Dealing with Distribution Mismatch in Semi-supervised Deep Learning for COVID-19 Detection Using Chest X-ray Images: A Novel Approach Using Feature Densities

Saul Calderon-Ramirez[1], Shengxiang Yang[1], David Elizondo[1], Armaghan Moemeni[2]

Abstract

In the context of the global coronavirus pandemic, different deep learning solutions for infected subject detection using chest X-ray images have been proposed. However, deep learning models usually need large labelled datasets to be effective. Semi-supervised deep learning is an attractive alternative, where unlabelled data is leveraged to improve the overall model's accuracy. However, in real-world usage settings, an unlabelled dataset might present a different distribution than the labelled dataset (i.e. the labelled dataset was sampled from a *target* clinic and the unlabelled dataset from a *source* clinic). This results in a distribution mismatch between the unlabelled and labelled datasets. In this work, we assess the impact of the distribution mismatch between the labelled and the unlabelled datasets, for a semi-supervised model trained with chest X-ray images, for COVID-19 detection. Under strong distribution mismatch conditions, we found an accuracy hit of almost 30%, suggesting that the unlabelled dataset distribution has a strong influence in the behaviour of the model. Therefore, we propose a straightforward approach to diminish the impact of such

[1] S. Calderon-Ramirez, D. Elizondo, S. Yang work at the Institute of Artificial Intelligence (IAI), School of Computer Science and Informatics, De Montfort University, United Kingdom (e-mails:sacalderon@itcr.ac.cr, elizondo@dmu.ac.uk, syang@dmu.ac.uk and simon.colreavy-donnelly@dmu.ac.uk). S. Calderon-Ramirez works also at the Instituto Tecnologico de Costa Rica, Costa Rica.

[2] Armaghan Moemeni works at the School of Computer Science, University of Nottingham, United Kingdom (e-mail:armaghan.moemeni@nottingham.ac.uk).

distribution mismatch. Our proposed method uses a density approximation of the feature space. It is built upon the target dataset to filter out the observations in the source unlabelled dataset that might harm the accuracy of the semi-supervised model. It assumes that a small labelled source dataset is available together with a larger source unlabelled dataset. Our proposed method does not require any model training, it is simple and computationally cheap. We compare our proposed method against two popular state of the art *out-of-distribution* data detectors, which are also cheap and simple to implement. In our tests, our method yielded accuracy gains of up to 32%, when compared to the previous state of the art methods. The good results yielded by our method leads us to argue in favour for a more data-centric approach to improve model's accuracy. Furthermore, the developed method can be used to measure data effectiveness for semi-supervised deep learning model training.

Keywords: Semi-supervised Deep Learning, MixMatch, Distribution Mismatch, Out of Distribution Detection, Chest X-Ray, Covid-19, Computer Aided Diagnosis.

1. Introduction

The COVID-19 disease is caused by the novel SARS-CoV2 coronavirus, discovered in 2019 [66]. The COVID-19 pandemic has caused thousands of human losses around the world, where even the most developed health systems have not been able to cope with the infection peaks [66]. Health practitioners are struggling with the detection and tracking of infected subjects, as the number of patients in need for medical assistance increases.

Therefore, accurately detecting patients infected with the SARS-CoV2 virus is a critical task to control the pandemic. Nevertheless, SARS-CoV2 detection methods like the Real-time Reverse Transcription Polymerase Chain Reaction (RT-PCR) test can be expensive and time consuming. As an alternative and/or complementary method, the usage of medical imaging based approaches can be less expensive and also accurate [15, 19]. Moreover, X-ray based imaging diagnosis can be considered cheaper. The usage of X-ray machines is more widespread when compared to other imaging technologies like computer tomography. This is specially the case in less industrialised countries [3]. However, a limitation of X-ray based diagnosing of COVID-19 is the need of highly trained clinical practitioners like radiologists, which in less industrialised countries are scarce [3].

The implementation of Computer Aided Diagnosis (CAD) systems for COVID-19 diagnosis can be a solution to mitigate the specialized staff shortage. Deep learning based CAD systems have been extensively explored for different medical imaging applications [7, 14, 1]. More specifically, several deep learning architectures for COVID-19 detection have been proposed recently in the literature [32, 33, 6]. These systems have been developed using publicly available X-ray images datasets, with COVID-19 positive [21] and negative cases [9].

Nevertheless, a short-coming of implementing a deep learning architecture for real-world usage is the need of a large labelled dataset from the specific target clinic or hospital where the system is intended to be used. Labeling images in the medical domain is time-consuming and requires expensive human effort from highly trained clinical practitioners, which makes building an extensive labelled dataset costly. Previous work on COVID-19 detection with deep learning has relied on large and heterogeneous datasets, where around 100-400 COVID-19 positive cases sampled from the dataset [21], and larger datasets of COVID-19 negative cases sampled from different sources [38, 31, 22]. Such testing conditions can be considered far from a real-world scenario, where usually in the target clinic/hospital a limited set of labelled observations is available. Using external datasets for training might harm the overall performance of the model. This is mainly due to the differences between patient features and imaging protocols. This affects the final data distribution between the test and training data [68].

Another short-coming of the aforementioned previous work, is the bias of the population between the positive and negative COVID-19 samples. For example, as reported in [58], negative COVID-19 observations in [38] were sampled from pediatric Chinese patients, while positive COVID-19 cases in [21] correspond to adult patients from different countries. This dataset combination has been extensively used for training Convolutional Neural Network (CNN) based models to detect COVID-19, and leads to deceptive bias in both the test and training model data [58].

To deal with the limited labelled datasets, different approaches have been implemented in literature [18]. In the context of COVID-19 detection, namely data augmentation and transfer learning [45, 25] have been used. In transfer learning, a source labeled dataset D_l^s is used to pre-train a model, and then fine-tune it in the target dataset D_l^t. However, as discussed in [79], fine-tuning might not be enough to improve the model's accuracy. The distribution mismatch between D_l^s and D_l^t due to different patient populations and imaging acquisition protocols, is frequently a reason for poor transfer learning performance.

Another approach to deal with scarce labelled data is the usage of Semi-supervised Deep Learning (SSDL). SSDL leverages cheaper and more widely available unlabelled data. Semi-supervised learning for COVID-19 detection have been explored in [9, 10] with positive results, where very small labelled datasets have been used. Such previous work suggests that using unlabelled data can increase the model's performance. The authors combined SSDL with common data augmentation and transfer learning approaches. However, to implement deep learning based solutions for extensive real-world usage, testing different model attributes like robustness and predictive uncertainty is crucial for its safe usage. A deep review on the importance of measuring different model attributes like robustness in medical applications of Artificial Intelligence (AI) can be found in [54]. In a real-world scenario, the use of unlabelled data sampled from different sources (hospitals or clinics) can be considered. However, the usage of unlabelled datasets with different distributions from the labelled test and training target data might harm the accuracy of the model. This leads to the need of analyzing model robustness to different data distributions in the unlabelled dataset. Therefore, in this work, we study the impact of different unlabelled data sources in a SSDL model. Specifically, the MixMatch algorithm, which previously yielded interesting accuracy gains with very small labelled datasets for COVID-19 detection using X-ray images [10, 9] is used. Moreover, we propose a simple approach to select and build an unlabelled dataset. This aims to improve the overall SSDL model accuracy.

1.1. Problem Definition

In this work, we evaluate a setting where the following datasets are available:

1. A labelled dataset in the target clinic/hospital D_t^l is available. The number of labelled observations n_t^l is very small. The target dataset is sampled from the clinic/hospital where the model is intended to be deployed.
2. A larger unlabelled dataset in a different source clinic/hospital D_s^u is available, with $n_s^u > n_t^l$.

Different deep learning applications in medical imaging face distribution mismatch situations between the different datasets used. This might be the case for SSDL, when using different unlabelled data sources. We argue that quantifying distribution mismatch with respect to the model behaviour is important for medical imaging applications, as different unlabelled data sources might be considered. Moreover, simple dataset transformation procedures to improve model robustness to data distribution mismatch between the labelled and unlabelled datasets, is also important. This helps to narrow the gap between machine learning research and its real-world usage.

The first contribution of this work aims to first explore the impact of distribution mismatch between the labelled and unlabelled dataset in SSDL in a real-world application: COVID-19 detection using chest X-ray images. We examine different distribution mismatch settings with data from the specific domain only (chest X-ray images), different than classic testing benchmarks where distribution mismatch is caused by adding images from different domains. We explore the influence of using unlabelled data from different data sources from the same domain, and measure its impact in SSDL. The second contribution consists in two novel methods based upon the feature space of a generic pre-trained CNN, to score unlabelled data according to its likelihood in the distribution of the labelled data. Such scores are used to filter possibly harming unlabelled data, and improve the performance of the SSDL model by using the filtered unlabelled data.

1.2. Manuscript Organization

This manuscript is organized as follows: Section 2 studies recent literature around SSDL methods, and more specifically SSDL techniques designed to be robust to unlabelled data with a considerable distribution mismatch with respect to labelled data. In such section we also study Out of Distribution (OOD) detection techniques, as they are closely related to distribution mismatch robustness. Given the research gap described in Section 2, in Section 4 we propose our novel method to increase distribution mismatch robustness in a SSDL setting. We test our proposed method using the state of the art MixMatch algorithm [8]. The datasets used to create the different distribution mismatch tested throughout the experiments are described in Section 3. The detailed description of the experimental design is depicted in Section 5. An analysis of the yielded results and the initial observations is developed in Section 6, to later address the conclusions and future work in Section 7.

2. State of the art

2.1. Semi-supervised Deep Learning

SSDL aims to deal with small labelled datasets, by leveraging unlabelled data. Supervised deep learning networks often require large labelled datasets. This is partially addressed with the usage of data augmentation and transfer learning [73]. However, the usage of cheaper and more widely available unlabelled data, can further lower the need for labelled data. With a formal notation, in SSDL both labelled and unlabelled datasets are used. Each labelled observation $X_l = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_l}\}$ is mapped to a label in the set $Y_l = \{y_1, \ldots, y_{n_l}\}$. The unlabelled dataset corresponds to a set of observations $X_u = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_u}\}$, with $S_u = X_u$.

