Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Correcting data imbalance for semi-supervised COVID-19 detection using X-ray chest images

Saul Calderon-Ramirez [a,b,*], Shengxiang Yang [a], Armaghan Moemeni [c], David Elizondo [a], Simon Colreavy-Donnelly [a], Luis Fernando Chavarría-Estrada [d], Miguel A. Molina-Cabello [e,f]

[a] Centre for Computational Intelligence (CCI), De Montfort University, United Kingdom
[b] Instituto Tecnologico de Costa Rica, Costa Rica
[c] School of Computer Science, University of Nottingham, United Kingdom
[d] Imágenes Médicas Dr Chavarría Estrada, La Uruca, San José, Costa Rica
[e] Department of Computer Languages and Computer Science, University of Málaga, Spain
[f] Instituto de Investigación Biomédica de Málaga (IBIMA), Spain

ARTICLE INFO

Article history:
Received 29 September 2020
Received in revised form 11 June 2021
Accepted 7 July 2021
Available online 13 July 2021

Keywords:
Coronavirus
COVID-19
Computer aided diagnosis
Data imbalance
Semi-supervised learning

ABSTRACT

A key factor in the fight against viral diseases such as the coronavirus (COVID-19) is the identification of virus carriers as early and quickly as possible, in a cheap and efficient manner. The application of deep learning for image classification of chest X-ray images of COVID-19 patients could become a useful pre-diagnostic detection methodology. However, deep learning architectures require large labelled datasets. This is often a limitation when the subject of research is relatively new as in the case of the virus outbreak, where dealing with small labelled datasets is a challenge. Moreover, in such context, the datasets are also highly imbalanced, with few observations from positive cases of the new disease. In this work we evaluate the performance of the semi-supervised deep learning architecture known as MixMatch with a very limited number of labelled observations and highly imbalanced labelled datasets. We demonstrate the critical impact of data imbalance to the model's accuracy. Therefore, we propose a simple approach for correcting data imbalance, by re-weighting each observation in the loss function, giving a higher weight to the observations corresponding to the under-represented class. For unlabelled observations, we use the pseudo and augmented labels calculated by MixMatch to choose the appropriate weight. The proposed method improved classification accuracy by up to 18%, with respect to the non balanced MixMatch algorithm. We tested our proposed approach with several available datasets using 10, 15 and 20 labelled observations, for binary classification (COVID-19 positive and normal cases). For multi-class classification (COVID-19 positive, pneumonia and normal cases), we tested 30, 50, 70 and 90 labelled observations. Additionally, a new dataset is included among the tested datasets, composed of chest X-ray images of Costa Rican adult patients.

1. Introduction

The COVID-19 disease is caused by the SARS-CoV2 coronavirus. Coronaviruses spread across the gastrointestinal and the respiratory tracks within a large variety of animal groups, with a high infectivity rate in the case of the SARS-CoV2, which has caused a virus outbreak at the end of 2019 [1]. As more and more people regularly travel across the world, the rapid spread is a lurking danger of a worldwide scale. A key priority for societies across the world, is to develop tools to enable the identification of virus outbreaks and to be able to diagnose them in a short time frame. The quick identification of potential virus carriers is vital to contain a virus outbreak. This is where state of the art Artificial Intelligence (AI) based techniques, such as deep learning, can play a key role, enabling pre-diagnostic and triage systems to effectively identify the presence of the virus in a subject. They offer quick diagnostic responses to enable health systems to cope with rapid spread of virus out-breaks. Deep learning based approaches have been proposed to tackle medical imaging problems [2–6]. However, deep learning models typically need large labelled datasets [7,8].

[*] Corresponding author at: Centre for Computational Intelligence (CCI), De Montfort University, United Kingdom.

E-mail addresses: sacalderon@itcr.ac.cr (S. Calderon-Ramirez), syang@dmu.ac.uk (S. Yang), armaghan.moemeni@nottingham.ac.uk (A. Moemeni), elizondo@dmu.ac.uk (D. Elizondo), simon.colreavy-donnelly@dmu.ac.uk (S. Colreavy-Donnelly), drchavarriaestrada@gmail.com (L.F. Chavarría-Estrada), miguelangel@lcc.uma.es (M.A. Molina-Cabello).

https://doi.org/10.1016/j.asoc.2021.107692
1568-4946/© 2021 Elsevier B.V. All rights reserved.

This research extends a novel SSDL framework known as MixMatch [9] for the detection of COVID-19 based on chest X-ray images. MixMatch is a semi-supervised learning method allows the combination of labelled and unlabelled data to train the model. Semi-supervised learning is more cost effective and accessible, as unlabelled data is cheaper than labelled data. Semi-supervised models can easily be adapted for mutations of the virus at a later stage, with relatively small labelled samples.

We propose a modification for the MixMatch architecture, designed to improve its accuracy under data imbalance settings. Added to smaller labelled datasets, in an outbreak situation, datasets can also be strongly imbalanced, as data available for the subjects manifesting symptoms of the new pathogen are more scarce than non-pathogenic patient records.

1.1. Use of X-ray images towards the diagnosis of COVID-19

A common, well established and robust method for the detection of COVID-19 virus is the Real-time Reverse Transcription Polymerase Chain Reaction (RT-PCR) test [10]. This is a molecular test, which uses respiratory tract samples to identify and confirm infection of COVID-19 [11]. Samples from symptomatic patients suspected of infection of the COVID-19 are gathered [12]. Nevertheless, the costs associated to the use of RT-PCR can be significant, since the facilities and trained personnel needed to perform these tests can be expensive. These severely limit the use of this technique in less industrialized countries, making urgent the need to develop more accessible methods, adding the possible need of testing asymptomatic patients [13].

Diagnosing COVID-19 based on medical imaging can be a reliable and accurate alternative, which is still under exploration. The accuracy and sensitivity levels of this approach as a first stage in COVID-19 detection using chest images, have been analysed in a number of studies [14,15]. The usage of X-ray images for COVID-19 diagnosis has been studied recently. In [16] the authors proposed a severity score using radiography chest images, with a dataset sample of 783 SARS-CoV-2 infected cases. The score was used to identify patients that could potentially acquire more life threatening symptoms. Several studies [14,17,18] have suggested that in a small number of people there is a low level of sensitivity towards the manual detection of alterations using medical images of the chest which can indicate the presence of COVID-19. The use of features extracted and learned by a machine might overcome the variable subjective evaluation of X-ray images. This leads us to explore the potential implementation of deep learning solutions using more widely available and less expensive chest X-ray images. As typical deep learning architectures require many labelled images, we aim to explore the usage of SSDL for COVID-19 detection using X-ray images, evaluating it under another frequent challenge; labelled data imbalance.

1.2. Contribution

In this work, we extensively test the SSDL technique known as MixMatch [9] in a variety of data imbalance situations, with a very limited number of labelled observations. We aim to assess MixMatch's performance under real-world usage scenarios, specifically medical imaging in the context of a virus out-break. Within such context, small labelled samples are available with a strong under-representation of the new pathology, leading to imbalanced datasets. An imbalanced dataset can frequently lead also to a distribution mismatch between the labelled and unlabelled dataset, as described in [19].

Moreover, in this work we propose a simple, yet effective approach for correcting data imbalance for the SSDL MixMatch architecture. We implement a loss based imbalance correction, giving more weight to the under-represented classes in the labelled dataset, a common approach for this aim. In the context of MixMatch, we make use of the pseudo-label and augmented labels predictions to choose the corresponding class-weight. The implemented SSDL solution for COVID-19 detection makes use of unlabelled data. Using unlabelled data can improve model's accuracy, in the absence of high quality and large labelled datasets.

The proposed method uses chest X-ray images. X-ray machines are commonly available, which results in a wealth of unlabelled datasets due to the shortage of radiologists and technicians who can label the images. As an example, India, with its current 1.44 billion population, has a ratio between radiologists and patients of 1:100,000 [20]. However, X-ray machines can be found even in remote areas in under-developed countries, compared to other medical devices like computer tomography scanners [21].

In the event of a viral outbreak, it becomes essential to help health practitioners to quickly identify and classify viral pathologies using digital X-ray images. Outbreaks create a large number of cases, which require the intervention of trained radiologists. Labelling data is time consuming, and in the context of a virus out-break gathering high quality and reliable labelled data can be challenging. SSDL can provide much needed key support for the diagnosis, trace and isolation of the COVID-19 infection and other future pandemics through an early, fast and cheap diagnosis, by using more widely available unlabelled data.

Unlike previous work on COVID-19 detection using deep learning as in [22], we focused in the usage of very small labelled datasets for training a semi-supervised model with wider available unlabelled data. In the context of a pandemic, a specific clinic/hospital might gather a very small labelled dataset, but a larger number of unlabelled observations might be available. Furthermore, given the different patient ethnicity's and characteristics, along with varying imaging protocols, using a model trained with data from another set of hospitals or clinics (from possibly different countries) might yield a distribution mismatch between the training and test datasets. This possibly would yield a very low performance [23,24]. Therefore, training the model with data from the specific clinic/hospital where the model is intended to be used (target data), is an urgent task, which faces the challenge of dealing with very limited labelled datasets [23–25].

In this work, we also make available a first sample of a chest-X ray dataset from the Costa Rican medical private clinic Imagenes Medicas Dr. Chavarria Estrada, with observations containing no findings, and test its usage for training the SSDL framework. If the reader is interested in using such dataset, please contact the main author.

2. Related work

2.1. Deep learning for chest X-ray based COVID-19 detection

The identification of COVID-19 infection based on X-ray images is a new challenge. Thus, up to date there is not much research available with regards to the use of deep learning models for automatically identifying COVID-19 infection. This is the reason why this paper presents mainly pre-published work in the area up-to-date. Since most pre-published articles have not been peer reviewed, it is used here as a general guide and not as a reference towards performance.

A classification model based on a support vector machine fed with deep features was presented in [26]. Different common deep learning architectures were used for feature extraction. These included: VGG16, AlexNet, GoogleNet, VGG19, several variations of Inception and Resnet, DenseNet201 and XceptionNet. The dataset used included a total of fifty observations with half representing COVID-19 images and the other half representing a combination of pneumonia and normal images. The COVID-19 images were acquired from the GitHub repository created by Dr. Joseph Cohen from the University of Montreal [27]. COVID-19 negative images were downloaded from the public repository on X-ray images presented in [28]. The highest level of accuracy was obtained with the ResNet50 model which was combined with a support vector machine as a top model. An accuracy of around 95%, with statistical significance, was obtained.

Several machine learning architectures were compared in [29]. Some of the tested methods by the authors included: support vector machines, random forests and Convolutional Neural Network (CNN) models. The results reported the CNN model as the best performing approach, with an accuracy of 95.2%. The dataset used in such work includes 48 Cases for COVID-19^+ and 23 for negative COVID-19 cases from Dr. Cohen's repository [27]. Data augmentation was used to deal with scarce labelled data.

Another study involving the use of CNNs along with transfer-learning for the automatic classification of pneumonia, COVID-19 and images presenting no lung pathology was presented in [30]. The authors used a 10-fold cross-validation, to test the following CNN architectures: VGG-19, MobileNet v2, Inception, Xception and Inception ResNet v2. An accuracy of around 93% was obtained in the identification of COVID-19, with the use of a VGG-19 model. No statistical significance tests were performed. As for the data used in [30], similar to related proposed solutions, positive COVID-19 cases were extracted from [27], while pneumonia and no lung pathology observations were taken from [28].

