IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier

Improving Uncertainty Estimation With Semi-supervised Deep Learning for COVID-19 Detection Using Chest X-ray Images

SAUL CALDERON-RAMIREZ[1,2], SHENGXIANG YANG (SENIOR MEMBER, IEEE)[1], ARMAGHAN MOEMENI[3], SIMON COLREAVY-DONNELLY[1], DAVID A. ELIZONDO[1] (Senior Member, IEEE), LUIS OALA[4], JORGE RODRÍGUEZ-CAPITÁN[5,7], MANUEL JIMÉNEZ-NAVARRO[5,7], EZEQUIEL LÓPEZ-RUBIO[6,7], and MIGUEL A. MOLINA-CABELLO[6,7]

[1]School of Computer Science and Informatics, De Montfort University, The Gateway, Leicester LE1 9BH, United Kingdom
[2]Instituto Tecnologico de Costa Rica, Costa Rica
[3]School of Computer Science, University of Nottingham, Nottingham NG8 1BB, United Kingdom
[4]XAI Group, AI Department, Fraunhofer Heinrich Hertz Institute, Germany
[5]Hospital Universitario Virgen de la Victoria CIBERCV, Málaga, Spain
[6]Department of Computer Languages and Computer Science. University of Málaga, Bulevar Louis Pasteur, 35, 29071, Málaga, Spain
[7]Instituto de Investigación Biomédica de Málaga - IBIMA, C/ Doctor Miguel Díaz Recio, 28, 29010, Málaga, Spain

Corresponding author: Saul Calderon-Ramirez (e-mail: sacalderon@itcr.ac.cr).

ABSTRACT In this work we implement a COVID-19 infection detection system based on chest X-ray images with uncertainty estimation. Uncertainty estimation is vital for safe usage of computer aided diagnosis tools in medical applications. Model estimations with high uncertainty should be carefully analyzed by a trained radiologist. We aim to improve uncertainty estimations using unlabelled data through the MixMatch semi-supervised framework. We test popular uncertainty estimation approaches, comprising Softmax scores, Monte-Carlo dropout and deterministic uncertainty quantification. To compare the reliability of the uncertainty estimates, we propose the usage of the Jensen-Shannon distance between the uncertainty distributions of correct and incorrect estimations. This metric is statistically relevant, unlike most previously used metrics, which often ignore the distribution of the uncertainty estimations. Our test results show a significant improvement in uncertainty estimates when using unlabelled data. The best results are obtained with the use of the Monte Carlo dropout method.

INDEX TERMS Uncertainty estimation, Coronavirus, Covid-19, Chest X-Ray, Computer Aided Diagnosis, Semi-Supervised Deep Learning, MixMatch.

I. INTRODUCTION

The COVID-19 pandemic is putting significant pressure on governmental health systems, as the number of cases grows exponentially [1]. Furthermore, the availability of medical staff is lowered as they also get infected by the virus, reducing the overall capacity of hospitals and clinics [1]. The accurate and widespread detection of infected subjects is of great importance to control the growth of the disease [2]. The usage of medical imaging can be an alternative tool when other methods like Real-time Reverse Transcription Polymerase Chain Reaction (RT-PCR) testing become more expensive as less resources are available to supply the growing demand [3]. The usage of computed tomography and X-ray based tests for COVID-19 detection has been studied in [4]–[6], reporting mixed sensitivity and accuracy in the case of X-ray imaging based solutions. However, the usage of X-ray images is ubiquitous, as this technology is usually cheaper and more widely available [7].

X-ray chest imaging is in general more widely accessible when compared to computed tomography imaging [7]. Furthermore, the low availability of medical staff to sample and analyze the medical images can increase the costs of this alternative solution, especially in low resource environments [8]. For example, in India, with a population of around 1.44 billion, approximately one radiologist for every 100,000 people is currently available [8]. This increases the need of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3085418, IEEE Access

X-ray based COVID-19 computer aided diagnosis tools.

The application of deep learning based models to estimate the prevalence of COVID-19 from X-ray images has recently been explored, with different deep learning architectures reporting high test accuracy [9], [10]. Given the lack of high quality labeled data, semi-supervised methods have also been implemented to perform COVID-19 detection, making use of cheaper unlabelled data to improve the model's accuracy [11], [12].

Along with high model accuracy, Artificial Intelligence (AI) based solutions should also provide explainable decisions to increase reliability, especially in the medical domain [13], [14]. Model uncertainty estimation is a common approach to increase model interpretability and safety in use [13], [15]. The estimation of model uncertainty allows the user to interpret how sure or confident is the model for a specific prediction. In the context of COVID-19 detection using X-ray images, an estimation with high uncertainty should justify further tests to be done in the subject. This enforces safety upon the usage of a computer aided diagnosis system, as low-confidence predictions are quantitatively estimated by the system itself.

In this work we focus in the measurement and improvement of uncertainty estimations for a deep learning model designed to identify COVID-19 infection using X-ray images. We aim to improve uncertainty estimations by using unlabelled data. Using unlabelled data is an useful approach when using datasets with a low number of high quality labelled data. This is a frequent setting during the onset of a pandemic. Moreover, for a statistically significant comparison of the tested uncertainty estimation methods, we propose a novel density function based divergence approach.

II. STATE OF THE ART

A. PREDICTIVE UNCERTAINTY ESTIMATION

Predictive uncertainty estimation (or simply referred as uncertainty estimation in this work) for machine learning models has been widely studied in the literature [16].In general, uncertainty sources can be categorized in aleatoric and epistemic. Aleatoric uncertainty refers to the uncertainty inherent in the measurements [17]. In conditional distribution terms, it refers to the distribution of the target variables with a given set of measured features. Aleatoric uncertainty cannot be reduced by taking a larger sample of features within the same distribution [17]. Epistemic uncertainty refers to the model's parameters uncertainty caused by the limited sample size used to build the model (or lack of knowledge of the feature space) [17]. Therefore, epistemic uncertainty can be diminished by sampling a larger dataset, specially collecting data in the sparser regions [17]. In the context of Semi-supervised Deep Learning (SSDL), epistemic uncertainty can be considered to be more important, as labeled data are usually very scarce when SSDL is used. Unlabelled data might lower epistemic uncertainty, usually less effectively as target information is missing [17].

In this work we analyze simple and straightforward uncertainty estimation methods. The tested methods were selected based on their post-hoc capacity, i.e. their ability to leave the original deep learning architecture intact and not require any re-training of the model.

The Softmax function, typically used as an activation function in the output layer of a neural network, is among the basic methods for uncertainty estimation. Assume a multi-class discrimination problem in $i = 1, ..., C$ classes [18]. Take the array of model outputs $\boldsymbol{y}_j = f_{\boldsymbol{\theta}}(\boldsymbol{x}_j)$ with network weights $\boldsymbol{\theta}$ for a given input $\boldsymbol{x}_j$. The Softmax function approximates a density function $\boldsymbol{p}$ as follows:

$$p_i = \frac{\exp\left(y_{i,j}\right)}{\sum_k \exp\left(y_{k,j}\right)} \tag{1}$$

Therefore, the output of the Softmax function for a specific output unit i can be interpreted as a proxy for model confidence for class i, given a specific input observation $\boldsymbol{x}_j$. Either the highest p_i for the estimated class or the entropy over $\boldsymbol{p}$ can be used for uncertainty estimation. However, authors in [19] highlight how neural networks are typically overconfident in their predictions, leading to poor uncertainty estimations.

To address this, authors in [20] propose to post-process the Softmax's confidence outputs, by implementing an additional temperature parameter T in the Softmax function:

$$p_i = \frac{\exp\left(y_{i,j}/T\right)}{\sum_k \exp\left(y_{k,j}/T\right)}. \tag{2}$$

To find the optimum T leading to better uncertainty estimates, the authors propose to minimize the negative log likelihood, encouraging the model to assign high confidence to correct classes only (ignoring incorrect classes). This means that an additional optimization step is needed.

Authors in [19] propose an alternative approach to avoid the Softmax based uncertainty estimates, known as Monte Carlo Dropout (MCD). In their method forward passes through M perturbed models $\boldsymbol{y}_{j,m} = f_{\boldsymbol{\theta'}_m}(\boldsymbol{x}_j)$ with perturbed weights $\boldsymbol{\theta'}_m$ are performed. This way, epistemic uncertainty is modeled with a distribution of the model's weights [21]. The approach estimates the dispersion $\sigma_{\text{model}}(\boldsymbol{x}_j)$ of M evaluations of the perturbed model, for the same input observation $\boldsymbol{x}_j$:

$$\sigma^2_{\text{model}}(\boldsymbol{x}_j) = \frac{1}{M}\sum_{m=1}^{M}\sum_{k=1}^{K}\left(y_{m,j,k} - \bar{y}_{j,k}\right)^2. \tag{3}$$

The calculation of the dispersion or the distribution of the outputs can be summed for all the output units $k = 1, ..., K$, or only the unit with the highest output can be taken into account.