SSDL architectures can be classified as: Pre-training [23], pseudo-labelled [24] and regularization based. Within regularization based approaches, consistency loss term and graph based regularization and generative based [18] regularization techniques can be distinguished. A detailed survey regarding SSDL can be found in [74, 39].

Concerning regularization based SSDL, a regularization term leveraging unlabelled data is implemented in the loss function S_u:

$$\mathcal{L}\left(S\right) = \sum_{(\boldsymbol{x}_i, \boldsymbol{y}_i) \in S_l} \mathcal{L}_l\left(\boldsymbol{w}, \boldsymbol{x}_i, \boldsymbol{y}_i\right) + \gamma \sum_{\overrightarrow{x}_j \in X_u} \mathcal{L}_u\left(\boldsymbol{w}, \boldsymbol{x}_j\right), \tag{1}$$

with $\boldsymbol{w}$ the model's weights array, $\mathcal{L}_l$ and $\mathcal{L}_u$ the labelled and unlabelled loss terms respectively. The coefficient γ weighs the influence of unsupervised regularization. As previously mentioned, a number of regularization based variations can be found in the literature. The main ones include: consistency loss based [69, 68], graph based [76, 44] and generative augmentation based [64, 60]. Consistency based methods make the assumption of clustered-data/low-density separation. Such assumption refers to how the observations corresponding to a class, are clustered together. This makes the decision manifold lie in very sparse regions [74]. A violation to this assumption might degrade the performance of the semi-supervised method [74].

In pseudo-label training, pseudo-labels are estimated for unlabelled data. These are used for later model refinement. A straightforward pseudo-label based approach is based in co-training two models [4]. The model is pre-trained with the limited size labelled dataset. Later, the pseudo-labels are estimated for the unlabelled data using two models trained with different views (features) of the data. A voting scheme is implemented for estimating the pseudo-labels.

MixMatch [8] combines both pseudo-label and consistency based SSDL, along with heavy data augmentation using the MixUp algorithm [77]. According to [8], MixMatch out-performs, accuracy wise, previous SSDL approaches. Given the recently state of the art performance demonstrated by MixMatch and also the good results yielded in [9, 10] for medical imaging applications, we chose it for the developed solution in this work. A detailed description of MixMatch

Model	**Category**	$n_l = 500$	$n_l = 1000$	$n_l = 2000$
Supervised only	Supervised	22.08 ± 0.73 [62]	14.46 ± 0.71 [62]	-
Pi Model (Pi-M)	Consistency based SSDL	6.83±0.66 [69]	4.82±0.17[69]	-
Temporal Ensemble Model (TEM)		5.12±0.13[69]	4.42±0.16[57, 69]	-
Virtual Adversarial Training with Entropy Minimization (VATM+EM)		-	3.86±0.22[50]	-
Virtual Adversarial Training Model (VATM)		-	5.42±0.22[50]	-
Mean Teacher Model (MTM)		4.18±0.5 [69]	3.95±0.19[57, 69]	-
Self Supervised network Model (SESEMI)		6.5±0.28[71]	5.59±0.12[71]	-
Mutual Exclusivity-Transformation Model (METM)		9.62±1.37[27]	4.52±0.4[27]	3.66±0.14[27]
Walker Model (WaM)		6.25±0.32[27]	5.14±0.17[27]	4.6±0.21[27]
Transductive Model (TransM)		4.32±0.3[62]	3.8±0.27[62]	3.35±0.27 [62]
Transductive Model with Mean Teacher (TransM+MTM)		4.09±0.42[62]	3.09±0.27 [62]	3.35±0.27 [62]
Memory based Model (MeM)		-	4.21±0.12[16]	-
MixMatch	Consistency and Pseudo-label based SSDL	-	3.5±0.28	-
ReMixMatch		-	2.65±0.08	-
FixMatch using Random Augmentation		-	2.28±0.11	-
FixMatch using CTA Augmentation		-	2.36±0.19	-
Tri-Net	Pseudo-label based SSDL	-	3.71±0.14[24]	-
Speed as a supervisor for SSDL (SaaSM)		-	3.82±0.09[20]	-
Tri-Net with the Pi-M		-	3.45±0.1[24]	-

Table 1: SSDL error rates (the lower the better) from literature of state of the art methods, using the SVHN dataset. As number of labels, $n_l = 500$, $n_l = 1000$ and $n_l = 2000$ were the most frequently used in the literature.

can be found in Section 4. Table 1 quantitatively summarizes the reported accuracy performance of some of the most recent SSDL approaches. The results suggest that MixMatch and similar methods yield the lowest error rates. The reported results used the Street View House Numbers dataset (SVHN) dataset. Based upon the good results of MixMatch compared to other state of the art methods, we selected it to test our proposed data-centric method to improve SSDL robustness to OOD data.

2.2. SSDL robustness to distribution mismatch

The distribution mismatch between S_u and S_l is also referred to as the identically and Independent and Identically Distributed (IID) assumption violation. It might have different degrees and causes, which are enlisted as follows [35]:

- Prior probability shift: The distribution of the labels in S_l can be different when compared to S_u. In a CAD system this can be exemplified when the labels of the medical images have different distributions between the two datasets S_l and S_u. A specific case would be the label imbalance of the labeled dataset S_l as discussed in [10].

- Covariate shift: A different distribution of the features in the input observations might be sampled, leading to a distribution mismatch. In a medical imaging application, this can be related to the difference in the frequencies of the observed features between S_l and S_u.

- Concept drift: It refers to the different features observed in a sample, with the same label. In the application at hand in this work, this might happen when different patients with different variations of the COVID-19 disease are sampled to build S_u with the same pathologies (classes) in S_l.

- Concept shift: It is associated to a shift in the labels, with the same features. In the aforementioned example, it would refer to labelling a medical image with similar features with a different pathology (a bias caused by the image labellers).

- Unseen classes: The dataset $S^{(u)}$ contains observations of unseen or unrepresented classes in the dataset $S^{(l)}$. One or more distractor classes are sampled in the unlabelled dataset. Therefore, a mismatch in the number of labels exist, along with a prior probability shift, and a feature distribution mismatch. For instance, the dataset $S^{(l)}$ might include only the classes *viral pneumonia* and *normal*, while the unlabelled dataset might include the classes *bacterial pneumonia*, *viral pneumonia* and *normal*.

In our tested setting, different data sources were used only to gather unlabelled data S_u. We recreate two of the aforementioned distribution mismatch causes: covariate and prior probability shift. The unlabelled datasets created and tested belong to normal (no pathology) chest X-ray images (COVID-19$^-$), from patients of different nationalities. As the labelled dataset S_l includes both classes (COVID-19$^+$ and COVID-19$^-$), a label distribution mismatch also occurs. The tested setting in this work simulates the case where different unlabelled data sources might be available (for instance from different hospitals), at the beginning of a pandemic. Furthermore, a small labelled dataset might be available in the target hospital/clinic.

The usage of different unlabelled datasets might potentially cause a violation of the aforementioned clustered-data/low-density separation assumption. Using unlabelled datasets with different distributions when compared to the labelled dataset, might create wrong sparse regions and/or less clustered groups of observations belonging to the same class. Therefore, in this work we explore data-oriented approaches to deal with potential violations of the clustered-data/low-density separation assumption. Unlabelled data can be considered significantly cheaper than labelled data. Thus, discarding potentially harmful observations with the aim to decrease the odds of violating the clustered-data/low-density separation assumption is viable and worthy to explore.

In [55], an extensive evaluation of different distribution mismatch settings and its impact in SSDL is developed. Authors concluded that distribution mismatch in SSDL is an important challenge to be addressed. Recently, different approaches for improving SSDL robustness to the distribution mismatch between S_u and S_l have been proposed. In [52], an OOD masking method is proposed, referred to as RealMix. It consists on weighting the observations likely to be OOD during semi-supervised training. The output of a softmax activation function after the raw model output, was used as OOD masking coefficient. A hard thresholding was applied to the unlabelled data, in order to discard OOD data. This works as an observation-wise masking during semi-supervised model training. The authors compared their proposed method with state of the art general-purpose SSDL approaches like MixMatch [8]. The test bed consisted in different unlabelled datasets with a varying degree of distribution mismatch. The contamination source consists of images with different labels and features (completely OOD), corresponding to the unseen class IID violation cause. Their method proved to improve model robustness against OOD data contamination in S_u, using general purpose datasets such as Canadian Institute For Advanced Research dataset with 10 classes (CIFAR-10) and SVHN. However, other types of distribution mismatch corruption such as concept drift or covariate shift were not tested.

Another approach to deal with distribution mismatch under OOD contamination (different labels and features), can be found in [17]. The proposed method also implements a weighting coefficient, calculated as the softmax output of a models ensemble. It is referred to as Uncertainty Aware Self-Distillation (UASD) by the authors. Similar to RealMix, a hard thresholding of the OOD data was proposed. However, more diverse distribution mismatch scenarios were tested, using different degrees of contamination using unseen classes as pollution source. In a similar trend, the work in [26] propose a weighted approach to deal with OOD observations (with different label, different features). The proposed method was named Deep Safe Semi-Supervised Learning (DS3L) by the authors. However, instead of using the softmax output, the observation-wise weight is estimated through an optimization step. The score or weight obtained for each observation, is used to weight it in the unlabelled loss term, instead for discarding the data. We refer to this approach as soft thresholding. Similar to [52], only general purpose datasets (CIFAR-10 and Modified National Institute of Standards and Technology dataset (MNIST), using approximately half of the dataset as unseen classes in the unlabelled dataset) were used, with no other variations of distribution mismatch settings. Another resembling approach and testing bed to [26], can be found in [78], where an optimization based approach to weight each observation is implemented, with a test-bed focused in OOD contaminated unlabelled datasets. To diminish the computational cost of estimating the observation-wise weights for the unlabelled data, a clustering step was implemented. The cluster centroids were used to calculate the weights for all the observations within the cluster. The method is referred to as Robust Semi-Supervised Learning (R-SSL) by the authors.

