Dealing with Scarce Labelled Data: Semi-supervised Deep Learning with Mix Match for Covid-19 Detection Using Chest X-ray Images

Saul Calderon-Ramirez [*†], Raghvendra Giri [*], Shengxiang Yang [*], Armaghan Moemeni [‖], Mario Umaña [‡]
David Elizondo [*], Jordina Torrents-Barrena [¶], Miguel A. Molina-Cabello [§]

[*] Centre for Computational Intelligence (CCI), De Montfort University, United Kingdom
[†]Instituto Tecnologico de Costa Rica, Costa Rica
[‖]School of Computer Science, University of Nottingham, United Kingdom
[‡] Comprehensive Health Care Center, Dr. Marcial Fallas Diaz, Costa Rica
[¶]Universitat Rovira i Virgili, Spain
[§]Department of Computer Languages and Computer Science. University of Málaga, Spain
Biomedic Research Institute of Málaga (IBIMA), Spain

[*]sacalderon@itcr.ac.cr, [*]raghvendra.giri.ai@gmail.com, [*]syang@dmu.ac.uk, [†]armaghan.moemeni@nottingham.ac.uk,
[‡]maumanav@ccss.sa.cr, [*]elizondo@dmu.ac.uk, [¶]jordina.torrents@urv.cat, [§]miguelangel@lcc.uma.es

***Abstract*—Coronavirus (Covid-19) is spreading fast, infecting people through contact in various forms including droplets from sneezing and coughing. Therefore, the detection of infected subjects in an early, quick and cheap manner is urgent. Currently available tests are scarce and limited to people in danger of serious illness. The application of deep learning to chest X-ray images for Covid-19 detection is an attractive approach. However, this technology usually relies on the availability of large labelled datasets, a requirement hard to meet in the context of a virus outbreak. To overcome this challenge, a semi-supervised deep learning model using both labelled and unlabelled data is proposed. We develop and test a semi-supervised deep learning framework based on the Mix Match architecture to classify chest X-rays into Covid-19, pneumonia and healthy cases. The presented approach was calibrated using two publicly available datasets. The results show an accuracy increase of around 15% under low labelled / unlabelled data ratio. This indicates that our semi-supervised framework can help improve performance levels towards Covid-19 detection when the amount of high-quality labelled data is scarce. Also, we introduce a semi-supervised deep learning boost coefficient which is meant to ease the scalability of our approach and performance comparison.**

***Index Terms*—Semi-supervised Deep Learning, Mix Match, Chest X-Ray, Covid-19, Computer Aided Diagnosis.**

I. Introduction

Coronavirus is a common type of virus which affects mammals, reptiles and birds, causing what is referred to as zoonotic infections [1]. The SARS-CoV2 virus belongs to the family of the single stranded Ribonucleic Acid (RNA) viruses known as coronaviridae [1]. Coronaviruses (COVs) infect the respiratory and gastrointestinal tract in a wide range of animal species. Even though most of the individual virus species appear to be restricted to narrow host range comprising single animal species, genome sequencing and phylogenetic analysis testify that COVs have successfully migrated to new host species [3].

Zoonotic infection outbreaks are explosive in nature, infecting a high number of subjects in a short period of time. An outbreak may cause the collapse of even state-of-the-art healthcare systems in developed countries within a few months. A recent example is the collapse of Italy's healthcare system due the Covid-19 infection [5]. Italy's public healthcare system is amongst one of the best world-wide [24].

It is important for global organizations such as the World Health Organization (WHO) and governments to implement cost-effective methods to reliably detect Covid-19 infection spread. Alternative solutions include the use of Artificial Intelligence (AI). AI based Computer Aided Diagnosis (CAD) systems can help identifying infected subjects quickly. In this work we implement a semi-supervised deep learning framework for Covid-19 detection using chest X-ray images. Semi-supervised learning makes use of unlabelled data, which is cheaper and more widely available. Effectively using this data can lead to quickly deploy cost-effective deep learning solutions for Covid-19 detection or later mutations of the virus. Making available AI solutions to deal with subject diagnosis in a fast manner, might help to develop a quick and effective response to rapidly evolving virus outbreaks.

A. Covid-19 diagnosis based on X-ray images

The Real-time Reverse Transcription Polymerase Chain Reaction (RT-PCR) test is the gold standard for robust Covid-19 virus detection [13]. This molecular based testing of respiratory tract samples aims to detect the nucleic acid from SARS-CoV-2 from upper and lower respiratory regions. However, the overall cost of the facilities for RT-PCR is rather high. The need for consumables and trained technicians increases further the costs. This makes mass testing frequently unfeasible even in developed countries [25].

Medical imaging is an alternative method for Covid-19 diagnosis. Computed Tomography (CT) of chest shows high accuracy and sensitivity towards Covid-19 detection [17]. Medical imaging studies are becoming important for the early detection and management of patients with Covid-19 [20]. In [6] the authors showed that the accuracy obtained using CT scans was 97%, which was significantly higher than that achieved with RT-PCR (75%). A database of 1014 patients was used in this study. Fang *et al.* [20] reached similar conclusions. However, CT machines are not widely available in less industrialised countries like India [31].

X-ray chest imaging is a less expensive and more accessible alternative than CT [31]. Nevertheless, X-ray can be considered expensive when human resources are considered. These include the availability of radiologists and medical imaging technicians. In India, with a current population of 1.44 billion, currently there is approximately one radiologist for every 100,000 people [10]. This makes X-ray based Covid-19 diagnosis attractive.

In [8], authors developed a severity score based on chest X-ray images. The study included 783 SARS-CoV-2 positive patients. The severity score allowed to screen patients who are likely to develop more severe symptoms. However, a low sensitivity in a small number of cases with alterations compatible with Covid-19 has been found by [34]. This draws the need for an additional validation of labelled data trough the diagnosis of multiple radiologists. With a high quality labelled dataset, AI solutions can be developed for CAD mass testing. However, building a large high quality dataset can be expensive and slow.