A deep learning model for the automatic detection of COVID-19 and pneumonia was proposed in [31]. The system proposed classifies images into three classes; COVID-19^+, viral pneumonia and normal readings. To increase the number of observations, the authors relied on data augmentation techniques including rotation, translation and scaling, along with transfer-learning. The architectures tested included: AlexNet, ResNet19, DenseNet201 and SqueezeNet. A combination of the datasets from [27] was used in this research. According to the results yielded by the authors, the SqueezeNet model outperformed all the other CNN networks. Regarding the data used in such work, a combination of two data repositories [28,32] was used for viral and normal image categories, and the data repository in [27] was used for positive COVID-19 cases.

Explainability for deep learning models is an important feature for medical imaging based systems [33]. Model uncertainty estimation is a common approach to enforce model explainability and usage safety [33]. A COVID-19 detection system with uncertainty assessment was proposed in [34]. By providing practitioners with a confidence factor of the prediction, the overall reliability of the system was improved. A high correlation between the prediction accuracy of the model and the level of uncertainty was reported [34]. The dataset used for positive COVID-19 cases also used Dr. Cohen's repository [27], and normal X-ray readings were collected from [28].

In [35], a semi-supervised approach for defining relevant features for COVID-19 detection was developed. The suspicious regions were extracted by training a semi-supervised auto-encoder architecture that minimizes the reconstruction error. This approach relied in the wider availability of COVID-19^- cases to learn relevant features. Such extracted features were used for classifying the input observations into three classes; COVID-19^+, pneumonia and normal, using a common supervised CNN approach. The extracted features were used to enforce model explainability. Similar to previous reviewed approaches, the datasets provided in [27,28] were used.

The work in [36] also used a feature extractor built from training a model to classify X-ray images in larger datasets with non COVID-19 observations. The model was trained for the regression of COVID-19 severity. Similar to [35], the built feature extractors simplified the extraction of further information from the model, improving the model's explainability. A wider range of datasets were used in such work for training the feature extractor [32,37–41].

In summary, the reviewed papers implemented transfer-learning and data augmentation to deal with limited labelled data. Fewer proposed methods trained more specific feature extractors [35,36]. The datasets in [27,28,32] have been used extensively in previous work. The frequently used dataset in [27] includes COVID-19^+ observations made available by Dr. Joseph Cohen, from the University of Montreal [27]. The images were collected from journal websites such as radiopaedia.org, the Italian Society of Medical and Interventional Radiology. The images were also collected from recent publications in this area such as [27]. The dataset is composed of chest X-ray images involving over 100 patients. Their ages range from 27 to 85 years old. The countries of origin include: Iran, China, Italy, Taiwan, Australia, Spain and the United Kingdom. A warning has been raised by the authors on [27] with regards to any diagnostic performance claims prior to doing a proper clinical study.

As for the dataset available in [28], frequently used in previous work for normal and pneumonia readings, all of them correspond to samples taken from paediatric Chinese patients. The usage of such data as negative COVID-19 cases can be less reliable, since different populations were sampled for COVID-19 and no COVID-19 cases. Observations of adults (with ages ranging between 20 and 86 years old) were used for COVID-19^+ cases, while for the normal and pneumonia cases in [28], the images were sampled from paediatric patients. The usage of biased datasets is a lurking danger in recent COVID-19 machine learning based detection systems [42]. Therefore, in this work we test a wider variety of sources for COVID-19^- cases, including a new dataset with Costa Rican adult patients.

We highlight the fact that both the test and training datasets are drawn from the same distribution in most of the aforementioned studies, with usually one data source for COVID-19 positive cases. Moreover, the test datasets are usually very small (for instance in [29] less than 50 test images were used). Little exploration on the benefits of using a fully SSDL model can be found in the literature, for COVID-19 detection using X-ray images. Furthermore, to our knowledge no work on the impact and correction of data imbalance in SSDL for COVID-19 detection has been developed so far in the literature.

2.2. Semi-supervised deep learning and data imbalance correction

In general, deep learning models require a large number of labelled observations to provide good levels of generalization. This limitation makes it hard to implement these techniques to medical applications. Given the lack of labelled data SSDL is gaining increasing popularity in the academic community. It is well suited to deal with datasets which are poorly labelled, or have few labels, making SSDL attractive for computer aided medical imaging analysis, as seen in [43,44]. Semi-supervised methods require the use of both labelled $S_l = (X_l, Y_l)$ and unlabelled samples $S_u = X_u = \left\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_u}\right\}$. Each labelled observation in $X_l = \left\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_l}\right\}$ has an associated label in the set $Y_l = \left\{y_1, \ldots, y_{n_l}\right\}$.

SSDL architectures can be classified as follows: Pre-training, self-training (also known as pseudo-labelled) and regularization based. Some of the regularization methods include generative based approaches, along consistency loss term as well as graph based. An extensive survey on SSDL approaches can be found in [45].

The MixMatch approach developed in [9] merged intensive data augmentation with unsupervised regularization and pseudo-labelled based semi-supervised learning. This method produced better results compared to other regularized, pseudo-labelled and generative based SSDL methods as shown in [9].

Data imbalance for supervised approaches, has been widely studied. The approaches range from data based transformations (data augmentation, over-sampling or under-sampling, generative methods) to architecture based (loss function or ensemble based) [46–48]. Scarce literature is to be found to our knowledge on data imbalance correction for modern SSDL architectures. Data imbalance in the labelled dataset, can be approached as a particularization of the data distribution mismatch problem outlined in [19], when the unlabelled dataset presents a different distribution. This is common under real-world usage conditions of SSDL techniques. In [19], authors made a first glance at the impact of Out of Distribution (OOD) data in the unlabelled dataset S_u, leading to a distribution mismatch between the distributions of S_l and S_u. The work in [49,50] goes deeper into the impact of distribution mismatch data in SSDL. Authors tested several distribution mismatch scenarios with different OOD data contamination degrees, and different OOD data sources. The results showed an important influence on the degree of OOD data in the unlabelled dataset S_u.

In [51], authors explored further the impact of the distribution mismatch, in the particular case of using imbalanced datasets. The results showed a classification error rate decrease, ranging from 2% to 10% for the SSDL model. Furthermore, the authors proposed a straightforward approach for correcting such accuracy degradation. The approach assigned weights to each unlabelled observation, depending on the number of observations per class. Higher weights were used for under-represented observations in the unlabelled loss term. To pick the right weight for each unlabelled observation, the highest label predicted with the model yielded for the current epoch, was used. The authors implemented and tested the approach in the mean teacher model [52]. The results demonstrated a significant accuracy gain by implementing the proposed approach. We base our contribution on these findings, and propose an extended data imbalance correction approach into MixMatch in the context of semi-supervised COVID-19 detection.

2.3. MixMatch

The proposed SSDL method is based on the MixMatch [9] architecture. It creates a set of pseudo-labels, and also implements an unsupervised regularization term. The consistency loss term used by the MixMatch method minimizes the distance between the pseudo-labels and predictions that the model makes on the unlabelled dataset X_u.

The average model output of a transformed input $\boldsymbol{x}_j$ was used to estimate pseudo-labels $\widehat{\boldsymbol{y}}_j = \frac{1}{K}\sum_{\eta=1}^{K} f_{\boldsymbol{w}}\left(\Psi^{\eta}\left(\boldsymbol{x}_j\right)\right)$. Here K corresponds to the number of transformations (like image flipping) Ψ^{η} performed. Based on the work done in [9], a value of $K = 2$ is recommended. The authors also mentioned that the estimated pseudo-label $\widehat{\boldsymbol{y}}_j$ usually presents a high entropy value. This can increase the number of non-confident estimations. Therefore, the output array $\widehat{\boldsymbol{y}}$ was sharpened with a temperature ρ, making up the modified Softmax activation function $s\left(\widehat{\boldsymbol{y}}, \rho\right)_i = \frac{\widehat{y}_i^{1/\rho}}{\sum_j \widehat{y}_j^{1/\rho}}$. The term $\widetilde{S}_u = \left(X_u, \widetilde{Y}\right)$ defines the dataset with the sharpened estimated pseudo labels. It is assumed here that $\widetilde{Y} = \left\{\widetilde{\boldsymbol{y}}_1, \widetilde{\boldsymbol{y}}_2, \ldots, \widetilde{\boldsymbol{y}}_{n_u}\right\}$

In [9] the authors argued that data augmentation is a key aspect when it comes to SSDL. The authors used the MixUp approach, as proposed in [53], to further augment data using both labelled and unlabelled observations, this can be represented as: $\left(S_l', \widetilde{S}_u'\right) = \Psi_{\text{MixUp}}\left(S_l, \widetilde{S}_u, \alpha\right)$. The MixUp method proposed to create new observations based on a linear interpolation of a combination of unlabelled (together with their pseudo-labels) and labelled data. More specifically, for two labelled or pseudo labelled data pairs $(\boldsymbol{x}_a, y_a)$ and $(\boldsymbol{x}_b, y_b)$, MixUp creates a new observation with its corresponding label $\left(\boldsymbol{x}', y'\right)$ based on the following steps:

1. Sample the MixUp parameter λ based on a Beta distribution $\lambda \sim \text{Beta}\left(\alpha, \alpha\right)$, with α chosen by the user.
2. Make sure that $\lambda > 0.5$. This is done by making $\lambda' = \max\left(\lambda, 1 - \lambda\right)$
3. Produce a new observation based on a lineal interpolation of the two observations: $\boldsymbol{x}' = \lambda'\boldsymbol{x}_a + \left(1 - \lambda'\right)\boldsymbol{x}_b$.
4. Generate the corresponding pseudo-label for the new observation $y' = \lambda' y_a + \left(1 - \lambda'\right) y_b$.

The augmented datasets $\left(S_l', \widetilde{S}_u'\right)$ were used by the MixMatch algorithm to train a model as specified in the training function T_{MixMatch}, resulting in the model f with weights $\boldsymbol{w}$:

$$f_{\boldsymbol{w}} = T_{\text{MixMatch}}\left(S_l, X_u, \alpha, \lambda\right) = \underset{\boldsymbol{w}}{\operatorname{argmin}} \mathcal{L}\left(S, \boldsymbol{w}\right) \tag{1}$$

$$\mathcal{L}\left(S, \boldsymbol{w}\right) = \sum_{\left(\boldsymbol{x}_i, \boldsymbol{y}_i\right) \in S_l'} \mathcal{L}_l\left(\boldsymbol{w}, \boldsymbol{x}_i, \boldsymbol{y}_i\right) + \gamma r(t) \sum_{\left(\boldsymbol{x}_j, \widetilde{\boldsymbol{y}}_j\right) \in \widetilde{S}_u'} \mathcal{L}_u\left(\boldsymbol{w}, \boldsymbol{x}_j, \widetilde{\boldsymbol{y}}_j\right) \tag{2}$$

For the labelled loss term, a cross-entropy loss was used in [9]; $\mathcal{L}_l\left(\boldsymbol{w}, \boldsymbol{x}_i, \boldsymbol{y}_i\right) = \delta_{\text{cross-entropy}}\left(\boldsymbol{y}_i, f_{\boldsymbol{w}}\left(\boldsymbol{x}_i\right)\right)$. As for the unlabelled loss term, an Euclidean distance was implemented $\mathcal{L}_u\left(\boldsymbol{w}, \boldsymbol{x}_j, \widetilde{\boldsymbol{y}}_j\right) = \left\|\widetilde{\boldsymbol{y}}_j - f_{\boldsymbol{w}}\left(\boldsymbol{x}_j\right)\right\|$ in [9]. The coefficient $r(t)$ was proposed as a ramp-up function that increases its value as the epochs t increase. In our implementation, $r(t)$ was set to $t/3000$. The γ factor was used as a regularization weight. In our work, we followed the same implementation of both loss functions. This coefficient controls the influence on unlabelled data. It is important to highlight that unlabelled data has also an effect on the *labelled* data term $\mathcal{L}_l$. The reason being that unlabelled data is used to artificially increase data observations by using the MixUp method for also the labelled term.

