



OPEN

Objective evaluation of deep uncertainty predictions for COVID-19 detection

Hamzeh Asgharnezhad^{1,6}, Afshar Shamsi^{1,6}✉, Roohallah Alizadehsani², Abbas Khosravi², Saeid Nahavandi², Zahra Alizadeh Sani³, Dipti Srinivasan⁴ & Sheikh Mohammed Shariful Islam⁵

Deep neural networks (DNNs) have been widely applied for detecting COVID-19 in medical images. Existing studies mainly apply transfer learning and other data representation strategies to generate accurate point estimates. The generalization power of these networks is always questionable due to being developed using small datasets and failing to report their predictive confidence. Quantifying uncertainties associated with DNN predictions is a prerequisite for their trusted deployment in medical settings. Here we apply and evaluate three uncertainty quantification techniques for COVID-19 detection using chest X-Ray (CXR) images. The novel concept of uncertainty confusion matrix is proposed and new performance metrics for the objective evaluation of uncertainty estimates are introduced. Through comprehensive experiments, it is shown that networks pertained on CXR images outperform networks pretrained on natural image datasets such as ImageNet. Qualitatively and quantitatively evaluations also reveal that the predictive uncertainty estimates are statistically higher for erroneous predictions than correct predictions. Accordingly, uncertainty quantification methods are capable of flagging risky predictions with high uncertainty estimates. We also observe that ensemble methods more reliably capture uncertainties during the inference. DNN-based solutions for COVID-19 detection have been mainly proposed without any principled mechanism for risk mitigation. Previous studies have mainly focused on on generating single-valued predictions using pretrained DNNs. In this paper, we comprehensively apply and comparatively evaluate three uncertainty quantification techniques for COVID-19 detection using chest X-Ray images. The novel concept of uncertainty confusion matrix is proposed and new performance metrics for the objective evaluation of uncertainty estimates are introduced for the first time. Using these new uncertainty performance metrics, we quantitatively demonstrate when we could trust DNN predictions for COVID-19 detection from chest X-rays. It is important to note the proposed novel uncertainty evaluation metrics are generic and could be applied for evaluation of probabilistic forecasts in all classification problems.

The COVID-19 pandemic has greatly increased the demand for fast and reliable screening of suspected cases. Real-time reverse transcription-polymerase chain reaction is the gold standard for the COVID-19 detection. While this diagnostic test has a high accuracy and sensitivity, it is time consuming, resource intensive, and expensive. These shortcomings and the rising positivity rates has led to a real need for auxiliary diagnostic tools which are fast, affordable, and available at scale. Common radiology images such as CXR or computed tomography (CT) contain salient information and visual indexes correlated with the COVID-19 infections¹. Accordingly, the detection can be inferred from these modalities of suspected individuals suffering from COVID-19 symptoms.

Several studies have been conducted since the onset of COVID-19 pandemic to automate its detection from chest radiology images using artificial intelligence techniques^{2,3}. Deep neural networks (DNNs) and transfer learning⁴ have been widely applied for this purpose due to their promising human-level or super-human level performance in object recognition tasks⁵⁻⁸.

DNN-based solutions for COVID-19 detection have been mainly proposed without any principled mechanism for risk mitigation. The focus of existing literature is mainly on generating single-valued predictions using

¹Individual researcher, Tehran, Iran. ²Institute for Intelligent Systems Research and Innovation (IISRI), Deakin University, Melbourne, VIC, Australia. ³Omid Hospital, Iran University of Medical Sciences, Tehran, Iran. ⁴Department of Electrical and Computer Engineering, National University of Singapore, Singapore, Singapore. ⁵Institute of Physical Activity and Nutrition, Deakin University, Melbourne, VIC, Australia. ⁶These authors contributed equally: Hamzeh Asgharnezhad and Afshar Shamsi. ✉email: afshar.shamsi.j@gmail.com

