Neural Computing and Applications (2023) 35:9819–9830
https://doi.org/10.1007/s00521-023-08219-3

ORIGINAL ARTICLE

Evaluating deep learning predictions for COVID-19 from X-ray images using leave-one-out predictive densities

Sergio Hernández[1] · Xaviera López-Córtes[1]

Received: 11 July 2022 / Accepted: 6 January 2023 / Published online: 6 February 2023

Abstract
Early detection of the COVID-19 virus is an important task for controlling the spread of the pandemic. Imaging techniques such as chest X-ray are relatively inexpensive and accessible, but its interpretation requires expert knowledge to evaluate the disease severity. Several approaches for automatic COVID-19 detection using deep learning techniques have been proposed. While most approaches show high accuracy on the COVID-19 detection task, there is not enough evidence on external evaluation for this technique. Furthermore, data scarcity and sampling biases make difficult to properly evaluate model predictions. In this paper, we propose stochastic gradient Langevin dynamics (SGLD) to take into account the model uncertainty. Four different deep learning architectures are trained using SGLD and compared to their baselines using stochastic gradient descent. The model uncertainties are also evaluated according to their convergence properties and the leave-one-out predictive densities. The proposed approach is able to reduce overconfidence of the baseline estimators while also retaining predictive accuracy for the best-performing cases.

Keywords Bayesian learning · Markov chain Monte Carlo · COVID X-ray

1 Introduction

Deep learning has become an essential tool for automated decision making in several domain applications including image classification, object detection and natural language processing, among others [2]. However, the impressive performance shown by this method in several large-scale benchmarks contrasts with its application to machine-assisted clinical decision making. There are several reasons for this reluctance, therefore, giving an artificial intelligence technique the power to take life-critical decision is still challenging [5].

Uncertainty refers to the lack of certainty due to imperfect or unknown information. In particular, aleatoric uncertainty is related to the notion of randomness itself and can be identified by running several experiments and observing their outcomes. Epistemic uncertainty in the other hand is related to the lack of knowledge and can only be reduced by introducing new observations or background knowledge [19]. Bayesian inference is a popular technique for incorporating domain knowledge into the model and evaluating both, the aleatoric and epistemic uncertainties. For deep learning models, the aleatoric uncertainty part can be well captured by fusing the standard neural network architecture with a probability distribution. Instead, for the epistemic uncertainty part, we must treat the model parameters as random variables and the predictive uncertainty is obtained by marginalizing the posterior distribution over the parameters [1]. Building predictive models for the COVID-19 pandemic is one such example of this lack of certainty. In particular, the detection of positive cases is usually performed using reverse transcription polymerase chain reaction (RT-PCR) tests. This technique is precise but costly in terms of human resources and infrastructure. Therefore, there have been significant efforts to develop COVID-19 detection procedures that can be used to complement or to provide faster and accurate alternatives. Computer tomography can be considered as being both fast

✉ Sergio Hernández
shernandez@ucm.cl

Xaviera López-Córtes
xlopez@ucm.cl

[1] Departamento de Computación en Industrias. Facultad de Ciencias de la Ingeniería, Universidad Católica del Maule, Av. San Miguel 3605, 100190 Talca, Maule, Chile

and accurate; however, it is expensive and its evaluation requires domain experts that can assess the disease onset. In the other hand, chest radiography (X-ray) uses lower radiation doses than computer tomography. Conversely, X-ray imaging is inexpensive and readily available in many hospitals and primary care health centers [23]. Several approaches for automatic COVID-19 detection using chest X-ray images have been published [3, 22]. These studies make use of different strategies for data handling and different neural network architectures. While most studies report classification accuracies above 90%, it is still unclear how the training data and the modeling assumptions affect the final results or the capability of the model to produce reliable predictions [21]. Narin et al. [24] used five pre-trained convolutional neural networks (ResNet50, ResNet101, ResNet152, InceptionV3 and Inception-ResNetV2) for the detection of COVID-19 infected patients from chest X-ray images. The authors performed three binary classifications with four classes (COVID-19, normal (healthy), viral pneumonia and bacterial pneumonia) and achieved the best accuracy (98%) for ResNet50. Also, Wang et al. proposed a tailored deep convolutional neural network [27]. The COVID-Net model was trained using publicly available data composed of 13975 X-ray images across 13870 patient cases. The authors reported 93.3% accuracy for the three class database using the COVID-Net model whose weights were pre-trained using the ImageNet database [9]. It is also important to notice that most models are trained using imbalanced databases. Conversely, data augmentation and oversampling play an important role on the final results. Chowdhury et al. evaluated different architectures and data augmentation schemes for binary and multi-class classification [6]. A variant of the DenseNet architecture named CheXNet that was previously trained on chest X-ray images outperformed other neural network models when no data augmentation was used. Nevertheless, the authors shown that a deeper neural network improved the classification results from CheXNet when using data augmentation techniques for training. Data imbalance is pervasive among most medical datasets [20]. Most of the research been done on automatic COVID-19 detection using data collected from multiple sources. Garcia et al. shown that this procedure cannot guarantee that a model can be built with low risk of bias [12]. Also, there are confidentiality issues and small number of labeled examples, which causes the number of positive cases being smaller than the number of control cases. A summary of deep learning-based COVID-19 detection from X-ray can be found in Ref. [17].

2 Related work

Dropout is a popular regularization technique for deep learning that randomly removes units from any base architecture. [11] demonstrated that using Monte Carlo sampling with dropout activations during test time produces samples from the posterior distribution. In Ref. [7], the authors proposed dropout and data augmentation schemes at test time in order to estimate aleatoric uncertainty for dermoscopic image classification. Reference [14] developed an uncertainty estimation framework for reporting confidence in medical image segmentation and diseases detection using deep learning. The authors used an ensemble of models trained with dropout at test time (MC-Dropout) to approximate the posterior distribution. This approach is not intended for producing state-of-the-art accuracy results, but for evaluating the usefulness of the predictive uncertainty to avoid overconfident predictions. Closely related to our approach, Gour and Jain used MC-Dropout with the EfficientNet-B3 architecture to evaluate predictive accuracy for detecting COVID-19 from chest X-ray images [15]. In order to evaluate predictive uncertainty, their model performs several forward passes using dropout activations and the mean entropy captures the model uncertainty. Reference [13] developed a cost-sensitive calibrated uncertainty estimation framework for COVID-19 detection. The model uses a variational posterior approximation with Monte Carlo drop-weights. Variational inference is a well-known technique for sampling from a posterior distribution; however, it suffers from mode collapse. Therefore, the authors also propose Jackknife resampling techniques to correct for sample bias. More recently, [4] evaluated three uncertainty quantification techniques for COVID-19 detection from X-ray images. The authors evaluated MC-Dropout, ensemble methods and a combination of ensembles a MC-Dropout. Their findings indicate that network pre-training using a chest X-ray dataset yields improved results when compared to the standard fine-tuning using ImageNet as a base model. Also, ensemble techniques were found to improve quantification of the predictive uncertainty. In Ref. [16], the author describes human-in-the-loop techniques for building trustworthy artificial intelligence. These methods are potentially capable to describe causal relationships that cannot be achieved with just supervised learning. In particular, most state-of-the-art deep learning architectures are prone to provide wrong outputs with high confidence when the input contains small perturbations.