In this work, we analyze the effect of distribution mismatch in SSDL within a real-world application: COVID-19 detection using chest X-ray images. Unlike previous work on SSDL under distribution mismatch, we test a real-world setting in the medical domain, and explore its implications within such context. As previously mentioned, we analyze the impact of a distribution mismatch caused by covariate and prior probability shift. Different unlabelled dataset sources within the same domain and features are used. We aim to evaluate dif-

Method name	**IID violation cause**	**Thresholding**	**OOD data filtering approach**
RealMix	Unseen classes	Hard	Output based
UASD	Unseen classes	Hard	Output based
DS3L	Unseen classes	Soft	Optimization based
R-SSL	Unseen classes	Soft	Optimization based

Table 2: State of the art SSDL methods robust to distribution mismatch. The *unseen classes* setting is the most tested cause for distribution mismatch. Our proposed method tests covariate and prior probability shift causes for distribution mismatch, and implements a feature space based method for scoring unlabelled data.

ferent approaches to weigh how harmful an unlabelled observation could be for SSDL training. We test different OOD detection approaches in this work. After calculating a *harm* coefficient for each unlabelled observation, different steps can be implemented to use such unlabelled dataset. For example, filtering the observations with high *harm* coefficients, select an unlabelled dataset upon its estimated benefit for SSDL, or weigh the unlabelled observation during SSDL training.

Moreover, we focus on a data-oriented approach to identify and/or build a good unlabelled dataset for SSDL. We propose a simple and very inexpensive method to evaluate the distribution mismatch between an unlabelled and labelled datasets, S_u and S_l respectively. Such method can be thought as an OOD scoring approach (*harm* coefficient), which leads us to compare our method to recent OOD detectors used in the context of OOD data filtering to improve the accuracy of an SSDL model. Unlike most recent SSDL methods which use output or optimization based scoring for the unlabelled data, our approach uses the feature space, as seen in very recent OOD detection approaches. This research gap can be inferred by the state of the art summary table for SSDL robust methods, in Table 2.

2.3. OOD data detection

OOD data detection refers to the general problem of detecting observations that are very unlikely given a specific data distribution (usually the training dataset distribution) [29]. The problem of OOD data detection can be thought as a generalization of the outlier detection problem, as it considers individual and collective outliers [63]. Specific scenarios of OOD data detection can be found in the literature. These include novel data and anomaly detection [56], with several applications like rare event detection [28, 2]. In classical pattern recognition literature different approaches to anomaly and OOD data detection are grounded in concepts such as density estimation [47], kernel representations [70], prototyping [47] and robust moment estimation [59].

Recent success of deep learning based approaches for image analysis [75] have motivated the development of OOD detection techniques for deep neural networks. OOD detection methods with deep learning architectures can be categorized in methods based upon the Deep Neural Networks (DNN)'s output, its input, or its learned feature space.

DNN's output based methods include the softmax based OOD detector proposed in [30]. In such work, OOD detection is framed as a confidence estimation using the model's raw output layer values and passing it through a softmax function. Its maximum softmax value is used as confidence. Authors claim that the highest softmax value of OOD observations meaningfully differ from in distribution observations.

However, as reported in [42], non calibrated models can be overconfident with OOD data. Therefore, in [42] a calibration methodology is introduced, implementing a temperature coefficient. OOD data detection in neural networks is implemented in [42] using input perturbations meant to maximize the softmax based separability. For this end, a gradient descent optimization is used, resulting in a preprocessed image. A *temperature* coefficient in the calculation of the softmax output is added and is estimated to make the true positive rate of 95% for in-distribution data detection, using the previously pre-processed images.

Another approach for OOD detection based on the model's output is the usage of Monte Carlo Dropout (MCD) based uncertainty estimations.MCD is a popular method for implementing predictive uncertainty estimation [43, 37]. It consists in analyzing the distribution of N predictions using the same input and adding noise to the model (drop-out in the context of DNNs). This idea has been ported to the OOD detection problem, where observations with high uncertainty are scored with high OOD likelihood [34, 61].

Regarding feature space (a latent space approximation in DNNs) based methods for OOD detection different approaches can be found in the literature. For example, in [41], the authors implemented the Mahalanobis distance in latent space of the dataset to the input observation, assuming a Gaussian distribution of the data. Both the mean and covariance are estimated for the in distribution dataset. For a new observation $\boldsymbol{x}$, the OOD score is estimated as the Mahalanobis distance for such given distribution. The authors also implemented the calibration approach used in [42]. A superior performance of their proposed method in generic OOD detection benchmarks is reported, when compared to the methods in [42, 30]. However, no statistical significance tests of the results were performed.

Another feature space based approach can be found in [72], known as deterministic uncertainty quantification. Such approach is also intended for uncertainty estimation, but also is tested as an OOD detection technique. It makes use of a centroid calculation of each category in the feature space, to later quantify the distance of a new observation to each centroid. Uncertainty quantification is estimated based in the kernel based distance to the category centroids. The approach is compared against an ensemble of deep neural networks (an output based approach for OOD detection). This is done in a simple OOD detection benchmark, where the CIFAR-10 is used as an in-distribution dataset and the SVHN as a OOD dataset. The authors reported the area under the Receiver Operator Characteristic (ROC) curve of their approach against other OOD methods. Their approach showed the highest area under the ROC curve index. However, no statistical analysis of the results were done.

In [12] the authors developed an extensive testing of the influence of distribution mismatch between unlabelled and labelled datasets. Moreover, they also developed an approach to estimate the accuracy hit of such distribution mismatch for a state of the art SSDL method. The proposed method estimates the distribution mismatch in the feature space between S_l and S_u, using what the authors referred as a Deep Dataset Dissimilarity Measure (DeDiM). Euclidean and Manhattan based DeDiMs were tested and compared against density based DeDiMs. All of them were applied within the feature space, built with an image net pre-trained network. The authors found a significant advantage of the density based distances. In [80], the authors proposed 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 In-Distribution (IOD) data is done based on the estimation of the approximated parametric model. Unfortunately, no comparison with other popular OOD methods was presented. Table 2.3 describes a summary of the state of the art methods and the benchmarks used to test them by the authors. This summary makes clear how most previous OOD detection methods have focused in the *unseen class* distribution mismatch cause. In this work we evaluate the covariate shift cause for a distribution mismatch between the labelled and unlabelled datasets in a real-world application, used by a SSDL method. Additionally we propose a simple feature based approach to improve SSDL performance under those circumstances, as few very recent OOD detection approaches have proposed.

2.3.1. Unsupervised Domain Adaptation

When using an unlabelled dataset S_u with a very different distribution to S_l, a solution would be to *correct* or *align* the feature extractor trained with labelled or unlabelled data from the source of the unlabelled dataset S_u, to the distribution of the labelled dataset S_l (target dataset, usually smaller). This is known as Unsupervised Domain Adaptation (UDA). For instance in [79], the authors proposed an UDA method to align the feature extractor from a source dataset to a specific target dataset. This is done within the context of COVID-19 detection using chest X-ray images. The feature extractor was originally trained with source data. Later, the feature extractor is aligned by using both labelled and unlabelled data from the target dataset. The feature

Method name	**IOD data**	**OOD data**	**Category**
Max. value of Softmax layer [30]	CIFAR-10 ^{1}	SUN^{1,2}	Output based
	CIFAR-100 ^{2}	Gaussian ^{1,2}	
	MNIST ^{3}	Omniglot ^{3}	
		notMNIST^{3}	
		Uniform noise^{3}	
Inhibited Softmax [51]	CIFAR-10^{1}	SVHN^{1}	
	MNIST^{2}	LFW-A^{1}	
		notMNIST^{2}	
		Omniglot^{2}	
ODIN [42]	CIFAR-10^{1}	TinyImageNet^{1,2}	
	CIFAR-100^{2}	LSUN^{1,2}	
		iSUN^{1,2}	
		Uniform^{1,2}	
		Gaussian^{1,2}	
Epistemic Uncertainty Estimation [67]	CIFAR *^{1}	CIFAR*^{1}	
	FashionMNIST*^{2}	FashionMNIST*^{2}	
	SVHN*^{3}	SVHN*^{3}	
	MNIST*^{4}	MNIST*^{4}	
Mahalanobis Latent Distance [41]	CIFAR-10^{1}	SVHN^{1,2}	Feature space based
	CIFAR-100^{2}	CIFAR-10^{3}	
	SVHN^{3}	TinyImageNet^{1,2,3}	
		LSUN^{1,2,3}	
Deterministic Uncertainty quantification	CIFAR-10	SVHN	
Deep Residual Flow [80]	CIFAR-10^{1}	CIFAR-10^{3}	
	CIFAR-100^{2}	TinyImageNet^{1,2,3}	
	SVHN^{3}	LSUN^{1,2,3}	
		SVHN^{1,23}	

Table 3: OOD test benchmarks for different techniques. Datasets with * were randomly cut by half for in-distribution training labelled data and the other half was used as OOD unlabelled data. The table reveals how arbitrary different testbeds have been used for benchmarking OOD detection algorithms, using the *unseen classes* cause for the IID assumption violation. IOD-OOD dataset pairs are indicated by number pairs in the table.

extractor alignment procedure basically consists in an adversarial training step using the aforementioned datasets. As a disadvantage of such method, the feature extractor needs to be trained with labelled source data (as usual in supervised learning). Hence a large number of labels is needed. Also, the feature extractor alignment process can be considered to be expensive, as an adversarial loss function needs to be optimized.