Another recent method in [22] relying on the feature representations of the training data, was proposed for uncertainty estimation and Out of Distribution (OOD) data detection. This method is known as Deterministic Uncertainty Quantification (DUQ). For a set of feature centroids $E = \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_C\}$ calculated using the training data, uncertainty

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3085418, IEEE Access

is calculated using the distance from each centroid to the input observation $\boldsymbol{x}_j$, with a radial basis kernel K_i:

$$K_i\left(f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_j\right), \boldsymbol{e}_i\right) = \exp\left(-\frac{\left\|W_i f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_j\right) - \boldsymbol{e}_i\right\|_2^2}{2\sigma^2}\right) \quad (4)$$

where W_i stands for a weight matrix, tuned to encourage feature insensitivity per class, thereby minimizing feature collapse [22]. The uncertainty is then estimated as the maximum class centroid distance in the feature space:

$$\arg\max_i K_i\left(f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_j\right), \boldsymbol{e}_i\right). \quad (5)$$

The authors of the DUQ method claim that their approach measures both the epistemic and aleatoric uncertainty. Epistemic uncertainty is modeled through the construction of the feature centroids and the kernel K_i, which can improve as more data is available. The measurement of the centroids also includes aleatoric uncertainty [22].

Other popular uncertainty approaches include deep ensembles [23] and interval networks [24]. These methods require additional training steps, increasing complexity, and are often impractical when no access to the original training data set is possible.

B. UNCERTAINTY ESTIMATION FOR RELIABLE MEDICAL IMAGING ANALYSIS AND COVID-19 DETECTION

Uncertainty estimation has been implemented in the literature to increase the reliability of medical imaging analysis systems. For example, in [25] uncertainty estimation is implemented for a diabetic retinopathy diagnosing system. A MCD based approach for uncertainty estimation was implemented. The system was evaluated using rejection plots, which calculate the average accuracy for the data rejected by using different uncertainty thresholds. Furthermore, the reliability was evaluated by measuring the impact of referring samples to further manual inspection during clinical usage.

In [26], a Bayesian deep learning approach was implemented to segment retinal optical coherence tomographies. The Bayesian model is able to estimate an uncertainty map, used to post-process the segmentation. Neither a comparison to other uncertainty methods nor the usage of uncertainty metrics was performed in the study.

As for COVID-19 detection, a system with uncertainty assessment was proposed in [27]. By providing practitioners with a confidence factor of the prediction, the overall reliability of the system is said to be improved. A high correlation between the prediction accuracy of the model and the level of uncertainty was reported in [27]. The data set used for positive COVID-19 cases uses the repository of [28], and normal X-ray readings were collected from [29].

Perhaps the most similar previous method to our proposed approach is the pre-published work of [30]. The authors write on the importance of measuring model uncertainty for COVID-19 detection from chest X-ray images. They compared three popular uncertainty estimation approaches, namely ensemble networks, Monte Carlo dropout and a combination of both approaches. An objective uncertainty estimation metric is also proposed, as the authors found a lack of metrics to compare uncertainty estimation methods. We agree on this gap in the literature, however we think that the metric should allow to compare not only different uncertainty estimation methods, but also several uncertainty estimations with different deep learning architectures, leading to different accuracy measurements, with statistical significance. [30] proposed a confusion matrix approach which does not hold statistical meaning by itself. Therefore, in our work, we propose an alternative metric to compare different uncertainty estimation methods and assess the impact of semi-supervised learning on uncertainty estimation.

C. SEMI-SUPERVISED LEARNING WITH MIXMATCH

In this work, we explore the recent and successful SSDL method referred to as MixMatch [31]. It creates a set of pseudo-labels, and also implements an unsupervised regularization term. The consistency loss implemented uses the pseudo-labels for the unlabelled dataset X_u to train the model. To calculate the pseudo-labels, the average model output of a perturbed input $\boldsymbol{x}_j$ is used:

$$\widehat{\boldsymbol{y}}_j = \frac{1}{K}\sum_{\eta=1}^{K} f_{\boldsymbol{w}}\left(\Psi^{\eta}\left(\boldsymbol{x}_j\right)\right). \quad (6)$$

Where K is the number of perturbations (like image flipping) Ψ^{η} done. A value of $K = 2$ is recommended by the authors. According to authors, the estimated pseudo-labels $\widehat{\boldsymbol{y}}_j$ might present high entropy, increasing low confidence estimations. To address this, the output array $\widehat{\boldsymbol{y}}$ is sharpened with a temperature coefficient ρ (with $\rho = 0.25$ recommended by the authors):

$$\widetilde{\boldsymbol{y}}_i = \frac{\widehat{y}_i^{1/\rho}}{\sum_j \widehat{y}_j^{1/\rho}}. \quad (7)$$

The set $\widetilde{S}_u = \left(X_u, \widetilde{Y}\right)$ corresponds to the data with the sharpened pseudo labels, where $\widetilde{Y} = \left\{\widetilde{\boldsymbol{y}}_1, \widetilde{\boldsymbol{y}}_2, \ldots, \widetilde{\boldsymbol{y}}_{n_u}\right\}$

Authors in [31] highlight how data augmentation is important to improve the SSDL performance. Therefore the authors proposed the MixUp approach [32], which consists on augmenting data using both labelled and unlabelled observations: $\left(S_l', \widetilde{S}_u'\right) = \Psi_{\text{MixUp}}\left(S_l, \widetilde{S}_u, \alpha\right)$, where $S_l = (X_l, Y_l)$ stands for the labelled data with a sample size of n_l. The MixUp algorithm generates new observations combining the unlabelled (with its pseudo labels) and labelled data through a linear interpolation. Specifically, for two labelled and/or pseudo labelled data pairs $(\boldsymbol{x}_a, y_a)$ and $(\boldsymbol{x}_b, y_b)$, the MixUp approach creates a new observation and its label $(\boldsymbol{x}' = \lambda' \boldsymbol{x}_a + (1-\lambda')\,\boldsymbol{x}_b, y' = \lambda' y_a + (1-\lambda')\,y_b)$ using a linear interpolation. The parameter α controls the Beta distribution where the MixUp coefficient is sampled from $\lambda \sim \text{Beta}\,(\alpha, \alpha)$. A value of $\alpha = 0.75$ is recommended

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3085418, IEEE Access

by the authors [31]. This results in the augmented data sets $\left(S_l', \widetilde{S}_u'\right)$, used by the MixMatch algorithm to train a model as specified in the training function T_{MixMatch}:

$$f_{\boldsymbol{\theta}} = T_{\text{MixMatch}}\left(S_l, X_u, \gamma\right) = \underset{\boldsymbol{w}}{\operatorname{argmin}} \mathcal{L}\left(S, \boldsymbol{w}\right) \tag{8}$$

$$\mathcal{L}\left(S, \boldsymbol{\theta}\right) = \sum_{\left(\boldsymbol{x}_i, \boldsymbol{y}_i\right) \in S_l'} \mathcal{L}_l\left(\boldsymbol{\theta}, \boldsymbol{x}_i, \boldsymbol{y}_i\right) + \gamma r(\tau) \sum_{\left(\boldsymbol{x}_j, \widetilde{\boldsymbol{y}}_j\right) \in \widetilde{S}_u'} \mathcal{L}_u\left(\boldsymbol{\theta}, \boldsymbol{x}_j, \widetilde{\boldsymbol{y}}_j\right) \tag{9}$$

In [31] the supervised loss term was implemented with a cross-entropy loss; $\mathcal{L}_l\left(\boldsymbol{w}, \boldsymbol{x}_i, \boldsymbol{y}_i\right) = \delta_{\text{cross-entropy}}\left(\boldsymbol{y}_i, f_{\boldsymbol{w}}\left(\boldsymbol{x}_i\right)\right)$. Regarding the unlabelled loss term, an Euclidean distance was implemented $\mathcal{L}_u\left(\boldsymbol{w}, \boldsymbol{x}_j, \widetilde{\boldsymbol{y}}_j\right) = \left\|\widetilde{\boldsymbol{y}}_j - f_{\boldsymbol{w}}\left(\boldsymbol{x}_j\right)\right\|$. Authors in [31], modelled the coefficient $r(\tau)$ as a ramp-up function that increases its value as the epochs τ increase. In our implementation, $r(\tau)$ was set to $\tau/3000$. The γ factor is used as a regularization coefficient. It regulates the influence of unlabelled data. It is important to remark how unlabelled data also affects the *labelled* data term $\mathcal{L}_l$, as unlabelled data is used to augment data observations by using the MixUp approach for the labelled term as well.