2.1 Contributions

Most studies have used either one of MC-Dropout, ensembles, variational inference techniques or a combination of them for estimating the uncertainty of deep learning predictions for COVID-19 detection. However, Stochastic-Gradient Markov Chain Monte Carlo (SG-MCMC) sampling techniques have received less attention. As opposed to MC-Dropout and variational inference, SG-MCMC produces samples from the posterior distribution. However, due to the sequential nature of the sampling mechanism, the samples are correlated and diagnosing convergence is notoriously difficult [25]. The main contribution of this paper can be summarized as:

- Baseline performance was obtained using four different convolutional neural networks that were fine-tuned using Bayesian optimization to detect COVID-19 from chest X-ray images.
- Stochastic-Gradient MCMC is used to obtain posterior samples from each one of the base architectures, and their convergence is diagnosed and evaluated.
- Predictive uncertainty is evaluated using a scoring function using Pareto-Smoothed Importance Sampling leave-one-out Cross-Validation (PSIS-LOO). The F_{LOO} metric is based on the leave-one-out predictive density and is compared to the predictive uncertainty obtained with an ensemble technique.

Figure 1 shows and schematic diagram of the proposed approach.

3 Materials and methods

Bayesian neural networks replace deterministic weights θ from standard neural networks with random variables. Conversely, deep learning architectures using stochastic weights can be used to quantify the uncertainty $p(y|X,\theta)$ in regression and classification for a given dataset $\mathcal{D} = \{(x_i, y_i)\}$ for $i = 1,\ldots,N$.

Given a joint density in the form $p(\mathcal{D},\theta)$, Bayesian inference aims to compute the posterior distribution $p(\theta) = \frac{p(\mathcal{D}|\theta)(\theta)}{p(\mathcal{D})}$. However, this requires a prior distribution $p(\theta)$ and a normalizing constant $p(\mathcal{D})$, which is usually intractable.

The choice of the prior for Bayesian deep learning models usually follows some eliciting mechanism that provides information about the neural network parameters. Such knowledge is usually vague or incomplete, so practitioners would normally select a convenient distribution (such as the isotropic Gaussian) that facilitates inference. In the Bayesian framework, the unknown parameter θ is considered as a random variable. The stochastic gradient Langevin Monte Carlo (SGLD) algorithm uses a stochastic gradient $\hat{\nabla} f(\theta)$ approximation to generate samples from the posterior distribution. The SGLD algorithm generates proposals using the following update rule:

$$\theta^k = \theta^{k-1} - \frac{\eta_k}{2}\hat{\nabla} f(\theta^{k-1}) + v_k \quad (1)$$

where η_k is a time-decaying learning rate, $v_k \sim \mathcal{N}(0,\eta_k)$ and $f(\theta) = -\frac{1}{B}\log p(\tilde{D}|\theta) - \log p(\theta)$.

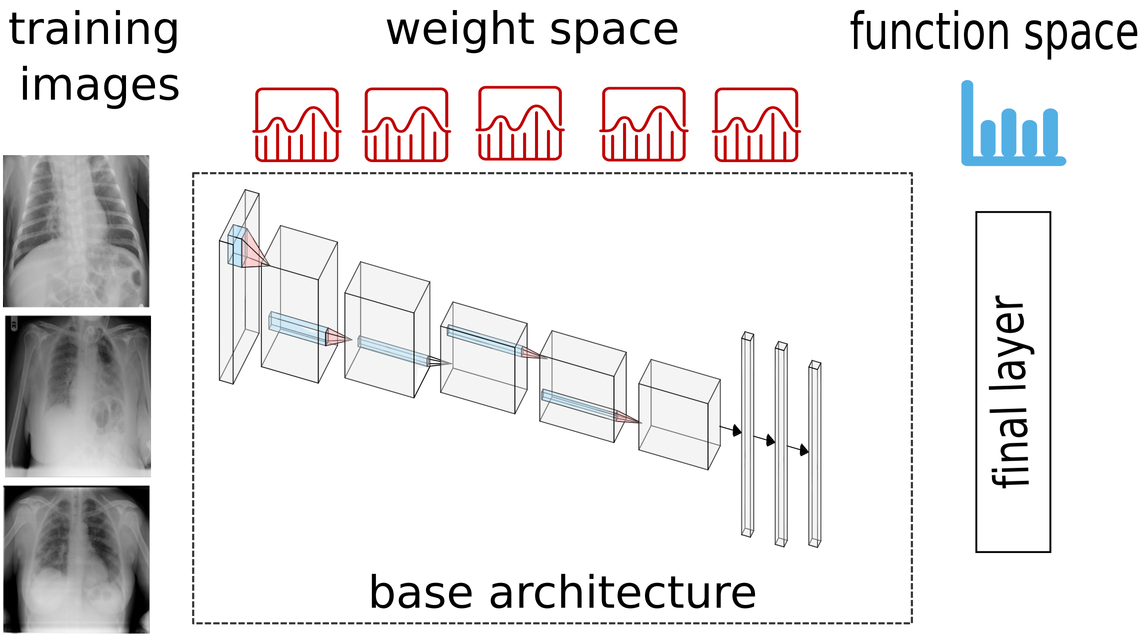


Fig. 1 Schematic diagram of the proposed approach. A base architecture trained on the ImageNet dataset is selected, and the top layer is replaced. SG-MCMC is used to obtain posterior samples that provides predictive uncertainty

3.1 Convergence diagnostics

The SGLD update rule in Eq. 1 is a discrete-time representation of a continuous-time stochastic process. SGLD has been successfully used to quantify uncertainty in Bayesian deep learning models for several problems such as age/gender estimation from facial images and plant diseases recognition. However, due to the discretization error, the algorithm converges weakly to the posterior distribution and can produce biased estimates depending on the specific choice of the learning rate η_k and the size of the mini-batch $\mathcal{B}$. Given the output of SGLD for a fixed number of iterations, we would like to assess the efficiency and accuracy of the samples to represent the posterior distribution. Given the sequential nature of MCMC methods, non-convergence of the sampler can be estimated from several parallel chains where the variance across the different simulations is higher than the variance of each one of the single chains. Let M be the number of chains and N the total number of samples. We can estimate the between B and within-chain W variances using Eqs. 2a and 2b.