B. Contributions

In this paper we propose the diagnosis of Covid-19 based on X-ray images for early diagnosis and detection by using Mix Match, a novel semi-supervised learning technique [12]. We highlight the difficulties to gather large high-quality labeled datasets in the medical imaging domain, specially in the context of a virus out-break. This makes the usage of unlabelled data an attractive alternative to improve the accuracy of deep learning architectures. To our knowledge, this is the first work implementing semi-supervised learning for Covid-19 detection.

The proposed model uses chest X-ray images for training and detection. X-ray equipment is widely available, easing the compilation of large unlabelled datasets, given the low availability of trained technicians or radiologists to label the data. It is vital to be able to quickly classify various types of pneumonia based on digital X-ray images when a virus outbreak occurs. Such outbreaks create very large volumes of cases which have to be manually analysed by radiologist. Early, fast, and cheap diagnosis of Covid-19 infection is key to trace, isolate and control the disease out-break. We stress that the use of semi-supervised deep learning can be a useful approach when dealing with the current Covid-19 out-break or the spread of similar viruses in future.

Finally, in this work we propose the usage of a normalized metric, the semi-supervised learning boost coefficient, for analyzing semi-supervised learning accuracy scalability under different evaluation, labelled and unlabelled data settings. This can be used as a more challenging and closer to real-world evaluation of deep learning solutions for the detection of Covid-19 infection.

II. Related Work

A. Semi-supervised deep learning

Semi-supervised deep learning is an increasingly popular approach to deal with scarcely labelled datasets. Typical deep learning architectures require large labelled datasets to generalize well. This requirement frequently makes its practical implementation in the medical domain hard, as high quality labelled data is expensive and scarce.

Formally, in a semi-supervised setting, combination of labelled and unlabelled samples is used. The labelled observations $X_l = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_l}\}$ include there corresponding labels in the set $Y_l = \{y_1, \ldots, y_{n_l}\}$. The unlabelled set includes all the observations with no labels $X_u = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_u}\}$.

We can categorize existing semi-supervised deep learning architectures as follows: Pre-training [19], self-training or pseudo-labelled [14] and regularization based. Regularization techniques include generative based approaches, along consistency loss term and graph based regularization [15].

Regularization based semi-supervised deep learning includes a regularization term using unlabelled data S_u, $\mathcal{L}(S) = \sum_{(\boldsymbol{x}_i, \boldsymbol{y}_i) \in S_l} \mathcal{L}_l(\boldsymbol{w}, \boldsymbol{x}_i, \boldsymbol{y}_i) + \gamma \sum_{\overrightarrow{x}_j \in X_u} \mathcal{L}_u(\boldsymbol{w}, \boldsymbol{x}_j)$, where $\boldsymbol{w}$ corresponds to the weights of the model to estimate, $\mathcal{L}_l$ and $\mathcal{L}_u$ correspond to the labelled and unlabelled loss terms, and γ corresponds to the unsupervised term weight, and controls the influence of the unlabelled data during training.

Different variations of the regularized approach have been developed, namely graph based [42], [23], generative augmentation based [35], [27], and consistency loss based [38], [37]. A deep review on semi-supervised deep learning, along key assumptions of popular approaches, can be found in [39].

More recently, Mix Match [12] combined regularization and pseudo-labelled based learning, with intensive data augmentation. Mix Match out-performed other regularized, pseudo-labelled and generative based semi-supervised deep learning techniques as described in [12]. Given the recently state of the art performance demonstrated by Mix Match, we chose it for the developed solution in this work.

B. Semi-Supervised Deep Learning for Medical Imaging

Availability of labelled data for supervised learning is cumbersome for narrow deep learning applications. In the medical domain, automatic pathology diagnosis requires clinicians to provide a consistent ground-truth for thousands of images. This is expensive and time consuming compared to the generation of weak image-level labels, or unlabelled data. Semi-supervised classification is an attractive alternative when strong annotations are hard to come by, enabling the use of unlabelled data to improve model accuracy.

A recent survey on semi-supervised, multi-instance and transfer learning for medical image analysis was published in [40]. Authors discussed several semi-supervised learning methods such as self-training, graph-based, co-training, and manifold regularization. Authors concluded that the usage of transfer learning was more frequent, given the still precarious advantage of using semi-supervised learning in real data.

However, recent successful implementations of semi-supervised deep learning can be found. In [30], breast masses were both localized and classified from ultrasound data using weakly and semi-supervised learning, self-training and VGG-16 network. Results obtained by training the method with 10 annotated images along with weakly annotated data were comparable to the ones achieved from 800 strongly annotated images. Additional weakly annotated data increased the performance from 80% to 84.50%.

Similarly, a graph-based semi-supervised deep learning scheme based on Convolutional Neural Network (CNN) co-training pseudo-labeling, for breast cancer diagnosis is presented in [36]. Authors obtained an accuracy of 82.43% using only 100 labelled observations and the rest of the dataset as unlabelled observations. Authors also highlighted how the fully-supervised model accuracy grows with the size of labelled data, but the accuracy difference decreases as the number of labelled observations becomes larger.

A self-ensembling CNN to leverage unlabelled data was also used for histopathology image analysis in [32]. The model reached an accuracy of 90.5% and 89.5% using only 20% of the labels in breast and lung cancer datasets, respectively. This performance was comparable to train with all labelled patients.

It is worth highlighting the lack of standardized testing methodologies to compare accuracy scalability of semi-supervised deep learning based solutions under different labelled and unlabelled data settings. This makes the comparison of semi-supervised deep learning frameworks less straightforward.

C. Previous work on Chest X-ray image analysis for Covid-19

Covid-19 diagnosis using X-ray images is a new challenge as previously discussed. Therefore, scarce work around this can be found in the use of deep learning models for its automatic detection. For this reason we include pre-published work, in order to provide an overview of the work in progress. We take pre-published work as a general guideline of the work in progress, but not as a performance reference.