D. SEMI AND SELF SUPERVISED LEARNING FOR IMPROVING UNCERTAINTY ESTIMATION

Recently, in [33] the authors analyze the use of unlabelled data to improve a model's calibration (defined by the authors as the correlation between accuracy and uncertainty). A regularization based approach was implemented, improving the calibration of the model for structured data. Moreover, in [34], authors explore the improvement of uncertainty estimations using self-supervised learning. Some popular semi-supervised approaches like MixMatch [31] use concepts implemented in self-supervised learning, namely consistency regularization. The results presented in [33] reveal the advantage of using unlabelled data for uncertainty estimation. Semi-supervised learning has recently been proven to enhance adversarial robustness, as argued in [35]. Moreover, in [36], the impact of MixUp data augmentation on the model uncertainty estimation (also known as model calibration) is assessed. Authors used the Softmax function to estimate the model's uncertainty, yielding better calibrations through the usage of MixUp. MixUp is also used in the MixMatch model [31].

E. COMPARING MODEL UNCERTAINTY RELIABILITY

To compare uncertainty reliability across different uncertainty estimation techniques, different approaches have been developed in the literature. Uncertainty reliability is related to the calibration error [37]. For a classification problem in a given data set D, intuitively, the calibration error refers to the difference between the total estimated probability (confidence) $\hat{p}$ for the observations of label y and the real proportion of the estimation of a label y, given in p.

Reliability histograms [38] are proposed to build a histogram, with bins defined for different uncertainty ranges. A reliability histogram plots the normalized confidence against the accuracy for each bin. Defining B_m as the set of indices of observations whose uncertainty prediction belongs to the interval $I_m = \left(\frac{m-1}{M}, \frac{m}{M}\right]$, the sample mean accuracy for the bin B_m is given by:

$$\overline{\text{acc}}\left(B_m\right) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbf{1}\left(\hat{y}_i = y_i\right), \tag{10}$$

where $\hat{y}_i$ corresponds to the model estimation for the observation i with label y_i. Similarly the average uncertainty for a bin B_m for an uncertainty density function $\hat{p}$ is given by:

$$\overline{\text{unc}}\left(B_m\right) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i. \tag{11}$$

An uncertainty estimator is considered better as the relationship of $\overline{\text{unc}}$ and $\overline{\text{acc}}$ reaches the identity and thus becomes less spiky. The Expected Calibration Error (ECE) measures this gap in one scalar, taking the average difference between the sample accuracy and confidence mean:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} |\overline{\text{acc}}(B_m) - \overline{\text{conf}}(B_m)|. \tag{12}$$

In [37] different downsides of the ECE are noted. One such downside is the sparseness that is frequently yielded by the computed confidence histogram. This is referred in [37] as the problem of fixed calibration ranges. Frequently used Softmax based uncertainty estimations are overconfident, making higher bins more populated. This makes the estimates of less populated bins potentially inaccurate. Other improvements added to the ECE include the root mean squared calibration error [39] and the static and adaptive calibration error [37].

However, an important downside of using the ECE is the assumption that it makes about the uncertainty measurement as a normalized measure between 0 and 1. Different approaches for uncertainty estimation as MCD and DUQ yield unbounded values (outside from the 0 to 1 interval), making the use of the ECE inappropriate. For instance in [26], MCD has been implemented for uncertainty estimation, with no normalized values reported. For instance, comparing uncertainty estimations of [25], [26] to the ones yielded in [40], is difficult as different uncertainty measures yield different uncertainty value ranges for different data sets. Using the ECE is only possible when a bounded uncertainty estimator such as the Softmax function is used (where its values are bounded from 0 to 1). This makes the comparison of uncertainty estimation approaches difficult as they can be normalized using the sampled values for the data set tested, but this leads to a data set bias.

However, a bigger issue when using a measure like the ECE is the limited statistical interpretation. The ECE relies on the sample mean per bin $\overline{\text{acc}}$, which ignores the distribution of the data and information from other statistical measurements like the variance.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3085418, IEEE Access

As an alternative, rejection-classification plots were used in [22]. Rejection-classification plots use as x-axis the proportion of data rejected based on the uncertainty score. The y-axis represents the level of accuracy. Similar to the rejection-classification plots, the accuracy vs. confidence curves were used in [23] to compare different uncertainty estimators graphically. For a quantitative comparison, the area under the curve of this plot can be used. However, such value is also unbounded and holds no statistical significance. For either the rejection plots or the ECE based metrics, a comparison problem arises when the compared curves present different accuracy levels. As the number of wrong estimations fluctuates for each model, the average accuracy per bin also changes, making it harder to compare the uncertainty estimation quality. This situation is faced in this work, where we compare the impact of a supervised to a semi-supervised model, which changes the model's accuracy.

Other common metrics to measure the error of a model have also been used for out of distribution data detection through uncertainty methods. In [41] for instance, the area under the precision-recall curve and the error rate have both been used for out of distribution detection to compare uncertainty estimation methods. However, the metric is also not statistically relevant as no distribution information is used to compare the evaluated methods. Using the outlined context, this work comprises the following contributions:

- We explore the impact of semi-supervised deep learning in the reliability of the uncertainty estimations for COVID-19 detection, using a common deep learning architecture.
- We evaluate and compare qualitatively as well as quantitatively the performance of three different uncertainty estimation techniques for both the supervised and semi-supervised models.
- We propose the use of the Jensen-Shannon divergence [42] as a probability density based metric to compare the performance of uncertainty estimation techniques.

We show that our proposed method is simple to implement and that it is often effective. The method takes advantage of unlabelled data to improve uncertainty estimations for COVID-19 detection using digital chest X-ray images. Unlabelled data is generally widely available, and in the context of a virus out-break, easier to obtain, when compared to labelled data.

III. PROPOSED METHOD

In this work, we propose the use of unlabelled data through MixMatch (as depicted in equations 8 and 9), to improve uncertainty estimation. We test the impact of using unlabelled data in three uncertainty estimation methods:

- Softmax as described in Equation 1, using the maximum Softmax value for the output layer. Therefore, the Softmax uncertainty estimation corresponds to $u_i = \arg\max_i p_i$.
- MCD as depicted in Equation 3, using the standard deviation of the distribution from the evaluation of the model with dropout for the same input observation $\boldsymbol{x}_j$ [19], making $u_i = \sigma_{\text{model}}(\boldsymbol{x}_j)$.
- DUQ as introduced in Equation 4. We used a generic weight matrix $W_i = 1$ for all classes $i = 1, ..., C$, implementing an Euclidean distance for the radial basis kernel K_i. The uncertainty estimation for this approach is implemented as

$$u_i = \arg\max_i K_i\left(f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_j\right), \boldsymbol{e}_i\right), \tag{13}$$

for an input observation $\boldsymbol{x}_j$.

In this work we also propose the comparison of the evaluated methods for uncertainty estimation, using the Jensen-Shannon divergence between the distribution of the uncertainty estimations for the correct and incorrect estimations. More specifically, take a model uncertainty estimation u_j for an input observation $\boldsymbol{x}_j$. For a given data set S, we group the uncertainties of the wrong estimations for the trained model, semi or supervised, as wrong or correct according to the labels in the test partition of the labelled dataset $S_l = (X_l, Y_l)$. This results in a set of uncertainties for the wrong estimations $U_{\text{wrong}} = \{u_1, \ldots, u_{n_{\text{wrong}}}\}$ and correct estimations $U_{\text{correct}} = \{u_1, \ldots, u_{n_{\text{incorrect}}}\}$, used to calculate the corresponding normalized histograms $\boldsymbol{p}_{u_{\text{correct}}}$ and $\boldsymbol{p}_{u_{\text{incorrect}}}$. We implement the Jensen-Shannon divergence $D_{\text{JS}}\left(\boldsymbol{p}_{u_{\text{correct}}}, \boldsymbol{p}_{u_{\text{incorrect}}}\right)$ to measure the divergence between the two non-parametric approximations of the density functions $\boldsymbol{p}_{u_{\text{correct}}}$ and $\boldsymbol{p}_{u_{\text{incorrect}}}$. Figure 1 summarizes the implemented pipeline in this work.

IV. EXPERIMENTS

A. DATASET

The COVID-19$^+$ data sample was downloaded from the publicly available github repository of Joseph Cohen [28]. The observations were gathered from journals such as radiopaedia.org and the Italian Society of Medical and Interventional Radiology. In this work we used only images labelled with COVID-19$^+$, discarding images labelled as Middle East Respiratory Syndrome (MERS), Acute Respiratory Distress Syndrome (ARDS) and Severe Acute Respiratory Syndrome (SARS). After applying this filtering, 99 observations of front chest X-rays were selected. The images were stored with resolutions ranging from 400×400 up to 2500×2500 pixels.