$$B = \frac{N}{M-1}\sum_{m=1}^{M}(\bar{\theta}_m - \bar{\theta}) \tag{2a}$$

$$W = \frac{1}{M}\sum_{m=1}^{M} s_m^2 \tag{2b}$$

where $\bar{\theta}_m = 1/N\sum_n \theta_{nm}$, $\bar{\theta} = 1/M\sum_m \bar{\theta}_m$ and $s_m^2 = \frac{1}{N-1}\sum_n(\theta_{nm} - \theta_m)^2$. Using the between and within-chain variances we can estimate the potential scale reduction factor $\hat{R}$, which can be thought as the overestimation of variance due to the finite number of samples. The $\hat{R}$ (see Eq. 3) diagnostic evaluates the benefit of sampling longer chains and $\hat{R} \approx 1$ indicates that increasing the number of samples will not reduce the variance of the estimator.

$$\hat{R} = \sqrt{\frac{N}{N-1} + \frac{M+1}{MN}\frac{B}{W}} \tag{3}$$

Apart from the variances, we can also take a look at the auto-correlation $\rho(l)$ for different lags l and estimate the amount of information contained in that sample. The effective sample size $\hat{N}_{\text{eff}}$ use variograms to extend the auto-correlation from a single chain to several chains using Eq. 4.

$$\hat{N}_{\text{eff}} = \frac{NM}{1 + 2\sum_l \hat{\rho}(l)} \tag{4}$$

In Ref. [18], the authors evaluated Bayesian deep learning models using full-batch Hamiltonian Monte Carlo (HMC). The $\hat{R}$ diagnostic was calculated for both the model parameters θ (weight space) and the model outputs $f(\theta)$ (function space). Since there is no indication of poor mixing (large $\hat{R}$ values) in function space, full-batch HMC is able to produce unbiased estimates from the posterior distribution. Moreover, the posterior estimates obtained with HMC are compared to the SG-MCMC counterpart and the authors report agreement and total variation metrics. Convergence in weight space in the other hand tends to be more elusive with differing values for the different parameters.

3.2 Leave-one-out predictive densities

Now, we would like to evaluate the different models based on predictive performance. A common approach would rely on posterior quantiles to deliver different point estimates such as the accuracy or the F score using leave-one-out (LOO) cross-validation. This method is computationally inefficient since it requires storing the model parameters θ from the NM simulations and then compute the required accuracy score for each on of the hold-out data point $d \in D^*$. In order to alleviate the computational complexity of performing LOO cross-validation, [26] proposed PSIS to estimate the predictive performance. For each one of the test examples $(x_i, y_i) \in D^*$, PSIS computes a smoothed importance sampling estimate from the existing posterior samples and fits a generalized Pareto distribution.

$$p(y_i|x_i, \mathcal{D}) \approx \frac{\sum_k p(y_k|\theta^k)w_i^k}{\sum_k w_i^k} \tag{5}$$

where $w_i^k = \frac{1}{p(y_i|\theta^k)} \propto \frac{p(\theta^k|y_{-i})}{p(\theta^k|y_i)}$.

In general, the importance weights w_i^k tend to have large or infinite variance. Therefore, the PSIS diagnostic computes a shape parameter $\hat{k}$ that regularizes the raw ratios. For Pareto values $\hat{k} < 0.5$, the predictive performance is guaranteed to be highly accurate. The values $0.5 \le \hat{k} < 0.7$ represent numerically stable, but inaccurate predictions and Pareto shape values $0.7 \le \hat{k}$ imply infinite variance. Here, we propose a simple diagnostic tool to correct the over-optimistic performance of the point estimates.

The area under the receiver operating characteristic curve (AUC) is a common performance measure for diagnosing performance of binary classifiers. However, its evaluation includes every possible decision threshold, including unrealistic ones. This choice makes the AUC too general and less informative. In the other hand, the F_1 score corresponds to the harmonic mean of precision and recall. Precision is a measure of the fraction of the detections $\hat{y}$ that are positive $p(y = 1|\hat{y} = 1)$ and recall measures the proportion of positive labels that were detected

$p(\hat{y} = 1|y = 1)$. As opposed to the AUC score, the dependence of the F_1 score on a single threshold makes it too specific.

Now, we derive F_{loo} as a weighted alternative to the F_1 score. Unlike the F_1 measure, the F_{loo} threshold is derived from a set of samples and automatically avoids unreliable test examples.

$$F_{\text{loo}} \equiv \frac{2\sum_i \hat{k}_{\text{loo}} y_i \hat{y}_i}{\sum_i \hat{y}_i + \sum_i \hat{y}_i} \tag{6}$$

where $\hat{k}_{\text{loo}} = 1 - \text{MAX}(\text{MIN}(\hat{k}_i, 1), 0)$

For multi-class problems, the proposed F_{loo} measure can be generalized using macro- and micro-averages for each one of the class instances.

4 Results and discussion

The experiments consider two deep learning models evaluated on a COVID-19 X-ray dataset.

4.1 Data

The dataset consists of X-ray images collected for COVID-19 positive along with normal, lung opacity and viral pneumonia cases and made publicly available from Kaggle.[1] The data examples were collected from multiple online sources and were made freely available for research purposes. Figure 2 shows one example per category from the database.

All images are grayscale, 299 × 299 pixels and stored using the PNG format. The dataset was collected from multiple online sources and may contain duplicated examples due to data augmentation or simple replication. There is no clear indication on whether any two particular examples come from the same person, which potentially makes the data non-independent nor identically distributed. Figure 3 shows the number of examples per category (COVID, lung opacity, normal and viral pneumonia) from the Kaggle dataset.

The dataset is randomly split using 80% of the data for training and /20% for testing purposes.

4.2 Deep learning models

Deep convolutional architectures are neural networks whose hidden layers apply convolution transformations to their inputs. In the case for 2D convolutions, these transformations have been successfully used to extract high and low-level features from images. Therefore, convolutional neural networks can be used to train image classification models.

In this paper, we consider two different convolutional architectures and two different variants for each one of them. The first one is the ResNet architecture, which is a deep convolutional neural network that contains residual connections to avoid the gradient to vanish. Residual connections propagate noiseless versions of the data before applying any transformation and therefore enable more stable gradient computations. Figure 4 shows and schematic of the residual block model behind the ResNet architecture.

Another popular technique that has shown good performance in deep neural networks is batch normalization. Standard data normalization is used to transform the original data to improve the model accuracy. Conversely, batch normalization is applied to the weights of the hidden layers and has shown to improve the model generalization. Batch normalization computes a running mean and variance of the current batch, which is used to normalize samples. Both, residual blocks and batch normalization have been successfully used to train the ResNet architecture for large-scale problems such as the ImageNet challenge. Batch normalization introduces data leakage that makes the likelihood principle difficult to interpret. Separable blocks are other type of operator that apply independent spatial (2D) (depthwise) convolutions to each one of the channels, before applying a pointwise convolution over all inputs. In practice, depth separable convolutions have fewer parameters than their plain convolutional counterpart and have also been successfully implemented for large-scale image classification tasks where the goal is also to perform inference in edge devices. MobileNet is a deep learning architecture that employs depthwise and pointwise separable convolutions. The MobileNet architecture is usually implemented using several depthwise separable convolutional blocks using a multiplier parameter that controls the actual number of channels per layer and batch normalization. In this case, fine-tuning is also implemented using pre-trained models from the ImageNet database. Figure 5 shows the depthwise separable block used in the MobileNet architecture.

