Dataset Similarity to Assess Semi-supervised Learning Under Distribution Mismatch Between the Labelled and Unlabelled Datasets

Saul Calderon-Ramirez*, Luis Oala*, Jordina Torrents-Barrena, Shengxiang Yang, David Elizondo, Armaghan Moemeni Simon Colreavy-Donnelly, Wojciech Samek, Miguel A. Molina-Cabello and Ezequiel López-Rubio

***Abstract*—Semi-supervised deep learning (SSDL) is a popular strategy to leverage unlabelled data for machine learning when labelled data is not readily available. In real-world scenarios, different unlabelled data sources are usually available, with varying degrees of distribution mismatch regarding the labelled datasets. It begs the question which unlabelled dataset to choose for good SSDL outcomes. Oftentimes, semantic heuristics are used to match unlabelled data with labelled data. However, a quantitative and systematic approach to this selection problem would be preferable. In this work, we first test the SSDL MixMatch algorithm under various distribution mismatch configurations to study the impact on SSDL accuracy. Then, we propose a quantitative unlabelled dataset selection heuristic based on dataset dissimilarity measures. These are designed to systematically assess how distribution mismatch between the labelled and unlabelled datasets affects MixMatch performance. We refer to our proposed method as deep dataset dissimilarity measures (DeDiMs), designed to compare labelled and unlabelled datasets. They use the feature space of a generic Wide-ResNet, can be applied prior to learning, are quick to evaluate and model agnostic. The strong correlation in our tests between MixMatch accuracy and the proposed DeDiMs suggests that this approach can be a good fit for quantitatively ranking different unlabelled datasets prior to SSDL training.**

***Impact Statement*—Semi-supervised deep learning is a technique for training a deep learning model when few labelled observations are available, leveraging unlabelled datasets. Different unlabelled data sources may be available, introducing the possibility for distribution mismatches between the labelled and unlabelled datasets. In this work we assess the impact of distribution mismatches on the outcomes of the semi-supervised MixMatch algorithm. We propose a set of simple feature-space density dataset distances, referred to as deep dataset dissimilarity measures (DeDiMs). In our extensive test-bed, the evaluated DeDiMs yield linear correlation coefficients of up to 96% to MixMatch accuracy.**

***Index Terms*—Semi-supervised deep learning, MixMatch, Out of distribution data, Deep learning, Distribution mismatch, Dataset similarity**

Manuscript received June 22, 2021; revised November 29, 2021 and May 3, 2022.

S. Calderon-Ramirez, D. Elizondo, S. Yang, and S. Colreavy-Donnelly are with the Institute of Artificial Intelligence (IAI), De Montfort University, Leicester LE1 9BH, United Kingdom, sacalderon@itcr.ac.cr, {delizondo,syang,simon.colreavy-donnelly}@dmu.ac.uk.

L. Oala and W. Samek are with the Department of Artificial Intelligence, Fraunhofer Heinrich Hertz Institute, Germany, {luis.oala,wojciech.samek}@hhi.fraunhofer.de.

A. Moemeni is with the School of Computer Science, University of Nottingham, Nottingham NG8 1BB, United Kingdom, armaghan.moemeni@nottingham.ac.uk.

J. Torrents-Barrena works with HP Inc., Brcelona Spain, 08174 jordina.torrents.barrena@hp.com.

M. Molina-Cabello and E. López-Rubio are with the University of Málaga, Spain, {miguelangel,ezeqlr}@lcc.uma.es.

This paragraph will include the Associate Editor who handled your paper.

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors.

*Equal contribution

I. Introduction

Training an effective deep learning solution typically requires a considerable amount of labelled data. In specific areas, like medical imaging technologies, high quality labelled data can be expensive to obtain, leading to a paucity of labelled data [4], [12]. Several approaches have been developed to address this data constraint, including data augmentation, transfer, weakly and semi-supervised learning, among others [34], [46]. Semi-supervised learning is an approach for learning problems where little labelled data is available, or a range of labels is lacking. It leverages the use of unlabelled data which is often cheap to obtain [44]. Formally, in a semi-supervised setting both labelled and unlabelled datasets are used. Labelled observations $X_l = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_l}\}$ and their corresponding labels $Y_l = \{y_1, \ldots, y_{n_l}\}$ make up the labelled dataset S_l. The set of unlabelled observations S_u is represented as $X_u = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_u}\}$, therefore $S_u = X_u$. Semi-supervised deep learning (SSDL) approaches can be grouped into pre-training [14], self-training or pseudo-labelled [15] and regularization-based. Regularization techniques include generative based approaches, along consistency loss term and graph based regularization [12]. A detailed survey on semi-supervised learning can be found in [44].

The practical implementation of SSDL techniques in different contexts has been limited; barring few exceptions [32]. As with other learning paradigms, the transfer of SSDL techniques from lab to real-world is complicated by, among other reasons, the violation of the Independent and Identically Distributed (IID) assumption. In principle, we would like to exploit available unlabelled data as flexibly as possible. In practice, distribution mismatches between the labelled and unlabelled data sets can lead to serious performance degradation [32]. The following example illustrates this problem. We can train a Convolutional Neural Network (CNN) to classify chest X-ray images between COVID-19 ill and healthy patients, as for example seen in [7]. The labelled dataset S_l can include a limited number of observations for each class. However, the unlabelled dataset S_u can include observations of patients with other lung pathologies

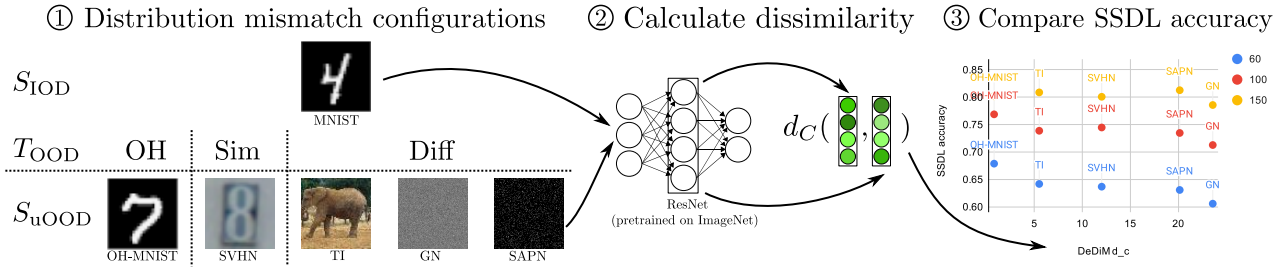

Fig. 1: A summary of the workflow presented in this paper. In step ① a labelled, inside-of-distribution dataset S_{IOD}, here MNIST, is paired with different potential unlabelled datasets for semi-supervised learning. The unlabelled data S_{uOOD} in our experiments is of the three types T_{OOD} other half (OH), similar (Sim) and different (Diff). In step ②, a pretrained ResNet is used to extract feature representations of the labelled and unlabelled datasets and a deep dataset dissimilarity measure (DeDiM) is applied. Finally, in step ③ the dissimilarity scores can be used as a proxy for SSDL accuracy to select unlabelled data. This example shows results from the MNIST S_{IOD} experiment. The colors in the last scatter plot designate the number of labelled samples.

not sampled in S_l, leading to a distribution mismatch between the labelled and unlabelled datasets. The mismatching data can be described as Out of Distribution (OOD) data [23] and it can harm the performance of a SSDL solution [32].

It begs the question how we can systematically select labelled and unlabelled data in non-IID settings such that performance on the downstream task is increased. A common recourse are what we call semantic matching heuristics. For example, Tiny ImageNet (TI) may be judged more similar to the Canadian Institute for Advanced Research dataset of 10 classes (CIFAR-10) than to Modified National Institute of Standards and Technology dataset (MNIST) because the first two datasets both contain object whereas the last dataset contains handwritten digits. Practices of semantic matching can be traced to other fields of machine learning, too, including out-of-distribution detection [52] or the domain adaptation literature [50], [47]. Insights from generative modelling should, at the very least, make us feel uneasy about such an approach to determine dataset similarity. Similarity can vary drastically depending on whether it is determined through semantic heuristics or quantified through the lens of a machine learning model [28].