Together with the COVID-19$^-$ observations we sampled a 5856 observations containing pneumonia and no lung pathologys as defined by [29]. The data set is composed of 4273 observations of viral and bacterial pneumonia and 1583 normal observations (with no lung pathology). We used the observations with no findings, for the COVID-19$^-$ class. The negative COVID-19 cases gathered in this dataset have been used in recent research related to COVID-19 detection using deep learning [43]–[45]. The images were stored with a resolution of 1300×600 pixels.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3085418, IEEE Access

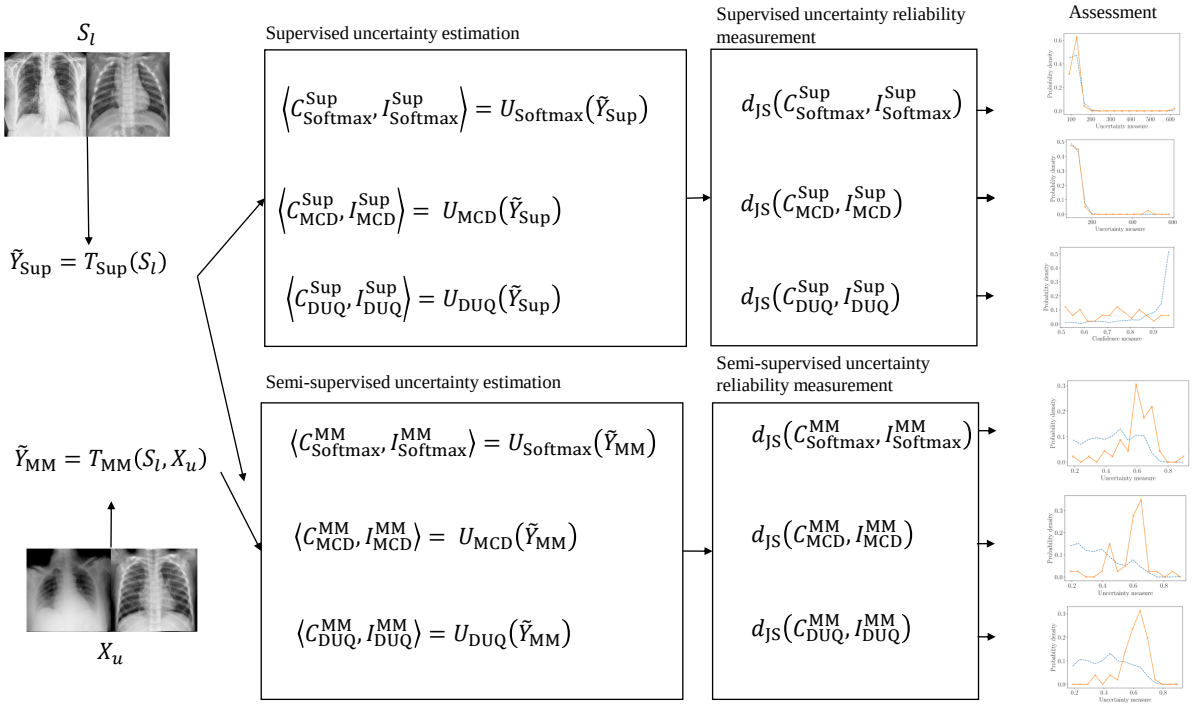


FIGURE 1. Description of the implemented work-flow: Training of the semi-supervised model MixMatch (MM) and the supervised model (Sup.). Calculation of the predictive uncertainties using the Softmax activation function, MonteCarlo Dropout (MCD) and Deterministic Uncertainty Quantification (DUQ). We propose to compare the distribution of the predictive uncertainties for correct (C) and Incorrect (I) estimations, using the Jensen-Shannon distance (JS).

We created a balanced base-line data set of 99 COVID-19^+ observations and also 99 observations for COVID-19^- cases, using the aforementioned data sets. Figure 2 shows a sample of the images used.

Both supervised and semi-supervised models were trained with $n_l = 20, 30, 60, 70, 100$ labelled observations, to study the impact of different labelled data sample sizes. We splitted the data set of 198 observations with 70% (138 observations) of the data for training and the remaining 30% (60 observations) for testing. The labelled observations were taken from the training dataset, and as for the SSDL model, we used the remaining as unlabelled data, always keeping the number of labels balanced. We chose to use the unlabelled data as a partition of the original labelled dataset, to avoid distribution mismatch related issues as suggested in [46]. This is out of the scope in this work, however testing unlabelled datasets from other sources with possibly more observations, is left for future work.

B. NEURAL NETWORK ARCHITECTURES AND METRICS

In this work we used a WideResNet model as a supervised model for binary classification (COVID-19^+ and COVID-19^- discrimination), with transfer learning from the ImageNet dataset. For the supervised model we used the cross entropy as loss function. The semi-supervised MixMatch framework implemented also used the WideResNet model with a $K = 2$ transformations, a sharpening coefficient $\rho =$

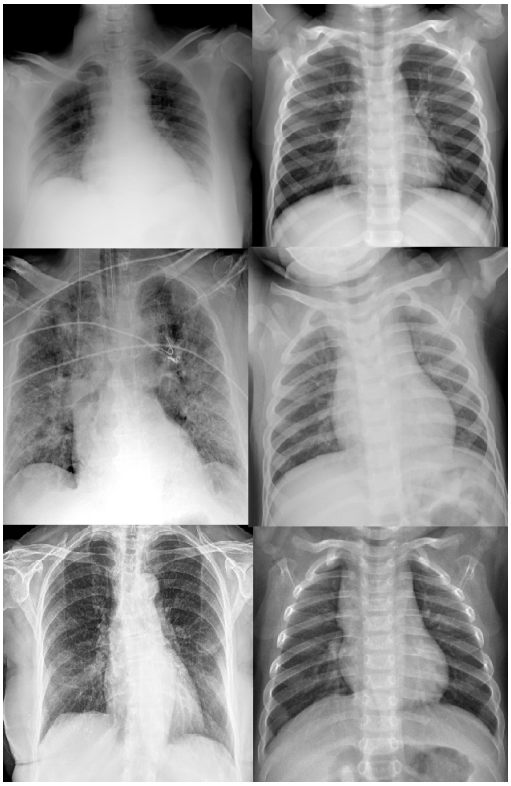


FIGURE 2. Left column, positive COVID-19 X-ray observations, right column, three negative COVID-19 observations. All of them were taken from the dataset used in this work.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3085418, IEEE Access

0.25, a MixUp parameter $\alpha = 0.75$, as recommended in [31], and a $\gamma = 200$ for the unsupervised coefficient, as advised in [11]. For both the supervised and semi-supervised model we used a learning rate of 0.00002 and a batch size of 10 observations, with 50 epochs per run. As a preprocessing stage, we implemented a standardization of the training dataset. All images were resized to 150×150 pixels. The model was implemented with the FastAI library, and optimized with the 1-cycle policy [47].

We evaluated the Softmax, MCD and DUQ uncertainty methods in the semi and supervised models to collect for each one of them a set of uncertainties U_{S}, U_{MCD}, and U_{DUQ}, respectively. As for the parameters of the tested uncertainty methods, for the MCD, we used $M = 100$ evaluations with the default dropout of WideResNet. Regarding the DUQ method, we used an Euclidian based kernel K for all the classes.

We first report the model's F1 score, to compare the accuracy gained when using SSDL, and use it as a reference for the uncertainty results analysis. This is depicted in Table 1. We also report the ρ_{lu} and δ_ρ (this last one for the SSDL model), as advised in [11] for assessing the accuracy gain for SSDL frameworks.

Secondly, we report the sample mean and standard deviation for the correct and incorrect estimations. We perform this comparison for all three tested uncertainty estimation methods (Softmax, MCD and DUQ). We also measure the Jensen-Shannon divergence between the distributions of the uncertainties $\boldsymbol{p}_{u_{\text{incorrect}}}$ and $\boldsymbol{p}_{u_{\text{correct}}}$, for the incorrect and correct estimations, respectively. Uncertainty for wrong and right estimations is expected to be higher and lower respectively. The reported descriptive statistics correspond to the results of 10 runs with 10 different test and training data partitions. The results yielded for the described experiment are displayed in tables 2, 3 and 7.

Finally, as a complementary qualitative test, we calculated the rejection-classification plots described in [22]. The average accuracy was calculated for each uncertainty bin. In general, for rejection plots, the less spiky and closer to an identity function, the better for an uncertainty estimator. Such plots are displayed in Table 9 for the three tested uncertainty estimators.

V. RESULTS

The F1-score and accuracy of the models trained with less than 70 labels reported a significant performance gain when using the tested SSDL model. The F1-score gain goes from around 0.18 with 20 labels to almost 0.01 when using 60 labels. With 70 labels, the sample mean accuracy and F1-score gets marginally better for the supervised model, making the impact of SSDL negligible. We also report the Δ_ρ to measure the accuracy gain under the specific SSDL data setting. The yielded results allow to evaluate uncertainty estimation performance under the setting of substantial ($n_l = 20, 30$), marginal ($n_l = 60$) and negative ($n_l = 70, 100$) accuracy and F1-score gains when using MixMatch.

