hierarchical set of clusters, with the initial cluster encompassing all available data. Through recursive division, smaller clusters are formed as one moves from the root to the leaves of the tree. The ensemble version employs PCTs as base models in the ensemble. The ensemble generates predictions for new examples by considering the collective predictions of all PCTs in the ensemble. The SSL PCTs and ensembles differ from these supervised counterparts in that they use both labeled and unlabeled examples in the training process.

The input to the algorithms comprises both labeled and unlabeled examples, in our case, the learned feature representations: $\mathbf{f} = \mathbf{f}_l \cup \mathbf{f}_u$, where \mathbf{f}_l and \mathbf{f}_u are the feature representations generated from the labeled and unlabeled examples, respectively. The tree construction algorithm uses the variance function to evaluate the candidate splits in the tree and choose the best one. The SSL version of the variance function considers both the descriptive and target attributes and can, thus, exploit both labeled and unlabeled examples [41], [42]

$$\text{Var}_f = w \cdot \text{Var}_f(Y) + (1 - w) \cdot \text{Var}_f(X) \quad (1)$$

where $w \in [0, 1]$ balances the contributions of the target space (Y) and the descriptive space (X) to the variance function Var_f (note that for unlabeled examples, only the descriptive space is available). The w parameter is key to the flexibility of the learned models, allowing learning to span the entire spectrum from fully supervised ($w = 1$) to fully unsupervised ($w = 0$) models. By controlling the weight given to unlabeled examples using the w parameter, the level of supervision for different datasets can be appropriately adjusted. This acts as a safety mechanism to guard against the performance degeneration

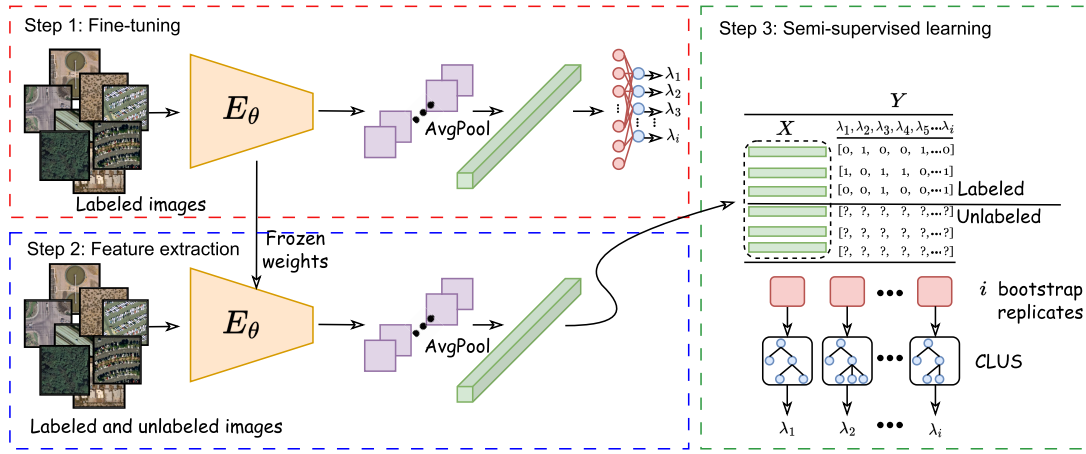


Fig. 2. Overview of the proposed framework for semi-supervised MLC of RSIs. In the first step, a backbone model is fine-tuned on the labeled training set, and in the second step, PCTs are utilized for SSL, where the inputs to the algorithm are labeled and unlabeled learned feature representations. Note that while tree ensembles are depicted, we also consider single PCTs.

caused by the use of unlabeled data [43]. The value of w is selected by inner cross-validation on the labeled set of data. In theory, if an optimal value of w is chosen, semi-supervised PCTs would never perform worse than supervised PCTs since supervised PCTs are a special case of SSL-PCTs, obtained for $w = 1$. However, since w is chosen by internal cross-validation on the training set, the chosen value can be suboptimal for the test set, and hence, the semi-supervised method can perform worse than the supervised one. This, however, rarely happens in practice [41], [42].

The variance of a set of examples E on the target space Y (i.e., a set of T binary labels Y_1, \dots, Y_T) is calculated as follows:

$$\text{Var}_f(E, Y) = \frac{1}{T} \cdot \sum_{i=1}^T \text{Gini}(E, Y_i). \quad (2)$$

The variance on the descriptive space X consisting of D attributes of a set of examples E is computed as follows:

$$\text{Var}_f(E, X) = \frac{1}{D} \cdot \sum_{i=1}^D \text{Var}(E, X_i) \quad (3)$$

where the variances and the Gini scores are calculated as follows, respectively:

$$\begin{aligned} \text{Var}(E, X_i) &= \frac{\frac{N-1}{K_i-1} \cdot \sum_{j=1}^{K_i} (x_{i,j})^2 - N \cdot \left(\frac{1}{K_i} \cdot \sum_{j=1}^{K_i} x_{i,j} \right)^2}{N} \\ \text{Gini}(E, Y_i) &= 1 - \left(\sum_{j=1}^{K_i} \frac{|\{e : e \in E \wedge y_{ij} = 1\}|}{K_i} \right)^2 \\ &= \left(\sum_{j=1}^{K_i} \frac{|\{e : e \in E \wedge y_{ij} = 0\}|}{K_i} \right)^2 \\ &= 2\hat{p}_i(1 - \hat{p}_i) \end{aligned}$$

where N is the number of examples (both labeled and unlabeled), K_i is the number of examples with nonmissing values of the i th label Y_i , and \hat{p}_i is the probability of label Y_i , estimated by using only examples for which the value for variable Y_i is known.

III. EXPERIMENTAL DESIGN

In this section, we provide a comprehensive overview of the experimental design employed in the comparative analysis of the performance of our methods for SSL on RSI data. First, we describe the multiclass and multi-label RSI datasets used to learn the predictive models. Next, we thoroughly explain the evaluation strategy, specifically focusing on the SSL evaluation protocol. We next provide a concise explanation of the metrics used to assess the predictive performance of the trained models for MCC and MLC of RSI. Furthermore, we briefly outline several state-of-the-art semi-supervised and self-supervised learning methods from the literature, which are used in our comparison to emphasize the significance of the methods that we propose. Finally, we specify in detail the parameters of our methods used in the training procedure.

A. Datasets

We assess the performance of the proposed methods on ten publicly available land use and land cover RSI datasets: five datasets for MCC and five datasets for MLC. We selected these datasets based on their diversity along several dimensions, including image resolution, number of labeled images available, and geographical location. The datasets are available at <http://eodata.bvlabs.ai>. A short description of the datasets is given in the following, while their quantitative details are given in Table I. Moreover, some typical images with their corresponding class labels for both the multiclass and multi-label datasets are given in Fig. 3.

1) *MCC Datasets*: For the MCC task, we used the datasets OPTIMAL-31 [44], RESISC45 [45], RSSCN7 [46], AID [47], and UCM [48]. OPTIMAL-31 is a dataset consisting of 31 distinct categories for scene classification, with images collected from Google Earth and sized at 256×256 pixels. The RESISC45 dataset, created by Northwestern Polytechnical University (NWPU), is a publicly available benchmark for RSI scene classification, comprising 31 500 images at 256×256 resolution and covering 45 scene classes. The RSSCN7 dataset contains 2800 scene classification images obtained from Google Earth, with a resolution of 400×400 .

TABLE I

DESCRIPTION OF THE USED MCC AND MLC RSI DATASETS. $|\mathcal{L}|$ DENOTES THE NUMBER OF POSSIBLE LABELS; CARD DENOTES LABEL CARDINALITY (I.E., AVERAGE NUMBER OF LABELS PER IMAGE); DENS DENOTES LABEL DENSITY (AVERAGE PROPORTION OF IMAGES LABELED WITH A GIVEN LABEL); N IS THE NUMBER OF IMAGES IN THE DATASET, OF WHICH N_{TRAIN} ARE IN THE TRAIN AND N_{TEST} IN THE TEST DATASETS; AND $w \times h$ IS THE DIMENSION OF THE IMAGES (IN PIXELS)

Task	Dataset	Image Type	$ \mathcal{L} $	<i>Card</i>	<i>Dens</i>	N	N_{train}	N_{test}	$w \times h$
MCC	OPTIMAL-31	Aerial RGB	31	1	0.032	1,860	1,488	372	256×256
	UCM	Aerial RGB	21	1	0.047	2,100	1,680	420	256×256
	RSSCN7	Aerial RGB	7	1	0.142	2,800	2,240	560	400×400
	AID	Aerial RGB	30	1	0.033	3,000	2,400	600	600×600
	RESISC45	Aerial RGB	45	1	0.022	31,500	25,200	6300	256×256
MLC	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	2,100	1,667	433	256×256
	AID Multilabel	Aerial RGB	17	5.152	0.468	3,000	2,400	600	600×600
	DFC-15 Multilabel	Aerial RGB	8	2.795	0.465	3,341	2,672	669	600×600
	MLRSNet	Aerial RGB	60	5.770	0.144	109,151	87,325	21,826	256×256




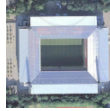

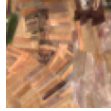

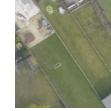


MCC datasets					MLC datasets				
OPTIMAL-31	UCM	RSSCN7	AID	RESISC45	Ankara	UCM	DFC-15	AID	MLRSNet
									
airplane	storagetanks	forest	stadium	beach	Bare Soil, Crop (Type-A), Crop (Type-B), Unpaved Road, Grass (Type-A)	bare-soil, buildings, cars, pavement, tanks	impervious, vegetation, building, tree	bare-soil, buildings, cars, court, pavement, trees	bare soil, buildings, grass, trail, wind turbine

Fig. 3. Illustrative RSIs and their class labels from different multiclass and MLC datasets.