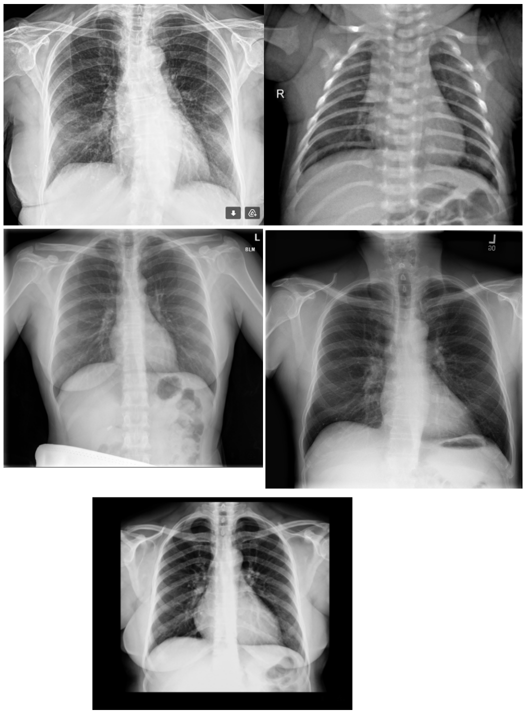
3. Datasets

In this work, we explore the sensitivity to distribution mismatch between S_u and S_l of a SSDL COVID-19 detection system using chest X-ray images. Therefore, we use different data sources for chest X-ray images for both COVID-19$^+$ (positive COVID-19) and COVID-19$^-$ (no pathology chest X-ray observations). For COVID-19$^+$ cases we use the open dataset made available by Dr. Cohen in [21]. This dataset is composed of 105 COVID-19$^+$ images at the time of writing this work. The observations were sampled from different journal websites like the Italian Society of Medical and Interventional Radiology and `radiopaedia.org`, and more recent publications in the field. In this work we used COVID-19$^+$ observations, discarding images related to Middle East Respiratory Syndrome (MERS), Acute Respiratory Distress Syndrome (ARDS) and Severe Acute Respiratory Syndrome (SARS).

The images present varying resolutions from 400 × 400 up to 2500 × 2500 pixels. As for COVID-19$^-$ observations, we used four different data-sources. Table 4 summarizes the COVID-19$^-$ cases data sources. Figure 1 shows observations for each one of the data sources used in this work. The datasets were randomly augmented with flips and rotations. No random crops were used to avoid discarding important regions in the images.

In this first set of experiments, we evaluate the impact of OOD on data with different unlabelled data sources and different degrees of *contamination*. We simulate the following scenario: A small labelled target dataset D_l^t (with $n_l = 20$ and $n_l = 40$ observations) is provided with a partition of the observa-

Figure 1: Row 1, column 1: a COVID-19^{+} observation from [21], row 1, column 2: a COVID-19^{-} observation from the Chinese dataset [38], row 2, column 1: ChestX-ray8 COVID-19^{-} image [31], row 2, column 2: Indiana dataset COVID-19^{-} sample image [22]. The bottom image corresponds to a sample image from the Costa Rica dataset [10]. As it can be seen, images from the Costa Rica dataset include a black frame.



tions of the COVID-19^{+} taken from Dr. Cohen's dataset and the COVID-19^{-} cases of the Indiana Chest X-ray dataset, described in Table 4. A larger number of 142 unlabelled observations is also available, to be used in the harm coefficient estimations methods. This can be thought as the target labelled dataset with limited labels which is accessible in a real-world application from the clinic/hospital where the model is intended to be deployed.

For the unlabelled dataset, different partitions of COVID-19^{-} cases the chest X-ray data sources described in Table 4. This simulates the usage of different sources of unlabelled datasets D_u^s, taken from different hospitals/clinics. All the unlabelled observations are COVID-19^{-}, to enforce a prior probability shift (label imbalance). As in our preliminary tests, the worst performing unlabelled dataset D_u^s dataset is the Costa Rican dataset described in Table 4, we used it to create different combinations with the rest of datasets. All of these are depicted in Table 7. A total of $n_u = 90$ unlabelled observations were picked from such datasets with different combinations. Using different data sources for the unlabelled dataset, can help to assess the impact of a distribution mismatch between S_u and S_l.

As for the test dataset, it consists in another partition of the target dataset which includes the COVID-19^{+} dataset, along with another partition of the Indiana Chest X-ray dataset (COVID-19^{-}). Both are the same size. This yields a completely balanced test setting. We used a total of $n_t = 62$ observations, drawn from the same target dataset (31 observations per class). The test data comes from the distribution of the labelled data with no contamination. This simulates the case where the labelled data comes from the target dataset distribution. Both unlabelled and labelled datasets were standardised, given that the authors in [13] found that normalisation is important in semi-supervised learning.

Table 4: COVID-19^{-} observation sources description used in this work.

Dataset	**CR**	**Chinese**	**ChestX-ray8**	**Indiana**
No. of patients	105	5856	65240	4000
Patient's age range (years)	7-86	children	0-94	adults
No. of obs.	105	5236	224316	8121
Hospital/clinic	Clinica Chavarria	No info.	Stanford Hospital	Indiana Network for Patient Care
Im. resolution	1907 × 1791	1300 × 600	1024 × 1024	1400 × 1400
Reference	[10]	[38]	[31]	[22]

4. Proposed method

4.1. SSDL with MixMatch

In this work, we explore the usage of MixMatch as an SSDL method, therefore, we describe it as follows. We selected MixMatch as a baseline method given its good performance compared to other state of the art methods, as described in Table 1. For more details please refer to [8]. As previously mentioned, MixMatch combines both pseudo label and consistency regularization SSDL. In such context, a pseudo-label $\widehat{\boldsymbol{y}}_j$ is estimated for each unlabelled observation $\boldsymbol{x_j}$ in X_u. It corresponds to the the mean model output of a transformed input $\boldsymbol{x'_j}$, using K number of different transformations, such as flips and rotations [8]. Each pseudo-label $\widehat{\boldsymbol{y}}$ is sharpened using a temperature parameter T [8]. Also, a simple data augmentation approach is implemented, by linearly combining unlabelled and labelled observations, through the usage of the MixUp algorithm [77].

The pseudo-labels are used in the MixMatch loss function, which combines a supervised and unsupervised loss terms. In this work, the well-known cross-entropy function is used as a supervised loss term. As for the unsupervised loss term, we used the previously implemented Euclidian distance loss in [8]. The Euclidian distance measures the distance between the current model output and its pseudo-label, for the unlabelled observations. This loss term is weighed by the unsupervised learning coefficient γ. In this work, we used the MixMatch hyper-parameters recommended in [8], of $K = 2$, and $T = 0.25$. As for the unsupervised coefficient, a value of $\gamma = 200$ is used, given our empirical test results.

4.2. Harm coefficient estimation for unlabelled observations

Interesting results were yielded in [12, 11], where the authors found an strong correlation between the feature-density based distances and the MixMatch's accuracy. Based upon it, we propose to estimate how harmful an individual unlabelled observation might be towards the MixMatch's level of accuracy. We refer to this operator as the SSDL harm coefficient $\mathcal{H}\left(\boldsymbol{x}_j^u\right)$, where $\boldsymbol{x}_j^u \in S_u$. We aim to implement a simple and computationally inexpensive method to filter OOD data in the unlabelled dataset, This is done in order to decrease the distribution mismatch between S_u and S_l.

As mentioned in Section 2, using different unlabelled data sources might increase the chance of violating the clustered-data/low-density separation assumption. This is particularly the case given the potential distribution mismatch between the labelled and unlabelled datasets. Therefore, our proposed method aims to discard harmful observations that might create wrong low density regions to build the manifold and/or sparser sample clusters for each category. In a real-world scenario for OOD filtering, DNNs are fed with high resolution images, frequently with images from the same domain (chest X-ray images in our case). This contrasts with the usual settings of the methods discussed in Section 2. As previously discussed, benchmarking in the literature have been usually performed with small resolution images and with relatively not very difficult OOD detection challenges (i.e distinguishing between CIFAR-10 and MNIST images). We aim to further test real-world distribution mismatch conditions in a medical imaging analysis application such as the COVID-19 detecion using chest X-ray images.

In this work, we propose to use the feature density of a labelled dataset S_l, to weigh how harmful could be to include an unlabelled observation $\boldsymbol{x}_j^u$ in the unlabelled dataset S_u. This is done witin the context of training a model using the SSDL algorithm known as MixMatch. This harmful coefficient is represented as $\mathcal{H}\left(\boldsymbol{x}_j^u\right)$. We test two different variations to estimate $\mathcal{H}\left(\boldsymbol{x}_j^u\right)$. The first one consists in a non-parametric estimation of the feature density through an histogram calculation. The second variation assumes a Gaussian distribution of the feature space, by using a Mahalanobis distance. We use a generic feature-space built from a pre-trained image-net model, to keep the computational cost of the proposed method low. For all the tested configurations, we only use the features of the final convolutional layer. Computational resource restrictions for solving a real-world problem in medical imaging makes very expensive to use all the features extracted in the different layers as done in [41]. The procedure to calculate the harm coefficient using both methods, is depicted as follows:

1. For all of the input observations $\boldsymbol{x}_j^l \in S_l$, with $\boldsymbol{x}_j^l \in \mathbb{R}^n$, being n the input space dimensionality, using the feature extractor f, we calculate its feature vector as $\boldsymbol{h}_j^l = f\left(\boldsymbol{x}_j^l\right)$.
2. The feature vector $\boldsymbol{h}_j^l \in \mathbb{R}^{n'}$ has dimension n' , with $n' < n$. For instance, a given feature extractor f using the Imagenet pretrained Wide-ResNet architecture, yields $n' = 512$ features. For architectures such as densenet that might yield larger feature arrays in its final convolutional layer, we sub-sampled it to keep it in $n' = 1024$ features, using an average pooling operation. This yields a feature set H_l.
3. For the Feature Histograms (FH) method, we perform the following steps:
   (a) For each dimension $r = 1, ..., n'$ in the feature space, we compute its normalized histogram to approximate the density functions $\widetilde{p}_r^l$, in the sample H_l. This yields the set of approximated feature density functions:
   $$\widetilde{P}^l = \left\{\widetilde{p}_1^l, \ldots, \widetilde{p}_{n'}^l\right\} \tag{2}$$
   (b) Using the approximated feature densities in $\widetilde{P}^l$, we estimate our SSDL harm coefficient $\mathcal{H}\left(\boldsymbol{x}_j^u\right)$, for an unlabelled observation in the following steps $\boldsymbol{x}_j^u$.
   (c) Calculate the features for each unlabelled observation as $\boldsymbol{h}_j^u = f\left(\boldsymbol{x}_j^u\right)$, for each dimension in $\boldsymbol{h}_j^u \in \mathbb{R}^{n'}$,