A. Problem statement

The central premise of this work is the quantitative impact assessment of distribution mismatch between labelled and unlabelled data on SSDL. This notion stipulates that a mismatch negatively affects the accuracy of models trained with SSDL algorithms [32]. Distribution mismatch occurs when the unlabelled data contains observations that do not correspond to or are too dissimilar to the observations of any of the classes present in the labelled data. It is not clear though what exactly the effect is when this mismatch occurs:

- Does it always harm the model accuracy in the context of SSDL?
- Does it help to use unlabelled data that is, supposedly, semantically more similar to the labelled data?
- Furthermore, if certain unlabelled datasets indeed harm accuracy of SSDL trained models, is there a reliable way to select the unlabelled data in an informed way prior to SSDL training?

We adopt the following definitions. Given a dataset S_1 emanating from the data generating process $y = f(\boldsymbol{x})$, with $y \in \mathcal{Y} := \{1, ..., K\}$ being a set of labels, and a second dataset S_2 emanating from the data generating process $y' = g(\boldsymbol{x})$, with $y' \in \mathcal{Y}' := \{1, ..., K'\}$, we define the following concepts:

Definition 1. *Inside of Distribution (IOD) data: Dataset S_2 is IOD relative to dataset S_1 if $f(\boldsymbol{x}) = g(\boldsymbol{x})$. In particular, we must have that $\mathcal{Y} = \mathcal{Y}'$.*

Definition 2. *OOD data: Dataset S_2 is OOD relative to dataset S_1 if $f(\boldsymbol{x}) \neq g(\boldsymbol{x})$. In particular, we may have that $\mathcal{Y} \neq \mathcal{Y}'$.*

Definition 3. *Distribution mismatch in SSDL: A distribution mismatch occurs if the unlabelled data S_u used for SSDL is OOD relative to the labelled data S_l.*

In practice, $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$ are typically not known explicitly. Thus, given two datasets S_1 and S_2 a definite formal verification of the distribution mismatch property is not possible. Instead, it is usually *assumed* that two *different* datasets, e.g., CIFAR-10 and MNIST, derive from different data generative processes. This working definition of OOD data follows the existing literature on distribution mismatch in SSDL [32] as well as OOD detection in deep learning [37]. We adopt this working definition for the OOD scenarios of our test bed. Note that different degrees of OOD contamination for S_u are possible as we describe in Section IV-A.

B. Contribution

In order to address the questions outlined in Section I-A we first study the effect of distribution mismatch on SSDL accuracy in systematic test-bed. Then, we present a set of Deep Dataset Dissimilarity Measure (DeDiM)s to assess, *prior to training*, the effectiveness of unlabelled datasets for MixMatch SSDL [5]. A visual summary of the process is provided in Figure 1. All code and experimental scripts, with automatic download of test bed data for ease of reproduction, is made publicly available[1]. It entails the following contributions:

- We present and make available a comprehensive simulation sandbox, called *non-IID-SSDL*, for stress testing SSDL

[1]https://github.com/luisoala/non-iid-ssdl

algorithms under various non-IID (distribution mismatch) configurations. We demonstrate that including OOD data in the unlabelled training dataset for the MixMatch algorithm can yield different degrees of accuracy degradation compared to the exclusive use of IOD data. However, in most cases, using unlabelled data with OOD contamination still improves the results when compared to the default fully supervised configuration.
- Markedly, unlabelled that is supposedly semantically similar to the IOD labelled data does not always lead to the highest accuracy gain. This counter-intuitive result suggests that using semantically similar unlabelled datasets does not always yield the best accuracy gain for SSDL.
- We propose and evaluate four DeDiMs that can be used to rank unlabelled datasets according to the expected accuracy gain *prior* to SSDL training. They can be considered to be less expensive to compute and model agnostic, which make them amenable for practical application.
- Our test results reveal a strong correlation between the tested DeDiMs and MixMatch accuracy, making them useful for unlabelled dataset selection. Therefore, we propose the usage of the tested DeDiMs to select the unlabelled dataset for improved MixMatch accuracy. The best performing DeDiMs use a non-parametric density function approximation of the feature space, which provides a method to quantitatively describe the distribution mismatch between two datasets.

II. Related work

In this work we address a combination of three overlapping problems that are often dealt with separately in the literature: OOD detection, distribution mismatch in SSDL, and dataset dissimilarity measures.

A. OOD data detection

In the context of machine learning, OOD data detection refers to the general problem of detecting observations that belong to a data distribution different from the distribution of the training data [18]. OOD detection can be considered as a generalization of outlier detection, since it considers individual and collective outliers [40]. Further variations of the OOD data detection problem are novel and anomaly data detection [33], with different applications such as rare event detection and artificial intelligence safety [17], [1]. Classical OOD and anomaly detection methods rely on density estimation, e.g., Gaussian Mixture Models [24], robust moment estimation, like the Minimum Covariance Determinant method [38], prototyping, e.g., k-nearest neighbor algorithm [24], as well as kernel based variants such as Support Vector Data Description [43]. Also, a variety of neural network based approaches for novelty detection can be found [24], implementing a more data-oriented approach.

With the success of deep learning, recent works have addressed the generic problem of discriminative detection of OOD data for deep learning architectures. In general, discriminative OOD detectors can be categorized in output- and feature-based. For instance, a simple output based OOD detection approach was proposed in [18]. The authors framed OOD detection as a prediction confidence estimation problem. The proposed method relies on the Softmax output, sampling the maximum value. [23] introduced OOD data detection in neural networks using input perturbations. A *temperature* coefficient T is used in the calculation of the Softmax output with a calibrated decision threshold δ for OOD data detection.

More recently, in [22] authors argue that deep neural networks with Softmax output layers are over-confident for inputs dissimilar from the training data and hence propose the usage of the Mahalanobis distance in latent space. Similarly [41] also exploit latent representations, defining what they refer to as learning certificates: neural networks that map feature vectors to zero for IOD data. A more challenging OOD detection setting was tested, where half of each tested dataset is used as IOD data, and the other half is used as OOD data, making OOD detection harder. [52] proposes an OOD detector using the feature space as well. The approach fits different parametric distributions in the feature space of the data. The decision to discriminate between OOD and IOD data is done based on the estimation of the approximated parametric model. Unfortunately, no comparison with other popular OOD methods was presented. A similar approach with a simpler linear model trained with the statistical moments of the feature space can be found in [35].

In this concise overview of OOD detection methods, two different main categories for OOD detection can be found: output and feature space based. The datasets selected for benchmarking OOD detection methods are usually different for each work, and quantitative evaluation of the *difficulty* of performing OOD detection is rare.

B. Distribution mismatch in SSDL

The distribution mismatch between S_u and S_l can be interpreted as a violation of the IID assumption. Different causes for this distribution mismatch can be distinguished, as discussed in [19]. We summarize them as follows:

- Prior probability shift: The density of the targets in S_l is different to the real target densities in S_u (increasing the possibility of sampling noise). Class imbalance in the labelled dataset S_l is a special case of this setting, as discussed in [8].
- Covariate shift: The labelled dataset S_l might sample a different density of the features when compared to the unlabelled dataset S_u, causing a distribution mismatch between the two datasets. For example, for handwritten digit recognition, the sample of S_l might capture different stroke widths, when compared to S_u. Concept drift is a similar setting where the change of features causes the concept to semantically change.
- Concept shift: It corresponds to a label change for a similar set of features. For instance, for sentiment analysis in audio, an observation might have different labels depending on the labeler (this is also related to label noise). In the context of distribution mismatch between S_l and S_u, as no label information is used from S_u during training.