For AID and UCM, we used the multiclass versions of the MLC datasets (as described in the following) with 21 and 30 scene classes, respectively.

2) *MLC Datasets*: For the MLC task, we used the datasets UC-Merced (UCM) Land Use [39], AID [49], Ankara HIS archive [50], DFC-15 [51], and MLRSNet [52]. UCM Land Use contains 2100 images grouped into 21 broad categories at a scene level, with a total of 17 object-level labels. The resolution of the images is 256×256 . AID is a more challenging dataset than UCM, containing 3000 aerial images grouped into 30 object categories, with 17 object-level labels. The image resolution of this dataset is 600×600 . Ankara HIS is a small hyperspectral image archive consisting of only 216 low-resolution images at 63×63 pixels, acquired by the NASA EO-1 satellite’s Hyperion sensor from the area surrounding the city of Ankara in Turkey. DFC-15 is built from a semantic segmentation dataset from the 2015 IEEE GRSS data fusion contest, containing a total of 3342 image patches with a resolution of 600×600 pixels. The image patches are grouped into eight different object categories. Finally, the MLRSNet dataset is the largest among the MLC datasets, containing 109 161 images annotated into 46 categories, with an image resolution of 256×256 pixels. Moreover, each image in the dataset is labeled with several of the 60 predefined class labels. In order to ensure that the dataset is comparable in size to the other datasets used in our study, we sampled (using stratification) approximately 5% of the dataset.

B. Evaluation Strategy

To evaluate the performance of the SSL methods, we use the inductive learning setting of SSL, where our goal is to learn models that can predict labels for new examples that are unseen during training. This means that a separate fixed-sized test set is utilized to assess the final model’s performance (see Fig. 10). We use different percentages of labeled data from the training sets to investigate the dependence of model performance on the amount of labeled data. Labeled samples were randomly subsampled from the available training set in the percentages $\{1, 5, 10, 25\}$, while the remaining training samples were used as unlabeled samples (with their labels removed). We repeated the experiments five times using different random initializations for each percentage of unlabeled data. The final predictive performance is averaged across the five runs.

C. Evaluation Measures

For MCC, we use six performance measures: microaveraged and macroaveraged $F1$, precision, and recall; accuracy; and micro-AUPRC. More details regarding the used measures are given by Wu and Zhou [53]. In the MLC scenario, we employed a range of 13 evaluation measures commonly used in this context. Specifically, we focused on two types of measures for the MLC task: bipartition-based measures (including example- and label-based measures) and ranking-based measures. Example-based measures involve computing the average difference between the true and predicted labels for each data point and then averaging these differences across all the examples in the dataset. We used Hamming loss

and subset accuracy from this group. In contrast, label-based measures evaluate each label individually and then average the performance across all labels. For these measures, we used microaveraged and macroaveraged $F1$, precision, and recall, the average area under the precision–recall curve (AUPRC), and weighted AUPRC. Finally, ranking-based measures compare the predicted ranking of the labels to the ground-truth ranking, where the present labels are ranked before the absent labels. From this group, we used ranking-loss, coverage, and one error. To calculate the measures mentioned above, we use `skit-learn` [54].

D. Methods for Comparison

To investigate whether semi-supervised methods benefit from the addition of unlabeled data, we compare the semi-supervised versions of PCTs (SSL-PCT) and random forests (SSL-RForest) methods with their supervised counterparts, i.e., supervised PCTs (SL-PCT) and random forest (SL-RForest). In our experiments, we also consider a CAE approach for end-to-end SSL. The supervised version of this approach is only the encoder part of this structure (denoted as baseline), which is trained on the labeled data only. The idea behind the CAE approach is relatively simple. It operates in two general steps to address the task of SSL: 1) self-supervised pretraining on the unlabeled data by using image reconstruction as a pretext task and 2) fine-tuning the encoder part of the CAE on a small portion of labeled examples for the downstream MCC and MLC tasks by using task dependent discriminative classifiers. This approach has been widely exploited in various computer vision tasks [14], [15], [16], where the main idea is also applicable to the RSI domain [55], [56].

In addition, to demonstrate the relevance of our SSL methods, we compare their predictive performance to that of several state-of-the-art self-supervised methods from the literature: BYOL [57], SimSiam [58], SimCLR [59], BarlowTwins [60], MoCo [61], DINO [62], and SwAV [63]. We adapt these methods for the SSL task by using only the pretrained encoder structure of the networks trained with linear evaluation, which is a common practice for utilizing such methods in an SSL setting [7], [64].

We also consider two state-of-the-art approaches specially designed for the SSL task. The first state-of-the-art method, referred to as HR-S²DML [35], is a deep metric learning architecture originally developed for RSI-MCC. The second method is called MSMatch [23]. MSMatch is an extension of the well-established FixMatch methodology [18], incorporating elements such as consistency regularization and pseudolabeling to effectively address the challenges of SSL. We used MSMatch only for MCC datasets. On the other hand, we modified the HR-S²DML method to operate in the MLC setting. First, instead of the categorical cross-entropy (CE) loss for MCC, we use the binary CE (BCE) loss, which is suitable for the MLC task, and second, for classification, we use the MLC version of the kNN classifier. Note that for the two previously published multi-label SSL methods for RSI [39], [40], the implementations are not publicly available; therefore, we did not include them in our experimental comparison.

E. Parameter Settings

For the feature extraction, we utilized different backbone CNNs implemented in the `PyTorch` framework [65], namely, VGG [66], ResNet [67], and EfficientNet [68], with weights pretrained on ImageNet. We used a total of eight backbone CNNs for comparison. The networks are modified to accept input images of arbitrary size by replacing the last max-pooling layer with an average-pooling operation using kernel size 1. The resulting d -dimensional feature representations are given as follows: 4096 features for VGG-16 and VGG-19, 512 features for ResNet-34, 2048 features for ResNet-50 and ResNet-152, 1280 features for EfficientNet-B0 and EfficientNet-B1, and 1408 features for EfficientNet-B2. The networks are fine-tuned for 25 epochs of training on an NVIDIA A100 (40-GB) GPU, by using the Adam optimizer with a fixed learning rate of 1×10^{-4} , a minibatch of 128, and a mixed precision of training to speed up the training process and reduce memory consumption. We also applied the same training procedure on the CAEs for end-to-end learning, which follows a symmetrical encoder–decoder structure.

The learned representations by the feature extractors are used as inputs to learn supervised and semi-supervised PCTs and random forests. Single trees are pruned using M5 pruning. The PCTs and the ensembles of PCTs are implemented in the `CLUS+` framework [69] (available at <https://github.com/knowledge-technologies/clus>). We use 100 unpruned trees to construct random forests, with the feature subset size set to the `sqrt` of the total number of features. For the SSL version, we optimize the parameter w using an internal threefold cross-validation procedure, applied to the labeled portion of the training set. We consider values of the parameter w in the range of 0–1, with 0.1 increments. For the state-of-the-art semi-supervised and self-supervised methods, we use the parameter values recommended by their respective authors.

The experiments were conducted on a computer with an AMD EPYC 7702 2 GHz 64-core processor and 1 TB of RAM.

F. Computational Complexity

The computational complexity of the proposed framework depends on the feature extractor part and the SSL part. The complexity introduced within the feature extractor part varies depending on the chosen architecture. For instance, VGG-16 and VGG-19 require approximately 15.3 and 20 GFLOPS. On the other hand, in the-ResNet based architectures, the complexities range from 3.9 GFLOPS for ResNet-34 to 11 GFLOPS for ResNet-152 [67]. Finally, in EfficientNets, renowned for their efficiency, the requirements are notably lower, with EfficientNet-B0 at 0.39 GFLOPS, EfficientNet-B1 at 0.68 GFLOPS, and EfficientNet-B2 at 1 GFLOPS [68].

The SSL part consists of semi-supervised PCTs and random forests. The complexity of building a single SSL-PCT tree is $\mathcal{O}(DN \log^2 N) + \mathcal{O}((D+T)ND \log N) + \mathcal{O}(N \log N)$, while the complexity of random forests of SSL-PCTs is bounded by $k(\mathcal{O}(D'N' \log^2 N') + \mathcal{O}((T+D)N'D' \log N'))$, where D is the number of descriptive variables, T is the number of labels,

N is the combined number of unlabeled and labeled training examples, N' is the number of bootstrap samples, D' is the number of features considered at each tree node in random forest, and k is the number of trees [42].

IV. RESULTS AND DISCUSSION

This section presents a comprehensive discussion of the experimental evaluation of the proposed methods for SSL on RSI. The evaluation aims to address several research questions that shed light on the advantages and limitations of the proposed methods. The outcomes of the evaluation are analyzed in relation to the following research questions.