3. Proposed method: Pseudo-label based balance correction

In this work an implementation of a data imbalance correction in the loss function of the MixMatch method is proposed. Positive results were yielded in [51] for correcting dataset imbalance by weighting the unsupervised loss function terms in a per observation basis. The authors in [51] developed a similar approach by modifying the SSDL framework known as mean teacher [52]. We extend this approach for the MixMatch architecture, but using both the pseudo-labels and augmented labels for selecting the appropriate weights for both the unlabelled and labelled loss terms. We refer to the proposed approach in this work as PBC, and is depicted as follows.

Let the number of observations per class is used to compute the array of correction coefficients $\mathbf{c}$. The actual computation is done by calculating the array $\mathbf{v}$ using the inverse of the amount of observations available in each class S_l: $v_i = \frac{1}{n_i}$. Here n_i corresponds to the total amount of observations for class i. The next step consists of the computation of the array with the normalized weights $\mathbf{c}$ as $c_i = \frac{v_i}{\sum_j^C v_j}$, where C corresponds to the total number of classes. The original and augmented/pseudo labels $\mathbf{y}_i$

and $\widetilde{\mathbf{y}}_j$ (respectively), are contained in the augmented labelled and unlabelled datasets, S_l' and $\widetilde{S}_u'$, respectively, after the MixUp method mentioned in Section 2.3 is executed. Such augmented labels are used to select its corresponding weight in **c**. To do so, the one-hot vector notation of the labels is converted to a numeric one; $b_i = \operatorname{argmax}_k y_{k,i}$, and $\widetilde{b}_j = \operatorname{argmax}_k \widetilde{y}_{k,j}$, for every b_i and $\widetilde{b}_j$ observation in S_l' and $\widetilde{S}_u'$, respectively.

Both the loss function and the calculated weights are used to weight both loss terms:

$$\mathcal{L}(S, \boldsymbol{w}) = \sum_{(\boldsymbol{x}_i, \boldsymbol{y}_i) \in S_l'} \mathcal{L}_l\left(\boldsymbol{w}, \boldsymbol{x}_i, \boldsymbol{y}_i, b_j\right) + \\ \gamma r(t) \sum_{(\boldsymbol{x}_j, \widetilde{\boldsymbol{y}}_j) \in \widetilde{S}_u'} \mathcal{L}_u\left(\boldsymbol{w}, \boldsymbol{x}_j, \widetilde{\boldsymbol{y}}_j, \widetilde{b}_j\right) \tag{3}$$

The chosen indices are used in the array of weights **c**. We used a cross-entropy and mean squared error loss for the labelled and unlabelled loss terms, respectively. Therefore, the modified cross-entropy and MSE functions are respectively described as follows: $\mathcal{L}_l(\boldsymbol{w}, \boldsymbol{x}_i, \boldsymbol{y}_i) = \delta_{\text{cross-entropy}}\left(c_{b_i}\boldsymbol{y}_i, c_{b_i} f_{\boldsymbol{w}}(\boldsymbol{x}_i)\right)$ and $\mathcal{L}_u\left(\boldsymbol{w}, \boldsymbol{x}_j, \widetilde{\boldsymbol{y}}_j\right) = \left\| c_{\widetilde{b}_j}\widetilde{\boldsymbol{y}}_j - c_{\widetilde{b}_j} f_{\boldsymbol{w}}(\boldsymbol{x}_j)\right\|$. The numerical estimated and real labels are then used for indexing the array **c**. The re-weighted loss functions are minimized as usual.[1]

4. Datasets

A system to classify x-ray images into: COVID-19^+ and no lung pathology (COVID-19^-) is presented in this work. We used different previously existing datasets, and add the usage of a new one, containing negative COVID-19 cases from Costa Rican patients. The following previously existing datasets were used in this work.

Cohen's COVID-19^+ dataset: Images containing COVID-19^+ observations were collected from the publicly available GitHub repository accessible from [27]. This repository was built by Dr. Joseph Cohen, from the University of Montreal [27], and is composed of around 100 images at the time of writing this work. The images were collected from journal websites such as radiopaedia.org and the Italian Society of Medical and Interventional Radiology. Images were also collected from recent publications in this area. Only images containing signs of COVID-19^+ were used in our work. All other images relating to Middle East Respiratory Syndrome (MERS), Acute Respiratory Distress Syndrome (ARDS) and Severe Acute Respiratory Syndrome (SARS) were discarded. This reduced the dataset to a subset containing 102 front chest X-ray containing COVID-19^+ observations. The grey-scaled observations were stored with varying resolutions from 400 × 400 up to 2500 × 2500 pixels.

Valencian Region Medical Image Bank COVID-19^+ dataset: An additional alternative source of COVID-19^+ readings, is the dataset depicted in [54], referred by the authors as Valencian Region Medical ImageBank (BIMCV). The dataset includes chest X-ray and Computed Tomography (CT) images. The dataset also contains detailed findings for the observations, covering different thoracic entities. A total of 1311 subjects were included in the dataset sample, with an age ranging from 25 to 100 years, with around 46% female patients. The dataset includes a total of 2427 chest X-rays. The images were stored in PNG format with an original resolution of 299 × 299 pixels.

Chinese paediatric patients dataset: A dataset of 5856 observations containing images of pneumonia and normal observations was defined in [28]. The patient sample used for the study correspond to Chinese children [28]. These images were divided into 4273 observations of pneumonia (including viral and bacterial) and 1583 of observations with no lung pathology (normal). We used the observations with no findings, and refer to it as the Chinese paediatric dataset. The negative and pneumonia observations from this dataset have been used extensively in recent related research to COVID-19 detection [30,55–58]. Most of the images were stored with a resolution of 1300 × 600 pixels.

ChestX-ray8 dataset: The ChestX-ray8 dataset, made available in [41], is also used for the category of no findings in this work. The dataset includes 224,316 chest radiographs from 65,240 patients from Stanford Hospital, US. The studies were done between October 2002 and July 2017. We picked a sample of this dataset available in its website[2] given the low labelled data setting used in this work. Patients sampled in this dataset were aged from 0 to 94 years old.

Indiana Chest X-ray dataset: The dataset published in [37] gathers 8121 images from the Indiana Network for Patient Care. The dataset can be accessed from its repository.[3] Images were stored with a resolution of 1400 × 1400 pixels. Only the observations with no pathologies were used in this work.

Costa Rican dataset: In this work we also used a dataset we gathered from a Costa Rican private clinic, Clinica Imagenes Medicas Dr. Chavarria Estrada. The data corresponds to chest X-rays from 153 different patients, with ages ranging from 7 to 86 years old. 63% of the patients were female and 37% are male. The images were taken using a Konica Minolta digital X-ray machine with 0.175 of pixel spacing. The images were stored with a resolution of 1907 × 1791 pixels. As the images were digitally sampled, no tags or manual labels are contained in the images.[4] As for ethical compliance of our procedure for gathering mammogram images data, we have an explicit permission from the Chavarria Clinic board to use it with academical purposes. Our data was gathered from the Clinica Chavarria's patients of 2020. Therefore, the data was already collected before this study. We declare that the data collection process of this study complies with the Helsinki's declaration for human based studies, as this study is entirely observational, and the data was already acquired during regular clinical practice.

RSNA dataset: For multi-class classification into common pneumonia (viral and bacterial) and COVID-19^+ positive cases (using the aforementioned Dr. Cohen's repository in [27]), we used the Radiological Society of North America (RSNA) dataset as described in [22,56]. A pool of 69 observations per class (pneumonia and normal observations) and 69 observations for COVID-19^+ cases was used for each batch, randomly picked from the original RSNA dataset.

5. Experiments definition

We implemented two test-beds for binary classification, a regular-sized dataset and an extended-size test dataset. For the regular-sized test dataset for binary classification, in each run, a random sample dataset of 204 observations was picked from both the evaluated COVID19^- dataset (Costa Rica, Indiana, ChestX-ray8 and Chinese paediatric dataset) and the COVID-19^+ dataset available in [27]. Therefore, a total of 10 different training and test samples were used. The same samples were used across all the tested architectures. A completely balanced test dataset comprising the 30% of the 204 observations was used (62 test observations), and the rest was used as labelled and unlabelled

[1] Upon paper publication, we are going to make it available through a public GitHub repository.

[2] https://www.kaggle.com/nih-chest-xrays/sample/data.

[3] https://www.kaggle.com/raddar/chest-xrays-indiana-university.

[4] The dataset will be available upon paper publication.

Table 1
Mean accuracy/F1-score/precision/recall for the Costa Rican dataset, for the oversampling (OS), the original MixMatch architecture and the proposed PBC imbalance correction method. LB stands for the label balancing usage (PBC in the case of SSDL).

COVID-19$^-$	COVID$^+$	LB	$n_l = 10$	$n_l = 15$	$n_l = 20$
			$\overline{x}$	$\overline{x}$	$\overline{x}$
80%	20%	PBC	0.943/0.905/0.937/0.879	0.943/0.894/0.944/0.853	0.936/0.913/0.951/0.883
		OS	0.877/0.718/1/0.562	0.881/0.727/1/0.572	0.882/0.698/1/0.5372
		No	0.891/0.726/1/0.573	0.877/0.718/1/0.56	0.874/0.696/1/0.535
70%	30%	PBC	0.941/0.889/0.931/0.853	0.946/0.907/0.948/0.872	0.953/0.918/0.965/0.875
		OS	0.91/0.793/0.982/0.671	0.903/0.828/1/0.707	0.906/0.798/0.996/0.669
		No	0.91/0.789/0.982/0.664	0.903/0.778/0.996/0.64	0.905/0.818/1/0.696

Table 2
Accuracy results with the COVID-19$^-$ from the Costa Rican dataset, the higher, the better. LB stands for label balancing, with usual weight correction for the supervised model, and the proposed PBC for the MixMatch model. A total of $n_l = 10$, $n_l = 15$ and $n_l = 20$ labelled observations were tested. Two data imbalance settings were tested, with 70%/30% and 80%/20%. The sample mean $\overline{x}$ and the sample standard deviation s are reported.

SSDL	COVID-19$^-$	COVID-19$^+$	LB	$n_l = 10$		$n_l = 15$		$n_l = 20$	
				$\overline{x}$	s	$\overline{x}$	s	$\overline{x}$	s
No	50%	50%	NA	0.871	0.039	0.912	0.049	0.951	0.025
	70%	30%	Yes	0.877	0.040	0.900	0.053	0.931	0.034
			No	0.877	0.040	0.924	0.056	0.931	0.044
	80%	20%	Yes	0.876	0.060	0.903	0.058	0.922	0.037
			No	0.876	0.079	0.907	0.072	0.938	0.035
Yes	50%	50%	NA	0.941	0.035	0.955	0.025	0.957	0.030
	70%	30%	Yes	0.955	0.027	0.947	0.035	0.950	0.029
			No	0.907	0.042	0.900	0.049	0.914	0.028
	80%	20%	Yes	0.957	0.025	0.964	0.021	0.960	0.020
			No	0.922	0.031	0.926	0.047	0.919	0.033

Table 3
Accuracy results with the COVID-19$^-$ cases gathered from the Chinese paediatric repository available in [28]. LB stands for label balancing, with usual weight correction for the supervised model, and the proposed PBC for the MixMatch model.