pretrained DNNs⁵. Proposed solutions are evaluated by point prediction-based performance metrics such as accuracy, sensitivity, specificity, and area under receiver operating characteristic (AUC). It is important to note that the transition from normal to COVID-19 is not always clear-cut. Difficult-to-diagnosis cases from radiology images could even lead to disagreement between experienced medical doctors. In fact, studies have shown that radiologists disagree with their colleagues 25% of the time and themselves 20% of the time⁹. As the model decision has a direct impact on patient's treatment, it is of critical to know how confident DNNs are about their predictions. This information could be used for identifying patients which may best benefit from a medical second opinion. DNNs flagging potentially erroneous predictions due to high uncertainty can be used to mimic the common practice of requesting a second opinion from another health practitioner in medial settings¹⁰. Such an uncertainty-aware decision-making pipeline could greatly improve the overall diagnosis performance.

In this paper, we comprehensively and quantitatively investigate the competency of DNNs for generating reliable uncertainty estimates for COVID-19 diagnosis. We first check the impact of pretraining using ImageNet and CXR image datasets on the network performance. Then MC-dropout (MCD), ensemble, and ensemble MC-dropout (EMCD) are implanted for quantifying uncertainties associated with point predictions of DNNs. Motivated by^{11,12}, we introduce a novel uncertainty-aware confusion matrix from which novel performance metrics (Uncertainty sensitivity, Uncertainty recall, Uncertainty specificity, and Uncertainty precision) can be computed. These performance metrics can be used for comprehensive and quantitative evaluation of uncertainty estimates. The uncertainty estimate evaluation is conducted in a similar manner to that of binary classification evaluation. Through experiments, we try to shed light on whether uncertainty quantification methods proposed in literature can provide high uncertainty for erroneous predictions. This investigation is done both qualitatively (visually) and quantitatively (using new uncertainty evaluation metrics). The proposed metrics can be used for proper evaluation of uncertainty generated by different types of machine learning models. As these have a high similarity to the traditional confusion matrix, users can easily understand and apply them in the process of fair evaluation. At the same time, the proposed metrics can be used for engineering novel uncertainty-aware training algorithms for neural networks. These uncertainty-aware algorithms not only improve the model accuracy, but also take care of its uncertainty estimates. This will result in more trustworthy models that can detect whether their prediction is reliable or not.

The rest of this paper is organised as follows. Section “[Related work](#)” briefly reviews papers reporting applications of deep uncertainty quantification for COVID-19 diagnosis. Uncertainty quantification techniques are described in section “[Uncertainty quantification techniques](#)”. Section “[Predictive uncertainty evaluation](#)” introduces metrics for quantitative evaluation of predictive uncertainty estimates. The dataset and experiments are described in sections “[Dataset](#)” and “[Experiments](#)” respectively. Section “[Simulations and results](#)” reports conducted simulations and obtained results. Finally, section “[Conclusion](#)” concludes the paper.

Related work

Several methods have been proposed in recent years for enabling DNNs to encompass uncertainty and generate probabilistic predictions. Many of the proposed solutions are based on the Bayesian theory¹³. Several approximate methods have been proposed to address the intractability of the exact Bayesian inference due to its massive computational burden. This include but not limited to variational inference^{14,15}, MCD¹⁶, ensemble¹⁷. All these methods could be put in the two-step category of *uncertainty via classification* as they first train the model (classifier here) and then postprocess predictions to generate an uncertainty score¹⁰. There have been also attempts to generate uncertainty estimates without resorting to sampling methods. *Direct uncertainty prediction* methods train DNNs to directly generate uncertainty scores or distribution parameters in one single round of scoring^{10,18,19}. Despite made progress in this field, reliable generation of uncertainty estimates is still an open question and subject to further investigation.