- 1) Can the proposed SSL methods for RSI classification effectively leverage information from unlabeled data to improve performance?
- 2) Does the ranking of the considered methods remain consistent across different evaluation measures for MLC and MCC?
- 3) Is the performance improvement of SSL methods consistent across different datasets, and to what extent does SSL outperform SL?
- 4) How robust is the performance of the proposed SSL methods to changes in the architecture of the underlying CNN feature extractors?
- 5) How do the results obtained with the proposed methods compare to those state-of-the-art semi-supervised and self-supervised methods for RSI classification from the literature?

A. Predictive Performance Using Increasing Amounts of Labeled Data

Here, we present the results in the form of learning curves showing the performance of the models learned with varying percentages of labeled data in order to investigate whether SSL methods can benefit from unlabeled data. The results are presented in terms of AUPRC (for MCC tasks) and $\overline{\text{AUPRC}}$ (for MLC tasks) evaluation measures. Note that although we considered multiple evaluation measures in our experiments, for simplicity and easier interpretation, we report the performance using a single evaluation measure for each task. As analyzed and discussed in Section IV-B, the results remain consistent across the different evaluation measures.

The results for the MCC and MLC datasets are presented in Figs. 4 and 5, respectively, where each row represents a specific backbone architecture, and the columns correspond to the different datasets used in the experiments. For each pair consisting of a dataset and a backbone architecture, the SSL methods are compared with their supervised learning (SL) counterparts (resulting in a single graph in the figure). Overall, the results clearly demonstrate that SSL-PCTs and SSL-RForests consistently outperform their supervised counterparts, SL-PCTs and SL-RForests, highlighting the importance of leveraging unlabeled data. Semi-supervised random forests yield overall the best performance, particularly when dealing with small percentages of labeled data.

Concerning the end-to-end CAEs, improvements in the performance of SSL over SL were observed only for specific architecture types, with the EfficientNet base architectures consistently displaying the largest performance gains. A more detailed analysis of these effects can be found in Section IV-D, where we present a more in-depth explanation of these observations.

B. Performance of Methods Across Different Evaluation Measures

We next investigate the methods' rankings in terms of the observed predictive performance across the six MCC evaluation measures and 13 MLC evaluation measures used throughout the experiments (as described in Section III-C). To assess the stability and consistency of the rankings, in Fig. 6, we present the distribution of ranks for both SSL and SL methods across the different evaluation measures, separately for MCC and MLC tasks.

An inspection of the results reveals that the ranks are fairly stable across the different percentages of labeled examples, particularly for the MCC task. Therefore, we chose to present the results of the remaining analyses using the AUPRC evaluation measure for MCC and the $\overline{\text{AUPRC}}$ evaluation measure for MLC, as these evaluation measures are independent of classification thresholds and provide a reliable evaluation of performance. We note that the Ankara dataset is an exception, where the results across different evaluation measures are not stable for small amounts of labeled data. This might be due to the fact that the Ankara dataset is the smallest among the datasets used, and experiments conducted with 1% and 5% labeled data use very few labeled images.

The SSL methods based on decision trees (SSL-PCTs and SSL-RFs) consistently outperform their SL counterparts. Among the SSL methods, the RForest approach exhibits the best predictive performance across all datasets and displays a more consistent trend when learning from different percentages of labeled examples. On the other hand, the ranks of the SL and SSL end-to-end methods are more stable at higher percentages of labeled examples, while they exhibit greater variability in ranks at lower percentages.

C. Relative Performance of SL and SSL Methods Across Different Datasets

In this section, our goal is to explore the extent of improvement achieved by SSL compared to SL methods, as well as to determine if the improvement of SSL over SL depends on the dataset at hand. To accomplish this, we calculate the distribution of the relative differences between the performance of SSL and SL methods across various network architectures. This analysis is performed separately for each dataset. The relative difference is defined as follows:

$$r\Delta_M = \frac{M_{\text{SSL}} - M_{\text{SL}}}{M_{\text{SSL}}} \quad (4)$$

where $r\Delta_M$ is positive if SSL outperforms SL and negative otherwise. In the case of MCC, M is AUPRC, while for MLC, M is $\overline{\text{AUPRC}}$.

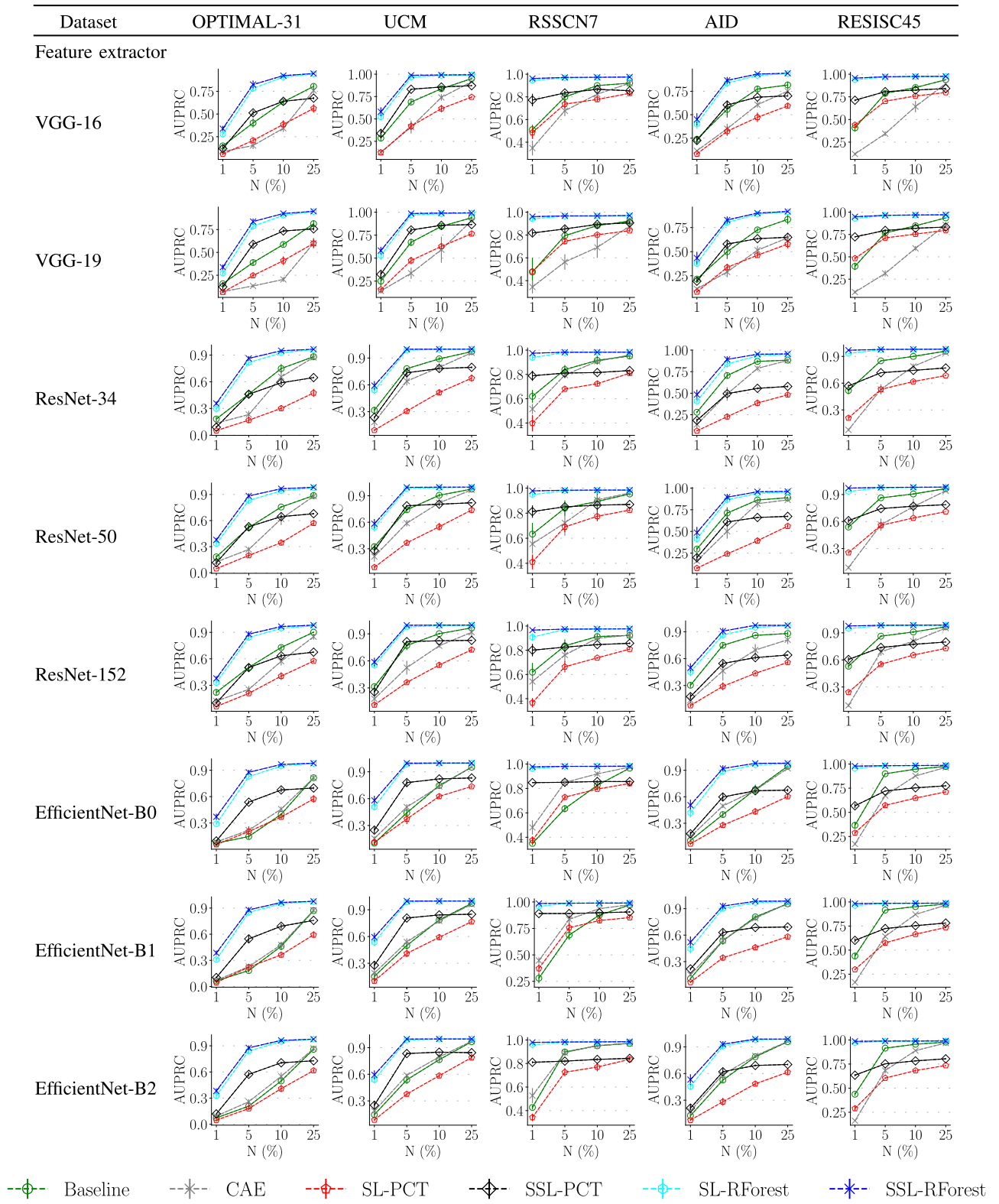


Fig. 4. Predictive performance of SL and SSL on MCC datasets of RSI. PCTs and RForests are applied to features obtained by backbone feature extractors using different CNN architectures. The same CNN architectures are used in end-to-end mode as SL baselines and as parts of the SSL-CAE approach. The learning curves depict $\mu \pm \sigma$ (mean and standard deviation) for AUPRC across five repeats, where higher values indicate better predictive performance. Note that the σ values are typically very small and are not visible in the graphs (especially at higher percentages of labeled data).