SSDL	COVID-19$^-$	COVID-19$^+$	LB	$n_l = 10$		$n_l = 15$		$n_l = 20$	
				$\overline{x}$	s	$\overline{x}$	s	$\overline{x}$	s
No	50%	50%	NA	0.882	0.077	0.868	0.080	0.925	0.039
	70%	30%	Yes	0.812	0.050	0.815	0.089	0.883	0.048
			No	0.823	0.048	0.815	0.087	0.868	0.064
	80%	20%	Yes	0.857	0.107	0.898	0.052	0.930	0.053
			No	0.823	0.125	0.872	0.066	0.930	0.037
Yes	50%	50%	NA	0.945	0.036	0.950	0.026	0.963	0.028
	70%	30%	Yes	0.925	0.042	0.930	0.053	0.943	0.034
			No	0.902	0.058	0.898	0.091	0.915	0.044
	80%	20%	Yes	0.947	0.037	0.957	0.022	0.962	0.028
			No	0.847	0.122	0.857	0.141	0.895	0.042

dataset (142 observations). As for the extended-size test dataset for binary classification, we used a total of 300 images from the BIMCV dataset as COVID-19$^+$ readings source, and for COVID-19$^+$ data source, we mixed the Chest-Xray8 and Indiana chest X-ray dataset in equal proportions, using also 300 images, accumulating 600 images in total. 400 of the images were used for test, and the remaining 200 images for training (with a varying number of labelled and unlabelled images and class imbalance settings, as we will detail later). Picking different data sources for COVID19$^-$ observations can be considered to raise discrimination complexity, a frequently skipped setting in previous work. We selected such datasets, as they present a similar patient age distribution to the COVID19$^+$ dataset used. For all the binary classification test datasets, the number of observations per-class are completely balanced.

Regarding the multi-class classification test-bed, we also implemented a regular-sized and a extended-size test datasets. For the regular-sized dataset, 90 test images are used, along with a total of 210 training images (either labelled or unlabelled, depending on the test-bed setting). COVID-19$^+$ observations were randomly picked from Dr. Cohen's dataset, with pneumonia and normal readings picked from the RSNA dataset. The extended-size multi-class classification dataset is composed as follows: For COVID-19$^+$ observations, we used the 102 images from the Dr. Cohen's dataset, and 98 images from BIMCV dataset. The normal and pneumonia observations were picked randomly from the RSNA dataset, with 200 observations for each class. Therefore, a total of 600 images compose the dataset. From there, 300 images were used for test, and the remaining 300 observations were used as either labelled or unlabelled observations, with a varying number of n_l labelled observations from 50 to 90, and class imbalance settings. This testing setting can be considered more challenging, as both COVID-19 positive and negative observations come from different distributions. For all the multi-class classification test datasets, the number of observations per-class are completely balanced.

To assess the data imbalance impact in binary classification, we evaluated both the supervised and the semi-supervised architectures using three balance configurations: 50%/50%, 80%/20% and 70%/30% for the labelled dataset S_l. The under-represented class corresponds to the COVID-19$^+$ class. We tested different sizes of labelled samples, $n_l = 10$, $n_l = 15$ and $n_l = 20$ (from

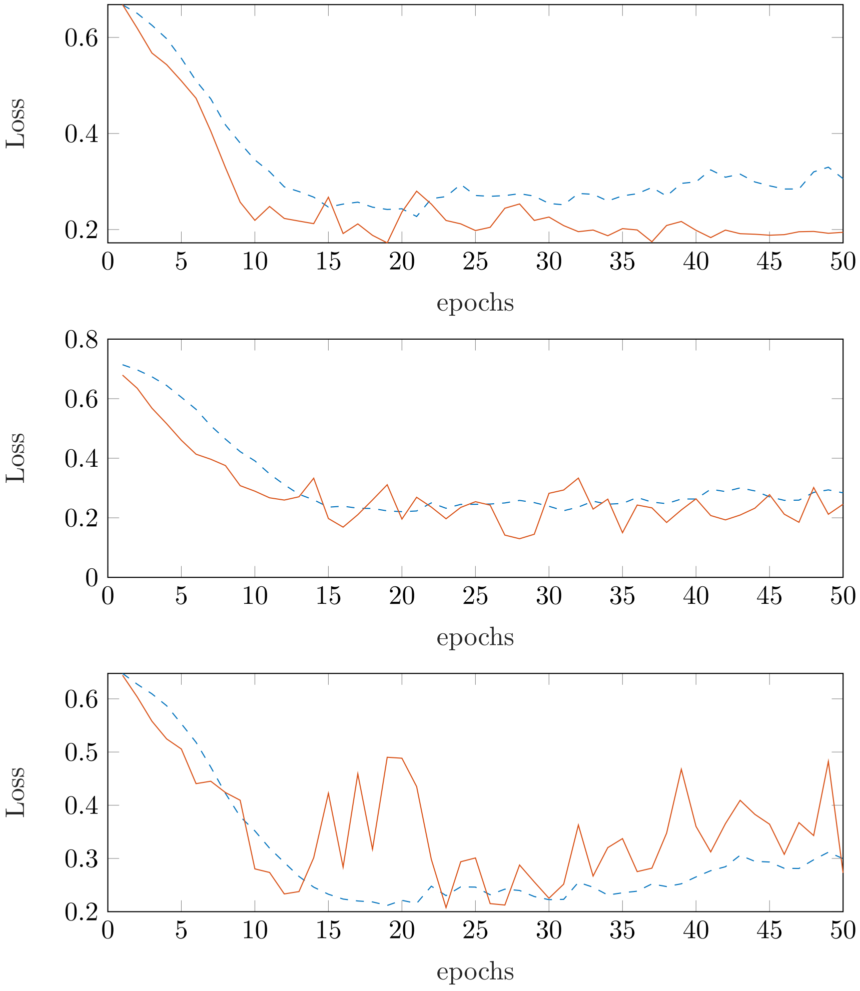

Fig. 1. Training and validation curves for the SSDL with PBC, the SSDL model with no label balancing and supervised models, respectively, from top to bottom. The blue dashed line corresponds to the training loss and the red continuous line to the validation loss.

the 142 observations for the regular-sized test dataset, and 400 observations for the extended-sized test dataset), using the rest as unlabelled data. The remaining data was used as unlabelled data, with close to a 50% data balance between the two classes. This leads to a distribution mismatch between S_u and S_l. Tables 2, 3, 5 and 4 show the evaluated setting and its results, for the Costa Rican, Chinese, Indiana and ChestX-ray8 datasets, for the regular-sized test-bed. Table 6 summarizes the results. As for the extended-size binary classification test-bed, Table 9 shows its results and the described test settings.

As for multi-class classification, we tested three different imbalance scenarios, with 10%/45%/45%, 20%/40%/40% and 30%/35%/35%, with COVID-19$^+$ as the under-represented class in all the three configurations, and balanced pneumonia (both viral and bacterial) and normal chest X-ray observations. We also tested different labelled sample sizes, with $n_l = 30$, $n_l = 50$, $n_l = 70$ and $n_l = 90$. The labelled sample sizes were higher than the binary classification setting, as a multi-classification problem often needs of more observations. Tables 7 and 8 show the described test layout (regular and extended size test datasets, respectively), with the averages and standard deviations reported for each configuration over 10 runs with randomly picked data partitions. To complement the results description of Table 7, we show the averaged confusion matrices over all the 10 runs for both the standard and extended size test datasets, in Tables 10 and 11, respectively. The confusion matrices were calculated from the final model yielded after the 50 epochs, and not the best one according to the validation dataset.

All the datasets have been preprocessed to exclude artefacts (manual labels), in the cases where one of them does not present any, to avoid artefact bias. Data augmentation using flips and rotations is implemented. No crops were used to avoid losing regions that might be important for image discrimination. Images stored with 8 bits were replicated by 3 to use the selected CNN architecture. We used the following hyper-parameters used for the MixMatch model for all the experiments performed: $K = 2$ transformations, $T = 0.5$ of sharpening temperature and $\alpha = 0.75$ for the beta distribution, as advised in [9].[5] A Wide-ResNet [59] model has been used for the binary classification experiments (regular-sized dataset) given its preliminar good results in our experiments, with an input image size of 110 × 110 pixels (limited by the graphics processor memory needed by MixMatch). For multi-class classification and the binary extended-sized dataset, we used a more efficient densenet model, which allowed us to use an input image size of 220 × 220 pixels, as more resolution might be necessary, given the higher number of classes to discriminate from.

The following training hyper-parameters were used: a weight decay of 0.0001, a learning rate of 0.00001, a batch size of 12 observations, a cross-entropy loss function and an Adam optimizer

[5] The MixMatch implementation used in this work is based on the implementation available in repository https://github.com/noachr/MixMatch-fastai.

Table 4
Accuracy results with the COVID-19^{-} cases gathered from the ChestX-ray8 repository available in [41]. LB stands for the label balancing usage (PBC in the case of SSDL).

SSDL	COVID-19^{-}	COVID-19^{+}	LB	$n_l = 10$		$n_l = 15$		$n_l = 20$	
				$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s
No	50%	50%	NA	0.756	0.062	0.727	0.062	0.756	0.050
	70%	30%	Yes	0.732	0.039	0.723	0.043	0.752	0.038
			No	0.739	0.051	0.744	0.053	0.773	0.049
	80%	20%	Yes	0.729	0.051	0.721	0.054	0.768	0.047
			No	0.735	0.052	0.739	0.070	0.777	0.050
Yes	50%	50%	NA	0.803	0.059	0.814	0.052	0.840	0.038
	70%	30%	Yes	0.816	0.048	0.815	0.038	0.839	0.049
			No	0.782	0.054	0.760	0.068	0.782	0.051
	80%	20%	Yes	0.798	0.050	0.818	0.044	0.824	0.039
			No	0.735	0.056	0.740	0.075	0.752	0.048

Table 5
Accuracy results with the COVID-19^{-} cases gathered from Indiana dataset [37]. LB stands for the label balancing usage (PBC in the case of SSDL).

SSDL	COVID-19^{-}	COVID-19^{+}	LB	$n_l = 10$		$n_l = 15$		$n_l = 20$	
				$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s
No	50%	50%	NA	0.845	0.044	0.853	0.053	0.879	0.038
	70%	30%	Yes	0.834	0.042	0.839	0.053	0.874	0.046
			No	0.845	0.058	0.860	0.050	0.869	0.061
	80%	20%	Yes	0.845	0.048	0.829	0.053	0.856	0.042
			No	0.840	0.041	0.827	0.045	0.853	0.066
Yes	50%	50%	NA	0.905	0.047	0.918	0.038	0.908	0.029
	70%	30%	Yes	0.882	0.067	0.902	0.046	0.902	0.042
			No	0.837	0.078	0.819	0.109	0.834	0.037
	80%	20%	Yes	0.860	0.076	0.889	0.056	0.885	0.035
			No	0.803	0.062	0.747	0.095	0.795	0.078

Table 6
Accuracy gain comparison when using no SSDL (No MM) vs. MixMatch with the proposed loss balancing correction (MM+PBC), and to using MixMatch with no balancing correction (MM) vs. MixMatch with the proposed loss balancing correction (MM+PBC). The accuracy gain is evaluated for the tested number of labelled observations (10, 15 and 20). Italic entries correspond to non statistically meaningful gains, after performing a Wilcoxon test, with $p > 0.1$.