In our experiments, we use pre-trained variants of ResNet with 18 and 50 layers (ResNet18v2 and ResNet50v2). For MobileNet, we also consider two variants of pre-trained models with multiplier parameter $\alpha = \{0.25, 1.0\}$ (MobileNetV2_0.25 and MobileNetV2_1.0).

[1] https://www.kaggle.com/tawsifurrahman/covid19-radiography-database.

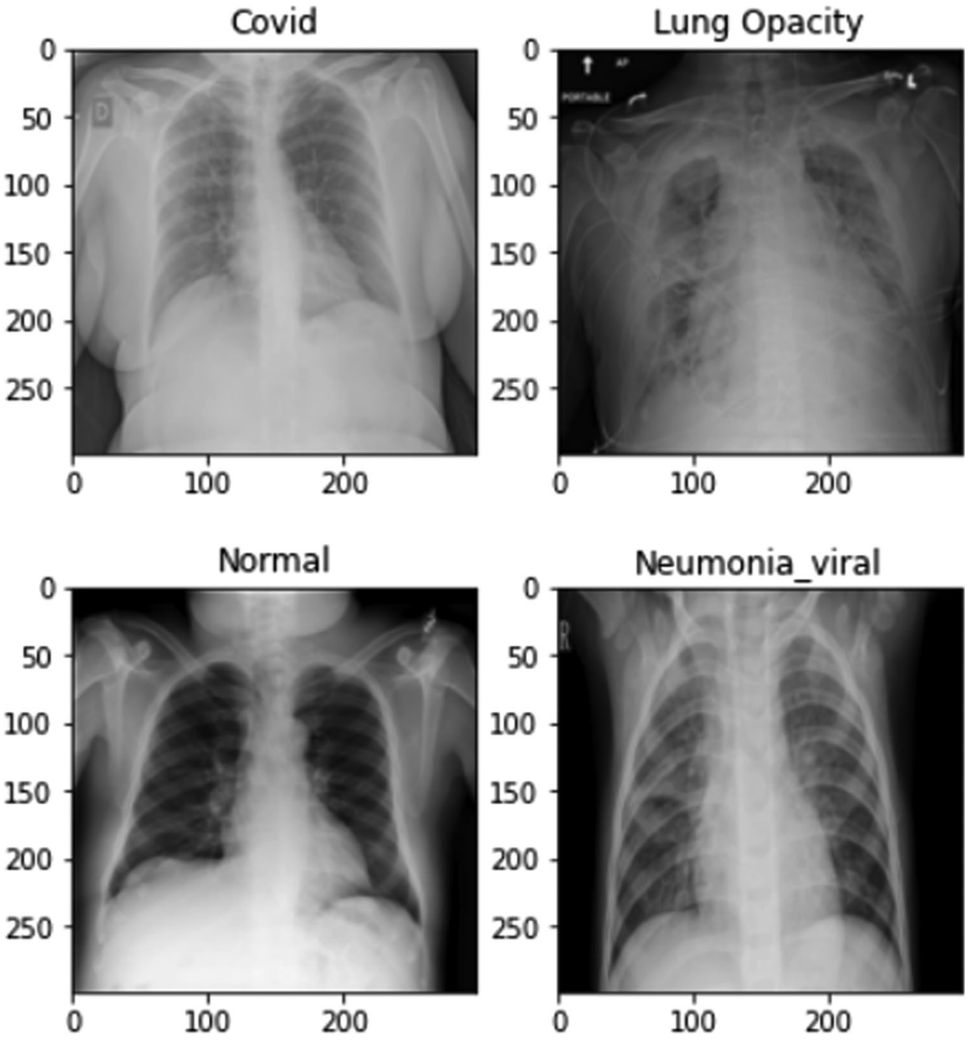


Fig. 2 Chest X-ray images from the Kaggle COVID-19 database

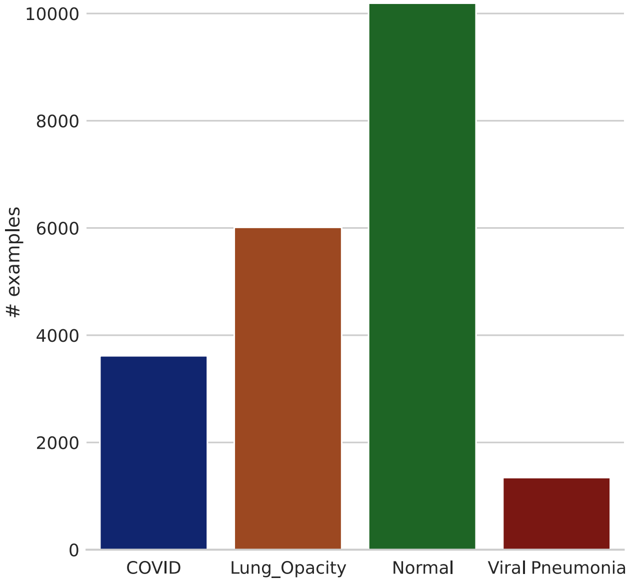


Fig. 3 Class distribution from the Kaggle COVID-19 database

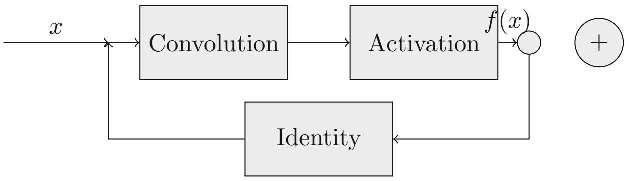


Fig. 4 Residual block from the ResNet architecture

5 Baseline performance

In order to estimate the predictive uncertainty, pre-trained versions of the ResNet and MobileNet models are fine-tuned using transfer learning. In all cases, the output layer is replaced to classify test images into the four new categories (COVID, lung opacity, normal and viral pneumonia). The fine-tuned models are trained using the stochastic gradient descent (SGD) where the learning rates are obtained using the hyper-parameter optimization (HPO) tuning found in AutoGluon. The entire training and pipelines were implemented as Python scripts executed in a