The distributions of $r\Delta_M$ values are depicted in Fig. 7. Our findings indicate that among the different methods tested, SSL-PCTs and SSL-RForest consistently exhibit improvements

over their supervised counterparts across all datasets. Specifically, the most significant improvement in the predictive performance of SSL over SL is observed for the single PCT

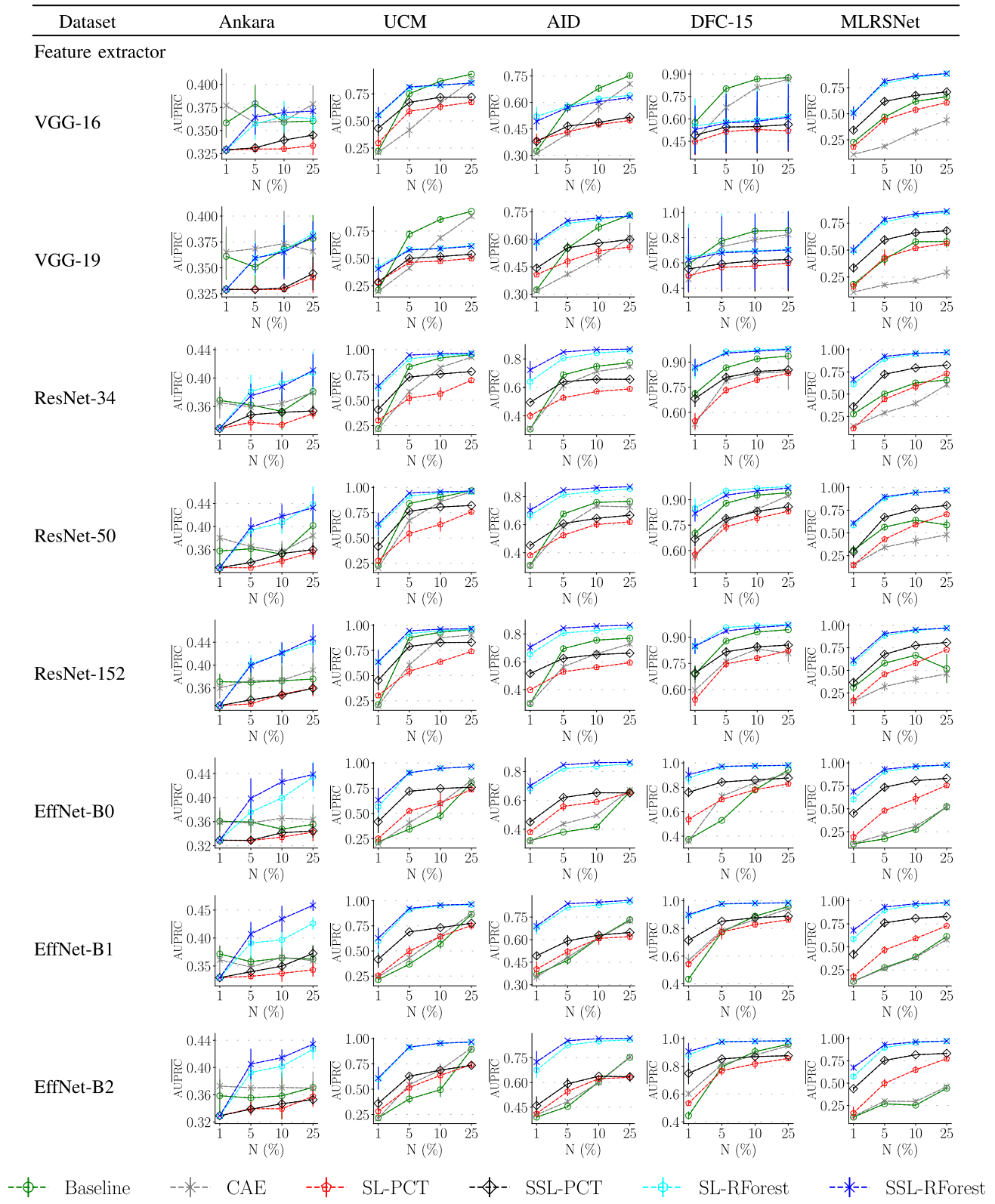


Fig. 5. Predictive performance of SL and SSL on MLC datasets of RSL PCTs and RForests are applied to features obtained by backbone feature extractors using different CNN architectures. The same CNN architectures are used in end-to-end mode as SL baselines and as parts of the SSL-CAE approach. The learning curves depict $\mu \pm \sigma$ (mean and standard deviation) for AUPRC across five repeats, where higher values indicate better predictive performance. Note that the σ values are typically very small and are not visible in the graphs (especially at higher percentages of labeled data).

method (SSL-PCT) compared to its supervised counterpart (SL-PCT), particularly for smaller percentages of labeled

data. While SSL-RForest also consistently enhances the predictive performance compared to SL-RForest, the degree of

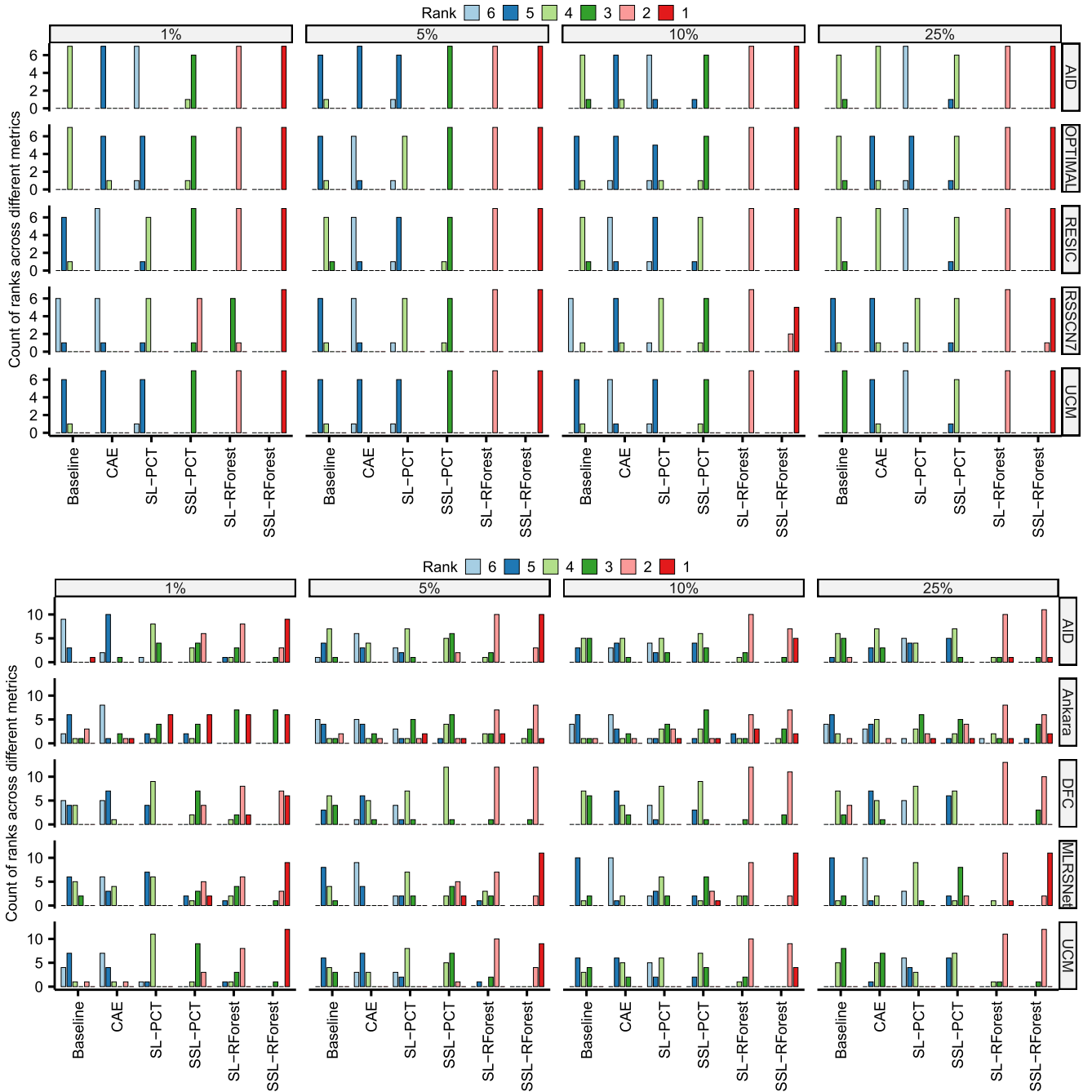


Fig. 6. Distribution of ranks of the newly proposed methods and the competing methods for SL and SSL, across the different MCC and MLC evaluation measures from Section III-C. Each graph represents the counts of ranks for a specific dataset and a given percentage of labeled data taken into consideration while learning the predictive models (ranks were rounded to the nearest integer). Overall, these graphs show the stability and robustness of the proposed methodology under the different evaluation scenarios. The distributions have clear peaks for the MCC datasets while being more spread for the MLC datasets. For the MCC datasets, our methodology is clearly top-ranked across the different datasets and percentages of labeled data. For the MLC datasets, our methodology is most frequently the top-ranked method, albeit not always the first.

improvement is lower than that of SSL-PCTs (although SSL-RForests demonstrate the best overall performance).

On the other hand, CAE does not demonstrate consistent performance improvement over the baseline method—they yield relatively similar results. These results suggest that PCTs and RForest methods are more versatile in using unlabeled data, and their SSL variants yield consistent improvement across different datasets.