In this work, we analyze the impact of distribution mismatch between S_l and S_u caused by a concept drift, as a mild

distribution mismatch cause (for instance using SVHN as S_u and MNIST as S_l). To create more significant distribution mismatch settings, we contaminate the unlabelled dataset S_u with different percentages of observations from completely different datasets (with different labels or features). For example, using MNIST as S_l and for S_u 50% Gaussian Noise (GN) images plus 50% MNIST images.

As previously highlighted, in [32] the authors call for the need of a more extensive testing of SSDL techniques in real-world testing scenarios. One of them is the possible data distribution mismatch between the labelled and unlabelled training data can adversely impact SSDL results. Real Mix was proposed [27] in response, implementing a masking coefficient to OOD data for the unlabelled dataset. The masking coefficient is used as a threshold of the Softmax output of the model, discarding unlabelled data used only in the unsupervised term. The authors performed limited testing on the significance of using OOD unlabelled data, with relatively few OOD contamination scenarios tested. The OOD dataset consisted of the splitted CIFAR-10 dataset, in two halves with different semantics. A total of four levels of OOD contamination were tested. We extend OOD datasets to more configurations.

More recently, the work in [11] proposes a simple approach to deal with OOD data, by using soft labels averaged by the output of the model along a number of epochs. The evaluation includes a benchmark with different proportions of distribution mismatch. The results yielded demonstrate an improved accuracy of the proposed method over other state of the art SSDL approaches when dealing with OOD data in the unlabelled dataset. However, MixMatch is not among the compared approaches. Moreover, the distribution mismatch scenarios were not extensive, testing only different degrees of mismatch contamination, and not evaluating the impact of different OOD data sources.

In [51] a SSDL robust framework to OOD data was proposed. Authors claim that OOD data far away from the decision boundaries affects SSDL performance less than OOD data lying very close to the decision boundaries. However, no explicit quantitative measure of distribution similarity was used. The authors also noted a high influence of data batch-normalization, where normalizing the data using far away OOD data can impact the accuracy of the model more. To address this issue, the authors proposed a dynamic approach to re-weight the observations in both batch-normalization and training time, using a gradient optimization approach for both. The model was tested using virtual adversarial training and the Π model, excluding the usage of MixMatch. The experiments included different degrees of OOD contamination and unlabelled datasets, however no comparison to other approaches explicitly designed for SSDL with OOD robustness was performed.

In [11] another approach for OOD robust SSDL was proposed, using also a per observation re-weighting and giving less weight to the observations that are most likely OOD. To calculate the per-observation weights, an uncertainty proxy, as in [16], was implemented, using an ensemble of models yielded during the past epochs. The model was tested with the CIFAR-10 dataset (6 classes) with a varying degrees of OOD contamination (the other 4 classes left from CIFAR-10). No other unlabelled contamination data-sources were used.

Unlike previous studies, in this work we aim to quantify the notion of OOD data, correlating it with the SSDL accuracy using different unlabelled datasets with varying degrees of OOD contamination and different data sources. This quantification can be used to select one unlabelled dataset among many, *prior* to SSDL training. This also allows us to analyze the influence of OOD data. Finally, the proposed method can be extended to weight how harmful an unlabelled observation can be for SSDL. Using the feature distribution to this end has not been fully explored in previous work.

C. Dataset dissimilarity measures

The need of comparing two datasets, in this case the labelled S_l and unlabelled datasets S_u to quantify the prior data mismatch between them, leads us to the need for dataset comparison measures. Computing a notion of dissimilarity between two sets of points (also known as shape matching [25]) is typically computationally more expensive than calculating the dissimilarity between a set of points and another single point. Strategies to reduce this burden are primarily centered around enriching the object space with a probability measure which helps guide attention to important areas of comparison [25]. When starting with raw datasets, as is typically the case when trying to decide which data to use for SSDL, additional pre-processing or modelling steps would be necessary to obtain this probability measure. Methods explicitly designed to compute dissimilarities between raw datasets for deep learning are, to the best of our knowledge, rare. In [42] authors define a dissimilarity measure based on the Euclidean distance between the frequency of a given feature function on two datasets, referred as the constrained measure distance. The calculation of the proposed measure can be efficiently performed using the covariance matrix of the feature function in the dataset.

More recently, authors in [6] proposed a distance dissimilarity index based on the statistical significance difference of the distance distributions between the two datasets. To calculate it, each data point in the test set is matched with the training data. After exchanging the associated observations, changes in the topology are assessed, using the distance distribution. The confidence p-value of the difference between the two distributions is calculated and used as a dissimilarity measure.

Note that our requirements differ from the above OOD detection and dissimilarity measure methods: we are interested in computationally inexpensive, prior-to-training and SSDL model agnostic quantification of the OOD degree between two *datasets*. Approaches that are computationally expensive or retrospective, applied after the model has been trained, are not feasible to address distribution mismatch before SSDL training. Closest to our work are the OOD detection ideas developed by [37]. The authors present introductory experiments on the correlation between OOD detection and the dataset dissimilarity using a genome distance [36]. We explore a similar comparison: the relationship between SSDL accuracy and OOD-IOD dissimilarity, which can be useful for a prior evaluation of unlabelled datasets for SSDL. This enables an

interesting quantitative insight on the real impact of OOD data to SSDL accuracy, which we explore in this work.

III. Proposed method

Our approach is based on a simple idea: if OOD data indeed affects MixMatch SSDL accuracy we would like to be able to select the unlabelled data *prior* to SSDL training such that resulting test accuracy of the model is maximized. To that end we propose and evaluate a number of DeDiMs. They provide a quantitative notion of similarity between the inputs of the IOD labelled data and the inputs of the OOD unlabelled data. The DeDiMs are based on dataset subsampling, as image datasets are usually large, following a sampling approach for comparing two populations, as seen in [21]. We compute the dissimilarity measures in the feature space of a generic Wide-ResNet pre-trained on ImageNet, making our proposed approach agnostic to the SSDL model to be trained. This enables an evaluation of the unlabelled data before training the SSDL model. The proposed measures in this work are meant to be simple and quick to evaluate with practical use in mind. We propose and test the implementation of two Minkowski based distance sets, $d_{\ell_2}(S_a, S_b, \tau, \mathcal{C})$ and $d_{\ell_1}(S_a, S_b, \tau, \mathcal{C})$, corresponding to the Euclidean and Manhattan distances, respectively, between two datasets S_a and S_b. Additionally, we implement and test two non-parametric density based dataset divergence measures; Jensen-Shannon (d_{JS}) and cosine distance (d_C). For all the proposed dissimilarity measures, the parameter τ defines the sub-sample size used to compute the dissimilarity between the two datasets S_a and S_b, and $\mathcal{C}$ the total number of samples to compute the mean sampled dissimilarity measure. The general procedure for all the implemented distances is detailed as follows.

- We randomly sub-sample each one of the datasets S_a and S_b, with a sample size of τ, creating the sampled datasets $S_{a,\tau}$ and $S_{b,\tau}$.
- We transform an input observation $\boldsymbol{x}_j \in S_i$, with $\boldsymbol{x}_j \in \mathbb{R}^n$, where n is the dimensionality of the input space, using the feature extractor f, yielding the feature vector $\boldsymbol{h}_j = f(\boldsymbol{x}_j)$.
- The feature vector $\boldsymbol{h}_i \in \mathbb{R}^{n'}$ has dimension n', with $n' < n$. For instance, the implemented feature extractor f uses the ImageNet pretrained Wide-ResNet architecture, extracting $n' = 512$ features. This yields the two feature sets $H_{a,\tau}$ and $H_{b,\tau}$.