There is an abundance of papers reporting applications of DNNs generating single-valued predictions for COVID-19 diagnosis from medical images^{5–8}. The input modalities are often chest X-ray and CT scan images which are processed using a wide variety of DNNs including pretrained networks for feature extraction, segmentation, and generative adversarial networks (GANs)²⁰. For example, Singh and Singh²¹ proposed an automated COVID-19 diagnosis method based on DNNs and Wavelet decomposition. Their model could classify input sample (i.e. patient X-Ray image) as Normal, Viral pneumonia, or COVID-19. Robust diagnostic tools are only part of the solution against COVID-19 pandemic. As pointed out by Pang et al.²², it is necessary to have a response plan for this pandemic. To this end, the concept of digital twin city and federated learning has been combined to achieve a collaborative paradigm in which multiple cities share their local status and strategies. Another study focused on the preventive measures to contain COVID-19 threat has been conducted in²³. The authors used correlation coefficients and multiple linear regression to determine key factors contributing to COVID-19 containment. Lockdown and social distancing have been reported as two key factors to suppress the spread rate of COVID-19.

A comprehensive review of medical imaging applications of DNNs for COVID-19 has been provided in⁵. While proposed solutions differ in terms of utilized networks and the task nature, they all focus on deterministic decisions generated by DNNs. This comprehensive review clearly shows that the literature is quite naive on applying deep uncertainty quantification techniques for processing COVID-19 datasets. Discussion about the reliability and confidence of proposed models has been often overlooked in these studies.

There are a few studies reporting the importance of predictive uncertainty estimates for reliable COVID-19 detection and growth rate monitoring. Alberti and Faranda²⁴ focused on statistical predictions of COVID-19 by fitting asymptotic distributions to actual data. They studied the epidemic evolution of COVID-19 infections in China and Italy. They reported that prediction about epidemic evolution is accompanied with high uncertainty at early stages of epidemic growth. As the epidemic reaches its peak, the uncertainty decreases. They found that

long term extrapolation of epidemics counts depend not only on the quality of data, but also on the stage of the epidemics. The MC-Dropweights method²⁵ was used in²⁶ to estimate uncertainties associated with COVID-19 predictions. It was shown that there is a strong correlation between model uncertainty and prediction correctness. The paper clearly highlights that availability of estimated uncertainties could potentially alert radiologists on false predictions and will accelerate the acceptance of deep learning-based solution in clinical practice. In another study²⁷, a variational inference with MC-Dropweights based Bayesian neural network model was proposed for COVID-19. The model could estimate cost-sensitive calibrated predictive uncertainty in the presence of asymmetric cost. In medical domain, the cost of different misclassification is different which makes the cost function asymmetric. Authors in²⁸ apply four DNNs pretrained on ImageNet dataset to process CXR and CT images. Extracted features are used to develop an ensemble of neural networks for epistemic uncertainty quantification. Obtained results clearly highlight the need for uncertainty estimate to build trust in DNNs for COVID-19 detection. That is why Alizadehsani et al.²⁹ proposed a semi-supervised approach for COVID-19 diagnosis using limited amount of labeled data. They utilized GAN to learn (unsupervised training) key features of CT images of COVID-19 patients. The trained GAN was then further trained in supervised manner from limited number of labeled CT images. The output of the discriminator network of GAN was then used for COVID-19 diagnosis (classification) based on CT images. They utilized the probabilistic output of the discriminator as some sort of uncertainty measure to reject classification of samples that the discriminator is uncertain about. RAMIREZ et al.³⁰ also took a semi-supervised approach to uncertainty estimation. They proposed a COVID-19 diagnosis system based on X-ray images with uncertainty estimates. To make use of unlabeled data, they utilized MixMatch (a semi-supervised deep learning method) which uses a hybrid loss function defined on labeled and unlabeled data. To compute the loss for unlabeled data, they were accompanied with pseudo-labels. Each pseudo-label was the average model output of a perturbed input sample.

In³¹, the choice of different cost functions to measure the uncertainty related to a segmentation task has been investigated. By using different architectures (MC dropout UNets, deep UNet ensembles, and conventional UNets without dropout), they tried to segment lung nodules on CT images using two different loss functions namely: weighted categorical cross-entropy (WCC) and soft Dice. In their study, more meaningful uncertainty information was achieved when they trained models by WCC loss function.