(d) The total likelihood calculation within the density function approximation set $\widetilde{P}^l$ assumes that each dimension is statistically independent. Thus:

$$\prod_{r=1}^{n'} p_r^l \left(h_{j,r}^u\right). \tag{3}$$

(e) To avoid under-flow, we calculate the negative logarithm of the likelihood, and use it as the harm coefficient:

$$\mathcal{H}\left(\boldsymbol{x}_j^u\right) = -\sum_{r=1}^{n'} \ln\left(p_r^l\left(h_{j,r}^u\right)\right). \tag{4}$$

4. For the Mahalanobis based filtering, we perform the following steps:
   (a) Calculate the covariance matrix Σ from the features set H_l, and the sample mean from the features set $\overline{h}_l$.
   (b) Calculate the features for each unlabelled observation as $\boldsymbol{h}_j^u = f\left(\boldsymbol{x}_j^u\right)$.
   (c) Compute the harm coefficient as:

$$\mathcal{H}\left(\boldsymbol{x}_j^u\right) = \left(\overline{\boldsymbol{h}}_l - \boldsymbol{h}_j^u\right)^T \Sigma^{-1}\left(\overline{\boldsymbol{h}}_l - \boldsymbol{h}_j^u\right). \tag{5}$$

The harm coefficient $\mathcal{H}\left(\boldsymbol{x}_j^u\right)$ can be used to discard the observations with high values, or to weigh them in case an online semi-supervised per-observation weighting is implemented. In this work, we test the impact of the distribution mismatch between the labelled target and unlabelled source datasets, D_t^l and D_s^u, respectively, in the accuracy of the SSDL MixMatch algorithm. Later, we test the impact of the proposed feature based *harm coefficient* to eliminate potentially harming observations from the unlabelled dataset. This was done to assess the accuracy of the model using the filtered unlabelled dataset D_s^u. This way, we can assess in a controlled setting the impact of the distribution rectification procedure, implemented through a data filtering process. Figure 2 summarizes both proposed methods.

5. Experiments

5.1. Experiment Design

Test-bed 1 (TB-1) is designed to assess the effect of on MixMatch's accuracy of using different unlabelled datasets D_u^s with a target labelled dataset D_l^t. As

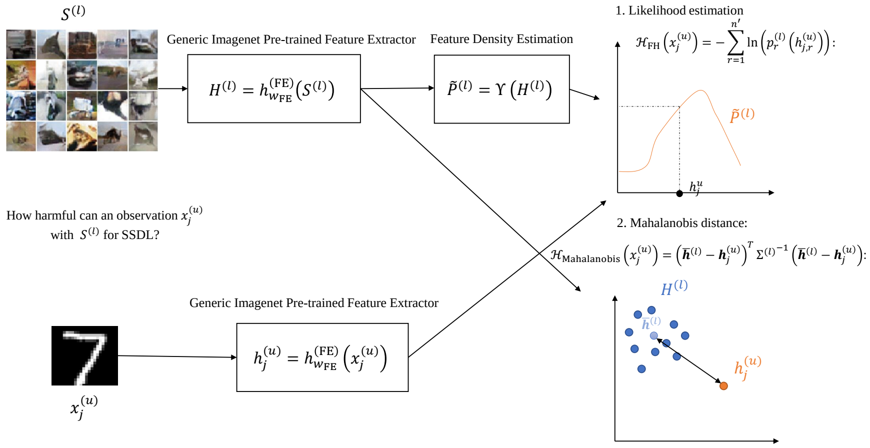

$S^{(l)}$
Generic Imagenet Pre-trained Feature Extractor
$H^{(l)} = h_{w_{\mathrm{FE}}}^{(\mathrm{FE})}\left(S^{(l)}\right)$
Feature Density Estimation
$\tilde{P}^{(l)} = \Upsilon\left(H^{(l)}\right)$
1. Likelihood estimation
$\mathcal{H}_{\mathrm{FH}}\left(x_j^{(u)}\right) = -\sum_{r=1}^{n'} \ln\left(p_r^{(l)}\left(h_{j,r}^{(u)}\right)\right)$:
$\tilde{P}^{(l)}$
h_j^u
How harmful can an observation $x_j^{(u)}$ with $S^{(l)}$ for SSDL?
2. Mahalanobis distance:
$\mathcal{H}_{\mathrm{Mahalanobis}}\left(x_j^{(u)}\right) = \left(\bar{\boldsymbol{h}}^{(l)} - \boldsymbol{h}_j^{(u)}\right)^T \Sigma^{(l)^{-1}}\left(\bar{\boldsymbol{h}}^{(l)} - \boldsymbol{h}_j^{(u)}\right)$:
$H^{(l)}$
$\bar{h}^{(l)}$
$h_j^{(u)}$
Generic Imagenet Pre-trained Feature Extractor
$h_j^{(u)} = h_{w_{\mathrm{FE}}}^{(\mathrm{FE})}\left(x_j^{(u)}\right)$
$x_j^{(u)}$

Figure 2: Summary of the proposed unlabelled data scoring methods for SSDL, $\mathcal{H}_{\mathrm{FH}}$ and $\mathcal{H}_{\mathrm{Mahalanobis}}$.

error measure we use the accuracy in a balanced test dataset. This test-bed recreates different distribution mismatch conditions between D_u^s and D_l^t. The Costa Rican dataset acts as a source of OOD data, as it yielded the lowest accuracy when used as D_u^s for MixMatch, among the empirically tested unlabelled data sources. We combine the aforementioned data sources with the Costa Rican dataset. This helps enforce different distribution mismatch settings.

In the Test-bed 1.1 (TB-1.1), the first sub-experiment defined within the TB-1, we measure MixMatch's accuracy using a densenet model, with feature extractor fine-tuning and without it. As error measure we also use the accuracy in a balanced test dataset. We aim to measure if there is a significant accuracy gain of fine-tuning the feature extractor during training. Table 5 shows the results of performing MixMatch's training without feature extractor fine-tuning, while Table 6 shows the results with it.

Additionally, we devised a Test-bed 1.2 (TB-1.2), where the baseline results obtained in this MixMatch accuracy baseline test-bed in Tables 5 and 7 are correlated with the cosine DeDiMs between each D_u^s and D_u^s. This is measured as proposed in [13], and represented as $d_C(D_u^s, D_l^t)$. We measure the linear correlation between the model's accuracy and its measured labelled-unlabelled dataset distance. For this experiment, we tested an alexnet's model feature extractor, given its low computational cost. We implemented the cosine dataset DeDiM with a batch dataset size of $n_b = 40$, with 10 batches of random samples. The same batches were used to test the different configurations. Similar to the proposed harm coefficient estimation methods, we used a generic Imagenet pre-trained feature extractor to build the feature density estimations, as proposed in [13]. The DeDiM results are linearly correlated using a Pearson coefficient in Table 9. We performed a Wilcoxon test to verify whether there is a statistically significance difference when comparing: feature extractor fine-tuning vs. no feature extractor fine-tuning, the two proposed methods to each one of the previous methods (softmax and MCD based), and the proposed Mahalanobis method vs. the also proposed FH approach, with $p < 0.05$.

Finally, Test-bed 2 (TB-2) aims to assess MixMatch's accuracy results when implementing the proposed methods in this work to filter the OOD observations, against two popular output based OOD filtering methods: the MCD and Softmax based OOD filters. In this test bed, we measure MixMatch's accuracy through the four different filtered datasets, testing both alexnet and densenet as a model. We also tested the model with $n_l = 20$ and $n_l = 40$ labels. The results using the proposed feature histograms and Mahalanobis distance for each generated unlabelled data source D_u^s are depicted in Tables 11 and 13, for the alexnet and the densenet models, respectively. To filter possible OOD observations, we eliminated the same percent of contaminated observations using the Costa Rican dataset (i.e, if the Chinese dataset was contaminated with 35% of observations with the Costa Rican dataset, we eliminated 35% of the observations with the highest harm coefficient, and so on). We leave the problem of defining the right harm coefficient threshold out of this study.

In all test beds, the MixMatch algorithm is tested with a densenet and alexnet models, using the recommended parameters in [8], along with an unsupervised regularization term coefficient of 200. As for model training, we use the one-cycle policy implemented in the FastAI library, with a weight decay

Table 5: TB-1.1 results: Accuracy of a Densenet model trained with MixMatch with different D_u^s datasets. The unlabelled datasets Chest-Xray8, Costa Rican and Chinese datasets include only COVID-19$^-$ observations. No use of a fine-tuned feature extractor.

Dataset	$n_l = 40$	$n_l = 20$
Supervised	0.851 ± 0.037	0.803 ± 0.039
Indiana (with COVID-19$^+$ [21])	0.891 ± 0.047	0.875 ± 0.04
China	0.735 ± 0.0621	0.722 ± 0.054
Costa Rica	0.493 ± 0.014	0.511 ± 0.029
ChestX-ray8	0.825 ± 0.061	0.795 ± 0.052
ChestX-ray8 65% - Costa Rica 35%	0.579 ± 0.115	0.582 ± 0.067
ChestX-ray8 35% - Costa Rica 65%	0.5 ± 0.001	0.503 ± 0.009
China 65% - Costa Rica 35%	0.588 ± 0.066	0.559 ± 0.067
China 35% - Costa Rica 65%	0.498 ± 0.004	0.508 ± 0.024
Indiana 65% - Costa Rica 35%	0.504 ± 0.014	0.553 ± 0.062
Indiana 35% - Costa Rica 65%	0.501 ± 0.004	0.5 ± 0.001

of 0.001, This way we can measure MixMatch's behaviour with models with different depth and architecture. For each configuration, we trained the model with 10 runs, using a different random data partition for training and test, for 50 epochs.

Finally, Table 14 shows the average and standard deviation of the execution time in seconds for the tested harmful data filters. As for the data load of the aforementioned tests, $n_l = 142$ and $n_u = 90$ observations were used. For these performance tests, a densenet backbone was used. The Mahalanobis based method is the fastest with an execution time of around 65.1 secs. in average and a standard deviation of 2.3 secs. (for a typical data load of the test bench), when compared to the histogram based approach. The Mahalanobis method was the fastest with statistical significance according to our Wilcoxon test, when compared to the rest of the evaluated methods.