D. Relative Performance of SL and SSL Methods Across Different Architectures

In this section, we examine the influence of the network architecture on the (relative) predictive performance of SSL and SL methods. To this end, we calculate the distribution of relative improvement of SSL methods over their supervised counterparts (similarly as in Section IV-C) across different datasets, for each network architecture

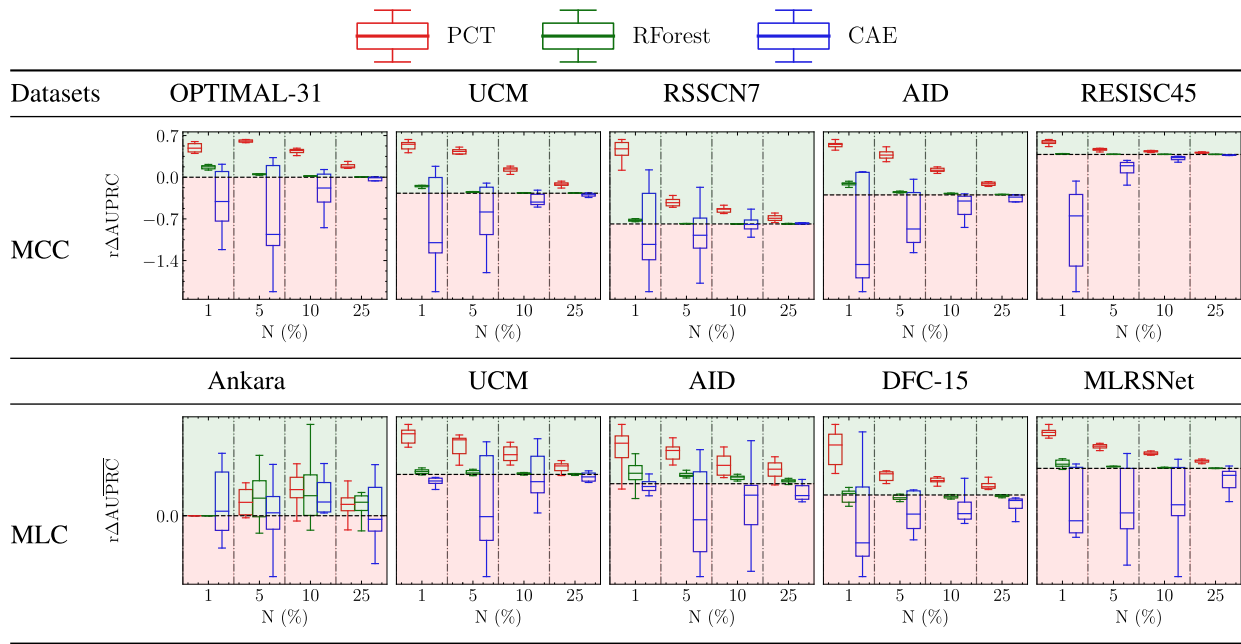


Fig. 7. Improvement of performance of SSL compared to SL in terms of $AUPRC/AUPRC_{\bar{}}$ for different MCC/MLC datasets. The distribution of the relative difference in performance between the SSL and SL methods across the different network architectures is summarized in the boxplots (one for each method and percentage of labeled data). Regions highlighted in green indicate that SSL outperforms SL, while red indicates the opposite.

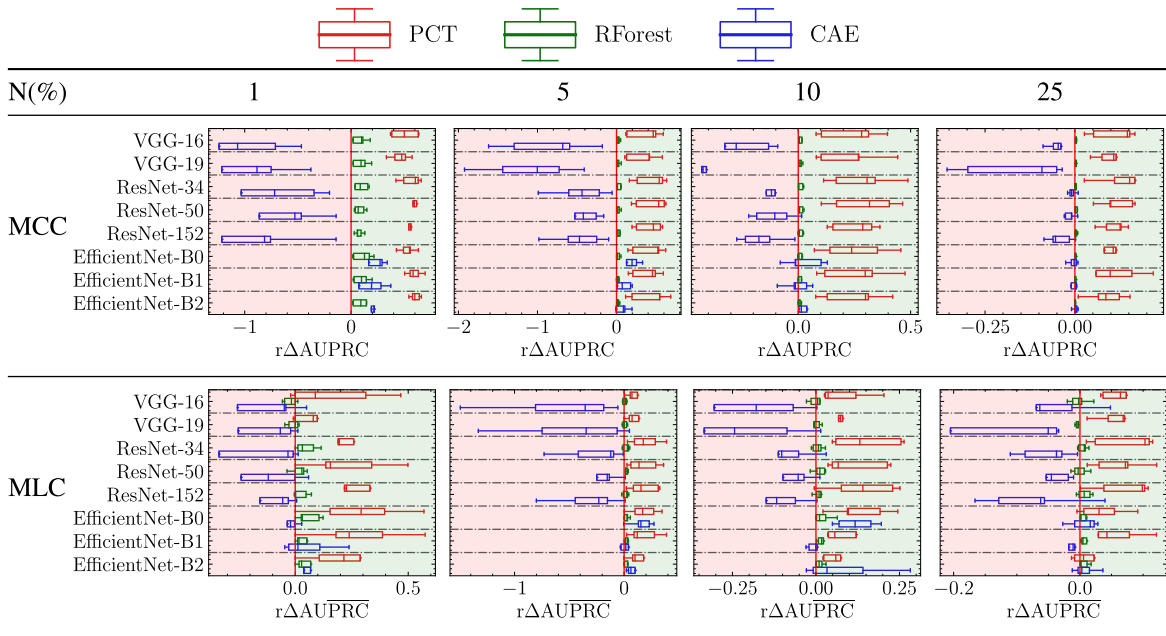


Fig. 8. Improvement of SSL performance compared to SL in terms of $AUPRC/AUPRC_{\bar{}}$ for various network architectures. The distribution of the relative difference in performance between the SSL and SL methods across the different MCC/MLC datasets is summarized in the boxplots (one for each method and percentage of labeled data). Regions highlighted in green indicate that SSL outperforms SL, while red indicates the opposite.

separately. The results of this analysis are visualized in Fig. 8.

The results demonstrate the consistent superiority of the tree-based semi-supervised methods, namely, SSL-PCTs and SSL-RForests, over their supervised counterparts, regardless of the specific network architecture employed. It is worth noting that SSL-PCTs exhibit a greater relative improvement compared to SSL-RForests.

In contrast, when considering the end-to-end learning method, the semi-supervised CAE method outperforms the supervised baseline only for the EfficientNet-based network architecture. For other network architectures, it actually degrades the performance of the supervised baseline. This suggests that the choice of network architecture plays an important role in achieving better predictive performance with an SSL end-to-end CAE.

TABLE II

PREDICTIVE PERFORMANCE OF THE PROPOSED SEMI-SUPERVISED METHODS (USING THE EFFICIENTNET-B2 NETWORK ARCHITECTURE) COMPARED AGAINST STATE-OF-THE-ART METHODS FROM THE LITERATURE. THE RESULTS ARE PRESENTED USING THE AUPRC MEASURE FOR MCC DATASETS AND THE AUPRC MEASURE FOR MLC DATASETS (HIGHER VALUES INDICATE BETTER PREDICTIVE PERFORMANCE). FOR EACH DATASET AND PERCENTAGE OF LABELED DATA, THE BEST PERFORMANCE IS HIGHLIGHTED IN BOLD