SSDL	COVID-19^{-}	COVID-19^{+}	Comparison	$n_l = 10$	$n_l = 15$	$n_l = 20$
				Acc. gain	Acc. gain	Acc. gain
Costa Rican dataset	70%	30%	MM+PBC vs. No MM	+0.07	+0.046	+0.018
			MM+PBC vs. MM	+0.048	+0.046	+0.036
	80%	20%	MM+PBC vs. No MM	+0.081	+0.06	+0.038
			MM+PBC vs. MM	+0.034	+0.038	+0.041
Chinese paediatric dataset	70%	30%	MM+PBC vs. No MM	+0.113	+0.115	+0.06
			MM+PBC vs. MM	+*0.023*	+*0.031*	+*0.028*
	80%	20%	MM+PBC vs. No MM	+0.09	+0.058	+0.031
			MM+PBC vs. MM	+0.1	+0.099	+0.066
Chest X-ray8 dataset	70%	30%	MM+PBC vs. No MM	+0.083	+0.092	+0.087
			MM+PBC vs. MM	+*0.033*	+0.055	+0.057
	80%	20%	MM+PBC vs. No MM	+0.069	+0.096	+0.056
			MM+PBC vs. MM	+0.063	+0.0774	+0.072
Indiana dataset	70%	30%	MM+PBC vs. No MM	+0.048	+0.063	+*0.027*
			MM+PBC vs. MM	+0.045	+0.082	+0.067
	80%	20%	MM+PBC vs. No MM	+0.014	+0.059	+*0.029*
			MM+PBC vs. MM	+*0.056*	+0.141	+0.09

with a 1-cycle policy [60]. For each configuration, we trained the model a total of 50 epochs, in 10 different runs. Fig. 1 shows validation and training loss curves for a Wide-ResNet model, trained with the MixMatch approach (with the proposed PBC and without it) and through a regular supervised fashion, in a particular batch. 10 labels were used, with the 30/70 percent imbalance scenario. For each epoch, the whole dataset was evaluated in batches of 10 observations. The curves show the regularization effect of semi-supervised learning, with fast convergence in less than 50 epochs for both training approaches. The proposed method improves even more the regularization effect of the SSDL.

A baseline experiment using the Costa Rican dataset was done, aiming to compare the performance of the proposed PBC approach to another simple technique frequently used to correct data imbalance in supervised models; over-sampling. Under-sampling was not used as the scarce labelling settings lead to model over-fitting (using the regular-sized test dataset). We skipped far more complex approaches in the comparison such as generative networks [48], to focus in more straightforward data imbalance correction approaches. We compare the testing accuracy, F1-score, precision and recall of these two methods with the non imbalance corrected MixMatch baseline in Table 1. Also, the ROC curves are plotted in Figs. 2 and 3 for both the standard and extended sized test datasets, respectively. Tables 2–5 show this layout. Given the low labelled setting, we report the highest validation accuracy, assuming the usage of early stopping

Table 7
Multi-class classification for accuracy measures, using the RSNA dataset (standard-sized test dataset). LB stands for the label balancing usage (PBC in the case of SSDL). PBC results with no statistical significance gains over the non-balanced SSDL implementation are written in italic.

COVID-19/Pneumonia/Normal	SSDL	LB	$n_l = 30$		$n_l = 50$		$n_l = 70$		$n_l = 90$	
			$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s
10%/45%/45%	No	No	0.561	0.052	0.575	0.062	0.588	0.056	0.603	0.039
	Yes	No	0.587	0.048	0.565	0.045	0.595	0.035	0.571	0.056
		Yes	0.736	0.054	0.748	0.043	0.757	0.045	0.744	0.027
20%/40%/40%	No	No	0.607	0.054	0.636	0.042	0.667	0.036	0.711	0.036
	Yes	No	0.67	0.063	0.691	0.043	0.695	0.048	0.7	0.044
		Yes	0.733	0.052	0.747	0.048	0.758	0.052	0.752	0.024
30%/35%/35%	No	No	0.656	0.048	0.666	0.057	0.698	0.041	0.707	0.039
	Yes	No	0.727	0.654	0.766	0.047	0.76	0.032	0.738	0.04
		Yes	*0.732*	*0.043*	*0.752*	*0.04*	*0.778*	*0.072*	*0.727*	*0.038*

Table 8
Multi-class classification for accuracy measures, using the BMIVC dataset (extended-sized dataset with 300 test observations). LB stands for the label balancing usage (PBC in the case of SSDL). PBC results with no statistical significance (with $p > 0.1$) gains over the non-balanced SSDL implementation are written in italic.

COVID-19/Pneumonia/Normal	SSDL	LB	$n_l = 50$		$n_l = 70$		$n_l = 90$	
			$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s
10%/45%/45%	No	No	0.55	0.041	0.58	0.04	0.61	0.035
	Yes	No	0.513	0.016	0.511	0.011	0.521	0.02
	Yes	Yes	0.618	0.047	0.647	0.027	0.67	0.027
20%/40%/40%	No	No	0.597	0.049	0.627	0.037	0.661	0.031
	Yes	No	0.5805	0.045	0.585	0.036	0.573	0.029
	Yes	Yes	0.652	0.036	0.677	0.03	0.686	0.035
30%/35%/35%	No	No	0.615	0.036	0.649	0.012	0.68	0.24
	Yes	No	0.671	0.037	0.675	0.025	0.695	0.021
	Yes	Yes	*0.67*	*0.03*	*0.671*	*0.02*	*0.677*	*0.027*

Table 9
Results for the extended test dataset binary classification setting. LB stands for the label balancing usage (PBC in the case for SSDL), using 400 test images. PBC results with no statistical significance (with $p > 0.1$) gains over the non-balanced SSDL implementation are written in italic.

COVID-19$^-$	COVID$^+$	SSDL	LB	$n_l = 10$		$n_l = 15$		$n_l = 20$	
				$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s
50%	50%	No	No	0.671	0.032	0.696	0.037	0.722	0.035
70%	30%	Yes	Yes	0.773	0.055	0.769	0.305	0.761	0.025
		Yes	No	0.651	0.065	0.678	0.075	0.652	0.09
		No	Yes	0.631	0.072	0.622	0.066	0.659	0.05
80%	20%	Yes	Yes	0.785	0.045	0.78	0.035	0.772	0.029
		Yes	No	0.7	0.062	0.667	0.058	0.695	0.053
		No	Yes	0.67	0.036	0.642	0.042	0.726	0.01

to avoid over-fitting. We trained the MixMatch model with both the uncorrected loss function and the proposed PBC modification for data imbalance correction. For reference, we also tested the supervised model with balance correction and without it, for binary classification.

Table 6 summarizes the accuracy gains when using MixMatch with PBC vs. not using MixMatch, and using MixMatch with no balance correction (under the same balance conditions) vs. using MixMatch with PBC. A non-parametric Wilcoxon test was performed to detect whether the accuracy gain is statistically significant (with $p > 0.1$) across the 10 runs (observations) sampled. Gains not statistically significant according such criteria are written in italic in Table 6. This was also done for the multi-class test results in Table 7.

Finally, as a qualitative experiment, we calculated the gradient activation maps using the technique proposed in [61].[6] For this qualitative experiment, we compared the supervised model and the MixMatch modification with the proposed PBC. The objective of this experiment was to spot the changes on the regions used by the model to output its decision, when trained with the semi-supervised approach. A sample with 20 labelled observations and around 180 unlabelled observations (for the MixMatch model with PBC) was used for training the model. A completely balanced dataset of 61 observations was used for validation. We trained a Densenet121 model for 50 epochs, for both the supervised and semi-supervised frameworks. Fig. 4 includes sampled heatmaps for the chest X-ray8 and Indiana datasets. The net weights in the final output layer for each entry, and the real and predicted labels are also shown for each output image in Fig. 4.

6. Results and analysis

6.1. Binary classification results

As for the first experiment comparing the proposed method against over-sampling for binary classification in the standard-sized dataset, the results depicted in Table 1 show a clear advantage of using the proposed PBC against over-sampling, with accuracy gains around 5%, and F1-score gains of almost 10%. The recall is heavily improved when using the PBC, since the false negative rate decreases. We think that the usage of specific information within unlabelled data is important for correcting data imbalance, as the PBC use the pseudo and augmented labels. The

[6] We used the FastAI implementation available of the gradient activation maps available in https://forums.fast.ai/t/gradcam-and-guided-backprop-intergration-in-fastai-library/33462.

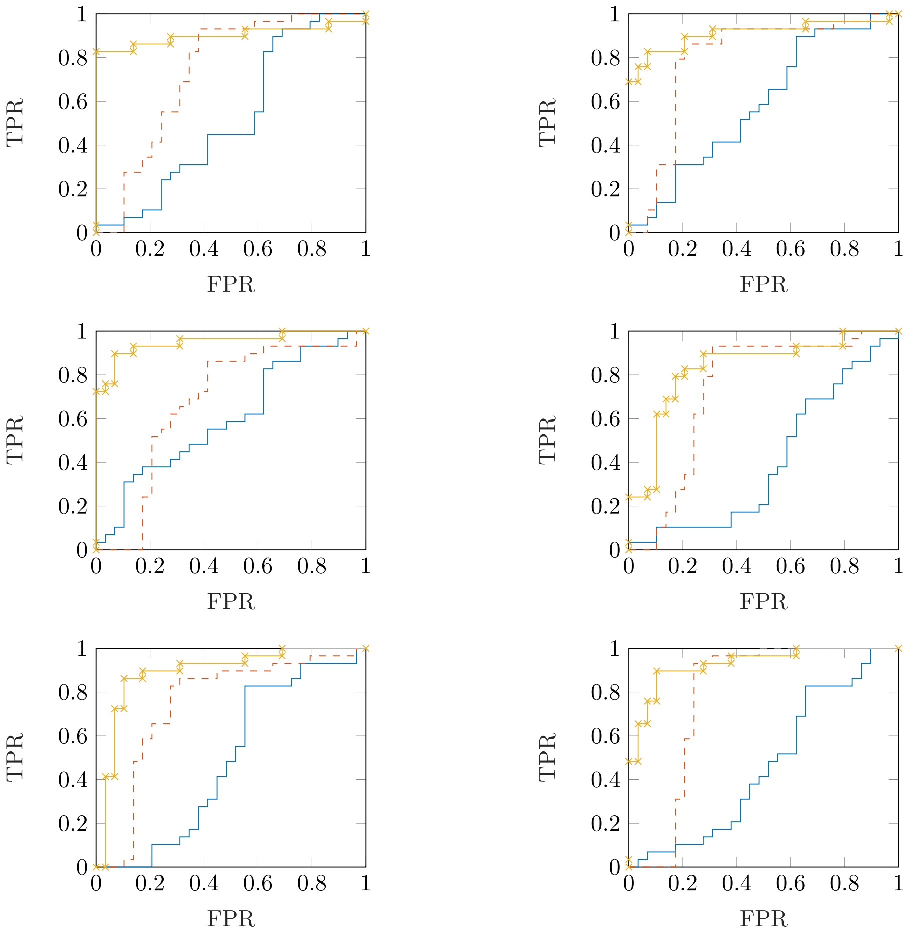

Fig. 2. Receiver Operator Curves for binary classification (regular-sized test dataset), for the semi-supervised and supervised models with $n_l = 10$, $n_l = 15$ and $n_l = 20$ (from top to bottom), for the 20/80 percent (left column) and 30/70 percent (right column) imbalance settings (COVID and non COVID classes). The yellow 'x' line corresponds to the SSDL with PBC ROC curve, the red dashed line to the SSDL with no imbalance correction, and the blue continuous curve to the supervised model ROC curve. As usual, the x-axis corresponds to the false positive ratio, and the y-axis to the true positive ratio.