Taking the accuracy gains into account for different number of labels n_l used for training we proceed to analyze the uncertainty estimation reliability by using the proposed Jensen-Shannon divergence between the uncertainty distribution of correct and wrong estimations.

For the Softmax function, the wrong-correct uncertainty distribution distances are depicted in Table 2. In this table, a significant Jensen-Shannon divergence gain is yielded when $n_l = 20$ and $n_l = 30$, with gains ranging from 0.32 to 0.2. However, when $n_l = 60$ and $n_l = 70$ the Jensen-Shannon (JS) divergence gets smaller between the supervised and SSDL model, with only 0.04 of difference. For $n_l = 100$, the supervised model gets a much higher JS divergence, suggesting a high correlation between the accuracy/f1-score gain and uncertainty reliability gain by using SSDL for the softmax uncertainty based approach. Table 6 shows how the distributions of the softmax uncertainties for the wrong and correct distributions are significantly different for both the SSDL and supervised models, however, the JS divergence makes easier to spot the difference between the distributions quantitatively.

As for the MCD for uncertainty estimation, a similar behavior can be observed, with decreasing uncertainty reliability gains when the number of labels go from $n_l = 20$ and $n_l = 100$ when using the SSDL model. Similarly, for $n_l = 20$ up to $n_l = 70$ labels, the reliability of the SSDL model uncertainty estimations outperform the supervised model by a larger margin. MCD obtains lower reliability gains for the SSDL model when compared to the Softmax approach, for the lowest number of labels $n_l = 20$ and $n_l = 30$ tested. Also for the SSDL model, when the number of labels increases from $n_l = 60$, the reliability of the MCD approach is better when compared to the softmax method. The uncertainty distribution plots for the correct and wrong estimations depicted in Table 5 show important differences between such distributions, but the improvement between the SSDL and supervised models is hard to discern visually.

Regarding the results for the DUQ uncertainty estimation method, the overall JS divergences are significantly lower than the MCD and softmax approaches. This suggests that both methods significantly outperform DUQ as an uncertainty estimation method. The plots in Table 4 qualitatively complement the small difference between the DUQ uncertainty distributions of correct and wrong estimations. However, similar to the softmax and MCD methods, the usage of SSDL makes a positive impact when n_l is between $n_l = 20$ up to $n_l = 60$.

A summary of the results is presented in Table 8. The use of the Jensen-Shannon divergence between the uncertainty distributions of the correct and wrong estimations allowed us to perform such analysis. We can see how the highest relative uncertainty estimation improvements are yielded when the models are trained with fewer labels. In such case, the gains range from 81 to 142 percent, for all the tested uncertainty estimation methods. In general, as the the number of labels increases, the reliability gain of the uncertainty estimations

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3085418, IEEE Access

n_l	Desc. stat.	No SSDL F1-Score/Accuracy	SSDL F1-Score/Acc	ρ_{lu}	Δ_ρ
20	$\overline{x}$	0.754/0.816	0.93/0.965	0.144	7.684
	s	0.069/0.07	0.0415/0.021		
30	$\overline{x}$	0.836/0.89	0.943/0.976	0.217	7.295
	s	0.07/0.048	0.032/0.013		
60	$\overline{x}$	0.902/0.96	0.917/0.971	0.434	0.641
	s	0.044/0.028	0.048/0.018		
70	$\overline{x}$	0.931/0.958	0.909/0.968	0.507	0.409
	s	0.032/0.025	0.0473/0.025		
100	$\overline{x}$	0.932/0.975	0.791/0.9	0.724	-2.287
	s	0.032/0.018	0.064/0.029		

TABLE 1. F1 score and accuracy statistics for batches tested with different number of labels n_l.

n_l	Desc. stat.	$s(\boldsymbol{y})$ correct NO-SSDL	$s(\boldsymbol{y})$ wrong NO-SSDL	JS div. No SSDL	$s(\boldsymbol{y})$correct SSDL	$s(\boldsymbol{y})$ wrong SSDL	JS div. No SSDL
20	$\overline{x}$	0.8216	0.7343	0.2407	0.9452	0.7481	0.5812
	s	0.1384	0.1457		0.0932	0.1298	
30	$\overline{x}$	0.8594	0.7207	0.3554	0.9597	0.7722	0.5497
	s	0.1337	0.1429		0.0755	0.1556	
60	$\overline{x}$	0.9159	0.7128	0.506587	0.9301	0.7315	0.5462
	s	0.1116	0.1504		0.0986	0.1384	
70	$\overline{x}$	0.9097	0.7387	0.4514	0.9232	0.7309	0.4976
	s	0.1167	0.1535		0.1066	0.1467	
100	$\overline{x}$	0.9324	0.7297	0.5067	0.8445	0.7254	0.2872
	s	0.1016	0.1471		0.1438	0.1505	

TABLE 2. Softmax results for the semi-supervised and supervised models with different numbers of labels n_l. Higher values indicate higher model confidence. The higher the better for correct estimations, and the lower the better for incorrect estimations.

n_l	Desc. stat.	σ correct No SSDL	σ wrong No SSDL	JS div. No SSDL	σ correct SSDL	σ wrong SSDL	JS div. SSDL
20	$\overline{x}$	0.5401	0.6222	0.271	0.3076	0.6334	0.4938
	s	0.1266	0.0852		0.2107	0.0925	
30	$\overline{x}$	0.5008	0.6370	0.367503	0.2890	0.6038	0.5204
	s	0.1552	0.0964		0.1904	0.1255	
60	$\overline{x}$	0.4041	0.6253	0.49075	0.3687	0.6317	0.5591
	s	0.1856	0.1150		0.1975	0.0730	
70	$\overline{x}$	0.4133	0.6253	0.4301	0.3968	0.6355	0.5105
	s	0.1882	0.1228		0.1825	0.0840	
100	$\overline{x}$	0.3540	0.6186	0.5534	0.5179	0.6313	0.3382
	s	0.2028	0.1348		0.1551	0.0903	

TABLE 3. MCD results for the semi-superivised and supervised models with different numbers of labels n_l. Lower values indicate higher model confidence. The lower the better for correct estimations, and the higher the better for incorrect estimations.

using SSDL tend to decrease. This correlates well with the average accuracy gains using SSDL depicted in Table 1.

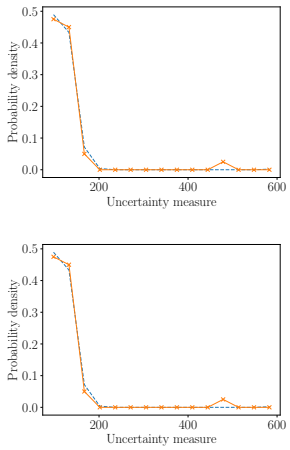
This tendency is more visible for the MCD and Softmax methods. The DUQ method is very unstable, as its capability for uncertainty estimation is more limited when compared to the first two methods, with lower JS divergences for all the tested configurations as seen in Table 7. Marginal uncertainty estimation improvements were obtained for the DUQ method, as seen in Table 8.

Finally, Figure 9 shows the rejection plots for the tested uncertainty estimation methods, with different numbers of labels n_l. In most cases the plots are rather similar, and also reveal a very high dispersion of the results for each bin, depicted by the blue (supervised model) and the orange areas (SSDL model). Such high dispersion suggests a possible statistically irrelevant comparison of results. Most of the plotted curves show higher accuracies per bin for the SSDL model, which corresponds to the results yielded in Table 1 where for most tested configurations the SSDL model outperforms the supervised one. This makes the comparison of the rejection plots between the supervised and the SSDL model harder.

VI. LIMITATIONS OF THE STUDY

This work used a limited sample of COVID-19 positive observations coming from a very different distribution when compared to the source of COVID-19 negative observations sampled from [29]. This causes a bias in the population of patients sampled for COVID-19 positive and negative cases related to age and ethnicity, as the data sources for both cases are completely different. The low availability of public repositories of COVID-19 chest X-rays with reliable labels at the time of writing poses a limitation to this work. Therefore, an additional validation of the proposed method in this work with other datasets with higher quality (with less age and ethnicity biases) is necessary. We plan to do this in the future. This work focused on measuring the impact of semi-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3085418, IEEE Access



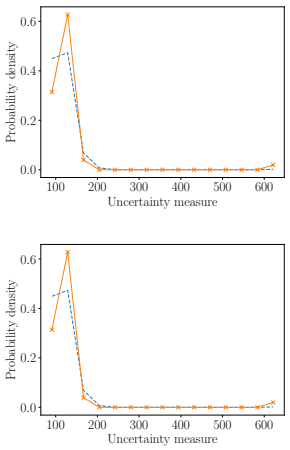
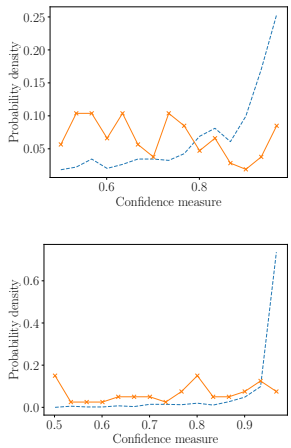


TABLE 4. DUQ uncertainty distributions for correct (blue dashed line) and incorrect estimations (orange dashed with 'x' line) using $n_l = 30, 70$ labels (from left to right). From top to bottom, the first row corresponds to the supervised model and the second row, to the SSDL model results.