The work in [29] describes the implementation of a support vector machine classifier fed with deep features. Popular deep learning architectures were tested for feature extraction. The dataset used in this work is composed of 25 observations for COVID-19 positive cases, and 25 COVID-19 negative cases. The positive observations were taken from the Github repository made available by Dr. Joseph Cohen from the University of Montreal [18], and the negative observations were obtained from the Kaggle public repository on X-ray images with pneumonia and no findings [22]. The model with the highest accuracy reported was ResNet50 with the proposed support vector machine as a top model, yielding a level of accuracy of around 95 %. The 50 images dataset was split into 60% of the images for training, 20% for the error evaluation during training and 20% for the model test. This makes up for a labelled to evaluation sample ratio of $30/(10+30) = 0.75$, with 30 images for training and 10 for testing.

In [7] authors compared different machine learning algorithms. They did a performance comparison between support vector machine, random forest and CNN models. The results showed a superior accuracy of the CNN model, with a test accuracy of 95.2%. The authors did not report the percentage of data used for the evaluation of the model.

The authors in [9] used a CNN along with transfer learning for the automatic classification of pneumonia, Covid-19 and normal cases. They achieved an overall average accuracy of 97.82% in the detection of Covid-19. A 10-fold cross validation was used, corresponding to a labelled to evaluation sample ratio of 0.9. The authors highlighted some of the limitations of deep learning including the need of very large amounts of high quality labelled data, which might be scarce in the case of a new virus out-break. The dataset is a compilation of data gathered from [2], [18], [4].

Authors in [16] developed an automatic Covid-19 pneumonia detection using deep learning. The proposed system classified between Covid-19+, viral and bacterial pneumonia. The authors implemented data augmentation techniques (namely rotation, translation and scaling) along transfer learning to boost model accuracy. Popular CNN models were tested, using a combination of the datasets found in [18]. The authors concluded that the SqueezeNet model outperforms other CNN networks with an accuracy of 98.3%. For evaluation, a 5-fold validation was used, corresponding to a labelled to evaluation ratio of 0.75.

With this brief state of the art overview we can easily distinguish the need of a high amount of labelled data of the proposed models. The data used in the studies remains to be validated by multiple experts. Furthermore, the dataset [22], frequently used in the previous work found, presents important biases towards pediatric and Chinese patients.

Using semi-supervised learning can alleviate the need for large high quality labelled datasets. Also, the evaluation under more challenging data scenarios, such as a low labelled to evaluation dataset size ratio, is still not covered in the literature. This includes non-peer reviewed work. Performing tests in more challenging data scenarios can help to distinguish better architectures for the problem at hand.

III. Proposed Method

In this work, we propose the use of semi-supervised deep learning to tackle the problem of scarcely high-quality labelled data for Covid-19 detection. We aim to evaluate the feasibility of a semi-supervised system with different proportions of labelled to unlabelled data, and their influence on the accuracy boost. It is conjectured that a semi-supervised model might boost the accuracy of Covid-19 early diagnosis from chest X-ray images, particularly when ground truth data is limited. Our

approach is formulated in a way that could be easily extended to other virus outbreak pathologies.

A. Mix Match for Semi-supervised deep learning

Our semi-supervised deep learning approach is based on Mix Match [12]. This technique estimates a set of pseudo-labels and implements an unsupervised regularization term. In Mix Match, the consistency loss term minimizes the distance of the pseudo-labels and the model predictions over the unlabelled dataset X_u. Pseudo-label estimation is performed with the average model output of a transformed input x_j: $\widehat{\boldsymbol{y}}_j = \frac{1}{K}\sum_{\eta=1}^{K} f_{\overrightarrow{w}}\left(\Psi^{\eta}\left(\boldsymbol{x}_j\right)\right)$, where K corresponds to the number of transformations Ψ^{η} applied. We used $K = 2$ as in [12]. Additionally, authors argued that the estimated pseudo-label $\widehat{\boldsymbol{y}}_j$ usually presents a high entropy, leading to unconfident estimations. To encourage confidence, the output array $\widehat{\boldsymbol{y}}$ was sharpened with a temperature T: $s\left(\widehat{\boldsymbol{y}}, T\right)_i = \frac{\widehat{y}_i^{1/T}}{\sum_j \widehat{y}_j^{1/T}}$. Similar to $T \to 0$, the sharpened distribution $\widetilde{\boldsymbol{y}} = s\left(\widehat{\boldsymbol{y}}, T\right)$ tends to become a Dirac function (assuming a one-hot vector representation). The dataset with the estimated and sharpened pseudo-labels was defined as $\widetilde{S}_u = \left(X_u, \widetilde{Y}\right)$, with $\widetilde{Y} = \left\{\widetilde{\boldsymbol{y}}_1, \widetilde{\boldsymbol{y}}_2, \ldots, \widetilde{\boldsymbol{y}}_{n_u}\right\}$.

Berthelot *et al.* [12] also found that data augmentation is a key aspect in semi-supervised deep learning. To further augment data using both labelled and unlabelled samples, they implemented the Mix Up algorithm developed in [44]: $\left(S_l', \widetilde{S}_u'\right) = \Psi_{\text{MixUp}}\left(S_l, \widehat{S}_u, \alpha\right)$

The Mix Up algorithm creates new observations from a linear interpolation of a mix of unlabelled (with its corresponding pseudo-labels) and labelled data. More specifically, it takes two labelled (or pseudo labelled) data pairs $(\boldsymbol{x}_a, y_a)$ and $(\boldsymbol{x}_b, y_b)$. The Mix Up method generates a new observation and its label $(\boldsymbol{x}', y')$ by following these steps:

1) Sample the Mix Up parameter λ from a Beta distribution $\lambda \sim \text{Beta}\left(\alpha, \alpha\right)$.
2) Ensure that $\lambda > 0.5$ by making $\lambda' = \max\left(\lambda, 1-\lambda\right)$
3) Create a new observation with a lineal interpolation of both observations: $\boldsymbol{x}' = \lambda' \boldsymbol{x}_a + \left(1-\lambda'\right)\boldsymbol{x}_b$.
4) Similarly, create the corresponding pseudo-label for such observation $y' = \lambda' y_a + \left(1-\lambda'\right) y_b$.