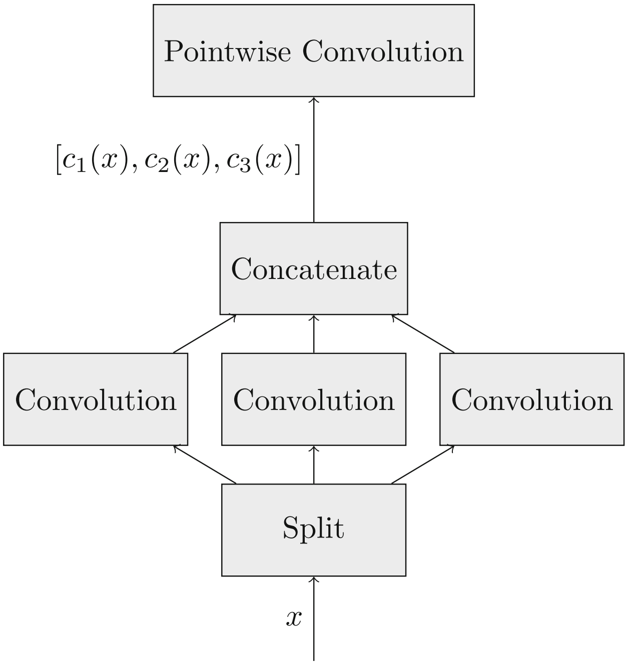


Fig. 5 Separable block from the MobileNet architecture

Table 1 AutoML hyper-parameter optimization settings

Model name	Learning rate (logarithmic)
ResNet18v2	[1e-5,1e-2]
ResNet50v2	[1e-5,1e-2]
MobileNetV2_0.25	[1e-5,1e-2]
MobileNetV2_1.0	[1e-5,1e-2]

Linux machine with a Intel Core I7-5930K CPU and an NVIDIA RTX 3080 GPU. Each one of the deep learning models is tuned with AutoGluon with a limited budget, measured in wall clock time. The search presets can be seen in Table 1.

Having obtained the hyper-parameters for each one of the models, SGD optimization is run for a fixed number of epochs ($n_{\text{epochs}} = 100$) and batch-size $\mathcal{B} = 16$. Data augmentation is also used to increase the dataset size. The image pre-processing steps during training include, data normalization, random resize (256×256) pixels and crop to 224×224 pixels and random left/right flips. For testing, data augmentation only includes center crop (224×224) and normalization. After training, performance is measured in the test dataset using the precision $P = \frac{\text{TP}}{\text{TP+FP}}$, recall $R = \frac{\text{TP}}{\text{TP+FN}}$ and $F_1 = 2\frac{P \times R}{P+R}$ metrics, where TP is the number of true positives, FP is the number of false positives, and FN is the number of false negatives.

ResNet18_v2 achieves the lowest performance for the COVID class (0.62 precision), which is improved when increasing the number of layers in the ResNet50_v2 layer (0.90 precision). Therefore, when the number of layers is increased, the false negatives and false positives rates are also reduced for this class. However, the number of false positives is increased for the viral pneumonia class in the larger ResNet50_v2 model, whose precision drops from 0.92 to 0.89 (Table 2).

The MobileNetv2_1.0 architecture achieves the highest performance on the precision metric but also a higher number of false negatives. The smaller-sized MobileNetv2_0.25 model achieves a good balance between precision and recall (as seen in the F_1 metric) along all four different classes. Now, we focus on the uncertainty quantification task. As already mentioned, data augmentation introduces a data leakage that cannot be interpreted using the likelihood function $f(\theta)$ (see Eq. 1).

The performance of the SGLD algorithm for each one of the models is lower than their SGD counterpart. Instead of estimating predictive accuracy from a single point estimate (such as the maximum a posteriori estimate), the Bayesian approach uses an approximate posterior density $p(\theta|D)$ from the SGLD samples. However, the posterior predictive accuracy tends to be lower than the point estimates [28]. Table 3 reports the predictive accuracy of SGLD for all deep learning models considered.

The prior for all models was an isotropic Gaussian $\mathcal{N}(0, \alpha^2 I)$, and the scale parameter was set to $\alpha^2 = 100$. This particular choice has been criticized in the literature as being inadequate and Wenzel et al. [28] proposed a posterior tempering technique [10]. In [18], the authors shown that vague priors (such as $\alpha^2 = 100$) lead to useful uncertainty estimates in function-space as measured with the $\hat{R}$ statistic. The potential scale reduction factor is a measure for the ratio of the average variance of samples to the pooled samples across different MCMC chains. Figure 6 shows the $\hat{R}$ statistic for the output layer (function-space) and the internal layers (weight-space) for all different models.

Now, we focus on the effective sample sizes for each one of the runs. Figure 7 shows the $\hat{N}_{\text{eff}}$ statistic for each one of the model runs. As opposed to the $\hat{R}$ statistic, we now see most of the MCMC runs having small sample sizes. Both ResNet models (ResNet50_v2 and ResNet18_v2 in Fig. 7a and b, respectively) show larger samples sizes in their internal layers when compared to the MobileNet models.

Table 2 Baseline performance for the X-ray COVID prediction dataset

Model name	Class	Precision	Recall	F_1 Score	Support
ResNet50_v2	COVID	0.90	0.95	0.92	724
	Lung opacity	0.93	0.93	0.93	1203
	Normal	0.95	0.92	0.93	2039
	Viral pneumonia	0.89	0.97	0.93	269
ResNet18_v2	COVID	0.62	0.86	0.72	724
	Lung opacity	0.89	0.83	0.86	1203
	Normal	0.90	0.81	0.85	2039
	Viral pneumonia	0.92	0.91	0.92	269
MobileNetv2_1.0	COVID	0.98	0.86	0.92	724
	Lung opacity	0.92	0.97	0.89	1203
	Normal	0.95	0.89	0.93	2039
	Viral pneumonia	0.96	0.91	0.94	269
MobileNetv2_0.25	COVID	0.95	0.97	0.96	724
	Lung opacity	0.92	0.91	0.92	1203
	Normal	0.95	0.94	0.94	2039
	Viral pneumonia	0.96	0.96	0.96	269

Table 3 SGLD performance for the X-ray COVID prediction dataset

Model name	Class	Precision	Recall	F_1 Score	Support
ResNet50_v2	COVID	0.29	0.95	0.45	724
	Lung opacity	0.59	0.45	0.51	1203
	Normal	0.96	0.26	0.41	2039
	Viral pneumonia	0.39	0.62	0.48	269
ResNet18_v2	COVID	0.20	0.97	0.33	724
	Lung opacity	0.66	0.11	0.19	1203
	Normal	0.77	0.02	0.03	2039
	Viral pneumonia	0.36	0.66	0.47	269
MobileNetv2_1.0	COVID	0.98	0.97	0.98	724
	Lung opacity	0.92	0.90	0.92	1203
	Normal	0.93	0.96	0.94	2039
	Viral pneumonia	0.98	0.95	0.97	269
MobileNetv2_0.25	COVID	0.99	0.98	0.98	724
	Lung opacity	0.94	0.91	0.93	1203
	Normal	0.94	0.97	0.95	2039
	Viral pneumonia	0.99	0.97	0.98	269