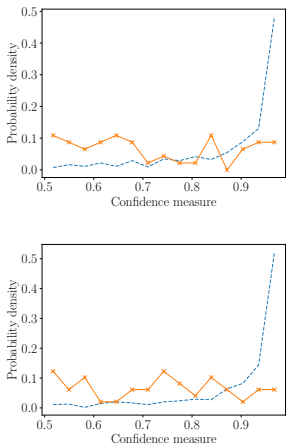
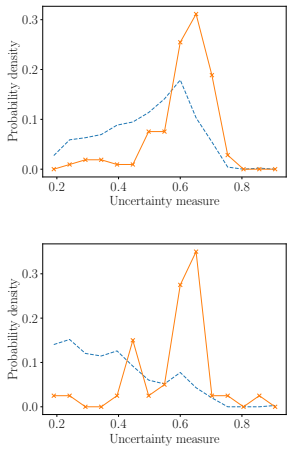


TABLE 6. Softmax confidence distributions for correct (blue dashed line) and incorrect (orange dashed with 'x' line) estimations using $n_l = 30, 70$, from left to right. From top to bottom, the supervised and the semi-supervised deep learning models results.



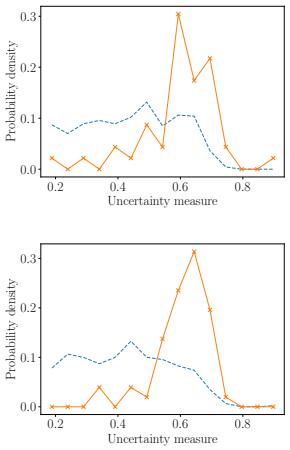


TABLE 5. MCD uncertainty distributions for correct (blue dashed line) and incorrect (orange dashed with 'x' line) estimations using $n_l = 30, 70$, from left to right. From top to bottom, the supervised and the semi-supervised deep learning models results.

supervised learning on uncertainty estimations for COVID-19 detection, and evidenced how predictive uncertainty estimations improve as model accuracy improves. However, the quality of the predictive uncertainty estimations can be improved through model calibration methods. Furthermore, other uncertainty estimation methods can be included in the comparison. We plan to test uncertainty estimation improvements in future work.

VII. CONCLUSIONS

In this work we have tested the impact of using unlabelled data to improve the reliability of uncertainty estimations through the implementation of the SSDL algorithm known as MixMatch. We tested three different uncertainty estimation methods (softmax, MCD and DUQ). The yielded descriptive statistics suggest an important reliability improvement of the uncertainty estimations when using SSDL for all the three uncertainty estimation methods. With low number of labels, the JS divergence is boosted by up to 142%, as seen in Table 8.

To ease the comparison of the tested uncertainty techniques, we proposed the use of the JS divergence, comparing the distributions of the wrong and correct estimations. The test is statistically relevant as it takes into account the whole results distribution, and it is easy to interpret, with values ranging from 0 to 1 (the higher the values the better). The use of the JS divergence index to compare the uncertainty estimations proved to be simple to analyze, with easy to map correspondence with the distribution plots. Its use is recommended when comparing different uncertainty methods under different models which cause fluctuations in the model accuracy.

When comparing the three tested uncertainty estimation methods, the MCD and the softmax techniques performed better than the DUQ approach. The comparison between the MCD and the softmax methods is rather mixed, with MCD performing better when n_l is higher. Results with the DUQ method yielded a significantly worse performance for uncertainty estimation. We speculate that this is due to the high similarity between the images of the two classes. This makes the averaged observations in the feature space similar for both classes and the comparison of new unseen observations less sensitive. In terms of the uncertainty source, the MCD approach seemed to be more sensitive to epistemic uncertainty, than the DUQ method. Epistemic uncertainty can be considered to be very high in models trained with very few labels, as the feature space sample is very limited. MCD takes into account the epistemic uncertainty of both the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3085418, IEEE Access

n_l	Desc. stat.	ℓ_2 correct No SSDL	ℓ_2 wrong No SSDL	JS div.No SSDL	ℓ_2 correct SSDL	ℓ_2 wrong SSDL	JS div. SSDL
20	$\overline{x}$	133.7481	143.8356	0.072	131.9169	137.8194	0.17024
	s	24.4915	55.7149		33.7233	21.8417	
30	$\overline{x}$	130.7032	147.3737	0.13599	132.2328	142.4301	0.108488
	s	32.4645	48.0116		31.6105	58.383	
60	$\overline{x}$	135.0814	137.0349	0.072	132.7359	142.3879	0.163092
	s	35.8013	20.6742		37.7423	20.4047	
70	$\overline{x}$	132.1344	149.7608	0.216469	132.5677	148.2541	0.14228
	s	24.8133	80.2325		32.204	75.853	
100	$\overline{x}$	134.1262	143.4125	0.1599	132.6018	140.198	0.0943
	s	29.4393	71.2851		34.5869	51.501	

TABLE 7. DUQ results for the semi-supervised and supervised models with different numbers of labels n_l. Lower values indicate higher model confidence. The lower the better for correct estimations, and the higher the better for incorrect estimations.

n_l	**MCD: JS div. gain SSDL vs. No SSDL**	**Softmax: JS div. gain SSDL vs. No SSDL**	**DUQ: JS div. gain SSDL vs. No SSDL**
20	+0.222/+81%	+0.34/+142%	+0.098/+136%
30	+0.153/+41.7%	+0.194/+54.6%	-0.027/-20%
60	+0.069/+14%	+0.04/+8%	+0.092/+126%
70	+0.08/+18%	+0.04/+10%	-0.073/-34%
100	-0.21/-38%	-0.22/-0.43%	-0.065/-41%

TABLE 8. Summary of the Jensen-Shannon divergence gains (uncertainty distributions divergence for the correct and incorrect estimations), for each tested uncertainty estimation method, using semi-supervised learning.

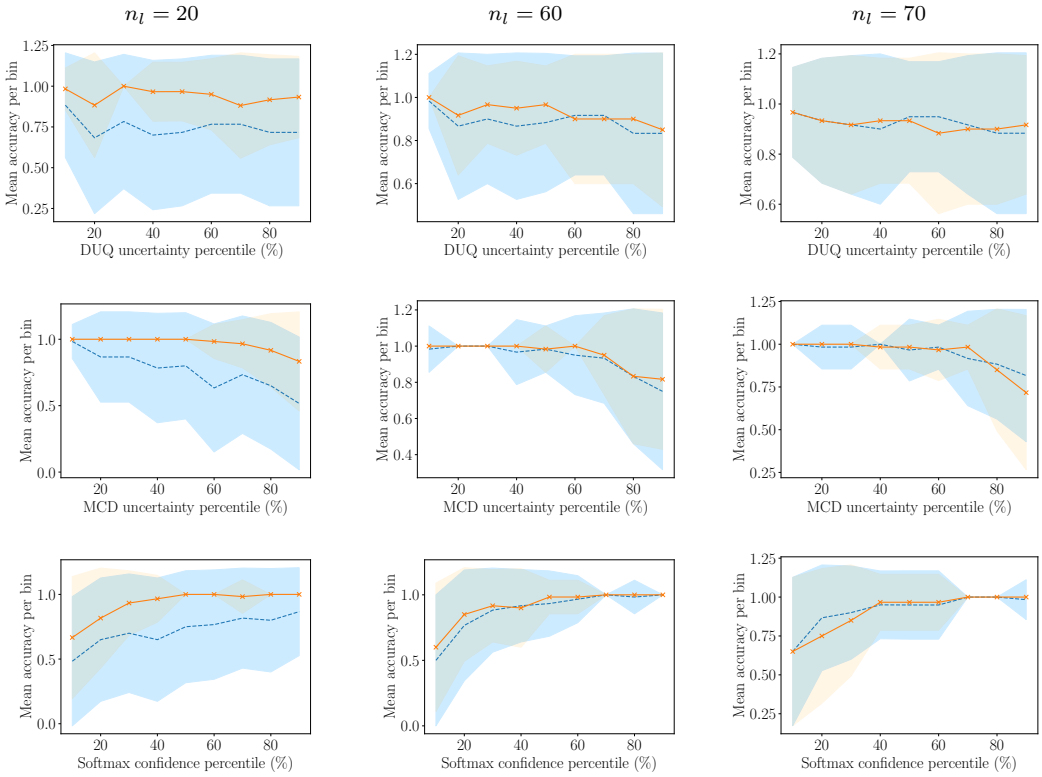


TABLE 9. Rejection plots for the three tested uncertainty approaches. The first row correspond to the DUQ estimations, the second one to the MCD uncertanties and the last one to the Softmax confindence scores. From left to right, models with different number of labels n_l. Orange and 'x' lines correspond to the semi-supervised model and the dashed and blue lines correspond to the supervised model.

feature extractor and the top model (fully connected network acting as classifier), unlike DUQ which only uses the feature extractor, and can be considered the only channel for the epistemic uncertainty for this method.