For the Minkowski based distance sets $d_{\ell_2}(S_a, S_b, \tau, \mathcal{C})$, $d_{\ell_1}(S_a, S_b, \tau, \mathcal{C})$, we perform the following steps for the sets of features obtained in the previous description $H_{a,\tau}$ and $H_{b,\tau}$:

- For each feature vector $\boldsymbol{h}_j \in H_{a,\tau}$, find the closest feature vector $\boldsymbol{h}_k \in H_{b,\tau}$, using the ℓ_p distance, with $p = 1$ or $p = 2$ for the Manhattan and Euclidean distances, respectively: $\widehat{d}_j = \min_k \|\boldsymbol{h}_j - \boldsymbol{h}_k\|_p$. We do this for a number of $\mathcal{C}$ samples, yielding a list of distance calculations $d_{\ell_p}(S_a, S_b, \tau, \mathcal{C}) = \left\{\widehat{d}_1, \widehat{d}_2, ..., \widehat{d}_{\mathcal{C}}\right\}$.
- We compute a reference list of distances for the same list of samples of the dataset S_a to itself (intra-dataset distance), thereby computing $d_{\ell_p}(S_a, S_a, \tau, \mathcal{C})$. This yields a list of reference distances $\check{d}_1, \check{d}_2, ..., \check{d}_{\mathcal{C}}$. In our case S_a corresponds to the labelled dataset S_l, as the distance to different unlabelled datasets S_u is to be computed. We highlight that this should result in values close to zero. However, as different samples are used for each distance computation, the results are not exactly zero.
- To ensure that the absolute differences between the reference and inter-dataset distances $d_c = \left|\widehat{d}_c - \check{d}_c\right|$ are statistically significant, we compute the p-value associated with a Wilcoxon test.
- After the distance set between two datasets $d_{\ell_p}(S_a, S_b, \tau, \mathcal{C})$ is obtained, its average reference subtracted distance $\overline{d}$ and its corresponding statistical significance p-value are computed.

As for the density based distances implemented we follow a similar sub-sampling approach, with these steps:

- For each dimension $r = 1, ..., n'$ in the feature space, we compute the normalized histograms to approximate the density functions $p_{r,a}$, in the sample $H_{a,\tau}$. Similarly, we compute the normalized histograms to yield the set of approximate density functions $p_{r,b}$ for $r = 1, ..., n'$, using the observations in the sample $H_{b,\tau}$.
- For the Jensen-Shannon divergence (d_{JS}) and the cosine distance (d_C), we compute the sum of the dissimilarities between the density functions $p_{r,a}$ and $p_{r,b}$, to yield the estimated dissimilarity for the sample j: $\widehat{d}_j = \sum_{r=1}^{n'} \delta_g(p_{r,a}, p_{r,b})$, where $g = JS$ and $g = C$ for the Jensen-Shannon divergence and the cosine distance, respectively. We do this for all the $\mathcal{C}$ samples, yielding the list of inter-dataset distances: $\widehat{d}_1, \widehat{d}_2, ..., \widehat{d}_{\mathcal{C}}$. To lower the computational burden, we assume that the dimensions are statistically independent. This assumption also simplifies the likelihood calculation, as seen in other methods [20].
- Similar to the Minkowski distances, we compute the intra-dataset distances for the dataset S_a, in this context the labelled dataset S_l, to obtain the list of reference distances $\check{d}_1, \check{d}_2, ..., \check{d}_{\mathcal{C}}$.
- Similarly, to verify that the inter- and intra-dataset distance differences $d_c = \left|\widehat{d}_c - \check{d}_c\right|$ are statistically significant, we compute the p-value associated with a Wilcoxon test. The distance computation yields the sample mean distance $\overline{d}$ and its statistical significance p-value.

The proposed dissimilarity measures do not fulfill the conditions of a mathematical metric or pseudo-metric since the distance of an object to itself is not strictly zero (but tends to be close) and symmetry properties are not fulfilled for the sake of evaluation speed [13]. Despite these relaxations, we will see that these dissimilarity measures, especially the two that are density based, are an effective proxy for estimating the $S_{\text{u,OOD}}$ accuracy gain.

To quantitatively measure the relationship between S_l and S_u distances and SSDL accuracy, we calculate the Pearson coefficient between them. This verifies the linear correlation between both. Table III describes the Pearson coefficient for each implemented dissimilarity measure and each SSDL configuration.

In summary, we propose to quantitatively rank a set of candidate unlabelled datasets $S_{u,1}, S_{u,2}, ..., S_{u,k}$ according to a dissimilarity measure $d(S_l, S_u)$, instead of using semantic matching heuristics. In all the tests of this work, we used

$n' = 512$, $\tau = 80$ and $\mathcal{C} = 10$.

IV. Experiments

A. Semi-supervised deep learning setup

The basis for all SSDL experiments in this paper is the MixMatch algorithm, a state of the art SSDL method [5]. MixMatch estimates pseudo-labels for unlabelled data X_u, and also implements an unsupervised regularization term. Pseudo-label $\widehat{\boldsymbol{y}}_j$ estimation is performed with the average model output of a transformed input x_j, with K number of different transformations. The pseudo-labels $\widehat{\boldsymbol{y}}$ are further sharpened with a temperature parameter ρ. To further augment the data using both labelled and unlabelled samples, MixMatch makes use of the MixUp algorithm by [49] which builds linear interpolations between labelled and unlabelled observations. For supervised and semi-supervised loss functions, the cross-entropy and the Euclidean distance, are used, respectively. The regularization coefficient γ controls the direct influence on unlabelled data. Unlabelled data also influences the labelled data term since unlabelled data is used also to artificially augment the dataset with the Mix Up algorithm. This loss term is used at training time, for testing, a regular cross entropy loss is implemented. For a detailed description of the MixMatch algorithm we refer to [5]. We use the recommended hyperparameters documented in the supplementary material.

B. SSDL with OOD data test bed

To assess the effect of OOD unlabelled data S_u on the accuracy of SSDL models trained with MixMatch, we construct the *non-IID-SSDL* test bed, with five variable parameters: (1) base data S_{IOD} which constitutes the original task to be learned, (2) the type of OOD data T_{OOD}, (3) the OOD data source $S_{\text{u,OOD}}$, (4) the relative amount of OOD data among the unlabelled data $\%_{\text{u,OOD}}$, (5) and the amount n_l of labelled observations. Each of the five axes is explored by varying only one of the variables at a time while keeping the others constant. This allows us to isolate the effect of the individual variables. We consider three configurations for S_{IOD} comprising MNIST, CIFAR-10 and FashionMNIST. A total of three configurations for T_{OOD} (Other-Half (OH), Similar (Sim) and Different (Dif)) are tested. We derived the possible types of OOD data from the existing literature cited in Section II. In the OH setting half of the classes and associated inputs are taken to be the S_{IOD} data, whereas the other half of classes are taken to be the $S_{\text{u,OOD}}$ data. *Similar* is a $S_{\text{u,OOD}}$ dataset that is assumed to be semantically related to S_{IOD}, e.g., MNIST and Street View House Numbers dataset (SVHN). *Different* is a $S_{\text{u,OOD}}$ dataset that is supposedly semantically unrelated to S_{IOD}, e.g., MNIST and TI. There are five configurations for $S_{\text{u,OOD}}$ as explained above: the other half OH, a similar dataset, and three different datasets including two noise baselines. They include SVHN, TI, GN, Salt and Pepper Noise (SAPN) and Fashion Product (FP). Please see Table I for the per task pairings. Each configuration represents a multi-class classification task with $|\mathcal{Y}| = 5$, that is a random subset of half of the classes of base data S_{IOD}.

We vary the relative amount of OOD data $\%_{\text{u,OOD}}$ between 0, 50 and 100 as well as the amount of labelled datapoints n_l between 60, 100 and 150. We study the behaviour of MixMatch under very limited number of labels settings, where the benefit of SSDL is usually higher. This makes the impact of distribution mismatch more evident. Note that for each result entry you can see in Table I we performed ten experimental runs and report the accuracy mean and standard deviation of the models performing best on the test data from each run, as overfitting is very likely to happen with a low n_l. For each run we sampled a disjunct subset of data from S_{IOD} and $S_{\text{u,OOD}}$ to obtain the required number of labelled n_l and unlabelled n_u samples for the run. Descriptive statistics (mean and standard deviation) for standardization of the neural networks inputs were only computed from these subsets to keep the simulation realistic and not use any information from the global training data. All other parameters (number of unlabelled observations $n_u = 3000$, neural network architecture , the set of optimization hyperparameters, number of training epochs) are kept constant across all experiments to enable direct comparison with respect to the variable parameters of the system. We clarify that the goal of this test-bed is to assess the impact of distribution mismatch for MixMatch, rather to achieve state of the art performance with MixMatch on the given data. Such hyper-parameters are described in the supplementary material. Note that it is possible to extend the test bed to other effects of interest. We address some of these ideas in Section VI.