Dataset	N(%)	Baseline	CAE	MSMatch	HR-S ² DML	DINO	SimCLR	MoCo	SimSiam	SwAV	BYOL	BarlowT	SL-PCT	SSL-PCT	SL-RForest	SSL-RForest
MCC datasets																
OPT.-31	1	0.074	0.094	0.184	0.783	0.188	0.227	0.173	0.135	0.217	0.195	0.165	0.048	0.119	0.327	0.383
	5	0.208	0.259	0.277	0.816	0.591	0.693	0.545	0.466	0.659	0.625	0.503	0.18	0.573	0.841	0.877
	10	0.497	0.55	0.676	0.859	0.685	0.784	0.639	0.452	0.785	0.713	0.597	0.409	0.705	0.948	0.962
	25	0.86	0.867	0.890	0.863	0.895	0.925	0.9	0.852	0.919	0.915	0.877	0.616	0.728	0.976	0.976
UCM	1	0.145	0.186	0.385	0.904	0.333	0.361	0.314	0.265	0.335	0.376	0.311	0.085	0.251	0.539	0.593
	5	0.537	0.588	0.645	0.926	0.814	0.85	0.74	0.595	0.833	0.801	0.733	0.377	0.834	0.986	0.997
	10	0.765	0.795	0.905	0.919	0.918	0.945	0.905	0.816	0.926	0.924	0.866	0.584	0.849	0.997	0.998
	25	0.967	0.972	0.966	0.94	0.981	0.979	0.977	0.956	0.977	0.983	0.975	0.789	0.844	0.998	0.998
RSSCN7	1	0.427	0.527	0.304	0.869	0.187	0.173	0.155	0.171	0.206	0.175	0.161	0.342	0.81	0.96	0.979
	5	0.898	0.893	0.784	0.860	0.911	0.924	0.861	0.908	0.89	0.929	0.916	0.726	0.82	0.981	0.983
	10	0.951	0.953	0.813	0.897	0.953	0.946	0.944	0.942	0.933	0.957	0.951	0.77	0.835	0.984	0.984
	25	0.971	0.97	0.917	0.897	0.974	0.973	0.97	0.974	0.97	0.977	0.973	0.836	0.843	0.983	0.985
AID	1	0.126	0.178	0.405	0.917	0.317	0.37	0.271	0.243	0.354	0.336	0.272	0.076	0.211	0.455	0.531
	5	0.527	0.589	0.862	0.931	0.706	0.8	0.675	0.534	0.758	0.737	0.64	0.282	0.62	0.908	0.931
	10	0.788	0.799	0.963	0.934	0.871	0.914	0.856	0.711	0.878	0.898	0.832	0.484	0.691	0.979	0.988
	25	0.958	0.961	0.983	0.960	0.955	0.953	0.95	0.915	0.941	0.959	0.943	0.615	0.702	0.986	0.988
RESISC45	1	0.437	0.162	0.565	0.870	0.778	0.821	0.578	0.069	0.613	0.789	0.683	0.29	0.634	0.97	0.985
	5	0.915	0.683	0.891	0.885	0.928	0.938	0.912	0.401	0.885	0.944	0.918	0.605	0.752	0.987	0.989
	10	0.955	0.889	0.929	0.898	0.955	0.96	0.95	0.93	0.94	0.963	0.955	0.682	0.783	0.988	0.989
	25	0.977	0.968	0.936	0.922	0.976	0.978	0.976	0.971	0.974	0.979	0.976	0.735	0.804	0.989	0.99
MLC datasets																
Ankara	1	0.358	0.373	-	0.394	0.358	0.370	0.386	0.37	0.365	0.378	0.348	0.329	0.329	0.329	0.329
	5	0.355	0.37	-	0.347	0.378	0.359	0.387	0.385	0.371	0.364	0.374	0.339	0.339	0.393	0.405
	10	0.358	0.371	-	0.350	0.378	0.369	0.376	0.383	0.363	0.38	0.363	0.339	0.347	0.402	0.414
	25	0.371	0.37	-	0.388	0.365	0.376	0.368	0.37	0.37	0.375	0.386	0.358	0.353	0.427	0.434
UCM	1	0.218	0.213	-	0.785	0.477	0.483	0.418	0.407	0.467	0.48	0.453	0.282	0.36	0.599	0.609
	5	0.404	0.545	-	0.801	0.808	0.844	0.759	0.695	0.836	0.802	0.741	0.516	0.628	0.912	0.915
	10	0.494	0.689	-	0.853	0.881	0.91	0.86	0.806	0.891	0.893	0.851	0.637	0.688	0.953	0.953
	25	0.893	0.907	-	0.893	0.949	0.953	0.948	0.931	0.95	0.952	0.94	0.742	0.736	0.964	0.966
AID	1	0.389	0.404	-	0.622	0.508	0.469	0.412	0.439	0.411	0.475	0.494	0.411	0.46	0.675	0.725
	5	0.454	0.485	-	0.648	0.719	0.695	0.663	0.683	0.615	0.718	0.692	0.546	0.592	0.826	0.854
	10	0.604	0.599	-	0.695	0.766	0.768	0.744	0.761	0.708	0.776	0.76	0.623	0.638	0.851	0.866
	25	0.752	0.755	-	0.719	0.804	0.799	0.806	0.801	0.791	0.812	0.802	0.631	0.635	0.859	0.869
DFC-15	1	0.445	0.602	-	0.881	0.71	0.708	0.663	0.672	0.652	0.687	0.705	0.534	0.749	0.881	0.907
	5	0.796	0.813	-	0.897	0.847	0.858	0.878	0.828	0.833	0.88	0.882	0.768	0.853	0.977	0.975
	10	0.906	0.88	-	0.904	0.916	0.905	0.925	0.913	0.897	0.924	0.924	0.818	0.87	0.981	0.981
	25	0.959	0.947	-	0.933	0.948	0.947	0.956	0.95	0.953	0.954	0.953	0.857	0.876	0.983	0.983
MLRSNet	1	0.119	0.128	-	0.348	0.113	0.11	0.103	0.112	0.113	0.111	0.107	0.165	0.438	0.574	0.675
	5	0.268	0.298	-	0.358	0.433	0.509	0.462	0.434	0.48	0.589	0.477	0.498	0.755	0.898	0.93
	10	0.253	0.295	-	0.443	0.65	0.688	0.665	0.619	0.63	0.73	0.66	0.65	0.818	0.953	0.962
	25	0.443	0.46	-	0.537	0.813	0.82	0.827	0.801	0.792	0.842	0.801	0.774	0.834	0.971	0.972

E. Comparison to State-of-the-Art Methods

We have conducted a comprehensive performance comparison of the proposed semi-supervised methods against seven state-of-the-art self-supervised methods and two SSL methods for RSI classification (HR-S²DML and MSMatch).

The results of the performance comparison, based on the AUPRC measure for MCC datasets and AUPRC measure for MLC datasets, are presented in Table II. It is evident that the SSL-RForest consistently outperforms all the other state-of-the-art methods, with only a few exceptions. Namely, on a few of the datasets, HR-S²DML is the best performer for

1% of labeled data. The results also demonstrate consistent improvements of SSL-RForest over SL-RForest and SSL-PCT over SL-PCT across all percentages of labeled data and all datasets.

In Fig. 10, we present example predictions for several randomly sampled images from the test sets of the UCM and DFC-15 MLC datasets, with their true labels and labels predicted with different methods. These results indicate that the SSL-RForest method is the most precise among the other methods; it clearly has the fewest false positives and false negatives.

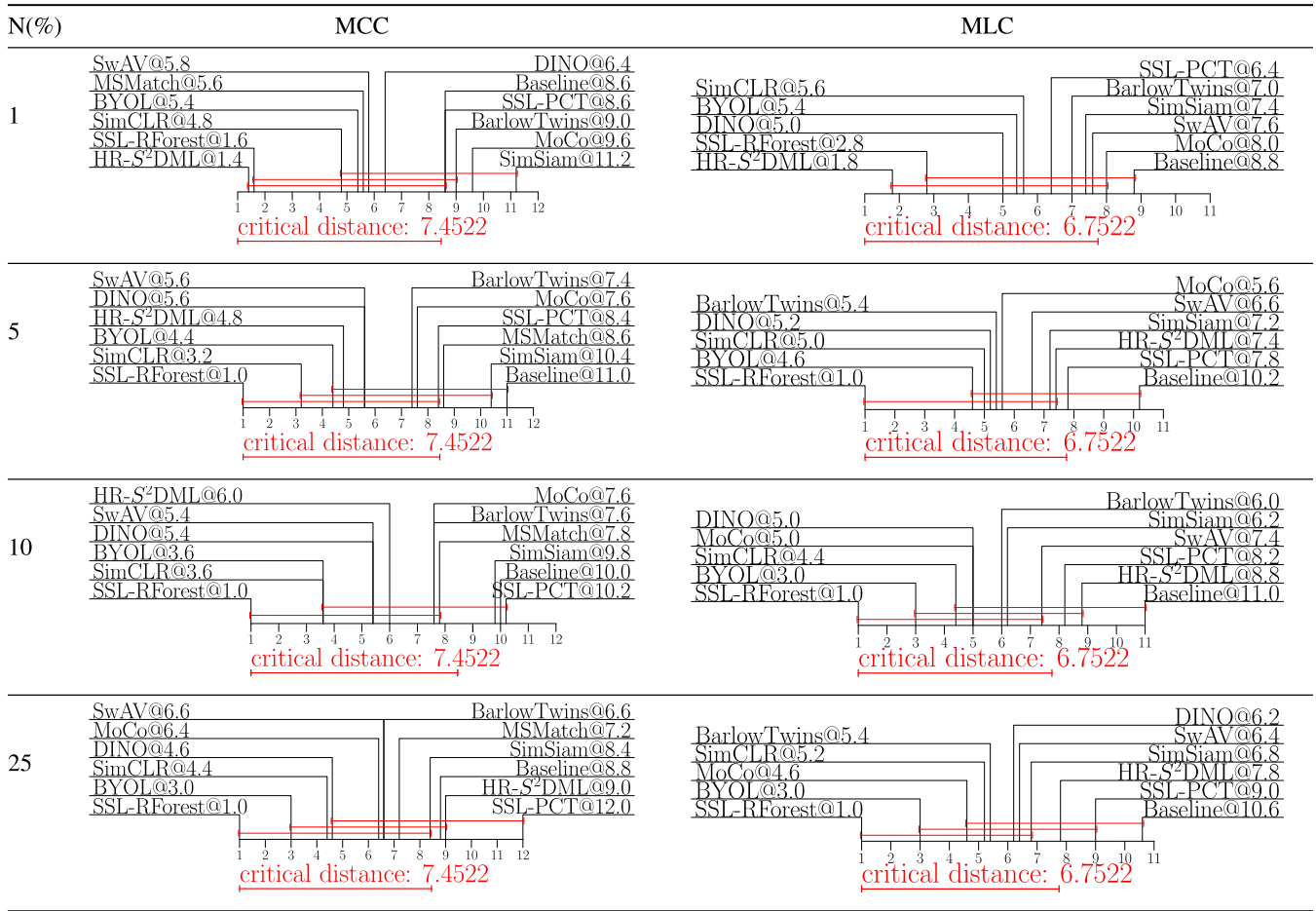


Fig. 9. Statistical analysis of the relative predictive performance of our proposed methods and state-of-the-art methods in terms of AUPRC for MCC datasets and AUPRC for MLC datasets considering the EfficientNet-B2 network architecture. The results are presented in the form of average rank diagrams, one for each percentage of labeled data. In the diagrams, the best ranking methods are positioned at the leftmost side, indicating their superior performance. The significance level is set at 0.05. If the methods are connected with a red line, the differences in performance among them are not statistically significant.