Using the augmented datasets $\left(S_l', \widetilde{S}_u'\right)$, the Mix Match training of a model $f_{\overrightarrow{w}}$ can be summarized as minimizing $\mathcal{L}\left(S, \boldsymbol{w}\right) = \sum_{\left(\boldsymbol{x}_i, \boldsymbol{y}_i\right) \in S_l'} \mathcal{L}_l\left(\boldsymbol{w}, \boldsymbol{x}_i, \boldsymbol{y}_i\right) + \gamma \sum_{\left(\boldsymbol{x}_j, \widetilde{\boldsymbol{y}}_j\right) \in \widetilde{S}_u'} \mathcal{L}_u\left(\boldsymbol{w}, \boldsymbol{x}_j, \widetilde{\boldsymbol{y}}_j\right)$. The supervised and semi-supervised loss terms were defined as the entropy $\mathcal{L}_l\left(\boldsymbol{w}, \boldsymbol{x}_i, \boldsymbol{y}_i\right) = \delta_{\text{entropy}}\left(\boldsymbol{y}_i, f_{\boldsymbol{w}}\left(\boldsymbol{x}_i\right)\right)$ and the Euclidean distances $\mathcal{L}_u\left(\boldsymbol{w}, \boldsymbol{x}_j, \widetilde{\boldsymbol{y}}_j\right) = \left\|\widetilde{\boldsymbol{y}}_j - f_{\boldsymbol{w}}\left(\boldsymbol{x}_j\right)\right\|$, respectively. The coefficient γ acts as a regularization weight, controlling the direct influence on unlabelled data. The ramp coefficient $r(t)$ is a scalar that increases at each epoch, as the confidence in unlabelled data naturally grows over training. We used a ramp coefficient of $r(t+1) = 1/3000 + r(t)$.

Note that unlabelled data also influences the labelled data term $\mathcal{L}_l$, as unlabelled data is used to artificially augment the dataset through the Mix Up algorithm.

B. Semi-supervised deep learning scalability measurement

To assess the scalability of our semi-supervised methodology, we propose the usage of the semi-supervised accuracy boost coefficient, based on the evaluation/labelled and labelled/unlabelled data ratios.

We define the labelled / evaluation data coefficient as $\rho_{\text{le}} = \frac{n_v}{n_v + n_l}$, where n_v and n_l are the number of validation (evaluation), and labelled observations, respectively. Similarly, the labelled/unlabelled coefficient is formulated as $\rho_{\text{lu}} = \frac{n_l}{n_u + n_l}$, where n_u stands for the number of unlabelled observations.

For semi-supervised learning we propose the usage of the semi-supervised boost coefficient based on the previously defined labelled/unlabelled coefficient ρ_{lu}. This coefficient summarizes the performance boost obtained with a specific pair of ρ_{lu} and ρ_{le}. Its formulation is depicted in Equation 1.

$$\Delta_{\rho} = \frac{\overline{a}_{\text{semi-supervised}} - \overline{a}_{\text{supervised}}}{\left(\rho_{\text{le}} + \rho_{\text{lu}}\right) s_{\text{semi-supervised}}} \tag{1}$$

The coefficients $\overline{a}_{\text{supervised}}$ and $\overline{a}_{\text{semi-supervised}}$ correspond to the reported sample mean accuracy of the supervised and semi-supervised learning framework, respectively. The sample standard deviation $s_{\text{semi-supervised}}$ is also added, to account for the results distribution. Lower ρ_{le} and ρ_{lu} increase the semi-supervised boost coefficient, as this corresponds to a more challenging data scenario. Reporting this coefficient can ease the comparison of semi-supervised deep learning solutions, which are very important in the medical domain.

IV. Dataset

In this work we implement a ternary classification of Covid-19+, pneumonia (bacterial and viral), and no lung pathology X-ray observations.

The observations for the Covid-19+ are gathered from the publicly available github repository available in [18]. Dr. Joseph Cohen, from the University of Montreal was the main author of such repository. A compilation from journal websites like radiopaedia.org, the Italian Society of Medical and Interventional Radiology and recent publications in the matter [18] was gathered by the authors in [18]. The dataset contains chest X-ray images from around 100 patients, with ages ranging from 27 to 85 years old. The patients nationalities include Iran, China, Italy, Taiwan, Australia, Spain and the United Kingdom. Authors warned researchers to avoid claiming diagnostic performance without a proper clinical study. Therefore, in this work we focus on exploring the possibility of using semi-supervised deep learning to improve diagnostic accuracy with small datasets. The need for a proper clinical study with more data to confirm the viability of computer aided diagnosis system for Covid-19, argued in [18]. From this dataset, we used only Covid-19+ images, discarding observations of Middle East Respiratory Syndrome (MERS),

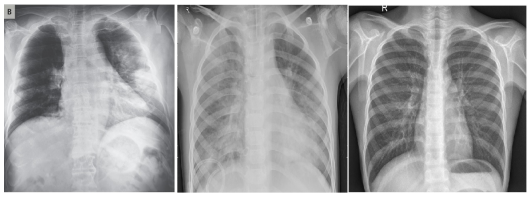

Fig. 1. From left to right: chest x-Ray of Covid-19 Patient, chest X-Ray of pneumonia Patient and normal chest X-Ray

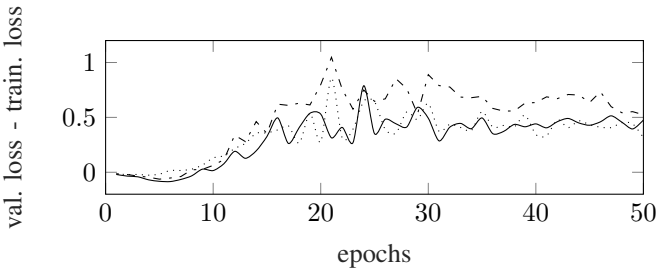

Fig. 2. Difference between both validation and training losses with: $\gamma = 1$ (continuos line, highest accuracy 74.6%), $\gamma = 25$ (dotted line, highest accuracy 76.1), $\gamma = 100$ (the line with the largest dashes, highest accuracy 79.3%). The lower and less spiky the better.