5.1 Evaluating predictive accuracy

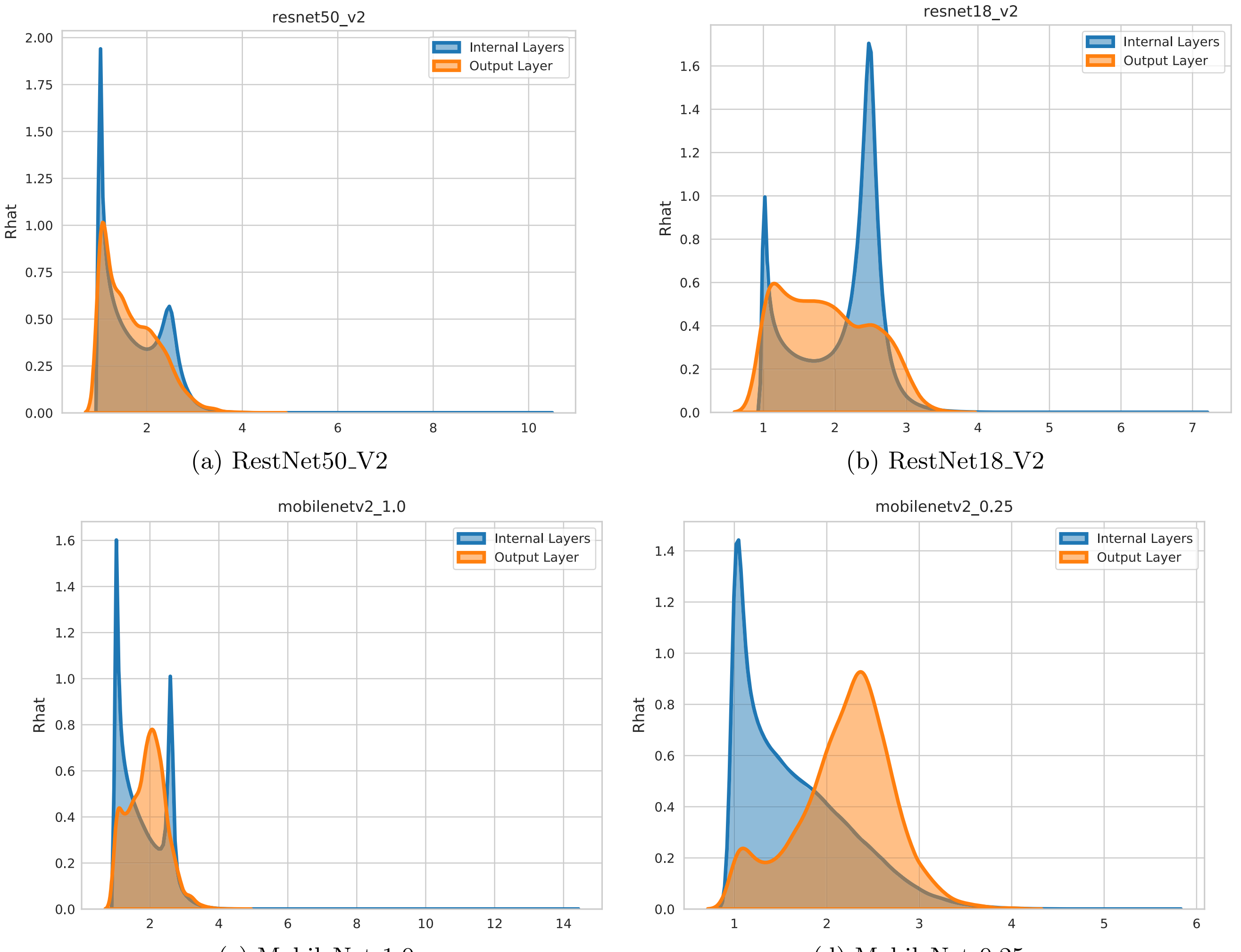
Traditionally, the performance of a model is measured using the out-of-sample predictive accuracy. This metric is useful when there is enough labeled data, so we can approximate the true data-generating process. This predictive distribution is not known, and therefore, we must approximate techniques to provide an estimate of the model accuracy.

$$\sum_i \log \int p(y_i|\theta)p(\theta|\mathcal{D})\mathrm{d}\theta \tag{7}$$

With SGLD, we obtained a finite set of samples whose effective sample size is usually smaller than the actual number of samples. Nevertheless, the actual out-of-sample predictions on unseen data (X_*) can use the full posterior distribution.

$$p(y_*|X_*) = \int p(y_*|x_*, \theta)p(\theta|\mathcal{D})\mathrm{d}\theta \tag{8}$$

These predictions can take into account the log-scoring rule (marginal likelihood of the model), although they could be biased toward the maximum aposteriori estimate. Therefore, different scoring weights w_i to evaluate predictive accuracy can be extracted from the existing posterior samples.



(a) RestNet50_V2

(b) RestNet18_V2

(c) MobileNet_1.0

(d) MobileNet_0.25

Fig. 6 $\hat{R}$ estimates from SGLD samples. The weights from the output layer tend to obtain smaller