Authors in³² also propose a probabilistic generalization of the non-parametric KNN approach for developing a deep uncertainty-aware classifier. The proposed probabilistic neighbourhood component analysis method maps samples to probability distributions in a latent space and then minimizes a form of nearest-neighbour loss for developing classifiers. It is shown that the proposed method generates less overconfident predictions for out of distribution samples compared to common DNNs and Bayesian neural networks. Despite that, the paper does not provide any quantitative and qualitative evaluation about predictive uncertainty estimates for correctly classified and misclassified samples.

Missing values may also cause uncertain predictions. In³³, a two-phase approach is proposed for dynamic quality of service (QoS) prediction. In the first phase, if the user has invoked the service in the previous time slice, it will be used to predict QoS value for the user in the next time slice. If the user has not invoked the service in the previous time slice, covering algorithm is used to predict the missing value.

Uncertainty quantification techniques

For a Bayesian network, finding the posterior distribution is intractable for real-world problems (since there are hundreds/thousands of parameters) though there are some alternative ways such as different sampling methods but they are computationally costly. Thus some methods such as Monte Carlo Dropout, Ensemble Bayesian Networks, and Ensemble Monte Carlo Dropout were designed to approximate posterior distribution and could be applied to different model architectures with millions of parameters. That is the reason they are used for uncertainty quantification in this paper. These three methods are reviewed below.

MCD. The most difficult part of the Bayesian network is finding the posterior distribution. This is often computationally intractable. One way to overcome this drawback is to use sampling methods. Gal¹⁶ showed that MC samples of the posterior can be obtained by performing several stochastic forward passes at test time (keeping dropout on). The output posterior distribution could be approximated this way with minimum computational burden. The predictive mean (μ_{pred}) of the model for a typical test input over MC iterations is estimated as below:

$$\mu_{pred} \approx \frac{1}{T} \sum_t p(y = c|x, \hat{\omega}_t) \quad (1)$$

where x is the test input. $p(y = c|x, \hat{\omega}_t)$ is the probability that y belongs to c (the output of softmax), and $\hat{\omega}_t$ is the set of parameters of the model on the t^{th} forward pass. T is the number of MC iterations (forward passes). The variance of the final distribution is also called predictive uncertainty. As per¹⁶, the predictive entropy (PE) can be treated as the uncertainty estimate generated by the trained model:

$$PE = - \sum_c \mu_{pred} \log \mu_{pred} \quad (2)$$

where c ranges over both classes. The smaller the PE, the more confident the model about its predictions. It is note that, in the uncertainty literature, for the classification task, the entropy is used as a metric for quantifying that how much a prediction is related to each individual class. In other words, it helps us to quantify how a prediction is far from its true label. The pseudo-code of MCD is given in Algorithm 1.

Algorithm 1 MC Dropout**Input** : Labeled data (CXR images)
 $(x_i, y_i)_{i=1}^{522}$ **Output** : Finding PEs

```

for 100 Iterations do
  Randomly split Dataset to train
  and test
  Train the model
  for  $M$  iterations do
    Perform forward pass with
    dropout  $\hat{y}_M = model(x_M)$ 
  Calculate the PEs

```

Ensemble Bayesian networks. Ensemble networks are a group of networks working together for a specific task. Each network predicts a probability and the mean of probabilities will resemble the final predictive probability (posterior). The PE measure is also defined as³⁴:

$$\hat{p}(y|x) = \frac{1}{N} \sum_{i=1}^N p_{\theta_i}(y|x) \quad (3)$$

$$PE = \sum_{i=0}^C \hat{p}(y_i|x) \log \hat{p}(y_i|x) \quad (4)$$

where θ_i represents the set of parameters of i_{th} network element, and C ranges over two classes. The PE value is small when predictions from all individual networks are similar. The pseudo-code of EBN is given in Algorithm 2.