ROC curves depicted in Fig. 2 for the tests with the Costa Rican dataset in the 20%/80% (left column) and 30%/70% (right column) imbalance scenarios, also show a strong gain of the proposed balance correction method over the semi-supervised model with no balance correction. This is correlated to the improvement on the true positive rate observed in the confusion matrices in Table 10 when using our proposed method. The statistical relevance of the results is evaluated for the rest of the experiments with more datasets.

The results using accuracy as a metric for the Costa Rican dataset are depicted in Table 2. The base-line accuracy is rather high for very limited labelled settings, even with the base-line supervised model, with accuracies ranging from 87% to 95%, using 10 and 20 labels, respectively. SSDL is more attractive when using 10 labels, with an accuracy gain of around 7%, as displayed in Table 6. The accuracy gain from implementing PBC vs. using the non-balanced MixMatch approach remained similar in disregard of the number of labels used, always with statistical significance. However, the accuracy gain of using MixMatch, even with the PBC modification, diminishes as the number of labels increases. The accuracy gain was rather similar for both of the data imbalance configurations tested. As seen in Table 2, the implemented PBC corrects the data imbalance impact, yielding similar results when using the completely balanced dataset.

Regarding the test results using the Chinese paediatric dataset, the base-line supervised accuracy results were initially low (from 86% to 92%), giving more room for SSDL accuracy gain, as seen in Table 3. The usage of MixMatch with the proposed PBC over regular supervised learning yielded an accuracy gain over +11% as seen in Table 6. Similar to the Costa Rican dataset, as the number of labels increases, the accuracy gain decreased. The benefit of using the PBC over the off-the-shelf MixMatch implementation is higher when facing a more imbalanced dataset scenario, as seen in Table 6 for the Chinese dataset. The accuracy gain was almost three times higher when using the 80%/20% configuration, increasing from around +3% to +10%, for the 70%/30% and 80%/20% imbalance scenarios, respectively. The PBC was able to almost correct the impact of data imbalance, as its accuracy shown in Table 3 often was similar to the base-line MixMatch accuracy with a balanced dataset.

Table 4 summarizes the results yielded for the Chest X-ray8 dataset. The base-line accuracy for the supervised model was the lowest from the tested datasets, sitting at around 75%. The accuracy gain of using MixMatch with PBC versus the usual supervised model ranged from +5% to +9.6%, as seen in Table 6, in the row for the Chest X-ray8 dataset. As for the accuracy gain of using MixMatch with PBC vs. MixMatch with no balance correction, it stayed around +3 to +5% for the 70%/30% imbalance configuration. Higher accuracy gains were obtained when dealing with the more challenging imbalance scenario of 80%/20%, with gains up to 14%. Similar to other datasets, the PBC was able to correct MixMatch's accuracy impact of data imbalance most of the times, as seen in Table 4.

The test results for the Indiana dataset are depicted in Table 5. The base-line accuracy for the Indiana chest x-ray dataset ranged

Table 10
Averaged and truncated confusion matrix for multi-class classification using the standard-sized test dataset, for 10 runs, using $n_l = 50$ labels. From left to right, using 10/35/35, 20/40/40 and 30/35/35 percent of the labels for COVID-19, Pneumonia and normal diagnostics, respectively. From top to bottom, the supervised model, the SSDL model with no PBC, and the SSDL model with the PBC.

	Covid-19	Pneumonia	Normal
Covid-19	6	13	13
Pneumonia	0	23	9
Normal	0	8	24

	Covid-19	Pneumonia	Normal
Covid-19	2	15	12
Pneumonia	0	22	7
Normal	0	8	21

	Covid-19	Pneumonia	Normal
Covid-19	23	4	2
Pneumonia	5	19	5
Normal	4	7	18

	Covid-19	Pneumonia	Normal
Covid-19	15	7	7
Pneumonia	2	19	8
Normal	2	6	21

	Covid-19	Pneumonia	Normal
Covid-19	11	10	7
Pneumonia	0	22	6
Normal	0	7	21

	Covid-19	Pneumonia	Normal
Covid-19	21	5	3
Pneumonia	2	20	7
Normal	3	7	19

	Covid-19	Pneumonia	Normal
Covid-19	20	4	5
Pneumonia	5	17	7
Normal	3	6	19

	Covid-19	Pneumonia	Normal
Covid-19	19	6	3
Pneumonia	1	21	6
Normal	2	7	20

	Covid-19	Pneumonia	Normal
Covid-19	20	6	3
Pneumonia	2	20	7
Normal	2	7	20

from 84% to 88%. The accuracy gain from implementing MixMatch with PBC ranged from 4% and to 5.6% versus the base-line supervised model. Implementing the PBC versus the original MixMatch yielded an accuracy gain from +4.5% to +14%. In the case of this dataset, data imbalance seems to further decrease MixMatch's accuracy, as we seen in Table 5 when comparing the accuracy results of the 50%/50% configuration to the 70%/30% and 80%/20% imbalance settings.

For the tested datasets in the binary classification setting, the accuracy can be considered to be very similar when evaluating the base-line supervised model under different data imbalance conditions, as seen in Tables 2–5, suggesting a higher sensitivity of MixMatch when trained with imbalanced data. The overall trend of the accuracy gain of using the proposed MixMatch with PBC over its original implementation was positive, as seen in 6, accross all the datasets tested. Most of the accuracy gains were higher than 3%, and also most of them are statistically significant, after performing a non parametric Wilcoxon test, with an acceptance criteria of the hypothesis of significant difference between the accuracies of both configurations of $p > 0.1$. There were some cases where the default MixMatch implementation did not bring any accuracy gain when facing an imbalanced dataset, as seen for instance in the test results of the Indiana dataset, detailed in Table 5. For example the accuracy of the supervised model with 10 labels was around 83%, and the accuracy of the MixMatch model with no PBC is no higher than 83%. This implies the mandatory need of correcting data imbalancing for the MixMatch model, given its high sensitivity to data imbalance.

Regarding the test-results for the extended-sized test dataset for binary classification, its results are depicted in Table 9. In general, the accuracy for all the tested model variations in this test-bed remains significantly lower than previous tested datasets for the binary classification setting. Such results were expected as the negative COVID-19 data sources were mixed. Nevertheless, in this challenging setting, our simple PBC method proves to significantly improve the model's accuracy (with statistical significance, according to our Wilcoxon test results), when compared to both the supervised model with balanced labelled data and the semi-supervised model with no imbalance correction. The accuracy gains go to up to +12%. No significant accuracy difference is perceived when increasing the number of labels in the tested settings. The sampled ROC curves show an important area under the curve gain for the semi-supervised model using the proposed PBC, as seen in Fig. 3.

Finally, regarding the qualitative experiments proposed, Fig. 4 show sample heatmaps for the Indiana and chest X-ray8 datasets, respectively. Both figures reveal how the neural network tend to focus more on lung areas when using the semi-supervised model trained with both datasets. The Densenet121 model trained with MixMatch including the PBC modification yielded an accuracy of 91.3% for the tested sample from the Indiana dataset, and 67.74% for the supervised model. For chest X-ray8 dataset, an accuracy of 93.4% was yielded for the MixMatch framework with PBC, and 77.4% for the supervised model. We can see in Fig. 4 how the hot pixels move towards lung regions when using the semi-supervised model. This tends to happen even when the resulting predictions in both models are correct.

Table 11
Averaged and truncated confusion matrix for multi classification using the Valencian-Cohen dataset for multi-class classification with 300 test images (extended-sized test dataset), for 10 runs, using 40/40/20 percent of imbalance setting (for SSDL). From left to right, using $n_l = 70$, and $n_l = 90$ labels respectively. From top to bottom, the supervised model (with completely balanced labels), the SSDL model with no PBC, and the SSDL model with the PBC.

	Normal	Pneumonia	Covid-19
Normal	50	24	18
Pneumonia	17	61	15
Covid-19	18	18	57

	Normal	Pneumonia	Covid-19
Normal	52	24	17
Pneumonia	15	61	16
Covid-19	17	10	65

	Normal	Pneumonia	Covid-19
Normal	69	20	3
Pneumonia	25	64	3
Covid-19	57	18	18

	Normal	Pneumonia	Covid-19
Normal	66	24	2
Pneumonia	22	68	2
Covid-19	58	20	15

	Normal	Pneumonia	Covid-19
Normal	48	20	24
Pneumonia	14	61	18
Covid-19	15	9	68

	Normal	Pneumonia	Covid-19
Normal	48	21	24
Pneumonia	13	61	19
Covid-19	14	5	75

6.2. *Multi-class classification test results*

Regarding the results depicted in Table 7 for multi-class classification using the standard-sized dataset (90 test images), the proposed PBC method also yielded significant accuracy gains. The highest accuracy boost (of around 18%) was yielded under the most imbalanced setting tested (10% of the labels for the COVID-19$^+$ class), when comparing the model with PBC to the semi-supervised model with no imbalance correction. In very imbalanced scenarios with few labels, the semi-supervised model tends to have similar results when compared to the supervised model. For the 20/40/40 percent imbalance scenario, the accuracy gain of the proposed PBC method decreased, yielding a boost of around 6% when compared to the semi-supervised model with no balance correction. The tendency of a decrease in the accuracy gain of the proposed balance correction method gets more clear for the 30/35/35 setting, with no statistically significant accuracy gain yielded over the semi-supervised model with no balance correction. To complement the analysis, the average confusion matrices for multi-class classification are depicted in Table 10, calculated across the tested imbalance configurations, with $n_l = 50$. For the 10/45/45 setting, the true positives for the COVID-19$^+$ class increased dramatically in the case of the semi-supervised model with the PBC, compared to the supervised and semi-supervised models with no balance correction. This occurred along with a very small decrease of the average of true positives for the rest of the classes. As the imbalance between the COVID-19$^+$ class and the rest of them gets smaller, the gain in the average true positives for the COVID-19$^+$ class decreases for the proposed method.

As for the multi-class classification test-bed, the results are depicted in Table 8. As expected, the yielded accuracy trend is lower when compared to the standard-sized dataset, as two different positive COVID-19 data sources were used. However, our proposed PBC method yields statistically significant accuracy gains for the 10%/45%/45% and 20%/40%/40% imbalance settings. When compared to the semi-supervised model with no imbalance correction, our method yields an accuracy gain of up to 9%. Increasing the number of labels decreases the advantage of using semi-supervised models (as also the number of unlabelled observations is decreased when using more labels). The averaged confusion matrices show a large accuracy gain for the COVID-19$^+$ class for the semi-supervised model using our proposed PBC, with a slight accuracy decrease for the remaining classes, as seen in Table 11. This is consistent with the improvement seen in the ROC curves in the case for the binary classification tests.