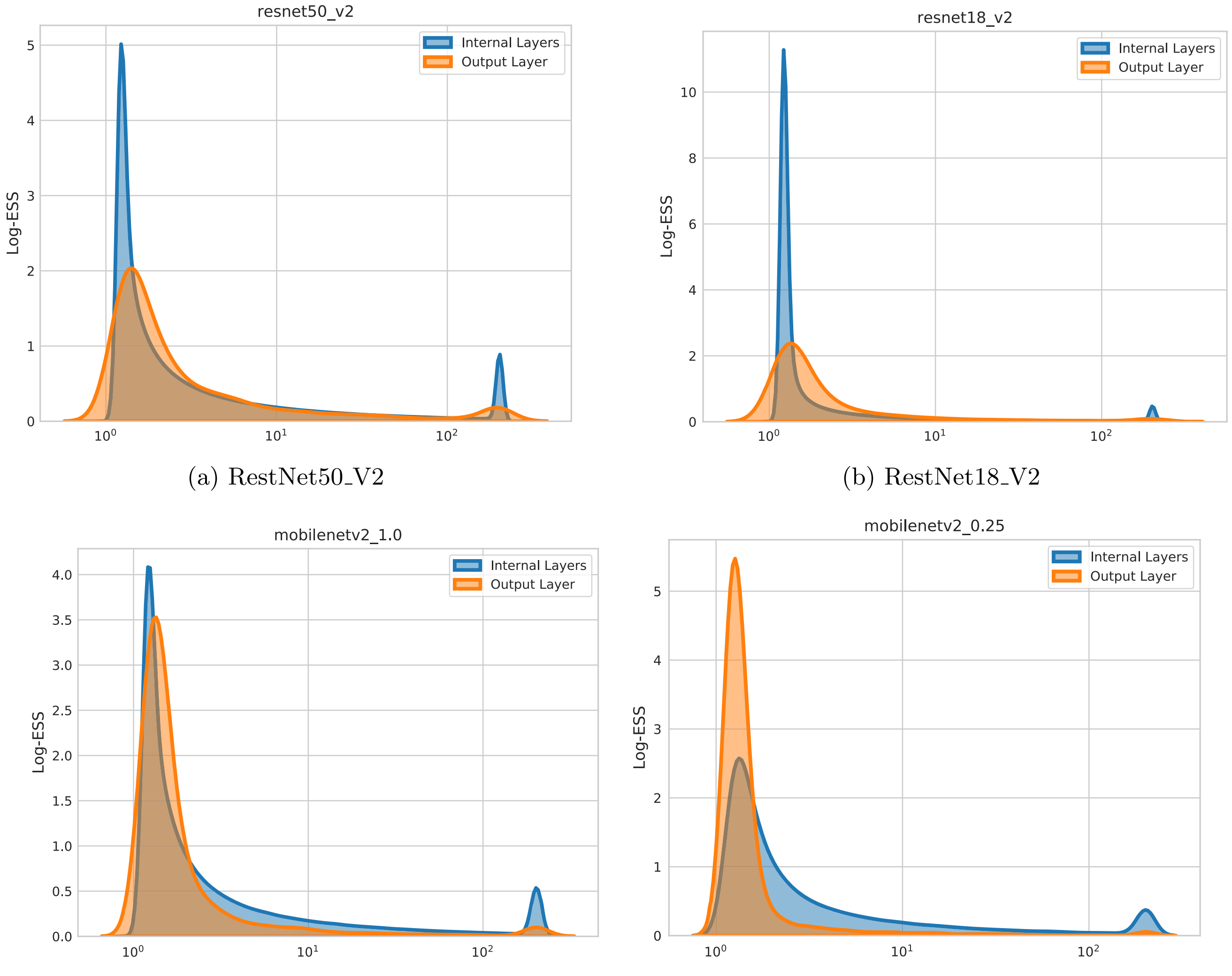
$$p(y_*|X_*) = \sum_i w_i p(y_*|x_*, \theta_i) \tag{9}$$

Now, we compare two different scoring rules. The first method is based on a popular ensembling technique called stacking. This method uses hold-out data to estimate the mixing weights w and has been successfully applied to improve predictive accuracy when the models are misspecified [29].

Alternatively, we also calculate predictive accuracy using PSIS-LOO $\hat{k}_{\text{loo}}$ as a scoring rule. In this case, there is no need for re-training the mixing weights. However, PSIS-LOO as a scoring rule automatically discards test samples far from the full distribution. As already mentioned, stacking is able to improve the model accuracy when the models are poorly specified (e.g., ResNet50_V2 and ResNet18_V2) and even improve the best-performing models (such as MobileNet_1.0 and MobileNet_0.25). PSIS-LOO in the other produces less confident predictive accuracy, decreasing the F-measure to zero for the COVID and viral pneumonia classes. Table 4 shows the predictive accuracy for both scoring functions.

Stacking is able to improve the precision and recall of the SGLD output. The ensemble technique requires an additional training step that takes a subset of the testing dataset and learns the mixing weights. Instead, the F_{loo} metric heavily penalizes both ResNet18_V2 and ResNet50_V2 architectures. F_{loo} based on PSIS-LOO does not perform re-training, so it can be seen as being more data efficient.

Also, while stacking improves the precision and recall across all classes, F_{loo} does not show any improvement and even worsens confidence on certain classes. The decrease in performance can be seen for the COVID-19 and viral pneumonia classes predicted with both ResNet models. The same effect is also achieved for lung opacity and viral pneumonia classes being predicted with MobileNet. The observed drop in performance is consistent with



(a) RestNet50_V2

(b) RestNet18_V2

(c) MobileNet_1.0

(d) MobileNet_0.25

Fig. 7 Effective sample size from SGLD

the results obtained by DeGrave et al. [8] who reported an area under the curve (AUC) of 0.76 and 0.70 when the model is tested using an external COVID-19 X-ray dataset. The authors argue that poor performance and the generalization gap can be attributed to models learning spurious correlations. By contrast, the F_{loo} metric accounts for such lack of certainty for predicting specific classes without the need of re-training or testing with another dataset.

6 Conclusion

Ensemble techniques allow to estimate model uncertainty, but there are no guarantees about the quality of the predictive distribution. Therefore, in this paper, we presented a novel method to quantify predictive uncertainty for COVID-19 detection from X-ray images. Stochastic-gradient MCMC techniques using the F_{loo} metric allow to estimate overconfidence on the model predictions. The method is able to evaluate models without re-training or testing with an additional dataset.

The results show a significant gap in accuracy from training and testing from fine-tuning a pre-trained image classifier to deliver reliable predictions for COVID-19. Firstly, it is difficult to obtain a large number of posterior samples with low auto-correlation. Secondly, it is also hard to evaluate predictive performance when the samples are biased and the predictions being overconfident.

In this study, a single dataset was used for model training and validation. Stacking posterior samples were shown to improve predictive accuracy and were then compared to F_{loo}. Future work will consider external validation with related datasets. Also, additional evaluation metrics could also be considered in order to gain a better

Table 4 Leave-one-out performance for the X-ray COVID prediction dataset. Stacking is trained on the SGLD output probabilities using a subset of the testing examples. The F_{loo} metric uses a weighting scheme based on the PSIS-LOO $\hat{k}$ parameter

Model name	Method	Class	Precision	Recall	F_1 Score	Support
ResNet50V_ 2	Stacking	COVID	0.67	0.79	0.73	195
		Lung opacity	0.85	0.78	0.81	322
		Normal	0.85	0.84	0.85	478
		Viral pneumonia	0.87	0.83	0.85	64
	F_{loo}	COVID	0.0	0.03	0.00	5
		Lung opacity	0.25	0.14	0.18	269
		Normal	0.85	0.15	0.26	394
		Viral pneumonia	0.0	0.0	0.0	2
ResNet18V_2	Stacking	COVID	0.67	0.34	0.45	195
		Lung opacity	0.74	0.73	0.74	322
		Normal	0.71	0.86	0.78	478
		Viral pneumonia	0.77	0.75	0.76	64
	F_{loo}	COVID	0.03	0.98	0.13	33
		Lung opacity	0.20	0.06	0.09	65
		Normal	0.67	0.0	0.01	486
		Viral pneumonia	0.0	1.0	0.0	0
MobileNetv2_1.0	Stacking	COVID	0.98	0.97	0.98	724
		Lung opacity	0.92	0.90	0.91	1203
		Normal	0.93	0.96	0.94	2039
		Viral pneumonia	0.98	0.95	0.97	269
	F_{loo}	COVID	0.82	0.70	0.76	20
		Lung opacity	0.24	0.17	0.20	43
		Normal	0.05	0.09	0.06	25
		Viral pneumonia	0.76	0.45	0.57	7
MobileNetv2_0.25	Stacking	COVID	0.99	0.98	0.98	195
		Lung opacity	0.96	0.90	0.93	322
		Normal	0.92	0.97	0.95	478
		Viral pneumonia	0.97	0.95	0.98	64
	F_{loo}	COVID	0.89	0.74	0.81	11
		Lung opacity	0.21	0.17	0.19	25
		Normal	0.08	0.11	0.09	20
		Viral pneumonia	0.90	0.90	0.90	10

perspective of the quality of the posterior samples and the model ability to generalize.