As future work, we plan to explore more recent uncertainty estimation approaches which have been originally designed for distribution mismatch measurement [48], [49]. Interchangeably, the quality of the unlabelled dataset and its impact in the model's accuracy and uncertainty estimations is also worth to explore. For this end, dataset quality metrics can be implemented [50]. Furthermore, we plan to explore the impact of unlabelled data in other engineering requirements of deep learning models such as model robustness. Little research has been done about the actual impact of semi or self supervised learning in important model properties such as robustness in practical applications like medical imaging analysis. For instance, we plan to further evaluate the improvement of model uncertainty reliability and robustness for COVID-19 detection using computed tomography as an alternative imaging technology which is also interesting to explore. The use of modern semi- and self-supervised techniques can do more than just improving the model accuracy under restricted number of labels. Therefore its impact should be studied in depth. In general, we highlight the need for evaluating other important model properties such as robustness and uncertainty reliability, specially for sensitive applications like medical imaging analysis.

ACKNOWLEDGMENTS

This work is partially supported by the Ministry of Economy and Competitiveness of Spain under grants TIN2016-75097-P and PPIT.UMA.B1.2017. It is also partially supported by the Ministry of Science, Innovation and Universities of Spain under grant RTI2018-094645-B-I00, project name Automated detection with low cost hardware of unusual activities in video sequences. It is also partially supported by the Autonomous Government of Andalusia (Spain) under grant UMA18-FEDERJA-084, project name Detection of anomalous behavior agents by deep learning in low cost video surveillance intelligent systems. All of them include funds from the European Regional Development Fund (ERDF). The authors thankfully acknowledge the computer resources, technical expertise and assistance provided by the SCBI (Supercomputing and Bioinformatics) center of the University of Málaga. The authors also thankfully acknowledge the grant of the Universidad de Málaga and the Instituto de Investigación Biomédica de Málaga (IBIMA).

REFERENCES

[1] J. Sun, W.-T. He, L. Wang, A. Lai, X. Ji, X. Zhai, G. Li, M. A. Suchard, J. Tian, J. Zhou *et al.*, "Covid-19: epidemiology, evolution, and cross-disciplinary perspectives," *Trends in molecular medicine*, vol. 26, no. 5, pp. 483–495, May 2020.

[2] M. N. Esbin, O. N. Whitney, S. Chong, A. Maurer, X. Darzacq, and R. Tjian, "Overcoming the bottleneck to widespread testing: a rapid review of nucleic acid testing approaches for covid-19 detection," *Rna*, vol. 26, no. 7, pp. 771–783, May 2020.

[3] 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-pcr assay validated in vitro and with clinical specimens," *Journal of clinical microbiology*, vol. 58, no. 5, Art. no. 00310, 2020.

[4] 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," *The Lancet*, vol. 395, no. 10223, pp. 507–513, Feb. 2020.

[5] 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*, vol. 295, no. 1, pp. 202–207, Feb. 2020.

[6] F. Song, N. Shi, F. Shan, Z. Zhang, J. Shen, H. Lu, Y. Ling, Y. Jiang, and Y. Shi, "Emerging 2019 novel coronavirus (2019-ncov) pneumonia," *Radiology*, vol. 295, no. 1, pp. 210–217, Feb. 2020.

[7] M. T. Shah, M. Joshipura, J. Singleton, P. LaBarre, H. Desai, E. Sharma, and C. Mock, "Assessment of the availability of technology for trauma care in india," *World journal of surgery*, vol. 39, no. 2, pp. 363–372, Oct. 2014.

[8] R. Arora, "The training and practice of radiology in India: current trends." *Quantitative imaging in medicine and surgery*, vol. 4, no. 6, pp. 449–44 950, Dec. 2014.

[9] L. Brunese, F. Mercaldo, A. Reginelli, and A. Santone, "Explainable deep learning for pulmonary disease and coronavirus covid-19 detection from x-rays," *Computer Methods and Programs in Biomedicine*, vol. 196, p. 105608, Nov. 2020.

[10] D. Das, K. Santosh, and U. Pal, "Truncated inception net: Covid-19 outbreak screening using chest x-rays," *Physical and engineering sciences in medicine*, vol. 43, no. 3, pp. 915–925, Jun. 2020.

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

[12] S. Calderon-Ramirez, A. Moemeni, D. Elizondo, S. Colreavy-Donnelly, L. F. Chavarria-Estrada, M. A. Molina-Cabello *et al.*, "Correcting data imbalance for semi-supervised covid-19 detection using x-ray chest images," *arXiv e-prints*, Aug. 2020. [Online]. Available: https://arxiv.org/abs/2008.08496

[13] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 4, Art. no. e1312, Apr. 2019.

[14] 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, Dec. 2020, pp. 280–317.

[15] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Müller, "Toward interpretable machine learning: Transparent deep neural networks and beyond," *arXiv e-prints*, Mar. 2020. [Online]. Available: https://arxiv.org/abs/2003.07631

[16] R. Alizadehsani, M. Roshanzamir, S. Hussain, A. Khosravi, A. Koohestani, M. H. Zangooei, M. Abdar, A. Beykikhoshk, A. Shoeibi, A. Zare *et al.*, "Handling of uncertainty in medical data using machine learning and probability theory techniques: a review of 30 years (1991–2020)," *Annals of Operations Research*, Mar. 2021. [Online]. Available: https://link.springer.com/article/10.1007/s10479-021-04006-2

[17] A. Der Kiureghian and O. Ditlevsen, "Aleatory or epistemic? does it matter?" *Structural safety*, vol. 31, no. 2, pp. 105–112, Mar. 2009.

[18] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *arXiv e-prints*, Oct. 2016. [Online]. Available: https://arxiv.org/abs/1610.02136

[19] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *International Conference on Machine Learning*, Jun. 2016, pp. 1050–1059.

[20] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International Conference on Machine Learning*, Aug. 2017, pp. 1321–1330.

[21] M. T. Le, F. Diehl, T. Brunner, and A. Knol, "Uncertainty estimation for deep neural object detectors in safety-critical applications," in *2018 21st*

International Conference on Intelligent Transportation Systems (ITSC). IEEE, Nov. 2018, pp. 3873–3878.

[22] J. van Amersfoort, L. Smith, Y. W. Teh, and Y. Gal, "Simple and scalable epistemic uncertainty estimation using a single deep deterministic neural network," *arXiv e-prints*, Jun. 2020.

[23] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances in neural information processing systems*, Nov. 2017, pp. 6402–6413.

[24] L. Oala, C. Heiß, J. Macdonald, M. März, W. Samek, and G. Kutyniok, "Interval neural networks: Uncertainty scores," *arXiv e-prints*, Mar. 2020. [Online]. Available: https://arxiv.org/abs/2003.11566

[25] C. Leibig, V. Allken, M. S. Ayhan, P. Berens, and S. Wahl, "Leveraging uncertainty information from deep neural networks for disease detection," *Scientific reports*, vol. 7, no. 1, pp. 17 816–17 816, Dec. 2017.

[26] P. Seeböck, J. I. Orlando, T. Schlegl, S. M. Waldstein, H. Bogunović, S. Klimscha, G. Langs, and U. Schmidt-Erfurth, "Exploiting epistemic uncertainty of anatomy segmentation for anomaly detection in retinal oct," *IEEE transactions on medical imaging*, vol. 39, no. 1, pp. 87–98, Jan. 2019.

[27] B. Ghoshal and A. Tucker, "Estimating uncertainty and interpretability in deep learning for coronavirus (covid-19) detection," *arXiv e-prints*, Mar. 2020. [Online]. Available: https://arxiv.org/abs/2003.10769

[28] J. P. Cohen, P. Morrison, and L. Dao, "Covid-19 image data collection," *arXiv e-prints*, Jun. 2020, data repository available at https://github.com/ieee8023/covid-chestxray-dataset. [Online]. Available: https://arxiv.org/abs/2006.11988

[29] 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*, vol. 172, no. 5, pp. 1122–1131, Feb. 2018.

[30] H. Asgharnezhad, A. Shamsi, R. Alizadehsani, A. Khosravi, S. Nahavandi, Z. Alizadeh Sani, and D. Srinivasan, "Objective evaluation of deep uncertainty predictions for covid-19 detection," *arXiv e-prints*, Dec. 2020. [Online]. Available: https://arxiv.org/abs/2012.11840

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

[32] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv e-prints*, Apr. 2018. [Online]. Available: https://arxiv.org/abs/1710.09412

[33] A. Chan, A. Alaa, Z. Qian, and M. Van Der Schaar, "Unlabelled data improves bayesian uncertainty calibration under covariate shift," in *International Conference on Machine Learning*. PMLR, Jun. 2020, pp. 1392–1402.