7. Conclusions

In this work we have analysed the impact of data imbalance for the detection of COVID-19 using chest X-ray images. This is a real-world problem, which can arise frequently in the context

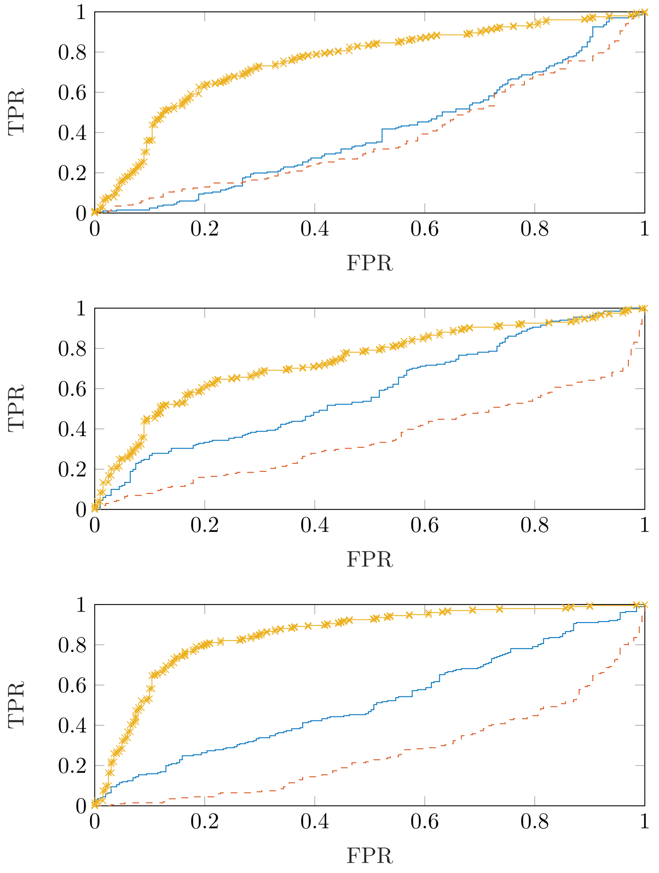

Fig. 3. Sample of Receiver Operator Curves (ROCs) for binary classification using the Valencia dataset (400 test images), for the semi-supervised and supervised models with $n_l = 10$, $n_l = 15$ and $n_l = 20$ (from top to bottom), for the 20/80 percent imbalance settings (COVID and non COVID classes). The yellow 'x' line corresponds to the SSDL with PBC ROC curve, the red dashed line to the SSDL with no imbalance correction, and the blue continuous curve to the supervised model ROC curve. As usual, the x-axis corresponds to the false positive ratio, and the y-axis to the true positive ratio.

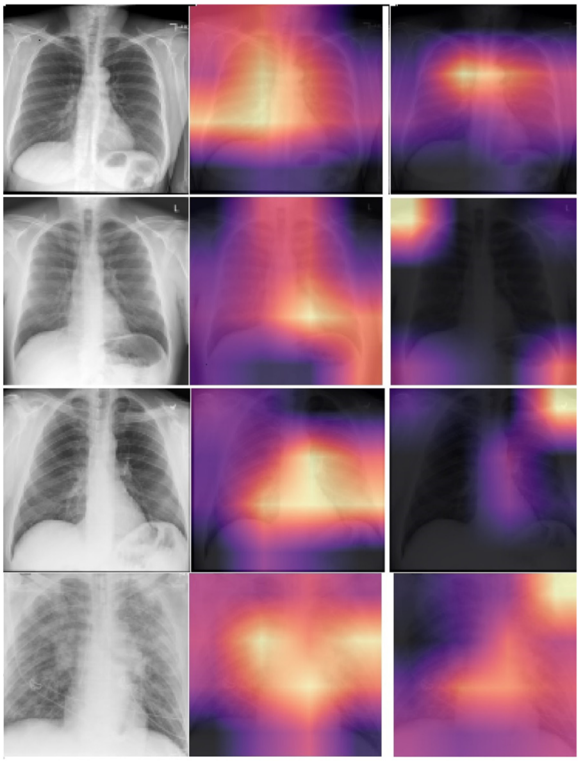

Fig. 4. From top to bottom: Two sample heatmaps for correct predictions using the Indiana dataset and two samples from the chest X-ray8 dataset, respectively. From left to right: the original image, the heatmap of the MixMatch trained model with the proposed PBC and the output of the supervised model.

of a pandemic, where few observations are available for the new pathology. To our knowledge, this is the first data imbalance analysis of a SSDL designed to perform COVID-19 detection using chest X-ray images, for both binary and multi-class classification. The experiment results suggest a strong impact of data imbalance in the overall MixMatch accuracy, since results in Table 6 reveal a stronger sensitivity of SSDL when compared to a supervised approach. The accuracy hit of training MixMatch with an imbalanced labelled dataset lies in the 2%–18% range, as seen in Tables 2, 3, 4, 5, 9, 8 and 7. Moreover, for the complex testbeds, mixing different data sources for a single class, for both binary and multi-class classification, the accuracy tends to be lower compared to the standard-sized datasets. This enforces the argument developed in [19,49] which draws the attention upon data distribution mismatch between the labelled and the unlabelled datasets, as a frequent real-world challenge when training a SSDL model.

Moreover, a simple and effective approach for correcting data imbalance by modifying Mix Match's loss function was proposed and tested in this work. The proposed method gives a smaller weight to the observations belonging to the under-represented class in the labelled dataset. Both the unlabelled and the labelled loss terms were re-weighted. This opposed to the unlabelled re-weighting developed for the mean teacher model in [51], which only modifies the weights of the unlabelled term. We implemented such approach since in our empirical tests the unlabelled term had less impact in the overall model accuracy. For the pseudo-labelled and MixUp augmented observations, we assigned the weights using the pseudo and augmented labels. The proposed method is computationally cheap, and avoids the need of complex and expensive generative approaches to correct data imbalance [47,48]. Our proposed method is simple and does not incur in an additional computational cost over the original Mix Match algorithm, as the weights are calculated once and assigned according to the pseudo-labels. A systematic accuracy gain is yielded when comparing the original MixMatch implementation with the proposed PBC for data imbalance correction, and also compared against data over-sampling. For the tested datasets, often the proposed PBC leads to significant accuracy gains from the supervised model, as data imbalance can even hinder any accuracy gain of using MixMatch, as seen in Tables 2–5. The accuracy gain ranges between 3% and 18%, with statistical significance for most of the datasets tested. In most of the datasets, the accuracy gain is higher for the more challenging 80%/20% 10%/45%/45% imbalance settings. Nevertheless, even in the more challenging extended-sized datasets with much larger test datasets than training and labelled datasets, with different data sources for the observations in the same class, a systematic accuracy gain was yielded using the proposed PBC method. The improvement of the ROC curves is usually achieved by class imbalance correction techniques commonly implemented for supervised methods [62]. Among the tested datasets, we included a new one with digital X-rays from healthy Costa Rican patients, which we will make available for the community. In our work, we have shown how the usage of pseudo labels for selecting the label imbalance correction weights is able to yield positive results also for the ROC curves, confirming a similar behaviour as seen previously for supervised models, as the minority class is better predicted.

As stated in [24], using the target dataset is vital for training a model, as using a different source dataset from other hospitals/clinics to train the model might yield poor test performance in the target dataset. Such distribution mismatch among different data sources is a frequent short-coming of deep learning solutions in the context of medical imaging. This is caused by data often presenting high heterogeneity due to patient diversity and different imaging protocols implemented [24]. Frequent low robustness distribution mismatch in deep learning systems raises the urgent need of training data from the specific clinic/hospital where the model is intended to be used. The challenge of labelling data becomes harder in the context of the pandemic, where a limited number of available high-quality labelled observations is usually available. Training a model with few labelled observations and an unlabelled dataset gathered from the target clinic/hospital, along with transfer learning and data augmentation as done in this work, might prove to be a practical solution in the context of a pandemic, where scarce labelled data is available. Moreover, we plan to test in the future the interaction between transfer learning from a source dataset with SSDL.

This work can be extended by using the customized feature extractors proposed in [36], as our architecture uses the more common transfer learning approach from a generic dataset (Imagenet), to later refine the feature extractor. The semantic relevance of the extracted features can be improved along with the model explainability, as seen in Fig. 4. Hence, the proposed solution in this work can be ported to use a more specific feature extractor. Therefore, we plan to test its usage under different customized feature extractors. Furthermore, it is interesting to investigate the impact of SSDL on deep learning explainability/uncertainty measures. We suspect that unlabelled data can improve models' uncertainty estimations and explainability accuracy.

CRediT authorship contribution statement

Saul Calderon-Ramirez: Software, Investigation, Data Curation, Writing - original draft. **Shengxiang Yang:** Conceptualization, Methodology, Formal analysis, Supervision. **Armaghan Moemeni:** Investigation, Writing - original draft, Validation. **David Elizondo:** Conceptualization, Writing - review & editing. **Simon Colreavy-Donnelly:** Investigation, Writing - original draft, Validation. **Luis Fernando Chavarría-Estrada:** Resources. **Miguel A. Molina-Cabello:** Investigation, Writing - original draft, Validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

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, Spain.

References

[1] J.F. Ludvigsson, Systematic review of COVID-19 in children shows milder cases and a better prognosis than adults, Acta Paediatr. 109 (6) (2020) 1088–1095.

[2] A. Bermudez, S. Calderon-Ramirez, T. Thang, P. Tyrrell, A. Moemeni, S. Yang, J. Torrents-Barrena, A first glance to the quality assessment of dental photostimulable phosphor plates with deep learning, in: 2020 International Joint Conference on Neural Networks (IJCNN), IEEE, 2020, pp. 1–6.

[3] I. Calvo, S. Calderon-Ramirez, J. Torrents-Barrena, E. Muñoz, D. 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, Springer, 2019, pp. 262–275.

[4] S. Calderon-Ramirez, F. Fallas, M. Zumbado, P. Tyrrell, H. Stark, Z. Emersic, B. Meden, M. Solis, Assessing the impact of the deceived non local means filter as a preprocessing stage in a convolutional neural network based approach for age estimation using digital hand x-ray images, in: 2018 25th IEEE International Conference on Image Processing (ICIP), IEEE, 2018, pp. 1752–1756.

[5] L. Oala, J. Fehr, L. Gilli, P. Balachandran, A.W. Leite, S. Calderon-Ramirez, D.X. Li, G. Nobis, E.A.M. Alvarado, G. Jaramillo-Gutierrez, et al., ML4h auditing: From paper to practice, in: Machine Learning for Health, PMLR, 2020, pp. 280–317.

[6] E. Alfaro, X.B. Fonseca, E.M. Albornoz, C.E. Martínez, S. Calderon-Ramirez, A brief analysis of U-net and mask R-CNN for skin lesion segmentation, in: 2019 IEEE International Work Conference on Bioinspired Intelligence (IWOBI), IEEE, 2019, pp. 000123–000126.

[7] M. Mendez, S. Calderon-Ramirez, P.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, Springer, 2019, pp. 307–319.