Data Availability The dataset analyzed during the current study is available in the COVID-19 Radiography Database repository, https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database.

Declarations

Conflict of interest This work was supported by Catholic University of Maule internal research grant UCM-IN-21202.

References

1. Abdar M, Pourpanah F, Hussain S et al (2021) A review of uncertainty quantification in deep learning: techniques, applications and challenges. Inf Fusion 76:243–297. https://doi.org/10.1016/j.inffus.2021.05.008
2. Alghamdi HS, Amoudi G, Elhag S et al (2021) Deep learning approaches for detecting Covid-19 from chest X-ray images: a survey. IEEE Access 9:20235–20254. https://doi.org/10.1109/ACCESS.2021.3054484
3. Arias-Londoño JD, Gómez-García JA, Moro-Velázquez L et al (2020) Artificial intelligence applied to chest X-ray images for the automatic detection of Covid-19. A thoughtful evaluation approach. IEEE Access 8:226811–226827. https://doi.org/10.1109/ACCESS.2020.3044858
4. Asgharnezhad H, Shamsi A, Alizadehsani R et al (2022) Objective evaluation of deep uncertainty predictions for Covid-19 detection. Sci Rep 12(1):815. https://doi.org/10.1038/s41598-022-05052-x

5. Begoli E, Bhattacharya T, Kusnezov D (2019) The need for uncertainty quantification in machine-assisted medical decision making. Nat Mach Intell 1(1):20–23
6. Chowdhury ME, Rahman T, Khandakar A et al (2020) Can AI help in screening viral and Covid-19 pneumonia. IEEE Access 8:132665–132676
7. Combalia M, Hueto F, Puig S, et al (2020) Uncertainty estimation in deep neural networks for dermoscopic image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops
8. DeGrave AJ, Janizek JD, Lee SI (2021) Ai for radiographic Covid-19 detection selects shortcuts over signal. Nat Mach Intell 3(7):610–619. https://doi.org/10.1038/s42256-021-00338-7
9. Deng J, Dong W, Socher R, et al (2009) Imagenet: a large-scale hierarchical image database
10. Fortuin V (2022) Priors in bayesian deep learning: a review. Int Stat Rev
11. Gal Y, Ghahramani Z (2016) Dropout as a bayesian approximation: representing model uncertainty in deep learning. In: Balcan MF, Weinberger KQ (eds) Proceedings of the 33rd International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 48. PMLR, New York, USA, pp 1050–1059
12. Garcia Santa Cruz B, Bossa MN, Sölter J et al (2021) Public Covid-19 X-ray datasets and their impact on model bias—a systematic review of a significant problem. Med Image Anal 74(102):225. https://doi.org/10.1016/j.media.2021.102225
13. Ghoshal B, Tucker A (2021) On cost-sensitive calibrated uncertainty in deep learning: an application on Covid-19 detection. In: 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS), pp 503–509
14. Ghoshal B, Tucker A, Sanghera B et al (2021) Estimating uncertainty in deep learning for reporting confidence to clinicians in medical image segmentation and diseases detection. Comput Intell 37(2):701–734. https://doi.org/10.1111/coin.12411
15. Gour M, Jain S (2022) Uncertainty-aware convolutional neural network for Covid-19 X-ray images classification. Comput Biol Med 140(105):047. https://doi.org/10.1016/j.compbiomed.2021.105047
16. Holzinger A (2021) The next frontier: Ai we can really trust. In: Kamp M, Koprinska I, Bibal A, et al (eds) Machine Learning and Principles and Practice of Knowledge Discovery in Databases, International Workshops of ECML PKDD 2021, Proceedings. Springer, Communications in Computer and Information Science, pp 427–440
17. Islam MM, Karray F, Alhajj R et al (2021) A review on deep learning techniques for the diagnosis of novel coronavirus (Covid-19). IEEE Access 9:30551–30572
18. Izmailov P, Vikram S, Hoffman MD, et al (2021) What are bayesian neural network posteriors really like? In: Meila M, Zhang T (eds) Proceedings of the 38th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 139. PMLR, pp 4629–4640
19. Kadane JB (2020) Principles of uncertainty. Chapman and Hall/CRC
20. Khushi M, Shaukat K, Alam TM et al (2021) A comparative performance analysis of data resampling methods on imbalance medical data. IEEE Access 9:109960–109975. https://doi.org/10.1109/ACCESS.2021.3102399
21. Maguolo G, Nanni L (2021) A critic evaluation of methods for Covid-19 automatic detection from X-ray images. Inf Fusion 76:1–7. https://doi.org/10.1016/j.inffus.2021.04.008
22. Mahmud T, Rahman MA, Fattah SA (2020) Covxnet: a multi-dilation convolutional neural network for automatic Covid-19 and other pneumonia detection from chest X-ray images with transferable multi-receptive feature optimization. Comput Biol Med 122(103):869. https://doi.org/10.1016/j.compbiomed.2020.103869
23. Murphy K, Smits H, Knoops AJG et al (2020) COVID-19 on chest radiographs: a multireader evaluation of an artificial intelligence system. Radiology 296(3):E166–E172. https://doi.org/10.1148/radiol.2020201874
24. Narin A, Kaya C, Pamuk Z (2021) Automatic detection of coronavirus disease (Covid-19) using X-ray images and deep convolutional neural networks. Pattern Anal Appl 24:1207–1220
25. Nemeth C, Fearnhead P (2021) Stochastic gradient Markov chain Monte Carlo. J Am Stat Assoc 116(533):433–450. https://doi.org/10.1080/01621459.2020.1847120
26. Vehtari A, Gelman A, Gabry J (2017) Practical Bayesian model evaluation using Leave-one-out Cross-validation and waic. Stat Comput 27(5):1413–1432
27. Wang L, Lin ZQ, Wong A (2020) Covid-net: a tailored deep convolutional neural network design for detection of Covid-19 cases from chest X-ray images. Sci Rep 10(1):1–12
28. Wenzel F, Roth K, Veeling B, et al (2020) How good is the bayes posterior in deep neural networks really? In: International Conference on Machine Learning, PMLR, pp 10248–10259
29. Yao Y, Vehtari A, Simpson D et al (2018) Using stacking to average Bayesian predictive distributions (with discussion). Bayesian Anal 13(3):917–1007. https://doi.org/10.1214/17-BA1091

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.