C. Deep Dataset Dissimilarity Measures

In Table II we show the dissimilarity results for the tested labelled and unlabelled dataset combinations. We tested the dissimilarity measures detailed in Section III, namely the Manhattan or ℓ_1 distance d_{ℓ_1}, the Euclidean distance d_{ℓ_2}, the cosine distance d_C and the Jensen-Shannon d_{JS} divergence. The distances and divergences are computed without the need for training a model, making the proposed approach appealing to choose unlabelled datasets before SSDL training.

As a complementary quantitative test, in Figure 2 we show the probability density function approximation plots of some of the features for the MNIST dataset, using both the similar dataset chosen (SVHN) and the different dataset selected (TI). We picked the features presenting the smallest divergences for the chosen datasets. The density functions were built using random samples for both data pairs. The probability density function approximation plots illustrate in a summarized manner the similarity computed between the two compared datasets, and its correlation with the measured Jensen-Shannon and cosine divergences.

V. Results

The experimental setup used in this work is detailed in the supplementary material. Table I shows the results of the distribution mismatch experiment described in Section IV-A. We make a number of observations.

We find in the majority of cases that using IOD unlabelled data or a 50-50 mix of IOD and OOD unlabelled data beats the fully supervised baseline. For instance take the results in

TABLE I: Results for the distribution mismatch experiment, best OOD performance in bold per configuration (mean ± standard deviation). Each result entry in the table represents the mean and variance of accuracy across ten random experimental runs per entry. For a detailed description of symbols and the experiment see Section IV-B.

S_{IOD}	T_{OOD}	S_{uOOD}	$\%_{uOOD}$	n_l = 60	n_l = 100	n_l = 150	row #
MNIST	Fully supervised baseline			0.457 ± 0.108	0.559 ± 0.125	0.645 ± 0.101	0
	SSDL baseline (no OOD data)			0.704 ± 0.096	0.781 ± 0.065	0.831 ± 0.062	1
	OH	OH-MNIST	50	$\mathbf{0.679 \pm 0.108}$	$\mathbf{0.769 \pm 0.066}$	0.802 ± 0.054	2
			100	0.642 ± 0.111	0.746 ± 0.094	0.798 ± 0.070	3
	Sim	SVHN	50	0.637 ± 0.098	0.745 ± 0.081	0.801 ± 0.069	4
			100	0.482 ± 0.113	0.719 ± 0.058	0.765 ± 0.072	5
	Dif	TI	50	0.642 ± 0.094	0.739 ± 0.074	0.809 ± 0.066	6
			100	0.637 ± 0.097	0.732 ± 0.074	0.804 ± 0.071	7
		GN	50	0.606 ± 0.0989	0.713 ± 0.087	0.786 ± 0.065	8
			100	0.442 ± 0.099	0.461 ± 0.073	0.542 ± 0.062	9
		SAPN	50	0.631 ± 0.102	0.735 ± 0.082	$\mathbf{0.813 \pm 0.057}$	10
			100	0.48 ± 0.0951	0.524 ± 0.09	0.563 ± 0.095	11
CIFAR-10	Fully supervised baseline			0.380 ± 0.024	0.445 ± 0.042	0.449 ± 0.022	12
	SSDL baseline (no OOD data)			0.453 ± 0.046	0.474 ± 0.019	0.501 ± 0.033	13
	OH	OH-CIFAR-10	50	$\mathbf{0.444 \pm 0.040}$	0.472 ± 0.039	0.525 ± 0.050	14
			100	0.443 ± 0.023	0.472 ± 0.047	0.499 ± 0.054	15
	Sim	TI	50	0.435 ± 0.054	0.473 ± 0.039	$\mathbf{0.543 \pm 0.040}$	16
			100	0.417 ± 0.020	$\mathbf{0.480 \pm 0.039}$	0.498 ± 0.042	17
	Dif	SVHN	50	0.419 ± 0.027	0.464 ± 0.044	0.469 ± 0.056	18
			100	0.385 ± 0.034	0.418 ± 0.035	0.440 ± 0.046	19
		GN	50	0.409 ± 0.047	0.454 ± 0.048	0.491 ± 0.032	20
			100	0.297 ± 0.029	0.306 ± 0.034	0.302 ± 0.038	21
		SAPN	50	0.438 ± 0.029	0.455 ± 0.037	0.485 ± 0.034	22
			100	0.236 ± 0.031	0.246 ± 0.032	0.232 ± 0.022	23
FashionMNIST	Fully supervised baseline			0.571 ± 0.073	0.704 ± 0.066	0.720 ± 0.093	24
	SSDL baseline (no OOD data)			0.715 ± 0.049	0.760 ± 0.044	0.756 ± 0.069	25
	OH	OH-FashionMNIST	50	$\mathbf{0.714 \pm 0.049}$	0.721 ± 0.104	0.765 ± 0.053	26
			100	0.660 ± 0.061	0.711 ± 0.090	0.747 ± 0.061	27
	Sim	FP	50	0.707 ± 0.039	0.724 ± 0.030	0.778 ± 0.078	28
			100	0.546 ± 0.101	0.542 ± 0.099	0.540 ± 0.105	29
	Dif	TI	50	0.690 ± 0.065	$\mathbf{0.745 \pm 0.093}$	0.792 ± 0.058	30
			100	0.690 ± 0.073	0.728 ± 0.066	$\mathbf{0.794 \pm 0.056}$	31
		GN	50	0.644 ± 0.061	0.689 ± 0.075	0.755 ± 0.055	32
			100	0.352 ± 0.025	0.366 ± 0.065	0.361 ± 0.057	33
		SAPN	50	0.671 ± 0.072	0.708 ± 0.095	0.729 ± 0.088	34
			100	0.276 ± 0.069	0.297 ± 0.046	0.283 ± 0.059	35

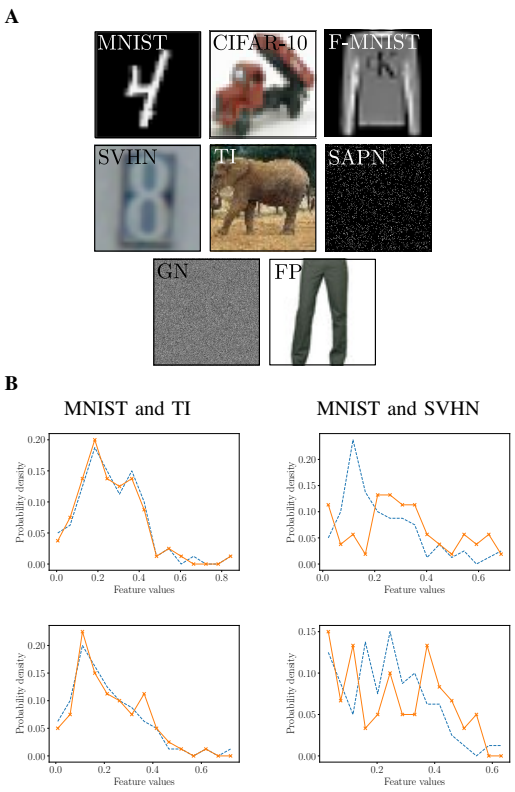

Fig. 2: **A**: Dataset samples. **B**: Feature distribution for a model trained with MNIST labelled data (continous orange line in both plots), and TI and SVHN unlabelled data (left and right column, respectively, blue dashed line in both). For each plot a different dataset partition was used. Top row: feature 372, bottom row: feature 159.

row 0 vs. the results yielded in rows 2-7 (for the SSDL model). A clear advantage of the SSDL model is revealed over the supervised model, even under distribution mismatch settings. The gains range from 15% to 25% for MNIST, 10% to 15% for CIFAR-10 and 7% to 13% for FashionMNIST across all $S_{\text{u,OOD}}$ and n_l. As expected, in most of the cases the accuracy is degraded when including OOD data in S_u, with a more dramatic hit when noisy datasets (SAPN, GN) are used as OOD data contamination.