Table 6: TB-1.1 results: Accuracy of a Densenet model trained with MixMatch with different D_u^s datasets. The unlabelled datasets Chest-Xray8, Costa Rican and Chinese datasets include only COVID-19$^-$ observations. Using the fine-tuned feature extractor.

Dataset	$n_l = 40$	$n_l = 20$
Supervised	0.852 ± 0.045	0.795 ± 0.005
Indiana (with COVID-19$^+$ [21])	0.892 ± 0.044	0.885 ± 0.039
China	0.733 ± 0.043	0.709 ± 0.059
Costa Rica	0.498 ± 0.004	0.501 ± 0.016
ChestX-ray8	0.804 ± 0.061	0.793 ± 0.044
ChestX-ray8 65% - Costa Rica 35%	0.598 ± 0.1	0.591 ± 0.105
ChestX-ray8 35% - Costa Rica 65%	0.501 ± 0.004	0.488 ± 0.033
China 65% - Costa Rica 35%	0.593 ± 0.057	0.614 ± 0.0926
China 35% - Costa Rica 65%	0.514 ± 0.055	0.496 ± 0.022
Indiana 65% - Costa Rica 35%	0.516 ± 0.048	0.535 ± 0.047
Indiana 35% - Costa Rica 65%	0.508 ± 0.016	0.501 ± 0.011

Table 7: TB-1.1 results: Accuracy of a Alexnet model trained with MixMatch with different D_u^s datasets. The unlabelled datasets Chest-Xray8, Costa Rican and Chinese datasets include only COVID-19$^-$ observations.

Dataset	$n_l = 40$	$n_l = 20$
Supervised	0.785 ± 0.038	0.809 ± 0.085
Indiana (with COVID-19$^+$ [21])	0.782 ± 0.039	0.75 ± 0.06
China	0.648 ± 0.0247	0.659 ± 0.033
Costa Rica	0.501 ± 0.001	0.5 ± 0.001
ChestX-ray8	0.72 ± 0.076	0.71 ± 0.074
ChestX-ray8 65% - Costa Rica 35%	0.711 ± 0.083	0.66 ± 0.11
ChestX-ray8 35% - Costa Rica 65%	0.516 ± 0.022	0.511 ± 0.016
China 65% - Costa Rica 35%	0.701 ± 0.055	0.688 ± 0.084
China 35% - Costa Rica 65%	0.53 ± 0.023	0.528 ± 0.019
Indiana 65% - Costa Rica 35%	0.532 ± 0.024	0.559 ± 0.059
Indiana 35% - Costa Rica 65%	0.501 ± 0.001	0.503 ± 0.009

Table 8: TB-1.2 results: Cosine DeDiM distance, using 10 different batches of 80 observations, between the labelled and unlabelled datasets, S_l and S_u, respectively. Using Alexnet, to keep computing cost low.

Dataset	$d(S_l, S_u)$
China	2.06 ± 0.11
Costa Rica	30.9 ± 0.4
ChestX-ray8	1.04 ± 0.27
ChestX-ray8 65% - Costa Rica 35%	3.95 ± 0.94
ChestX-ray8 35% - Costa Rica 65%	11.84 ± 0.94
China 65% - Costa Rica 35%	5.74 ± 0.79
China 35% - Costa Rica 65%	14.85 ± 0.0
Indiana 65% - Costa Rica 35%	6.33 ± 0.3
Indiana 35% - Costa Rica 65%	16.61 ± 0.3

Table 9: TB-1.2 test results: Pearson coefficient between the accuracy and the calculated divergences.

SSDL model	n_l	**Pearson coefficient**
Alexnet	20	-0.798
	40	-0.75
Densenet	20	-0.665
	40	-0.662

Table 10: Accuracy of a Alexnet model trained with MixMatch, with the filtered datasets using the harm coefficient with the two output-based methods: MCD and Softmax. The percentage of discarded observations is the same of the amount of Costa Rican observations.

	$n_l = 40$		$n_l = 20$	
Dataset	Acc. Softmax	Acc. MCD	Acc. Softmax	Acc. MCD
ChestX-ray8 35% - Costa Rica 65%	0.532 ± 0.059	0.506 ± 0.012	0.52 ± 0.038	0.5 ± 0.002
ChestX-ray8 65% - Costa Rica 35%	0.582 ± 0.096	0.567 ± 0.067	0.579 ± 0.096	0.558 ± 0.067
China 35% - Costa Rica 65%	0.514 ± 0.04	0.503 ± 0.009	0.525 ± 0.077	0.509 ± 0.02
China 65% - Costa Rica 35%	0.591 ± 0.096	0.579 ± 0.076	0.585 ± 0.096	0.567 ± 0.051
Indiana 35% - Costa Rica 65%	0.503 ± 0.009	0.503 ± 0.006	0.506 ± 0.019	0.509 ± 0.014
Indiana 65% - Costa Rica 35%	0.574 ± 0.078	0.544 ± 0.032	0.551 ± 0.054	0.543 ± 0.042

Table 11: Accuracy of a Alexnet model trained with MixMatch, with the filtered datasets using the harm coefficient with the two proposed feature density based methods: FH and the Mahalanobis based filter. The percentage of discarded observations is the same of the amount of Costa Rican observations.

	$n_l = 40$		$n_l = 20$	
Dataset	Acc. FD	Acc. Maha.	Acc. FD	Acc. Maha.
ChestX-ray8 35% - Costa Rica 65%	0.709 ± 0.084	0.727 ± 0.078	0.682 ± 0.09	0.685 ± 0.089
ChestX-ray8 65% - Costa Rica 35%	0.732 ± 0.064	0.7612 ± 0.049	0.717 ± 0.08	0.709 ± 0.09
China 35% - Costa Rica 65%	0.683 ± 0.065	0.708 ± 0.07	0.667 ± 0.078	0.667 ± 0.09
China 65% - Costa Rica 35%	0.693 ± 0.044	0.695 ± 0.079	0.687 ± 0.078	0.674 ± 0.072
Indiana 35% - Costa Rica 65%	0.732 ± 0.052	0.711 ± 0.032	0.703 ± 0.1	0.719 ± 0.09
Indiana 65% - Costa Rica 35%	0.719 ± 0.058	0.748 ± 0.059	0.709 ± 0.093	0.711 ± 0.09

Table 12: Accuracy of a Densenet model trained with MixMatch, with the filtered datasets using the harm coefficient with the two output-based methods: MCD and Softmax. The percentage of discarded observations is the same of the amount of Costa Rican observations.

	$n_l = 40$		$n_l = 20$	
Dataset	Acc. Softmax	Acc. MCD	Acc. Softmax	Acc. MCD
ChestX-ray8 35% - Costa Rica 65%	0.5 ± 0.001	0.5 ± 0.001	0.488 ± 0.025	0.529 ± 0.077
ChestX-ray8 65% - Costa Rica 35%	0.543 ± 0.09	0.537 ± 0.11	0.543 ± 0.095	0.498 ± 0.004
China 35% - Costa Rica 65%	0.498 ± 0.004	0.5 ± 0.001	0.49 ± 0.04	0.496 ± 0.009
China 65% - Costa Rica 35%	0.517 ± 0.029	0.501 ± 0.004	0.5 ± 0.007	0.504 ± 0.01
Indiana 35% - Costa Rica 65%	0.499 ± 0.001	0.5 ± 0.001	0.48 ± 0.036	0.496 ± 0.009
Indiana 65% - Costa Rica 35%	0.5 ± 0.001	0.501 ± 0.008	$0.497 \pm 0.$	0.503 ± 0.0173

Table 13: Accuracy of a Densenet model trained with MixMatch, with the filtered datasets using the harm coefficient with the two proposed feature density based methods: FH and the Mahalanobis based filter. The percentage of discarded observations is the same of the amount of Costa Rican observations.

	$n_l = 40$		$n_l = 20$	
Dataset	Acc. FD	Acc. Maha.	Acc. FD	Acc. Maha.
ChestX-ray8 35% - Costa Rica 65%	0.691 ± 0.10	0.769 ± 0.048	0.683 ± 0.105	0.779 ± 0.025
ChestX-ray8 65% - Costa Rica 35%	0.717 ± 0.091	0.811 ± 0.049	0.695 ± 0.1	0.783 ± 0.049
China 35% - Costa Rica 65%	0.794 ± 0.036	0.795 ± 0.053	0.787 ± 0.048	0.769 ± 0.076
China 65% - Costa Rica 35%	0.788 ± 0.056	0.812 ± 0.05	0.774 ± 0.053	0.798 ± 0.036
Indiana 35% - Costa Rica 65%	0.758 ± 0.047	0.729 ± 0.035	0.727 ± 0.0512	0.714 ± 0.046
Indiana 65% - Costa Rica 35%	0.737 ± 0.049	0.762 ± 0.055	0.703 ± 0.055	0.722 ± 0.032

Harmful data filter	**Time (secs.)**
Mahalanobis	65.1 ± 2.3
Feature Histograms	269.7 ± 2.7
Softmax	1246.7 ± 22.2
Monte Carlo Dropout	1089.6 ± 10.8

Table 14: Average and standard deviation of the execution time, in seconds, of the different unlabelled harmful data techniques tested in this work. The execution time of using 10 random data batches was measured.

5.2. Experiment setup

Regarding hardware resources, most of the experiments were run at the DIGITS computer, De Montfort University, equipped with a 12GB NVIDIA TITAN V GPU, 24 Intel(R) Xeon(R) E5-2620 0 @ 2.00GHz CPU and 32GB of RAM memory. Software wise, this system was used with Ubuntu 18.04 LTS, with Python version 3.7.0. The Pytorch library used to develop the algorithms in this thesis, with version 1.4.0 in both systems. We also used the FastAI library (version 1.0.61) to develop some sections of this work [3]. The repository with the code used in this work can be found in `https://gitlab.com/saul1917/mixmatch_with_ood`.