Acute Respiratory Distress Syndrome (ARDS) and Severe Acute Respiratory Syndrome (SARS). Therefore a subset of 102 front chest X-ray Covid-19+ observations were used.

For the pneumonia and normal observations, we used the data availaible in [22]. From such dataset, we selected 5856 chest X-ray images all of them from individual children. The images represent 4273 observations of pneumonia (including viral and bacterial) and 1583 of normal patients. All the pediatric patients in this study were Chinese [22]. The base-line dataset used in this work comprises 5958 observations. This includes 102 observations for Covid-19+, 4273 for pneumonia and 1583 with no lung pathology.

The aforementioned dataset combination have been extensively used in recent works [26], [45], [41], [21], [28], [9]. However, we warn about a practical short-coming of this dataset; the very different populations sampled for Covid-19 with adults (with ages between 21 and 85 years old), while for the normal and pneumonia cases, the images were sampled from pediatric patients. Furthermore, the nationalities of the sampled population are also widely skewed, as for the normal and pneumonia cases Chinese subjects were sampled. We warn for the need of a more balanced dataset, sampling different sub-populations equally. Formally, the diagnostic procedure does not change for pediatric patients, but this biased data might harm its generalization for everyday clinical use.

To avoid a class bias, in most of this work we use an under sampled dataset, containing 102 images for each class, randomly sampling the over-represented classes. Figure 1 shows sample observations from the dataset used in this work. As pre-processing of the data, we standardized the observations using the mean and standard deviation of the whole dataset

V. Experiments

The most important hyper-parameter to tune in Mix Match is the semi-supervised loss term coefficient γ, which weights the importance of unlabelled data as stated in [12]. Our implementation also includes a ramp coefficient to augment the weight of the unsupervised signal. This was recommended by Berthelot *et al.* [12], since pseudo-labels $\boldsymbol{y_j}$ can be misleading at the beginning of the training process. The chosen and empirically optimized Mix Match parameters used in our experiments are: $K = 2$ number of augmentations, $T = 0.5$ sharpening temperature and the distribution parameter $\alpha = 0.5$.

Our empirical study shows an important regularization effect of the unsupervised loss term $\mathcal{L}_u$. Figure 2 depicts the subtraction of the training and validation losses for a specific data partition. Lower values indicate better generalization. Moreover, we employed the Wide-ResNet architecture [43] for the calibration experiment. Wide-ResNet yielded around 96% of accuracy when using the entire dataset with 25% of the data for validation, as seen in Table I. Our experiments aim to explore the effect of γ for semi-supervised classification accuracy.

In the preliminar testing performed, we noted a strong influence of labelled data balance. A very imbalanced labelled dataset practically nullifies the advantage of using unlabelled data. Thus, we used an under-sampled baseline dataset with 102 observations for Covid-19, pneumonia (including viral and bacterial observations) and normal cases.

For our experiments we used different number of randomly chosen labelled observations to train the fully supervised and the Mix Match models, using Wide-resnet in both. To test the fully and semi-supervised models with variable number of labelled observations, we used the undersampled dataset with 102 observations for each of the three classes, comprising a total of 306 observations. We used 25% of the dataset for testing, choosing randomly $306 \times 0.25 \approx 78$ observations across all the tests performed in this section, regardless the amount of labelled data.

Using different number of labelled observations for training allows to explore the performance of the compared models with different ρ_{le} and ρ_{lu} coefficients. The chosen amount of labels and the data coefficients are depicted in Table I, in its first column. The specific values for ρ_{le} and ρ_{lu} coefficients are also described in the first column. We argue that most of the evaluations done in the literature regarding Covid-19 detection use a fixed 25% to 30% of test or validation data proportion. In a CAD system like the one at hand in this work, this proportion might not be adequate for real-world settings, given the likely intensive usage in a short time.

The hyper-parameters of the Wide-resnet model for both the fully and semi-supervised modes are defined as follows: an input image size of 100×100, an Adam optimizer with a 1-cycle policy [33], with a weight decay of 0.0001, a learning rate of 0.0001, a batch size of 12 and a cross entropy loss function.

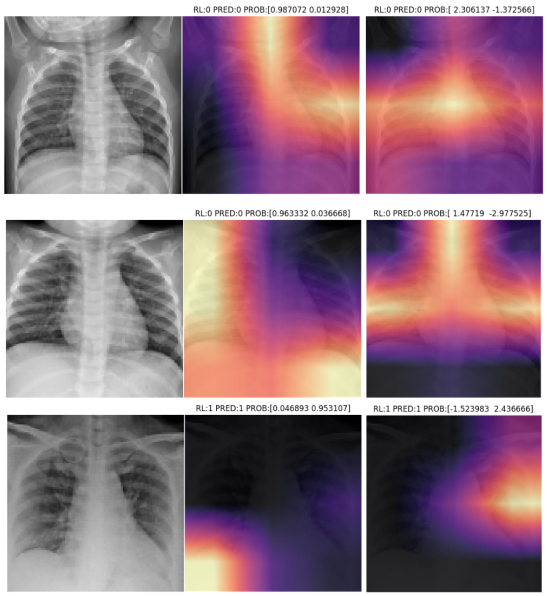

Fig. 3. From top to bottom: A three sample of the class activation maps for the tested dataset. From left to right: the original image, the heatmap of the usual supervised model, and the heatmap for the semi-supervised model. The legend RL corresponds to the real label, PRED to the model prediction and the array of two values is related to the output net values.