Another interesting observation from the experiment results is related to semantic matching heuristics and the yielded SSDL accuracy. Sometimes, using an unlabelled dataset that is semantically supposedly less similar can result in greater accuracy. This is observed for example in Table I, when S_l =CIFAR-10, $n_l = 100$ and $n_l = 150$, where OOD unlabelled data from TI (row 16) results in a similar accuracy (with no statistical significance gain, according to the Wilcoxon test performed) than using the other half of CIFAR-10 as $S_{\text{u,OOD}}$ (row 14). It is interesting that an $S_{\text{u,OOD}}$ dataset of type *different* can have a similar benefit than a $S_{\text{u,OOD}}$ dataset of type *similar*. A clearer case of this tendency is found for FashionMNIST and TI (row 31) versus FP at $n_l = 150$ (row 29). In such case using the TI (different) dataset, brings a higher SSDL accuracy, than using the FP (similar) dataset. This contradicts the common heuristic that unlabelled data that appears semantically more related to the labelled data is always the better choice for SSDL. Rather, as we demonstrate in the second set of results below, a notion of distance in the feature space between labelled and unlabelled data offers a more consistent and quantifiable proxy for the expected benefit of an unlabelled dataset.

As for qualitative illustration, Figure 2 shows an example of the density functions approximated for randomly selected samples for the MNIST-TI and MNIST-SVHN dataset pairs. The plots reveal a stronger density based similarity between the MNIST and ImageNet than the MNIST and SVHN datasets. This in spite of the higher semantic similarity of SVHN to MNIST (both represent numbers, the first one in natural scenes, and the second one in handwritten images). This correlates well with the quantitative figures yielded in Table II. For instance, in row 3, the MNIST dataset is more dissimilar to the SVHN dataset (MNIST contaminated by 100% with the SVHN dataset), than the TI dataset (MNIST contaminated by 100% with the TI dataset), revealed in row 5. This also highly correlates with the final SSDL accuracy yielded with both unlabelled datasets (MNIST contaminated by 100% with SVHN, in row 5, and TI, in row 7) shown in Table I.

MixMatch shows a *marginally higher accuracy* (with no statistical significance, after performing a Wilcoxon test) when using TI as an unlabelled dataset compared to using SVHN as unlabelled data.

TABLE II: Distance measures between the labelled and unlabelled datasets S_l and S_u (mean $\pm$ standard deviation). Numbers in italics correspond to results with $p > 0.05$ for the Wilcoxon test.

$\mathbf{S_l}$	$\mathbf{S_u}$	$\%_{\mathbf{uOOD}}$	$\mathbf{d}_{\ell_2}$	$\mathbf{d}_{\ell_1}$	$\mathbf{d}_{\mathbf{JS}}$	$\mathbf{d}_{\mathbf{C}}$	row #
MNIST	OH	50	*0.011 ± 0.006*	*0.459 ± 0.28*	*0.266 ± 0.221*	*0.811 ± 0.512*	0
		100	*0.014 ± 0.019*	*0.38 ± 0.507*	1.001 ± 0.725	1.263 ± 0.665	1
	SVHN	50	*0.09 ± 0.017*	1.569 ± 0.504	6.789 ± 0.924	12.021 ± 1.757	2
		100	0.25 ± 0.053	4.702 ± 1.04	52.349 ± 3.292	42.026 ± 4.311	3
	TI	50	0.008 ± 0.023	1.519 ± 0.223	3.663 ± 0.772	5.512 ± 0.767	4
		100	0.217 ± 0.04	4.3 ± 0.636	10.305 ± 1.667	15.18 ± 2.698	5
	GN	50	0.11 ± 0.0219	1.958 ± 0.534	14.785 ± 1.052	23.59 ± 1.859	6
		100	0.357 ± 0.081	5.987 ± 1.091	52.349 ± 4.253	86.21 ± 3.471	7
	SAPN	50	0.089 ± 0.0311	2.479 ± 0.7433	15.116 ± 1.475	20.151 ± 1.619	8
		100	0.323 ± 0.07	6.308 ± 1.366	53.397 ± 4.253	77.456 ± 4.474	9
CIFAR-10	OH	50	*0.056 ± 0.023*	*0.915 ± 0.934*	*0.338 ± 0.325*	0.892 ± 0.402	10
		100	*0.061 ± 0.04*	0.769 ± 0.461	*0.451 ± 0.41*	0.648 ± 0.407	11
	TI	50	0.082 ± 0.037	*0.928 ± 0.815*	*0.388 ± 0.243*	*0.423 ± 0.362*	12
		100	*0.056 ± 0.048*	*0.992 ± 0.517*	*0.469 ± 0.426*	0.415 ± 0.232	13
	SVHN	50	*0.055 ± 0.032*	*0.948 ± 0.699*	*0.665 ± 0.565*	*0.414 ± 0.357*	14
		100	*0.075 ± 0.036*	1.291 ± 0.925	0.736 ± 0.658	0.581 ± 0.343	15
	GN	50	*0.107 ± 0.083*	1.344 ± 1.156	1.708 ± 0.421	3.001 ± 0.696	16
		100	0.127 ± 0.087	1.531 ± 0.767	5.855 ± 0.552	8.703 ± 0.926	17
	SAPN	50	0.1146 ± 0.044	1.854 ± 0.894	2.299 ± 0.691	2.561 ± 0.762	18
		100	0.208 ± 0.05	5.502 ± 1.156	8.225 ± 0.866	9.554 ± 0.489	19
FashionMNIST	OH	50	*0.02 ± 0.012*	*0.34 ± 0.162*	*0.669 ± 0.566*	*0.575 ± 0.423*	20
		100	0.059 ± 0.032	0.801 ± 0.402	0.305 ± 0.237	0.774 ± 0.343	21
	FP	50	0.105 ± 0.0526	2.168 ± 0.774	7.263 ± 0.622	5.305 ± 0.405	22
		100	0.195 ± 0.0457	4.819 ± 1.077	9.056 ± 0.462	11.286 ± 0.751	23
	TI	50	*0.04 ± 0.03*	*0.798 ± 0.542*	*0.897 ± 0.516*	0.897 ± 0.516	24
		100	0.065 ± 0.03	1.66 ± 0.45	1.4 ± 0.488	1.912 ± 0.683	25
	GN	50	*0.047 ± 0.03*	*0.533 ± 0.347*	2.819 ± 0.703	3.843 ± 0.704	26
		100	0.074 ± 0.041	1.325 ± 0.631	9.042 ± 0.699	15.511 ± 0.445	27
	SAPN	50	0.036 ± 0.022	0.52 ± 0.303	2.799 ± 0.497	2.799 ± 0.497	28
		100	0.076 ± 0.044	1.411 ± 0.548	8.464 ± 0.553	8.464 ± 0.553	29

TABLE III: Correlation results for the dissimilarity measures between S_l and S_u with OOD contamination and SSDL accuracy.