6. Results Analysis

In this section we develop the interpretation of the obtained results. As for the results in TB-1.1, depicted in Table 5, we can see a very strong influence of the unlabelled data source D_u^s in the accuracy of the SSDL MixMatch algorithm. Training the model with the Indiana dataset including also COVID-19$^+$ observations, yields the highest accuracy, with around 0.89, higher than the supervised model. From there, using the ChestX-ray8 as D_u^s, yields an accuracy of 0.825, followed by the usage of the Chinese dataset as D_u^s, accuracy wise.

[3] The Pytorch/FastAI MixMatch implementation is based on the repository available at `https://mc.ai/a-fastai-pytorch-implementation-of-mixmatch/`

Using the Costa Rican dataset as D_u^s yields the lowest accuracy, with close to 0.493. *Contaminating* the ChestXray8, Chinese and Indiana dataset with the Costa Rican dataset, yields a lower accuracy with an increasing degree of contamination. As for the impact of fine-tuning the feature extractor, there is no statistical significant difference of performing it, when comparing the results in Tables 5 and 6. This suggests that using an image-net pre-trained feature extractor for harm coefficient estimation is justifiable.

Regarding TB-2 results, when comparing the accuracy yielded by MixMatch for each tested D_u^s with the calculated inter-dataset cosine DeDiMs in Table 8, we can see an interesting relationship. The Costa Rican dataset and heavily contaminated D_u^s data sources present the highest distances. For instance, the Chinese dataset contaminated with a degree of 65% with the Costa Rican dataset, presents a distance of 50.93 with the labelled dataset D_u^s, similar to the inter-dataset distance to the Costa Rican dataset of 57.19 (the D_u^s with the highest distance to D_l^t). We can see how using both of the aforementioned D_u^s datasets, yield very low MixMatch accuracy. This behaviour is summarized in the obtained Pearson coefficients depicted in Table 9, with a very high lineal correlation, of around 78% for the tested variations. The correlation is still high for the semi-supervised densenet model behaviour with the dataset distances, using a generic Imagenet pre-trained alexnet model. This suggests that the usage of the feature density can bring useful information to preserve or discard an unlabelled observation in a D_u^s.

Regarding the results of TB-2, Tables 13 and 11 show the accuracy of MixMatch yielded when filtering the unlabelled datasets with the proposed FH and Mahalanobis methods, for both tested models (alexnet and densenet, respectively). For both proposed methods, we can see how filtering potentially harming observations from the unlabelled dataset increases MixMatch's accuracy significantly, when compared to the baseline accuracies in Tables 7 and 5, for both tested models. For instance, when using the densenet model with $n_l = 40$, the ChestX-ray8 dataset contaminated with 35% and 65% with the *Costa Rica* dataset, increases its accuracy from 0.579 to 0.78 and 0.5 to 0.79, respectively, when filtering harmful observations with the Mahalanobis method (both with statistical significance, according to our Wilcoxon tests). This can be seen in both Tables 5 and 13. The usage of the FH method yields also an important accuracy gain. In this case however, it is lower than the gains obtained with the Mahalanobis method. The accuracy of the model trained with D_u^s using the ChestX-ray8 dataset with no contamination is almost restored, as MixMatch originally yielded 0.825. We have to consider that the filtered dataset is always smaller than the original unlabelled dataset. Despite this, the accuracy ends very close. Similarly, for the alexnet model with $n_l = 40$, the accuracy of using an *Indiana* unlabelled dataset contaminated with 65% of the *Costa Rica* dataset is close to 50%, according to Table 7. However, after filtering out harmful unlabelled observations ends close to the 71%, using both the FH or the Mahalanobis method.

When comparing the accuracy gain of using the feature histograms against the Mahalanobis distance based method, we can see a similar behaviour across almost all the tested unlabelled datasets D_u^s. This since according to our statistical analysis test using the Wilcoxon method, there is no statistically significant difference between the FH and Mahalanobis method. However, this behaviour is broken for the ChestX-ray8 dataset, when using the densenet model, where the Mahalanobis based method yields statistically significant accuracy gains the FH approach, as seen in Table 13. This suggests that the feature distribution of the labelled dataset D_l^t fits well with a Gaussian distribution, given the similar and sometimes slightly better results of the Mahalanobis method. The Mahalanobis based method is faster, as it only needs to compute a covariance matrix, when compared to the histogram based approach, which needs to build a feature histogram. This proved to be significantly slower in our tests as seen in Table 14.

As for the tested MCD and Softmax baseline methods, popular in OOD detection and uncertainty estimation, the results depicted in Tables 10 and 12, for the alexnet and densenet models, show a very poor performance. The accuracy gains are negligible and sometimes the accuracy is diminished, when compared to the baseline results shown in Tables 7 and 5. Therefore, the usage of the feature density based methods for filtering potentially harmful unlabelled observations prove to be a significantly better approach. Accuracy gains of up to 25% with statistical significance in all the tested settings were obtained (using a Wilcoxon test with $p < 0.05$), when using the feature density approaches over the tested output based ones. This can be seen when comparing the results for the proposed feature density techniques in Tables 11 and 13, with Tables 10 and 12, for the both tested architectures alexnet and densenet, respectively.

7. Conclusions

In this work, we have analyzed the impact of the distribution mismatch between the labelled and the unlabelled dataset for training a SSDL model, using the MixMatch algorithm. The setting assessed used medical imaging data, for COVID-19 detection. Measuring the impact of distribution mismatch between the unlabelled and labelled dataset for medical imaging applications is still an under-reported problem in the literature.

In the first test-bed, we have assessed the impact of using different unlabelled data sources D_u^s, and quantitatively analyzed the distribution mismatch between them using DeDiMs as a metric. The high linear correlation between the measured DeDiMs and the MixMatch accuracy, suggests a strong influence of the feature distribution mismatch between D_u^s and D_l^t. In contexts where a decision must be made about what unlabelled data source D_u^s must be used, from a set of possible unlabelled datasets, the DeDiMs might be used as a quantitative prior method. Implementing the tested DeDiMs requires no model training, as a generic pre-trained ImageNet model seems to be good enough to estimate the benefit of using a specific unlabelled dataset D_u^s, according to our results. Data quality metrics for deep learning models as argued in [48, 5] is an interesting path to develop further, as it might help to narrow the gap between research and real-world implementation of deep learning systems. For instance, building high quality datasets for training a semi-supervised model, or assess the safety of using a deep learning model before hand, can benefit from quantitative data quality measures. We argue for the community to include robust data quality metrics in the deployment of deep learning solutions.

To increase the robustness of the SSDL model to the distribution mismatch, we tested different approaches to discard potentially harming unlabelled observations from the unlabelled dataset D_u^s. The tested setting can be considered to be closer to real-world settings, as images within the same domain were used as OOD data contamination sources. This contrasts to the frequent OOD detection benchmarks where images from very different dataset were used as OOD data sources [80]. Our approach is data-oriented, as it modifies the original dataset in an explicit way by removing potentially harming unlabelled observations. We tested output based OOD filtering techniques against our proposed feature density based approaches.

Our proposed methods based on the feature densities built upon a pre-trained model with Imagenet, showed a large and significantly advantage over previous output based OOD filtering methods. In the context of SSDL, some approaches have relied in weighing each unlabelled observation using the output of the model, as in [52]. According to our results, we argue that using the model's output might yield over-confident results to filter or weigh unlabelled observations. This is widely known in OOD detection literature [40]. Even ensemble based approaches like the tested MCD method are not able to filter harming unlabelled observations, according to our test results. However, both feature density based approaches demonstrated a good performance on detecting harming unlabelled observations, almost recovering the original accuracy of the no contaminated datasets. The proposed methods can be deployed to correct and create more effective unlabelled datasets. Moreover both proposed methods do not require any deep learning model training, making it cheap and reducing the carbon footprint of its implementation [65]. Research of computationally efficient methods to identify potentially harmful data for deep learning systems remains as an interesting future research path.

Recently, the renowned deep learning researcher, Andrew Ng, has urged the community to focus in data-centric based AI solutions, that are able to tackle the main challenges faced by AI systems during its everyday usage [49]. As argued in [36], most of development effort of AI solutions for real-world usage is invested in data manipulation tasks. Nevertheless, data-oriented operations are often overlooked in the deep learning research community. Also different dataset testing settings (scarcely labelled datasets, datasets with distribution mismatch settings), are frequently omitted. This often obscures the actual accuracy gain of using a specific methodology. Therefore, we agree with Andrew's call on focusing in more data-centric methods and more sophisticated dataset settings evaluations to develop deep learning and AI technology, along with stronger data quality and evaluation standards for data-driven AI systems.

In the context of the currently active COVID-19 pandemic, these shortcomings for deep learning based solutions have hindered its path to solve urgent challenges to face the pandemic. It can be argued that the AI and deep learning community mostly focused on developing model-centric solutions that delivered questionable accuracy gains, often using datasets under unrealistic assumptions (same distribution of the test and training datasets) and hidden biases (age and other types of biases have been found in popular datasets used in recent publications) [49]. This has led to a poor and almost null impact of AI tools in the struggle against the COVID-19 pandemic [46, 53]. The lack of high quality data standards and regulations to obtain them (data bias acknowledgement, data standardisation and sharing, data quality and robustness metrics, etc) in the AI research community, is an obstacle to develop robust models for daily clinical usage.

References

[1] Erick Alfaro, Ximena Bolanos Fonseca, Enrique M Albornoz, César E Martínez, and Saúl Calderón Ramrez. A brief analysis of u-net and mask r-cnn for skin lesion segmentation. In *2019 IEEE International Work Con-*

ference on Bioinspired Intelligence (IWOBI), pages 000123–000126. IEEE, 2019.

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

[3] Richa Arora. The training and practice of radiology in India: current trends. *Quantitative imaging in medicine and surgery*, 4(6):449–44950, Dec. 2014.

[4] Maria-Florina Balcan and Avrim Blum. 21 an augmented pac model for semi-supervised learning. 2006.

[5] Indranil Balki, Afsaneh Amirabadi, Jacob Levman, Anne L Martel, Ziga Emersic, Blaz Meden, Angel Garcia-Pedrero, Saul C 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.

[6] Sanhita Basu, Sushmita Mitra, and Nilanjan Saha. Deep learning for screening covid-19 using chest x-ray images. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 2521–2527. IEEE, 2020.