Samples	Ground Truth	Baseline	CAE	HR-S ² DML	SSL-PCT	SSL-RForest	Samples	Ground Truth	Baseline	CAE	HR-S ² DML	SSL-PCT	SSL-RForest
	bare-soil, pavement, trees, water	bare-soil, pavement, trees, water	bare-soil, pavement, trees, water	bare-soil, pavement, trees, water, grass	bare-soil, pavement, trees, water, grass	bare-soil, pavement, trees, water, grass		impervious, water, clutter	impervious, water, clutter, vegetation	impervious, water, clutter, vegetation	impervious, water, clutter, vegetation	impervious, water, clutter, vegetation	impervious, water, clutter
	bare-soil, buildings, cars, pavement, trees	bare-soil, buildings, cars, pavement, trees	bare-soil, buildings, cars, pavement, trees	bare-soil, buildings, cars, pavement, trees, grass	bare-soil, buildings, cars, pavement, trees, grass	bare-soil, buildings, cars, pavement, trees		impervious, vegetation, building, car	impervious, vegetation, building, car, car	impervious, vegetation, building, car, car	impervious, vegetation, building, car, car	impervious, vegetation, building, car, car	impervious, vegetation, building, car
	airplane, grass, pavement	airplane, grass, pavement, cars	airplane, grass, pavement, cars	airplane, grass, pavement, cars	airplane, grass, pavement, cars	airplane, grass, pavement		impervious, clutter	impervious, clutter, car	impervious, clutter, car	impervious, clutter, car	impervious, clutter, car	impervious, clutter

Fig. 10. Qualitative results of different methods based on 10% of labeled data on: (a) UCM dataset and (b) DFC-15 dataset for MLC of RSIs. The predictions shown in red indicate false positives. Cases shown in blue are false negatives (should have been predicted, but were not).

To determine the statistical significance of the performance differences among the SSL methods across the RSI datasets, we employed the Friedman test and Nemenyi posthoc analysis [70] with the significance level set at $\alpha = 0.05$. The

results are presented using average rank diagrams in Fig. 9, where the best-performing methods are positioned at the leftmost side of the diagram (with a rank of 1). Methods with performances that do not differ significantly are connected

with a red line. As depicted in the diagrams, the SSL-RForest is a clear best performer for all percentages of labeled data, except for 1%, where HR-S²DML is ranked the best. The HR-S²DML method loses its advantage as the amount of labeled data increases, with SimCLR and BYOL becoming the closest competitors to SSL-RForest. For >1% of labeled data, SSL-RForest statistically significantly outperforms the supervised end-to-end baseline.

V. CONCLUSION

In this study, we have proposed a framework for semi-supervised multi-label and MCC of RSIs. It consists of two main steps: feature representation learning using a backbone CNN model, to capture the semantically rich content of RSI, followed by semi-supervised PCTs and ensembles.

We have conducted an extensive experiment evaluation covering various aspects of the proposed approach. We use different RSI datasets for MLC and MCC. First, we evaluate the performance of the methods with different amounts of labeled data. Next, we analyze the dependence of their predictive performance on the different network architectures used as backbone feature extractors. Finally, we rigorously evaluate our method against several state-of-the-art self-supervised and SSL methods from the literature.

The results of our experiments demonstrate that the proposed semi-supervised approaches, SSL-PCT and SSL-RForest, clearly benefit from unlabeled data as they consistently outperform their supervised counterparts across different datasets, network architectures, and amounts of labeled data. This was not the case for the semi-supervised end-to-end CAE method, which exhibited improvement over its supervised learning counterpart, only for the EfficientNet-B2 network architecture. Next, the SSL-RForest method emerged as a clear winner among nine state-of-the-art self-supervised and semi-supervised deep learning methods (including BYOL [57], SimSiam [58], SimCLR [59], BarlowTwins [60], MoCo [61], DINO [62], SwAV [63], HR-S²DML [35], and MSMatch [23]). The simpler method, SSL-PCT, even if it cannot compete with the other methods in terms of predictive performance, efficiently builds interpretable models, i.e., classification trees (the only such method among the ones considered).

The findings of this study imply that our semi-supervised framework can effectively reduce the reliance on extensive supervision in RSI multiclass and MLC. This is particularly valuable in real-world scenarios where the cost of labeling is high. Building upon these promising results, future extensions of this work will focus on exploring the incorporation of hierarchical label information to further enhance the predictive performance of semi-supervised MLC of RSIs.

REFERENCES

- [1] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, "BigearthNet: A large-scale benchmark archive for remote sensing image understanding," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2019, pp. 5901–5904.
- [2] D. Hou, S. Wang, X. Tian, and H. Xing, "PCLUDA: A pseudo-label consistency learning-based unsupervised domain adaptation method for cross-domain optical remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5600314.
- [3] Y. Feng, J. Jiang, H. Xu, and J. Zheng, "Change detection on remote sensing images using dual-branch multilevel intertemporal network," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4401015.
- [4] J.-X. Wang, S. Chen, C. H. Q. Ding, J. Tang, and B. Luo, "Semi-supervised semantic segmentation of remote sensing images with iterative contrastive network," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
- [5] Y. Du, L. Du, Y. Guo, and Y. Shi, "Semisupervised SAR ship detection network via scene characteristic learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5201517.
- [6] I. Dimitrovski, I. Kitanovski, D. Kocov, and N. Simidjievski, "Current trends in deep learning for Earth observation: An open-source benchmark arena for image classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 197, pp. 18–35, Mar. 2023.
- [7] Y. Wang, C. M. Albrecht, N. A. A. Braham, L. Mou, and X. X. Zhu, "Self-supervised learning in remote sensing: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 4, pp. 213–247, Dec. 2022.
- [8] D. Hong et al., "Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks," *Remote Sens. Environ.*, vol. 299, Dec. 2023, Art. no. 113856.
- [9] D. Hong et al., "SpectralGPT: Spectral remote sensing foundation model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5227–5244, Aug. 2024.
- [10] A. Safonova, G. Ghazaryan, S. Stiller, M. Main-Knorn, C. Nendel, and M. Ryo, "Ten deep learning techniques to address small data problems with remote sensing," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 125, Dec. 2023, Art. no. 103569.
- [11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. CVPR*, Jun. 2009, pp. 248–255.
- [12] M. Stojimchev, D. Kocov, and S. Džeroski, "Deep network architectures as feature extractors for multi-label classification of remote sensing images," *Remote Sens.*, vol. 15, no. 2, p. 538, Jan. 2023.
- [13] Y. Huang et al., "Two-branch attention adversarial domain adaptation network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5540813.
- [14] S. Chen and W. Guo, "Auto-encoders in deep learning—A review with new perspectives," *Mathematics*, vol. 11, no. 8, p. 1777, Apr. 2023.
- [15] J. Xiao, Y. Bai, A. Yuille, and Z. Zhou, "Delving into masked autoencoders for multi-label thorax disease classification," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 3577–3589.
- [16] P. Ma, Z. He, W. Ran, and H. Lu, "A transferable generative framework for multi-label zero-shot learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 5, pp. 3409–3423, May 2024.
- [17] Q. Liu, J. Peng, G. Zhang, W. Sun, and Q. Du, "Deep contrastive learning network for small-sample hyperspectral image classification," *J. Remote Sens.*, vol. 3, p. 25, Jan. 2023.
- [18] K. Sohn et al., "FixMatch: Simplifying semi-supervised learning with consistency and confidence," 2020, *arXiv:2001.07685*.
- [19] B. Chen, P. Li, X. Chen, B. Wang, L. Zhang, and X.-S. Hua, "Dense learning based semi-supervised object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4805–4814.
- [20] J. Zhuang, Z. Wang, and Y. Gao, "Semi-supervised video semantic segmentation with inter-frame feature reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 3253–3261.
- [21] Q. Liu, M. He, Y. Kuang, L. Wu, J. Yue, and L. Fang, "A multi-level label-aware semi-supervised framework for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–12, 2023.
- [22] D. Mishra and O. Hadar, "CLSR: Contrastive learning for semi-supervised remote sensing image super-resolution," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.
- [23] P. Gómez and G. Meoni, "MSMatch: Semisupervised multispectral scene classification with few labels," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 11643–11654, 2021.