$\mathbf{S_l}$	$\mathbf{n_l}$	$\mathbf{d}_{\ell_1}$	$\mathbf{d}_{\ell_2}$	$\mathbf{d}_{\mathbf{JS}}$	$\mathbf{d}_{\mathbf{C}}$
MNIST	60	-0.876	-0.898	-0.969	-0.944
	100	-0.805	-0.83	-0.786	-0.948
	150	-0.794	-0.822	-0.81	-0.944
CIFAR-10	60	-0.823	-0.853	-0.944	-0.921
	100	-0.826	-0.878	-0.966	-0.947
	150	-0.808	-0.838	-0.952	-0.927
FashionMNIST	60	-0.2	-0.268	-0.735	-0.789
	100	-0.264	-0.326	-0.781	-0.824
	150	-0.286	-0.347	-0.785	-0.827

The second set of results demonstrate the potential of using distance measures as a systematic and quantitative ranking heuristic when selecting unlabelled datasets for the MixMatch algorithm. The exact distances, as described in Section III, for all OOD configurations from the ablation study can be found in Table II. We can observe that these distances trace the accuracy results found in Table I, as confirmed by the Pearson correlation. This correlation is quantified in Table III with the cosine based density measure d_c correlating particularly well with the accuracy results of Table I. Also, the p-values are consistently lower for the density based distances (with fewer p-values that exceed 0.05, as shown by the italicized entries in Table II), meaning that density based distances present more confidence. We suspect that this is related to the quantitative approximation of the feature distribution mismatch implemented both in the d_{JS} and d_{C} distances. In Table I we indicate the distance-based preference ranking in parentheses. The OOD configurations resulting in the best SSDL accuracy are contained in the top two selections seven out of nine times. Note that with our proposed approach we can do this selection *before* SSDL training and thus improve the overall result.

VI. Conclusions and Recommendations

In this work we extensively tested the behavior of the MixMatch algorithm under various OOD unlabelled data settings. We introduced a set of quantitative data selection heuristics, DeDiMs, to rank unlabelled datasets prior to model training according to their expected benefit to SSDL. Our results lead us to the following conclusions:

1) In the experiments conducted in this study the implemented DeDiMs correlate strongly with SSDL accuracy. In particular, density based measures yield high correlation with MixMatch accuracy. This suggests that DeDiMs can be applied in SSDL prior to learning, aiding the unlabelled data selection process and mitigate the distribution mismatch problem. The proposed method is agnostic to the downstream SSDL algorithm, simple and fast to compute making it particularly suitable for practical application in SSDL. Different OOD detectors [52] use the feature space for building a discrimination criteria to filter OOD data. Our results suggest that online OOD data filtering approaches for SSDL as the ones developed in [27], [11] might benefit from using the feature space for OOD detection. Other criteria for online OOD detection during training as the model Softmax output used in [27] might discard data that might be useful for learning. This is tested in [10] for a practical application.
2) In real-world usage scenarios of SSDL the unlabelled dataset S_u may contain observations of classes not present in labelled dataset S_l . We simulated a similar scenario with the OH setting which resulted in a subtle accuracy degradation in most cases. However, the accuracy gain obtained vis-a-vis the fully supervised baseline is still substantial, making the application of SSDL attractive in such a setting.
3) Another plausible real-world scenario for SSDL is the inclusion of widely available unlabelled datasets, e.g., built with web crawlers, where shifts in crawl queries can lead to different unlabelled datasets. This scenario has been simulated with the OOD types similar and different. We can observe that notions of semantic similarity between labelled and unlabelled dataset pairings, e.g., (MNIST-SVHN) or (FashionMNIST-FP), do not necessarily imply an SSDL

accuracy gain. The quantitative comparison of the density function plots in Figure 2 suggest a higher similarity for dataset pairs with less semantic similarity, for some of the tested dataset setups. Distance measures, in particular d_C, seem to be an accurate and systematic proxy for SSDL accuracy, according to our test results. This is visible when comparing the accuracy and distance results of the previous pairings to (MNIST-TI) and (FashionMNIST-TI) which have higher accuracies and, also, surprisingly, lower distance measurements. We speculate that using more diverse data for pre-training might yield an even better feature extractor, similar to results in self-supervised learning methods [48].

4) As suggested, our method can be used to rank different unlabelled datasets. The proposed DeDiMs can be considered efficient to implement, requiring only small samples, and with no need for model training, as a pre-trained ImageNet feature extractor is used. According to our tests, a ResNet model pre-trained on ImageNet without further fine-tuning works surprisingly well for quantifying unlabelled-to-labelled dataset affinity. As preliminary studies show a growing concern for the carbon footprint of training deep learning models [2], inexpensive and quantified data selection heuristics like DeDiMs can help to avoid unnecessary computation loads. Further studying our method to decrease training time and resources is an interesting future research path.
5) The claim in [51] regarding the impact of OOD data close to the decision boundary compared to OOD data far from it, relies on an Euclidean space projection of the data. In this work, we have gathered evidence that Euclidean based similarity measures correlate worse with SSDL accuracy than the density function based measures tested. Using a density based divergence like Jensen-Shannon might not correlate well with semantic similarity, but according to our tests, it better explains the obtained SSDL accuracy. This shows how the feature extractor and the consequent feature space projections play a more important role in the final model performance than the original input space, as the feature space is built through non-linear convolution operations that significantly change the input representation.

Based on these results, we can draw a number of recommendations for the researchers in the field, which we enlist as follows:

1) Our results shift the attention to data-oriented approaches to improve the model performance. Similar to [26], where dataset sparsity is related to downstream model accuracy, our method allows the use of DeDiMs to assess the impact of unlabelled datasets on SSDL training. This enables the exclusion of datasets that are not beneficial for a given SSDL task.
2) The use of SSDL can also improve other model properties like uncertainty [9]. Hence, exploring the impact of OOD data in other aspects of SSDL performance, such as robustness [30], explainability [39] and confidence [3], as recommended in [29], [31], is a promising next step for distribution mismatch analysis. For instance, in [3] the impact of OOD data is tested in the overall model robustness and explainability. Evaluating the impact of distribution mismatch between S_l and S_u in other performance aspects opens up further questions for research.
3) In unsupervised domain adaptation we find similar challenges where the target domain presents a different distribution than the source domain. Using SSDL for such setting can leverage unlabelled data in the target domain. For instance, in [50], an SSDL approach is proposed for unsupervised domain adaptation. Quantifying the degrees of OOD for the unlabelled data could improve the analysis of the test results and estimate the performance for unsupervised domain adaptation.
4) Finally, the proposed test bed and distance measures can be used for a more systematic quantitative evaluation of SSDL algorithms. Counterintuitively, datasets with a high perceived semantic similarity can be less beneficial for SSDL than other unlabelled datasets with less perceived semantic similarity, adding further evidence that we should be wary to conflate human and machine perception.

In future work, we plan to extend the test bed to other SSDL variants, depth-first analyses (e.g., fewer tasks with more training epochs), additional axes of test bed variables (e.g., n_u) and more testing around the appropriate dissimilarity measures parameters. Investigating the relationship between generic feature similarity and SSDL downstream performance further is a promising topic in data-centric machine learning. The fact that feature dissimilarity scores can be calculated before SSDL training and independent of the SSDL model offers an interesting profile for application. Connections to OOD detection [52], concept drift [45] and distribution mismatch [11] could be explored further. Efficient and effective quantitative dataset evaluation prior to training a deep learning model offers an opportunity to push the envelope in computationally efficient deep learning further and to narrow the gap between deep learning research and its real-world application.

REFERENCES

[1] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

[2] Lasse F Wolff Anthony, Benjamin Kanding, and Raghavendra Selvan. Carbontracker: Tracking and predicting the carbon footprint of training deep learning models. *arXiv preprint arXiv:2007.03051*, 2020.

[3] Maximilian Augustin, Alexander Meinke, and Matthias Hein. Adversarial robustness on in-and out-distribution improves explainability. *arXiv preprint arXiv:2003.09461*, 2020.

[4] Indranil Balki, Afsaneh Amirabadi, Jacob Levman, Anne L Martel, Ziga Emersic, Blaz Meden, Angel Garcia-Pedrero, Saul Calderon-Ramirez, Dehan Kong, Alan R Moody, et al. Sample-size determination methodologies for machine learning in medical imaging research: A systematic review. *Canadian Association of Radiologists Journal*, 2019.