[7] Ariana Bermudez, Saul Calderon-Ramirez, Trevor Thang, Pascal Tyrrell, Armaghan Moemeni, Shengxiang Yang, and Jordina Torrents-Barrena. Quality assessment of dental photostimulable phosphor plates with deep learning. Institute of Electrical and Electronics Engineers.

[8] 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, Dec. 2019.

[9] 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. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5294–5301. IEEE, Jan. 2021.

[10] 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 e-prints*, Aug. 2020.

[11] Saul Calderon-Ramirez and Luis Oala. More than meets the eye: Semi-supervised learning under non-iid data. *arXiv e-prints*, Apr. 2021.

[12] Saul Calderon-Ramirez, Luis Oala, Jordina Torrents-Barrena, Shengxiang Yang, Armaghan Moemeni, Wojciech Samek, and Miguel A. Molina-Cabello. Mixmood: A systematic approach to class distribution mismatch in semi-supervised learning using deep dataset dissimilarity measures, 2020.

[13] Saul Calderon-Ramirez, Luis Oala, Jordina Torrents-Barrena, Shengxiang Yang, Armaghan Moemeni, Wojciech Samek, and Miguel A Molina-Cabello. Mixmood: A systematic approach to class distribution mismatch in semi-supervised learning using deep dataset dissimilarity measures. *arXiv e-prints*, Jun. 2020.

[14] Iván Calvo, Saul Calderon-Ramirez, Jordina Torrents-Barrena, Erick Muñoz, and Domenec Puig. Assessing the impact of a preprocessing stage on deep learning architectures for breast tumor multi-class classification with histopathological images. In *Latin American High Performance Computing Conference*, pages 262–275. Springer, 2019.

[15] Nanshan Chen, Min Zhou, Xuan Dong, Jieming Qu, Fengyun Gong, Yang Han, Yang Qiu, Jingli Wang, Ying Liu, Yuan Wei, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in wuhan, china: a descriptive study. *The Lancet*, 395(10223):507–513, Feb. 2020.

[16] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Semi-supervised deep learning with memory. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 268–283, 2018.

[17] Yanbei Chen, Xiatian Zhu, Wei Li, and Shaogang Gong. Semi-supervised learning under class distribution mismatch. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3569–3576, 2020.

[18] 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.

[19] Michael Chung, Adam Bernheim, Xueyan Mei, Ning Zhang, Mingqian Huang, Xianjun Zeng, Jiufa Cui, Wenjian Xu, Yang Yang, Zahi A Fayad, et al. Ct imaging features of 2019 novel coronavirus (2019-ncov). *Radiology*, 295(1):202–207, Feb. 2020.

[20] Safa Cicek, Alhussein Fawzi, and Stefano Soatto. Saas: Speed as a supervisor for semi-supervised learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 149–163, 2018.

[21] Joseph Paul Cohen, Paul Morrison, and Lan Dao. Covid-19 image data collection. *arXiv e-prints*, Jun. 2020. Data repository available at `https://github.com/ieee8023/covid-chestxray-dataset`.

[22] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.

[23] 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.

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

[25] Mohamed Elgendi, Muhammad Umer Nasir, Qunfeng Tang, David Smith, John-Paul Grenier, Catherine Batte, Bradley Spieler, William Donald Leslie, Carlo Menon, Richard Ribbon Fletcher, et al. The effectiveness of image augmentation in deep learning networks for detecting covid-19: A geometric transformation perspective. *Frontiers in Medicine*, 8, 2021.

[26] Lan-Zhe Guo, Zhen-Yu Zhang, Yuan Jiang, Yu-Feng Li, and Zhi-Hua Zhou. Safe deep semi-supervised learning for unseen-class unlabeled data. In *International Conference on Machine Learning*, pages 3897–3906. PMLR, 2020.

[27] Philip Haeusser, Alexander Mordvintsev, and Daniel Cremers. Learning by association–a versatile semi-supervised training method for neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 89–98, 2017.

[28] 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.

[29] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv e-prints*, Oct. 2016.

[30] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *CoRR*, abs/1610.02136, 2016.

[31] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie

Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597, 2019.

[32] Aras M Ismael and Abdulkadir Şengür. Deep learning approaches for covid-19 detection based on chest x-ray images. *Expert Systems with Applications*, 164:114054, 2021.

[33] Rachna Jain, Meenu Gupta, Soham Taneja, and D Jude Hemanth. Deep learning based detection and analysis of covid-19 on chest x-ray images. *Applied Intelligence*, 51(3):1690–1700, 2021.

[34] Baihong Jin, Yingshui Tan, Yuxin Chen, and Alberto Sangiovanni-Vincentelli. Augmenting monte carlo dropout classification models with unsupervised learning tasks for detecting and diagnosing out-of-distribution faults. *arXiv preprint arXiv:1909.04202*, 2019.

[35] 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.

[36] Shivani Kapania, Nithya Sambasivan, Kristen Olson, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora Aroyo. Data desiderata: Reliability and fidelity in high-stakes ai. 2020.

[37] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 5580–5590, 2017.

[38] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, Feb. 2018.

[39] Gyeongho Kim. Recent deep semi-supervised learning approaches and related works. *arXiv preprint arXiv:2106.11528*, 2021.

[40] Michael Truong Le, Frederik Diehl, Thomas Brunner, and Alois Knol. Uncertainty estimation for deep neural object detectors in safety-critical applications. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 3873–3878. IEEE, Nov. 2018.

[41] 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.

[42] 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.

[43] Antonio Loquercio, Mattia Segu, and Davide Scaramuzza. A general framework for uncertainty estimation in deep learning. *IEEE Robotics and Automation Letters*, 5(2):3153–3160, 2020.

[44] Yucen Luo, Jun Zhu, Mengxi Li, Yong Ren, and Bo Zhang. Smooth neighbors on teacher graphs for semi-supervised learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8896–8905, 2018.

[45] Halgurd S Maghdid, Aras T Asaad, Kayhan Zrar Ghafoor, Ali Safaa Sadiq, Seyedali Mirjalili, and Muhammad Khurram Khan. Diagnosing covid-19 pneumonia from x-ray and ct images using deep learning and transfer learning algorithms. In *Multimodal Image Exploitation and Learning 2021*, volume 11734, page 117340E. International Society for Optics and Photonics, 2021.

[46] Yashpal Singh Malik, Shubhankar Sircar, Sudipta Bhat, Mohd Ikram Ansari, Tripti Pande, Prashant Kumar, Basavaraj Mathapati, Ganesh Bal-

asubramanian, Rahul Kaushik, Senthilkumar Natesan, et al. How artificial intelligence may help the covid-19 pandemic: Pitfalls and lessons for the future. *Reviews in medical virology*, 31(5):1–11, 2021.

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

[48] 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, Feb. 2020.

[49] Lester James Miranda. Towards data-centric machine learning: a short review. *ljvmiranda921. github. io.*

[50] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018.

[51] Marcin Możejko, Mateusz Susik, and Rafał Karczewski. Inhibited softmax for uncertainty estimation in neural networks. *arXiv preprint arXiv:1810.01861*, 2018.

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

[53] Wim Naudé. Artificial intelligence vs covid-19: limitations, constraints and pitfalls. *AI & society*, 35(3):761–765, 2020.

[54] 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, Dec. 2020.

[55] 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.

[56] 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.

[57] Siyuan Qiao, Wei Shen, Zhishuai Zhang, Bo Wang, and Alan Yuille. Deep co-training for semi-supervised image recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 135–152, 2018.

[58] Michael Roberts, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans. *Nature Machine Intelligence*, 3(3):199–217, 2021.

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

[60] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016.

[61] Andreas Sedlmeier, Thomas Gabor, Thomy Phan, and Lenz Belzner. Uncertainty-based out-of-distribution detection in deep reinforcement learning. *Digitale Welt*, 4(1):74–78, 2020.

[62] Weiwei Shi, Yihong Gong, Chris Ding, Zhiheng MaXiaoyu Tao, and Nanning Zheng. Transductive semi-supervised deep learning using min-max features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 299–315, 2018.

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

[64] Jost Tobias Springenberg. Unsupervised and semi-supervised learning with categorical generative adversarial networks. *arXiv preprint arXiv:1511.06390*, 2015.

[65] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for modern deep learning research. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13693–13696, 2020.

[66] Jiumeng Sun, Wan-Ting He, Lifang Wang, Alexander Lai, Xiang Ji, Xiaofeng Zhai, Gairu Li, Marc A Suchard, Jin Tian, Jiyong Zhou, et al. Covid-19: epidemiology, evolution, and cross-disciplinary perspectives. *Trends in molecular medicine*, 26(5):483–495, May 2020.

[67] Natasa Tagasovska and David Lopez-Paz. Frequentist uncertainty estimates for deep learning. *arXiv preprint arXiv:1811.00908*, 2018.

[68] Jeremy Tan, Anselm Au, Qingjie Meng, and Bernhard Kainz. Semi-supervised learning of fetal anatomy from ultrasound. In *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*, pages 157–164. Springer, 2019.

[69] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems*, pages 1195–1204, 2017.

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

[71] Phi Vu Tran. Semi-supervised learning with self-supervised networks. *arXiv preprint arXiv:1906.10343*, 2019.

[72] Joost van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Simple and scalable epistemic uncertainty estimation using a single deep deterministic neural network. *arXiv e-prints*, Jun. 2020.

[73] David A Van Dyk and Xiao-Li Meng. The art of data augmentation. *Journal of Computational and Graphical Statistics*, 10(1):1–50, 2001.

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

[75] Ritika Wason. Deep learning: Evolution and expansion. *Cognitive Systems Research*, 52:701–708, 2018.

[76] Jason Weston, Frédéric Ratle, Hossein Mobahi, and Ronan Collobert. Deep learning via semi-supervised embedding. In *Neural Networks: Tricks of the Trade*, pages 639–655. Springer, 2012.

[77] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv e-prints*, Apr. 2018.

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

[79] Jieli Zhou, Baoyu Jing, Zeya Wang, Hongyi Xin, and Hanghang Tong. Soda: Detecting covid-19 in chest x-rays with semi-supervised open set domain adaptation. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2021.

[80] 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.