The computational hardware used in the experiments includes an NVIDIA [TITAN V] GPU memory of 12 GB, 32 GB of main system memory and an Intel(R) Xeon(R) CPU E5-2620 0 @ 2.00GHz. Python programming language was used for coding. The Pytorch/FastAI MixMatch implementation is based on the repository available at https://mc.ai/a-fastai-pytorch-implementation-of-mixmatch/.

We also implemented transfer learning based on the imagenet weights and data augmentation with random flips and rotations. For all the experiments, the model is trained for 50 epochs, with 10 replicas for each model configuration, with randomly selected training and validation datasets.

As a preliminar qualitative experiment, we trained a binary classification model based on the densenet201 architecture, to discriminate between positive COVID-19 cases and normal (no lung pathology) observations, to analyze the change in the class activation heatmaps of both models. Figure 3 shows a sample of the heatmaps obtained for the supervised and semi-supervised models. Most of the heatmaps obtained reveal a tendency on the heatmaps extracted from the semi-supervised model to focus on more consistent features from the lung area. As seen in Figure 3, the heatmaps of the supervised model tend to focus on less semantically meaningful areas (namely the corners). The semi-supervised and supervised model in this case have been trained with 70 labels, and 68 unlabelled observations for the semi-supervised model. We used transfer learning (with imagenet weights), and flip and rotation data augmentation. For the tested batch, the semi-supervised model yielded an accuracy of 96.6% while the supervised model 91.6%, with balanced test dataset of 60 observations. We noted in some images, that the shoulder joints present high activation values, a feature often used to discriminate children and adult samples. Also in Figure 3, the net raw outputs are depicted for both semi-supervised and supervised models.

VI. Results analysis

The yielded accuracy results for different γ values are shown in Table I. To statistically compare the yielded results, a Wilcoxon non-parametric test has been carried out, as 10 replicas comparing the results of the supervised model against Mix Match with $\gamma = 200$, which yielded the highest sample mean values.

As an initial observation, a rather low accuracy is reported for an otherwise well performing model (that yielded around 96% percent when using the entire dataset, as seen in Table I), with an evaluation/labelled data coefficients from $\rho_{\text{le}} = 0.24$ up to $\rho_{\text{le}} = 0.39$. As expected, the accuracy of the fully supervised model increases while ρ_{le} increases.

The obtained results also show how with a lower ρ_{lu}, a wider accuracy boost is obtained with the semi-supervised model. The results reveal a strong and statistically significant accuracy boost of any of the semi-supervised models tested (with $\gamma = 1, 100, 200$) over the fully supervised model when the labelled/unlabelled data coefficient ρ_{lu} is low. The highest accuracy difference with statistical significance yielded comes when $\gamma = 200$ and $\rho_{\text{lu}} = 0.11$, with an increase of almost 15%. The last column in Table I describes the confidence obtained when performing a Wilcoxon test comparing the results of the semi-supervised model (with $\gamma = 200$) against the fully supervised one. When $p < 0.05$, a statistically meaningful accuracy boost is obtained when using the implemented semi-supervised model. As seen in Table I, the Wilcoxon test returned a $p = 0.000236$, confirming a statistically significant accuracy boost of using the implemented semi-supervised approach when $\rho_{\text{lu}} = 0.1$.

The difference of using different γ values becomes apparently wider as ρ_{lu} is lower. However, by performing a Wilcoxon test of comparing the case when $\gamma = 1$ and $\gamma = 200$ when $\rho_{\text{lu}} = 0.11$, we obtained $p = 0.3182$, making $p > 0.05$, rejecting the hypothesis of significance difference between them. With this we conclude that tweaking the γ value does not have a statistically significant impact in our tests. This suggests a stronger contribution from the mix up data augmentation guided by the unlabelled observations, implemented in the first term $\mathcal{L}_l$ of the loss function. The effect of the γ correlates with the preliminary results plotted in Figure 2, demonstrating a mild effect in the semi-supervised performance. Increasing the value of γ marginally improves the results, by fully leveraging the information in the unlabelled dataset.

As for the values of the proposed Δ_ρ coefficient, it approaches zero when the benefit of the semi-supervised model has no statistically significant advantage. From the executed experiments, we can define a threshold of $\Delta_\rho = 1.9$ to achieve a significant accuracy boost when using unlabelled data.

Figure 4 plots the Δ_ρ for the tested models with $\gamma = 1, 100, 200$. The results for Mix Match using $\gamma = 100$ overall

TABLE I

SEMI-SUPERVISED LEARNING ACCURACY (MEAN AND STD.) USING MIX MATCH (MM) FOR DIFFERENT UNSUPERVISED COEFFICIENTS VS. A FULLY SUPERVISED MODEL (F.S). ALWAYS $\rho_{\text{LU}} = 1$ FOR THE FULLY SUPERVISED MODEL. THE SIXTH COLUMN DENOTES THE CONFIDENCE P-VALUE OF THE ACCURACY DIFFERENCE BETWEEN MIX MATCH AND THE SUPERVISED MODEL.