[5] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 5050–5060, 2019.

[6] Federico Cabitza and Andrea Campagner. Who wants accurate models? arguing for a different metrics to take classification models seriously. *arXiv preprint arXiv:1910.09246*, 2019.

[7] Saul Calderon-Ramirez, Raghvendra Giri, Shengxiang Yang, Armaghan Moemeni, Mario Umana, David Elizondo, Jordina Torrents-Barrena, and Miguel A Molina-Cabello. Dealing with scarce labelled data: Semi-supervised deep learning with mix match for covid-19 detection using chest x-ray images. IEEE Press, 2020.

[8] Saul Calderon-Ramirez, Armaghan Moemeni, David Elizondo, Simon Colreavy-Donnelly, Luis Fernando Chavarria-Estrada, Miguel A Molina-Cabello, et al. Correcting data imbalance for semi-supervised covid-19 detection using x-ray chest images. *arXiv preprint arXiv:2008.08496*, 2020.

[9] Saul Calderon-Ramirez, Diego Murillo-Hernandez, Kevin Rojas-Salazar, Luis-Alexander Calvo-Valverde, Shengxiang Yang, Armaghan Moemeni, David Elizondo, Ezequiel Lopez-Rubio, and Miguel Molina-Cabello. Improving uncertainty estimations for mammogram classification using semi-supervised learning. In *Institute of Electrical and Electronics Engineers*, 2021.

[10] Saul Calderon-Ramirez, Shengxiang Yang, David Elizondo, and Armaghan Moemeni. Dealing with distribution mismatch in semi-supervised deep learning for covid-19 detection using chest x-ray images: A novel approach using feature densities. *arXiv preprint arXiv:2109.00889*, 2021.

[11] Yanbei Chen, Xiatian Zhu, Wei Li, and Shaogang Gong. Semi-supervised learning under class distribution mismatch. In *AAI*, pages 3569–3576, 2020.

[12] Veronika Cheplygina, Marleen de Bruijne, and Josien PW Pluim. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical image analysis*, 54:280–296, 2019.

[13] Daniel Cremers and Stefano Soatto. A pseudo-distance for shape priors in level set segmentation. In *2nd IEEE workshop on variational, geometric and level set methods in computer vision*, pages 169–176. Citeseer, 2003.

[14] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1422–1430, 2015.

[15] WeiWang Dong-DongChen and Zhi-HuaZhou WeiGao. Tri-net for semi-supervised deep learning. IJCAI, 2018.

[16] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA, 20–22 Jun 2016. PMLR.

[17] Ryuhei Hamaguchi, Ken Sakurada, and Ryosuke Nakamura. Rare event detection using disentangled representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9327–9335, 2019.

[18] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

[19] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

[20] Josef Kittler and John Illingworth. Minimum error thresholding. *Pattern recognition*, 19(1):41–47, 1986.

[21] WJ Krzanowski. Non-parametric estimation of distance between groups. *Journal of Applied Statistics*, 30(7):743–750, 2003.

[22] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, pages 7167–7177, 2018.

[23] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *6th International Conference on Learning Representations, ICLR 2018*, 2018.

[24] Markos Markou and Sameer Singh. Novelty detection: a review—part 1: statistical approaches. *Signal processing*, 83(12):2481–2497, 2003.

[25] Facundo Mémoli. Gromov–wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11(4):417–487, 2011.

[26] Mauro Mendez, Saul Calderon-Ramirez, and Pascal N Tyrrell. Using cluster analysis to assess the impact of dataset heterogeneity on deep convolutional network accuracy: A first glance. In *Latin American High Performance Computing Conference*, pages 307–319. Springer, 2019.

[27] Varun Nair, Javier Fuentes Alonso, and Tony Beltramelli. Realmix: Towards realistic semi-supervised deep learning algorithms. *arXiv preprint arXiv:1912.08766*, 2019.

[28] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, and Balaji Lakshminarayanan. Detecting out-of-distribution inputs to deep generative models using a test for typicality. *arXiv preprint arXiv:1906.02994*, 5:5, 2019.

[29] Luis Oala, Jana Fehr, Luca Gilli, Pradeep Balachandran, Alixandro Werneck Leite, Saul Calderon-Ramirez, Danny Xie Li, Gabriel Nobis, Erick Alejandro Muñoz Alvarado, Giovanna Jaramillo-Gutierrez, et al. Ml4h auditing: From paper to practice. In *Machine Learning for Health*, pages 280–317. PMLR, 2020.

[30] Luis Oala, Cosmas Heiß, Jan Macdonald, Maximilian März, Gitta Kutyniok, and Wojciech Samek. Detecting failure modes in image reconstructions with interval neural network uncertainty. *International Journal of Computer Assisted Radiology and Surgery*, 16(12):2089–2097, 2021.

[31] Luis Oala, Andrew G Murchison, Pradeep Balachandran, Shruti Choudhary, Jana Fehr, Alixandro Werneck Leite, Peter G Goldschmidt, Christian Johner, Elora DM Schörverth, Rose Nakasi, et al. Machine learning for health: Algorithm auditing & quality control. *Journal of medical systems*, 45(12):1–8, 2021.

[32] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *Advances in Neural Information Processing Systems*, pages 3235–3246, 2018.

[33] Pramuditha Perera and Vishal M Patel. Deep transfer learning for multiple class novelty detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11544–11552, 2019.

[34] Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*, 2017.

[35] Igor M Quintanilha, Roberto de ME Filho, José Lezama, Mauricio Delbracio, and Leonardo O Nunes. Detecting out-of-distribution samples using low-order deep features statistics. 2018.

[36] Jie Ren, Xin Bai, Yang Young Lu, Kujin Tang, Ying Wang, Gesine Reinert, and Fengzhu Sun. Alignment-free sequence analysis and applications. *Annual Review of Biomedical Data Science*, 1:93–114, 2018.

[37] Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In *Advances in Neural Information Processing Systems*, pages 14680–14691, 2019.

[38] Peter J. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79(388):871–880, 1984.

[39] Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. *Explainable AI: interpreting, explaining and visualizing deep learning*, volume 11700. Springer Nature, 2019.

[40] Karanjit Singh and Shuchita Upadhyaya. Outlier detection: applications and techniques. *International Journal of Computer Science Issues (IJCSI)*, 9(1):307, 2012.

[41] Natasa Tagasovska and David Lopez-Paz. Single-model uncertainties for deep learning. In *Advances in Neural Information Processing Systems*, pages 6414–6425, 2019.

[42] Nikolaj Tatti. Distances between data sets based on summary statistics. *Journal of Machine Learning Research*, 8(Jan):131–154, 2007.

[43] David M. J. Tax and Robert P. W. Duin. Support vector data description. *Mach. Learn.*, 54(1):45–66, 2004.

[44] Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440, 2020.

[45] Geoffrey I Webb, Roy Hyde, Hong Cao, Hai Long Nguyen, and Francois Petitjean. Characterizing concept drift. *Data Mining and Knowledge Discovery*, 30(4):964–994, 2016.

[46] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3(1):9, 2016.

[47] Kurt Willis and Luis Oala. Post-hoc domain adaptation via guided data homogenization, 2021.

[48] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE international conference on computer vision*, pages 1476–1485, 2019.

[49] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[50] Yabin Zhang, Bin Deng, Kui Jia, and Lei Zhang. Label propagation with augmented anchors: A simple semi-supervised learning baseline for unsupervised domain adaptation. In *European Conference on Computer Vision*, pages 781–797. Springer, 2020.

[51] Xujiang Zhao, Killamsetty Krishnateja, Rishabh Iyer, and Feng Chen. Robust semi-supervised learning with out of distribution data. *arXiv preprint arXiv:2010.03658*, 2020.

[52] Ev Zisselman and Aviv Tamar. Deep residual flow for out of distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13994–14003, 2020.