Algorithm 2 Ensemble**Input** : Labeled data (CXR images)
 $(x_i, y_i)_{i=1}^{522}$ **Output** : Finding PEs

```

for 100 Iterations do
  Randomly split Dataset to train
  and test
  for 1 to 30 do
    Generate the random models
    Train the models
    Average the outputs of all mod-
    els
  Calculate the PEs

```

EMCD. A combination of Ensemble networks and MCD algorithms produces EMCD. Here the ensemble is consist of DNNs with different architectures. The evaluation of each network is done using the MCD algorithm by performing several stochastic forward passes. A single Gaussian distribution will be estimated by averaging all posterior probabilities. For PE metric, the algorithm is similar to the ensemble and just differs in the way of finding the posterior:

$$\hat{p}(y|x) \approx \frac{1}{T} \sum_{t=1}^T p(\hat{y}|\hat{x}, \hat{\omega}_t) \quad (5)$$

$$PE = \sum_{i=0}^C \hat{p}(y_i|x) \log \hat{p}(y_i|x) \quad (6)$$

where $\hat{\omega}_t$ are the parameters of the model and C ranges over both classes. The pseudo-code of EMCD is given in Algorithm 3.

		Confidence	
		Certain	Uncertain
Correctness	Correct	True Certainty (TC)	False Uncertainty (FU)
	Incorrect	False Certainty (FC)	True Uncertainty (TU)

Figure 1. The uncertainty confusion matrix.

Algorithm 3 EMCD

Input : Labeled data (CXR images)
 $(x_i, y_i)_{i=1}^{522}$

Output : Finding PEs

```

for 100 Iterations do
  Randomly split Dataset to train
  and test
  for 1 to 30 do
    Generate the random models
    Train the models
    for M iterations do
      Perform forward pass with
      dropout for each model
    Average the outputs of all mod-
    els
  Calculate the PEs

```

Predictive uncertainty evaluation

Similar to the idea of confusion matrix, here we define quantitative performance metrics for predictive uncertainty estimates. In contrast to the research done in²⁶, the purpose is to do an objective and quantitative evaluation of the predictive uncertainty estimates. Predictions are first compared with ground truth labels and put into two groups: correct and incorrect. Predictive uncertainty estimates are also compared with a threshold and cast into two groups: certain and uncertain. The combination of correctness and confidence groups results in four possible outcomes as shown in Fig. 1: (i) correct and certain indicated true certainty (TC), (ii) incorrect and uncertain indicated by true uncertainty (TU), correct and uncertain indicated by false uncertainty (FU), and (iv) incorrect and certain indicated by false certainty (FC). TC and TU are the diagonal and favourite outcomes. These correspond to TN and TP outcomes in the traditional confusion matrix respectively. FU is a fortunate outcome as an uncertain prediction is correct. FC is the worst outcome as the network has confidently made an incorrect prediction.

According to these, we define multiple quantitative performance metrics to objectively quantify predictive uncertainty estimates:

- Uncertainty sensitivity (USen): USen is calculated as the number of incorrect and uncertain predictions divided by the total number of incorrect predictions:

$$USen = \frac{TU}{TU + FC}. \quad (7)$$

USen or uncertainty recall (URec) corresponds to sensitivity (recall) or true positive rate of the conventional confusion matrix. USen is of paramount importance as it quantifies the power of the model to communicate its confidence in misclassified samples.

- Uncertainty Specificity (USpe): USpe is calculated as the number of correct and certain predictions (TC) divided by the total number of correct predictions:

$$USpe = \frac{TC}{TC + FU}. \quad (8)$$

USpe or correct certain ratio is similar to the specificity performance metric.

- Uncertainty precision (UPre): UPre is calculated as the number of incorrect and uncertain predictions divided by the total number of uncertain predictions:

$$UPre = \frac{TU}{TU + FU}. \quad (9)$$

UPre has the same concept of precision in traditional binary classification.

- Uncertainty accuracy (UAcc): Similar to the accuracy of classifiers, the UAcc is calculated as the number of all diagonal outcomes divided by the total number of outcomes:

$$UAcc = \frac{TU + TC}{TU + TC + FU + FC}. \quad (10)$$

A reliable model will generate a high UAcc.