- [24] W. Huang, Y. Shi, Z. Xiong, Q. Wang, and X. X. Zhu, "Semi-supervised bidirectional alignment for remote sensing cross-domain scene classification," *JSPRS J. Photogramm. Remote Sens.*, vol. 195, pp. 192–203, Jan. 2023.
- [25] W. Miao, J. Geng, and W. Jiang, "Semi-supervised remote-sensing image scene classification using representation consistency Siamese network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5616614.
- [26] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. Raffel, "MixMatch: A holistic approach to semi-supervised learning," 2019, *arXiv:1905.02249*.
- [27] Y. Wang et al., "FreeMatch: Self-adaptive thresholding for semi-supervised learning," 2023, *arXiv:2205.07246*.
- [28] Y. Oh, D.-J. Kim, and I. S. Kweon, "DASO: Distribution-aware semantics-oriented pseudo-label for imbalanced semi-supervised learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 9786–9796.
- [29] B. Zhang et al., "FlexMatch: Boosting semi-supervised learning with curriculum pseudo labeling," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 18408–18419.
- [30] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves ImageNet classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 10687–10698.
- [31] I. Nassar, S. Herath, E. Abbasnejad, W. Buntine, and G. Haffari, "All labels are not created equal: Enhancing semi-supervision via label grouping and co-training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7237–7246.
- [32] S. Qiao, W. Shen, Z. Zhang, B. Wang, and A. Yuille, "Deep co-training for semi-supervised image recognition," in *Computer Vision—ECCV*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., Cham, Switzerland: Springer, 2018, pp. 142–159.
- [33] M. Zheng, S. You, L. Huang, F. Wang, C. Qian, and C. Xu, "SimMatch: Semi-supervised learning with similarity matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14471–14481.
- [34] J. Li, C. Xiong, and S. C. Hoi, "Semi-supervised learning with contrastive graph regularization," in *Proc. ICCV*, 2021, pp. 9475–9484.
- [35] J. Kang, R. Fernández-Beltrán, Z. Ye, X. Tong, P. Ghamisi, and A. Plaza, "High-rankness regularized semi-supervised deep metric learning for remote sensing imagery," *Remote Sens.*, vol. 12, no. 16, p. 2603, Aug. 2020.
- [36] C. Hu, X.-J. Wu, and J. Kittler, "Semi-supervised learning based on GAN with mean and variance feature matching," *IEEE Trans. Cognit. Develop. Syst.*, vol. 11, no. 4, pp. 539–547, Dec. 2019.
- [37] Z. Dai, Z. Yang, F. Yang, W. W. Cohen, and R. R. Salakhutdinov, "Good semi-supervised learning that requires a bad GAN," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
- [38] R. Sharma, S. Mashkaria, and S. P. Awate, "A semi-supervised generalized VAE framework for abnormality detection using one-class classification," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 1302–1310.
- [39] B. Chaudhuri, B. Demir, S. Chaudhuri, and L. Bruzzone, "Multilabel remote sensing image retrieval using a semisupervised graph-theoretic method," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 1144–1158, Feb. 2018.
- [40] T. Üstüncök and M. Karakaya, "SS-MLA: A semisupervised method for multi-label annotation of remotely sensed images," *J. Appl. Remote Sens.*, vol. 15, no. 3, Aug. 2021, Art. no. 036509.
- [41] J. Levatić, M. Ceci, D. Kocev, and S. Džeroski, "Semi-supervised classification trees," *J. Intell. Inf. Syst.*, vol. 49, no. 3, pp. 461–486, Dec. 2017.
- [42] J. Levatić, M. Ceci, D. Kocev, and S. Džeroski, "Semi-supervised predictive clustering trees for (hierarchical) multi-label classification," *Int. J. Intell. Syst.*, vol. 2024, Apr. 2024, Art. no. e5610291. [Online]. Available: <https://www.hindawi.com/journals/ijis/2024/5610291/>
- [43] Y.-F. Li and D.-M. Liang, "Safe semi-supervised learning: A brief introduction," *Frontiers Comput. Sci.*, vol. 13, no. 4, pp. 669–676, Aug. 2019.
- [44] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1155–1167, Feb. 2019.
- [45] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.
- [46] Q. Zou, L. Ni, T. Zhang, and Q. Wang, "Deep learning based feature selection for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 11, pp. 2321–2325, Nov. 2015.
- [47] G.-S. Xia et al., "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017.
- [48] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, Nov. 2010, pp. 270–279.
- [49] Y. Hua, L. Mou, and X. X. Zhu, "Relation network for multilabel aerial image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4558–4572, Jul. 2020.
- [50] F. Ömrüüzun, B. Demir, L. Bruzzone, and Y. Y. Çetin, "Content based hyperspectral image retrieval using bag of endmembers image descriptors," in *Proc. 8th Workshop Hyperspectral Image Signal Process., Evol. Remote Sens. (WHISPERS)*, Aug. 2016, pp. 1–4.
- [51] Y. Hua, L. Mou, and X. X. Zhu, "Recurrently exploring class-wise attention in a hybrid convolutional and bidirectional LSTM network for multi-label aerial image classification," *JSPRS J. Photogramm. Remote Sens.*, vol. 149, pp. 188–199, Mar. 2019.
- [52] X. Qi et al., "MLRSNet: A multi-label high spatial resolution remote sensing dataset for semantic scene understanding," *JSPRS J. Photogramm. Remote Sens.*, vol. 169, pp. 337–350, Nov. 2020.
- [53] X.-Z. Wu and Z.-H. Zhou, "A unified view of multi-label performance measures," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3780–3788.
- [54] F. Pedregosa, S. Varoquaux, A. Gramfort, V. Michel, and B. Thirion, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Dec. 2011.
- [55] S. Cheng, H. Chen, P. Yao, and L. Song, "MLAE: A pretraining method for automatic identification of urban public space," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.
- [56] C. Qiu, A. Yu, X. Yi, N. Guan, D. Shi, and X. Tong, "Open self-supervised features for remote-sensing image scene classification using very few samples," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.
- [57] J.-B. Grill et al., "Bootstrap your own latent: A new approach to self-supervised learning," 2020, *arXiv:2006.07733*.
- [58] X. Chen and K. He, "Exploring simple Siamese representation learning," 2020, *arXiv:2011.10566*.
- [59] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," 2020, *arXiv:2002.05709*.
- [60] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," 2021, *arXiv:2103.03230*.
- [61] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," 2019, *arXiv:1911.05722*.
- [62] M. Caron et al., "Emerging properties in self-supervised vision transformers," 2021, *arXiv:2104.14294*.
- [63] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," 2020, *arXiv:2006.09882*.
- [64] X. Zhai, A. Oliver, A. Kolesnikov, and L. Beyer, "S⁴L: Self-supervised semi-supervised learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1476–1485.
- [65] (2023). *PyTorch*. Accessed: Jun. 1, 2023. [Online]. Available: <https://pytorch.org/>
- [66] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
- [67] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [68] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.
- [69] M. Petković, J. Levatić, D. Kocev, M. Breskvar, and S. Džeroski, "CLUSplus: A decision tree-based framework for predicting structured outputs," *SoftwareX*, vol. 24, Dec. 2023, Art. no. 101526.
- [70] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, Dec. 2006.



Marjan Stoimchev received the M.S. degree from the Faculty of Electrical Engineering, University of Ljubljana, Ljubljana, Slovenia, in 2021. He is currently pursuing the Ph.D. degree with the Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana.

His main research interests include computer vision, specifically in deep learning, classical machine learning, the combination of both, and multilabel classification of remote sensing images with supervised and semisupervised learning.



Jurica Levatić received the Ph.D. degree in semisupervised learning for structured output prediction from the Jožef Stefan Institute Postgraduate School, Ljubljana, Slovenia, in 2017.

From 2018 to 2021, he was a Marie Skłodowska-Curie Post-Doctoral Fellow with the Institute for Research in Biomedicine, Barcelona, Spain. He is currently a Post-Doctoral Researcher with the Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana.

His research interests include cancer biology semisupervised learning and the application of machine learning to solve practically relevant problems.



Dragi Kocev received the Ph.D. degree in learning ensemble models for predicting structured outputs from the Jožef Stefan Institute Postgraduate School, Ljubljana, Slovenia, in 2011.

He was a Visiting Research Fellow at the University of Bari, Bari, Italy, from 2014 to 2015. He is currently a Researcher with the Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana. He was involved with the EU-funded projects TRUSTroke, IQ, PHAGOSYS, and HBP. He was a Co-Coordinator of the FP7 FET Open

Project MAESTRA and a Coordinator of ESA-funded activities AiTLAS and GalaxAI.

Dr. Kocev is a member of the PC for several conferences (e.g., DS, ECML PKDD, AAAI, IJCAI, and KDD). He served as the PC Co-Chair for DS 2014 and the Journal Track Co-Chair for ECML PKDD 2017. He is an Action Editor of *Expert Systems with Applications, Machine Learning, and Data Mining and Knowledge Discovery*.



Sašo Džeroski is currently the Head of the Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia. His department develops artificial intelligence methods for machine learning and decision support and uses them to solve practical problems from agriculture and environmental sciences, medicine and life sciences, and space operations/Earth observation. His research focuses on artificial intelligence for science, including machine learning from complex data and in the presence of domain knowledge, learning models

of dynamical systems, and developing ontologies for describing different branches of computer science research.

Dr. Džeroski is a member of the Macedonian Academy of Sciences and Arts and Academia Europea (European Academy of Humanities, Letters and Sciences). He is a fellow of EurAI, European Association of Artificial Intelligence, awarded in recognition for his "Pioneering Work in AI and Outstanding Service for the European AI Community." He is currently the Vice President of Slovenian Artificial Intelligence Society.