[34] D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song, "Using self-supervised learning can improve model robustness and uncertainty," in *Advances in Neural Information Processing Systems*, Dec. 2019, pp. 15 663–15 674.

[35] Y. Carmon, A. Raghunathan, L. Schmidt, J. C. Duchi, and P. S. Liang, "Unlabeled data improves adversarial robustness," in *Advances in Neural Information Processing Systems*, Dec. 2019, pp. 11 192–11 203.

[36] S. Thulasidasan, G. Chennupati, J. A. Bilmes, T. Bhattacharya, and S. Michalak, "On mixup training: Improved calibration and predictive uncertainty for deep neural networks," in *Advances in Neural Information Processing Systems*, Dec. 2019, pp. 13 888–13 899.

[37] J. Nixon, M. W. Dusenberry, L. Zhang, G. Jerfel, and D. Tran, "Measuring calibration in deep learning." in *CVPR Workshops*, Jun. 2019, pp. 38–41.

[38] J. Bröcker and L. A. Smith, "Increasing the reliability of reliability diagrams," *Weather and forecasting*, vol. 22, no. 3, pp. 651–661, Jun. 2007.

[39] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "Augmix: A simple data processing method to improve robustness and uncertainty," *arXiv e-prints*, Feb. 2020. [Online]. Available: https://arxiv.org/abs/1912.02781

[40] J. I. Orlando, P. Seeböck, H. Bogunović, S. Klimscha, C. Grechenig, S. Waldstein, B. S. Gerendas, and U. Schmidt-Erfurth, "U2-net: A bayesian u-net model with epistemic uncertainty feedback for photoreceptor layer segmentation in pathological oct scans," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, Apr. 2019, pp. 1441–1445.

[41] A. Malinin and M. Gales, "Predictive uncertainty estimation via prior networks," in *Advances in Neural Information Processing Systems*, Dec. 2018, pp. 7047–7058.

[42] P. W. Lamberti and A. P. Majtey, "Non-logarithmic jensen–shannon divergence," *Physica A: Statistical Mechanics and its Applications*, vol. 329, no. 1-2, pp. 81–90, Nov. 2003.

[43] L. Wang, Z. Q. Lin, and A. Wong, "Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images," *Scientific Reports*, vol. 10, no. 1, Nov. 2020. [Online]. Available: https://www.nature.com/articles/s41598-020-76550-z

[44] E. E.-D. Hemdan, M. A. Shouman, and M. E. Karar, "Covidx-net: A framework of deep learning classifiers to diagnose covid-19 in x-ray images," *arXiv e-prints*, Mar. 2020. [Online]. Available: https://arxiv.org/abs/2003.11055

[45] I. D. Apostolopoulos and T. A. Mpesiana, "Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks," *Physical and Engineering Sciences in Medicine*, vol. 43, no. 2, pp. 635–640, Apr. 2020.

[46] Q. Oliveau and H. Sahbi, "Semi-supervised deep attribute networks for fine-grained ship category recognition," in *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, Jul. 2018, pp. 6871–6874.

[47] L. N. Smith, "A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay," *arXiv e-prints*, 2018. [Online]. Available: https://arxiv.org/abs/1803.09820

[48] S. Calderon-Ramirez and L. Oala, "More than meets the eye: Semi-supervised learning under non-iid data," *arXiv e-prints*, Apr. 2021. [Online]. Available: https://arxiv.org/abs/2104.10223

[49] S. Calderon-Ramirez, L. Oala, J. Torrents-Barrena, S. Yang, A. Moemeni, W. Samek, and M. A. Molina-Cabello, "Mixmood: A systematic approach to class distribution mismatch in semi-supervised learning using deep dataset dissimilarity measures," *arXiv e-prints*, Jun. 2020. [Online]. Available: https://arxiv.org/abs/2006.07767

[50] M. Mendez, S. Calderon, and 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, Feb. 2020, pp. 307–319.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3085418, IEEE Access

SAUL CALDERON-RAMIREZ received his BSc in Computer Science, and his MSc in Electrical Engineering from the University of Costa Rica, Costa Rica, in 2012 and 2015, respectively. He has previously worked in the private industry as a consultant in big data, and as a researcher at Intel Costa Rica. He is now lecturer at the Computing Engineering school, Costa Rica Institute of Technology, Costa Rica. He also is currently a PhD student, at the Institute of Artificial Intelligence, De Montfort University, United Kingdom. His research interests include deep learning semi and self supervised learning, robustness and uncertainty analysis for deep learning models in medical imaging applications.

SHENGXIANG YANG (M'00–SM'14) received the Ph.D. degree from Northeastern University, Shenyang, China in 1999. He is currently a Professor in Computational Intelligence and Director of the Centre for Computational Intelligence, School of Computer Science and Informatics, De Montfort University, Leicester, U.K. He has over 330 publications with an H-index of 57 according to Google Scholar. His current research interests include evolutionary computation, swarm intelligence, artificial neural networks, data mining and data stream mining, and relevant real-world applications. He serves as an Associate Editor/Editorial Board Member of a number of international journals, such as the IEEE Transactions on Cybernetics, IEEE Transactions on Evolutionary Computation, Information Sciences, and Enterprise Information Systems.

ARMAGHAN MOEMENI Armaghan is a senior fellow of higher education academy (SFHEA) and and an assistant professor in computer science at the University of Nottingham. Armaghan's research interests are in computer vision and machine learning, human computer interaction, ambient intelligence, and applied artificial intelligence. She received her BSc in Electronics and Electrical Engineering from Shiraz University, MSc in Multimedia Computing and PhD in Computer Science from De Montfort University.

SIMON COLREAVY-DONNELLY A Lecturer in Computer Games Programming at De Montfort University. He obtained a PhD in Digital Arts and Humanities from the National University of Ireland, Galway in 2016. His research interests include Intelligent Systems and Computational Intelligence, Graphics and Image Processing.

DAVID A. ELIZONDO is a Professor in Intelligent Transport Systems at the Department of Computer Technology at De Montfort University, United Kingdom. He has a BA in Computer Science from Knox College, Galesbourg, Illinois, USA, an MS in Artificial Intelligence at the Department of Artificial Intelligence and Cognitive Computing of the University of Georgia, Athens, Georgia, USA and a PhD in computer science from the University of Strasbourg , France in cooperation with the Swiss Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP).

LUIS OALA is a PhD research associate in the XAI Group at the AI Department of Fraunhofer HHI. He works on reliable and trustworthy machine learning with a particular focus on uncertainty quantification and robustness. Further, he chairs an ITU/WHO working group on data and AI quality which develops methods, standards and software for auditing AI systems. He received his BA in Liberal Arts and Sciences from University College Maastricht and his MA in 2018 studying econometrics and machine learning at HTW, HU and TU Berlin.

JORGE RODRÍGUEZ-CAPITÁN acquired his medical degree at University of Malaga (Spain), and obtained his specialization in Cardiology at the Hospital Virgen de la Victoria CIBERCV (Malaga, Spain). He received his PhD degree in Medicine from the University of Malaga (Spain) in 2016. Currently, he develops his career both as a clinical cardiologist and as a clinical researcher. His research concerns include tricuspid valvular disease, heart failure, ischemic heart disease, hypertrophic cardiomyopathy, and recently COVID-19 disease.

MANUEL JIMÉNEZ-NAVARRO acquired his medical degree at University of Sevilla (Spain), and obtained his specialization in Cardiology at the Hospital Virgen de la Victoria (Malaga, Spain). He received his PhD degree in Medicine from the University of Malaga (Spain) in 1999. Currently, he works as full professor of Medicine in University of Malaga; as clinical cardiologist in the same hospital and he is the director of the CIBERCV node research group (Carlos III Health Institute,Spain) which is integrated into the Biomedical Research Institute of Malaga (IBIMA).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3085418, IEEE Access

EZEQUIEL LÓPEZ-RUBIO (born 1976) received his MSc and PhD (honors) degrees in Computer Engineering from the University of Málaga, Spain, in 1999 and 2002, respectively. He joined the Department of Computer Languages and Computer Science, University of Málaga, in 2000, where he is currently a Full Professor of Computer Science and Artificial Intelligence. His technical interests are in deep learning, pattern recognition, image processing and computer vision.

MIGUEL A. MOLINA-CABELLO received his MSc and PhD degrees in Computer Engineering from the University of Málaga, Spain, in 2015 and 2018. He joined the Department of Computer Languages and Computer Science, University of Málaga, in 2015, where he has a teaching and researching position. He also keeps pursuing research activities in collaboration with other Universities. His technical interests include visual surveillance, image/video processing, computer vision, neural networks and pattern recognition.