Number of labels/coefficients	**Fully supervised**	$\gamma = 1$	$\gamma = 100$	$\gamma = 200$	**F.S vs. MM** ($\gamma = 200$)	Δ_ρ ($\gamma = 200$)
25 ($\rho_{\text{le}} = 0.24$,$\rho_{\text{lu}} = 0.11$)	0.683 ± 0.056	0.808 ± 0.053	0.816 ± 0.051	0.829 ± 0.057	$p = 2.36e-04$	7.318
40 ($\rho_{\text{le}} = 0.33$,$\rho_{\text{lu}=0.17}$)	0.729 ± 0.048	0.828 ± 0.04	0.848 ± 0.048	0.846 ± 0.048	$p = 0.0016$	4.875
50 ($\rho_{\text{le}} = 0.39$,$\rho_{\text{lu}=0.21}$)	0.785 ± 0.046	0.834 ± 0.038	0.843 ± 0.047	0.843 ± 0.049	$p = 0.0163$	1.972
70($\rho_{\text{le}} = 0.47$,$\rho_{\text{lu}=0.3}$)	0.808 ± 0.046	0.848 ± 0.053	0.864 ± 0.039	0.858 ± 0.041	$p = 0.1155$	1.5838
100 ($\rho_{\text{le}} = 0.56$,$\rho_{\text{lu}=0.43}$)	0.851 ± 0.049	0.853 ± 0.033	0.856 ± 0.051	0.854 ± 0.047	$p = 0.5194$	0.0648
All-undersampled (229)	0.896 ± 0.035					
All-imbalanced (4468)	0.966 ± 0.003					

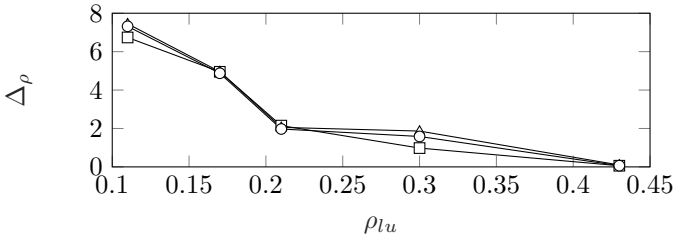

Fig. 4. Scalability curves using Δ_ρ against the ρ_{lu}. $\gamma = 200$ (triangle), $\gamma = 100$ (circle) and $\gamma = 1$ (square).

scale slightly better. However, as previously mentioned, there is no statistically significant difference when using different γ values. This is reflected in how close the curves are. The series of Δ_ρ values summarize well the behaviour of the Mix Match variations tested. To summarize semi-supervised scalability behaviour in a scalar value, we advise the use of the area under the curve Δ_{AUC}. For this experiment $\Delta_{\text{AUC},\gamma=1} = 14.87$, $\Delta_{\text{AUC},\gamma=100} = 16.43$ and $\Delta_{\text{AUC},\gamma=200} = 15.81$, confirming a very slightly advantage of Mix Match with $\gamma = 100$ taking into account all the data settings used.

VII. CONCLUSIONS

In this work we proposed and tested the use of a novel semi-supervised learning framework based on the recently proposed Mix Match technique. A virus outbreak like the COVID-19 draws the need for quickly available and reliable AI solutions for computer aided diagnosis. In the context of a virus outbreak, a strong lack of high quality labelled data i.e., very low number of high quality labelled observations causes severe limitations on the development of computer based diagnosis systems. Semi-supervised deep learning makes use of more widely available unlabelled data, which can help to boost the accuracy of these systems.

As a contribution in this work, we proposed the usage of the semi-supervised accuracy boost coefficient, to measure model scalability under different proportions of evaluation using labelled and unlabelled data. With the tested prototypical dataset (which we warned about the fact that it is still not of acceptable quality to be considered for real-world clinical use given its age and race biases), a significant increase in accuracy is achieved when the labelled/unlabelled data coefficient ρ_{lu} is set to a low value.

We highlight how, in previous work, typical deep convolutional architectures yield high accuracy performances, when using the typical 75%/25% training/evaluation data split. We argue however that this evaluation setting might not be accurate to estimate the real-world performance of a deep learning CAD solution. This is specially the case for a CAD system used in a virus out-break, where a large amount of test data will be fed in a short-time before including new high quality labelled data to re-train the model. For evaluating the system scalability in different labelled/evaluation data scenarios, we proposed the usage of the Δ_ρ coefficient. As expected, our tests revealed an important accuracy decrease as ρ_{el} decreases, making the usage of semi-supervised deep learning more attractive in such setting.

As future work, we plan to test semi-supervised learning approaches with more data for Covid-19 detection. We are building our own chest X-ray dataset from Costa Rican clinics. We aim to extend the usage of the proposed metric Δ_ρ under different real-world settings as unbalanced labelled and unlabelled datasets, and out of distribution unlabelled data. We stress that scalability testing is important to estimate the model performance under real-world operation settings. Most of the test beds used so far in previous work can be thought as *saturated*, since many of the CNN models tested yield accuracies higher than 90% with typical testing strategies. More extensive and demanding testing approaches can be developed, to further assess the accuracy of the model under different training and evaluation scenarios. This is of special relevance given that the definition of high quality and large enough data is still an open question for deep learning solutions, as argued in [11].

ACKNOWLEDGMENTS

This work is partially supported by the following Spanish grants: TIN2016-75097-P, RTI2018-094645-B-I00 and UMA18-FEDERJA-084. All of them include funds from the European Regional Development Fund (ERDF). The authors acknowledge the funding from the Universidad de Málaga.

REFERENCES

[1] Coronavirinae - an overview — ScienceDirect Topics.
[2] COVID-19 DATABASE — SIRM.
[3] Deltacoronaviruses - an overview — ScienceDirect Topics.
[4] RSNA Pneumonia Detection Challenge — Kaggle.

[5] WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020.

[6] Tao Ai, Zhenlu Yang, Hongyan Hou, Chenao Zhan, Chong Chen, Wenzhi Lv, Qian Tao, Ziyong Sun, and Liming Xia. Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases. *Radiology*, page 200642, 2020.

[7] Ali Alqudah, Shoroq Qazan, Hiam Alquran, Isam Abuqasmieh, and Amin Alqudah. Covid-2019 Detection Using X-Ray Images And Artificial Intelligence Hybrid Systems, March 2020.

[8] Roberto Masciullo Andrea Borghesi, Angelo Zigliani. Radiographic severity index in covid-19 pneumonia: relationship to age and sex in 783 italian patients. *World Journal of Surgery*, pages 125: 461–464, https://doi.org/10.1007/s11547–020–01202–1, 2020.