[8] I. Balki, A. Amirabadi, J. Levman, A.L. Martel, Z. Emersic, B. Meden, A. Garcia-Pedrero, S.C. Ramirez, D. Kong, A.R. Moody, et al., Sample-size determination methodologies for machine learning in medical imaging research: a systematic review, Canad. Assoc. Radiol. J. 70 (4) (2019) 344–353.

[9] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, C.A. Raffel, Mixmatch: A holistic approach to semi-supervised learning, in: Advances in Neural Information Processing Systems, 2019, pp. 5050–5060.

[10] J.F.-W. Chan, C.C.-Y. Yip, K.K.-W. To, T.H.-C. Tang, S.C.-Y. Wong, K.-H. Leung, A.Y.-F. Fung, A.C.-K. Ng, Z. Zou, H.-W. 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, J. Clin. Microbiol. (2020).

[11] Advice on the use of point-of-care immunodiagnostic tests for COVID-19. https://www.who.int/news-room/commentaries/detail/advice-on-the-use-of-point-of-care-immunodiagnostic-tests-for-covid-19, 0000.

[12] S.K. Vashist, In vitro diagnostic assays for covid-19: recent advances and emerging trends, 2020.

[13] K. Narayanan, I. Frost, A. Heidarzadeh, K.K. Tseng, S. Banerjee, J. John, R. 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).

[14] M. Chung, A. Bernheim, X. Mei, N. Zhang, M. Huang, X. Zeng, J. Cui, W. Xu, Y. Yang, Z.A. Fayad, et al., CT Imaging features of 2019 novel coronavirus (2019-nCoV), Radiology 295 (1) (2020) 202–207.

[15] Y. Fang, H. Zhang, J. Xie, M. Lin, L. Ying, P. Pang, W. Ji, Sensitivity of chest CT for COVID-19: comparison to RT-PCR, Radiology (2020) 200432.

[16] R.M.e.a. Andrea Borghesi, Radiographic severity index in COVID-19 pneumonia: relationship to age and sex in 783 Italian patients, World J. Surg. (2020).

[17] N. Chen, M. Zhou, X. Dong, J. Qu, F. Gong, Y. Han, Y. Qiu, J. Wang, Y. Liu, Y. Wei, et al., Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in wuhan, China: a descriptive study, Lancet 395 (10223) (2020) 507–513.

[18] F. Song, N. Shi, F. Shan, Z. Zhang, J. Shen, H. Lu, Y. Ling, Y. Jiang, Y. Shi, Emerging 2019 novel coronavirus (2019-nCoV) pneumonia, Radiology 295 (1) (2020) 210–217.

[19] A. Oliver, A. Odena, C.A. Raffel, E.D. Cubuk, I. Goodfellow, Realistic evaluation of deep semi-supervised learning algorithms, in: Advances in Neural Information Processing Systems, 2018, pp. 3235–3246.

[20] R. Arora, The training and practice of radiology in India: current trends., Quant. Imaging Med. Surg. 4 (6) (2014) 449–44950, http://dx.doi.org/10.3978/j.issn.2223-4292.2014.11.04.

[21] M.T. Shah, M. Joshipura, J. Singleton, P. LaBarre, H. Desai, E. Sharma, C. Mock, Assessment of the availability of technology for trauma care in India, World J. Surg. 39 (2) (2015) 363–372.

[22] Y. Oh, S. Park, J.C. Ye, Deep learning covid-19 features on cxr using limited training data sets, IEEE Trans. Med. Imaging 39 (8) (2020) 2688–2700.

[23] S. Niu, M. Liu, Y. Liu, J. Wang, H. Song, Distant domain transfer learning for medical imaging, IEEE J. Biomed. Health Inf. (2021).

[24] A. Choudhary, L. Tong, Y. Zhu, M.D. Wang, Advancing medical imaging informatics by deep learning-based domain adaptation, Yearb. Med. Inform. 29 (1) (2020) 129.

[25] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J.N. Chiang, Z. Wu, X. Ding, Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation, Med. Image Anal. 63 (2020) 101693.

[26] P.K. Sethy, S.K. Behera, Detection of coronavirus disease (COVID-19) based on deep features, 2020, Preprints.

[27] J.P. Cohen, P. Morrison, L. Dao, COVID-19 image data collection, 2020, arXiv:2003.11597 Available at https://github.com/ieee8023/covid-chestxray-dataset URL https://github.com/ieee8023/covid-chestxray-dataset.

[28] D.S. Kermany, M. Goldbaum, W. Cai, C.C. Valentim, H. Liang, S.L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, et al., Identifying medical diagnoses and treatable diseases by image-based deep learning, Cell 172 (5) (2018) 1122–1131.

[29] A. Alqudah, S. Qazan, H. Alquran, I. Abuqasmieh, A. Alqudah, Covid-2019 detection using X-Ray images and artificial intelligence hybrid systems, 2020.

[30] I.D. Apostolopoulos, T. Bessiana, Covid-19: Automatic detection from X-Ray images utilizing transfer learning with convolutional neural networks, 2020, http://dx.doi.org/10.1007/s13246-020-00865-4, URL http://arXiv.org/abs/2003.11617 arXiv:2003.11617.

[31] M.E.H. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M.A. Kadir, Z.B. Mahbub, K.R. Islam, M.S. Khan, A. Iqbal, N. Al-Emadi, M.B.I. Reaz, Can AI help in screening viral and COVID-19 pneumonia?, 2020, arXiv:2003.13145 URL http://arXiv.org/abs/2003.13145.

[32] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R.M. Summers, Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2097–2106.

[33] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, H. Müller, Causability and explainability of artificial intelligence in medicine, Wiley Interdiscip. Rev.: Data Min. Knowl. Discov. 9 (4) (2019) e1312.

[34] B. Ghoshal, A. Tucker, Estimating uncertainty and interpretability in deep learning for coronavirus (COVID-19) detection, 2020, arXiv:2003.10769 URL http://arXiv.org/abs/2003.10769.

[35] S. Khobahi, C. Agarwal, M. Soltanalian, Coronet: A deep network architecture for semi-supervised task-based identification of COVID-19 from chest X-ray images, MedRxiv (2020).

[36] J.P. Cohen, L. Dao, P. Morrison, K. Roth, Y. Bengio, B. Shen, A. Abbasi, M. Hoshmand-Kochi, M. Ghassemi, H. Li, et al., Predicting covid-19 pneumonia severity on chest x-ray with deep learning, 2020, arXiv preprint arXiv:2005.11856.

[37] D. Demner-Fushman, M.D. Kohli, M.B. Rosenman, S.E. Shooshan, L. Rodriguez, S. Antani, G.R. Thoma, C.J. McDonald, Preparing a collection of radiology examinations for distribution and retrieval, J. Amer. Med. Inform. Assoc. 23 (2) (2016) 304–310.

[38] A. Bustos, A. Pertusa, J.-M. Salinas, M. de la Iglesia-Vayá, Padchest: A large chest x-ray image dataset with multi-label annotated reports, 2019, arXiv preprint arXiv:1901.07441.

[39] A.E. Johnson, T.J. Pollard, S. Berkowitz, N.R. Greenbaum, M.P. Lungren, C.-y. Deng, R.G. Mark, S. Horng, MIMIC-CXR: A large publicly available database of labeled chest radiographs, 2019, arXiv preprint arXiv:1901.07042.

[40] A. Majkowska, S. Mittal, D.F. Steiner, J.J. Reicher, S.M. McKinney, G.E. Duggan, K. Eswaran, P.-H. Cameron Chen, Y. Liu, S.R. Kalidindi, et al., Chest radiograph interpretation with deep learning models: assessment with radiologist-adjudicated reference standards and population-adjusted evaluation, Radiology 294 (2) (2020) 421–431.

[41] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, et al., Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 590–597.

[42] M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, A.I. Aviles-Rivero, C. Etmann, C. McCague, L. Beer, et al., Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans, Nat. Mach. Intell. 3 (3) (2021) 199–217.

[43] S. Calderon-Ramirez, R. Giri, S. Yang, A. Moemeni, M. Umana, D. Elizondo, J. Torrents-Barrena, M.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), IEEE, 2021, pp. 5294–5301.

[44] S. Calderon-Ramırez, D. Murillo-Hernandez, K. Rojas-Salazar, L.-A. Calvo-Valverde, S. Yang, A. Moemeni, D. Elizondo, E. Lopez-Rubio, M. Molina-Cabello, Improving Uncertainty Estimations for Mammogram Classification Using Semi-Supervised Learning, Institute of Electrical and Electronics Engineers, 2021.

[45] J.E. Van Engelen, H.H. Hoos, A survey on semi-supervised learning, Mach. Learn. 109 (2) (2020) 373–440.

[46] S. Ando, C.Y. Huang, Deep over-sampling framework for classifying imbalanced data, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2017, pp. 770–785.

[47] A. Taherkhani, G. Cosma, T. McGinnity, Adaboost-CNN: an adaptive boosting algorithm for convolutional neural networks to classify multi-class imbalanced datasets using transfer learning, Neurocomputing (2020).

[48] S.S. Mullick, S. Datta, S. Das, Generative adversarial minority oversampling, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 1695–1704.

[49] S. Calderon-Ramirez, L. Oala, J. Torrents-Barrena, S. Yang, A. Moemeni, W. Samek, M.A. Molina-Cabello, Mixmood: A systematic approach to class distribution mismatch in semi-supervised learning using deep dataset dissimilarity measures, 2020, arXiv preprint arXiv:2006.07767.

[50] S. Calderon-Ramirez, L. Oala, More than meets the eye: Semi-supervised learning under non-IID data, 2021, arXiv preprint arXiv:2104.10223.

[51] M. Hyun, J. Jeong, N. Kwak, Class-imbalanced semi-supervised learning, 2020, arXiv preprint arXiv:2002.06815.

[52] A. Tarvainen, H. Valpola, Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, in: Advances in Neural Information Processing Systems, 2017, pp. 1195–1204.

[53] H. Zhang, M. Cisse, Y.N. Dauphin, D. Lopez-Paz, Mixup: Beyond empirical risk minimization, 2017, arXiv preprint arXiv:1710.09412.

[54] M.d.l.I. Vayá, J.M. Saborit, J.A. Montell, A. Pertusa, A. Bustos, M. Cazorla, J. Galant, X. Barber, D. Orozco-Beltrán, F. Garcia, et al., BIMCV Covid-19+: a large annotated dataset of RX and CT images from COVID-19 patients, 2020, arXiv preprint arXiv:2006.01174.

[55] A. Narin, C. Kaya, Z. Pamuk, Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks, 2020, arXiv preprint arXiv:2003.10849.

[56] L. Wang, A. Wong, COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-Ray images, 2020, arXiv preprint arXiv:2003.09871.

[57] E.E.-D. Hemdan, M.A. Shouman, M.E. Karar, Covidx-net: A framework of deep learning classifiers to diagnose covid-19 in x-ray images, 2020, arXiv preprint arXiv:2003.11055.

[58] F.M. Salman, S.S. Abu-Naser, E. Alajrami, B.S. Abu-Nasser, B.A.M. Ashqar, COVID-19 Detection using Artificial Intelligence, Tech. Rep., 2020, URL www.ijeais.org/ijaer.

[59] S. Zagoruyko, N. Komodakis, Wide residual networks, 2016, arXiv preprint arXiv:1605.07146.

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

[61] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.

[62] L. Lusa, et al., Class prediction for high-dimensional class-imbalanced data, BMC Bioinformatics 11 (1) (2010) 1–17.