[9] Ioannis D. Apostolopoulos and Tzani Bessiana. Covid-19: Automatic detection from X-Ray images utilizing transfer learning with convolutional neural networks. *Phys Eng Sci Med*, page 43: 635–640, 2020.

[10] Richa Arora. The training and practice of radiology in India: current trends. *Quantitative Imaging in Medicine and Surgery*, 4(6):449–44950, 2014.

[11] 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*, 70(4):344–353, 2019.

[12] 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.

[13] Jasper Fuk-Woo Chan, Cyril Chik-Yan Yip, Kelvin Kai-Wang To, Tommy Hing-Cheung Tang, Sally Cheuk-Ying Wong, Kit-Hang Leung, Agnes Yim-Fong Fung, Anthony Chin-Ki Ng, Zijiao Zou, Hoi-Wah Tsoi, et al. Improved molecular diagnosis of covid-19 by the novel, highly sensitive and specific covid-19-rdrp/hel real-time reverse transcription-polymerase chain reaction assay validated in vitro and with clinical specimens. *Journal of Clinical Microbiology*, 58(5):e00310–20, 2020.

[14] Dong-Dong Chen, Wei Wang, Wei Gao, and Zhi-Hua Zhou. Tri-net for semi-supervised deep learning. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 2014–2020, 2018.

[15] 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.

[16] Muhammad E. H. Chowdhury, Tawsifur Rahman, Amith Khandakar, Rashid Mazhar, Muhammad Abdul Kadir, Zaid Bin Mahbub, Khandakar R. Islam, Muhammad Salman Khan, Atif Iqbal, Nasser Al-Emadi, and Mamun Bin Ibne Reaz. Can AI help in screening Viral and COVID-19 pneumonia? *IEEE Access*, (8):132665–132676, 2020.

[17] 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, 2020.

[18] Joseph Paul Cohen, Paul Morrison, and Lan Dao. Covid-19 image data collection. *arXiv 2003.11597*, 2020. Available at https://github.com/ieee8023/covid-chestxray-dataset.

[19] 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.

[20] Yicheng Fang, Huangqi Zhang, Jicheng Xie, Minjie Lin, Lingjun Ying, Peipei Pang, and Wenbin Ji. Sensitivity of chest CT for COVID-19: comparison to RT-PCR. *Radiology*, 296(2):E115–E117, August, 2020.

[21] Ezz El-Din Hemdan, Marwa A Shouman, and Mohamed Esmail Karar. Covidx-net: A framework of deep learning classifiers to diagnose covid-19 in x-ray images. *arXiv preprint arXiv:2003.11055*, 2020.

[22] 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, 2018.

[23] 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.

[24] Vittorio Maio, Lamberto Manzoli, et al. The Italian health care system: W.H.O ranking versus public perception. *P&T*, 27(6):301–308, 2002.

[25] Krishna Narayanan, Isabel Frost, Anoosheh Heidarzadeh, Katie K Tseng, Sayantan Banerjee, Jacob John, and Ramanan Laxminarayan. Pooling rt-pcr or ngs samples has the potential to cost-effectively generate estimates of covid-19 prevalence in resource limited environments. *medRxiv*, 2020.

[26] Ali Narin, Ceren Kaya, and Ziynet Pamuk. Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks. *arXiv preprint arXiv:2003.10849*, 2020.

[27] 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.

[28] Fatima M. Salman, Samy S. Abu-Naser, Eman Alajrami, Bassem S. Abu-Nasser, and Belal A. M. Alashqar. Covid-19 detection using artificial intelligence. *International Journal of Academic Engineering Research (IJAER)*, 2020.

[29] Prabira Kumar Sethy and Santi Kumari Behera. Detection of coronavirus Disease (COVID-19) based on Deep Features. *Preprints.org*, 2020.

[30] Il Dong Yun Sun Mi Kim Seung Yeon Shin, Soochahn Lee and Kyoung Mu Lee. Joint weakly and semi-supervised deep learning for localization and classification of masses in breast ultrasound images. *IEEE Transactions on Medical Imaging*, 38(3):762–774, 2019.

[31] Mihir Tejanshu Shah, Manjul Joshipura, Jered Singleton, Paul LaBarre, Hem Desai, Eliza Sharma, and Charles Mock. Assessment of the availability of technology for trauma care in India. *World Journal of Surgery*, 39(2):363–372, 2015.

[32] Xiaoshuang Shi, Hai Su, Fuyong Xing, Yun Liang, Gang Qu, and Lin Yang. Graph temporal ensembling based semi-supervised convolutional neural network with noisy labels for histopathology image analysis. *Medical Image Analysis*, 60:101624, 2020.

[33] Leslie N Smith. A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay. *arXiv preprint arXiv:1803.09820*, 2018.

[34] Fengxiang Song, Nannan Shi, Fei Shan, Zhiyong Zhang, Jie Shen, Hongzhou Lu, Yun Ling, Yebin Jiang, and Yuxin Shi. Emerging 2019 novel coronavirus (2019-ncov) pneumonia. *Radiology*, 295(1):210–217, 2020.

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

[36] Wenqing Sun, Tzu-Liang Bill Tseng, Jianying Zhang, and Wei Qian. Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data. *Computerized Medical Imaging and Graphics*, 57:4–9, 2017.

[37] 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.

[38] 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.

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

[40] Marleen de Bruijne Veronika Cheplygina and Josien P. W. 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.

[41] Linda Wang and Alexander Wong. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *arXiv preprint arXiv:2003.09871*, 2020.

[42] 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.

[43] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *Proceedings of the British Machine Vision Conference (BMVC)*, 8:87.1–87.12, September 2016.

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

[45] Jianpeng Zhang, Yutong Xie, Yi Li, Chunhua Shen, and Yong Xia. Covid-19 screening on chest x-ray images using deep learning based anomaly detection. *arXiv preprint arXiv:2003.12338*, 2020.