The best USen, USpe, UPre, and UAcc values are one, whereas the worst are zero. It is always desirable to have these metrics as close as possible to one. Having USen and USpe, and UPre close to one means that the network is self-aware of what it knows and what it does not know. Such a network can tell us when the user can trust its predictions as it reliably gauges and communicates its lack of confidence (as captured in predictive uncertainty estimates).

Dataset

In this study, CXR images sourced from two databases (COVID and Non-COVID) are used for model training and testing. It is noted that Institutional approval was granted for the use of the patient datasets in research studies for diagnostic and therapeutic purposes. Approval was granted on the grounds of existing datasets. Informed consent was obtained from all of the patients in this study. All methods were carried out in accordance with relevant guidelines and regulations. Ethical approval for the use of these data was obtained from the Tehran Omid hospital.

Non-COVID dataset. Cohen et al.³⁵ developed a CXR image database by combining seven existing non-covid datasets: RSNA Pneumonia Challenge³⁶, CheXpert—Stanford University³⁷, ChestX-ray8—National Institutes of Health (NIH)³⁸, ChestX-ray8—NIH with labels from Google³⁹, MIMIC-CXR—MIT⁴⁰, PadChest—University of Alicante⁴¹, and OpenI⁴². This dataset is mainly used for training a DenseNet121 network which will be later on used as a pretrained network for uncertainty-aware COVID-19 detection.

COVID-19 dataset. The main COVID-19 dataset used in this study for model development and evaluation contains 522 CXR images from 391 COVID-19 patients and 131 normal subjects (The dataset is collected in Iran). It is important to note that the normal class represents patients that did not have COVID-19. The term normal here does not imply that these patients do not have any emerging disease.

Experiments

The main COVID-19 dataset has a limited number of images. This makes developing reliable DNNs from scratch impractical. To address this issue, we develop the deep model in a transfer learning setting⁴. The common research practice is to pick an existing deep network pretrained on natural image datasets such as ImageNet and then finetune its weights on the medical images. It has been recently shown that this approach is not optimal for medical imaging⁴³. Motivated by these findings, we pretrain a DenseNet121⁴⁴ using thousands of CXR images of the non-COVID datasets described in section “Non-COVID dataset”.

The whole dataset is split to 75–25% between training and testing subsets. All images are resized to 224×224 and standardised before being fed to convolutional layers of DenseNet121. This results in 50,176 convolutional features which are then processed by fully connected layers with a softmax on top of them. Relu activation function, 300 epochs, and dropout rate of 0.25 are used for model development using three uncertainty quantification techniques. The Adam algorithm with a learning rate of 0.001 is applied to optimize the cross entropy loss function. For the MCD model, the number of neurons in three fully connected layers is set to 512, 256 and 64 respectively. The ensemble model consists of 30 individual networks in which hidden layers are randomly chosen between 2 and 3. Also the number of neurons in fully connected layers is randomly chosen between (512, 1024), (128, 512), (8, 128) respectively. The ensemble MCD is designed similar to the ensemble. The only difference is that the evaluation of each network is done by the MCD algorithm. To implement the experiments, Python language and libraries such as Numpy, PyTorch, and Pandas have been used. Each experiment has been repeated 100 times for different random seeds. The experiments were run on Google Colab with its default settings (GPU: 1xTesla K80, compute 3.7, 2496 CUDA cores, 12GB GDDR5 VRAM).

Simulations and results

We first compare the performance of DenseNet121 networks pretrained using ImageNet and CXR datasets. Figure 2 shows the violin plot of the accuracy and AUC performance metrics for these two networks trained and evaluated 100 times. Both performance metrics are greater and more consistent for networks pretrained using CXR datasets. A paired t-test is also run at 95% confidence level to determine whether the mean of performance metrics are statistically different. The obtained p-values for accuracy and AUC are 10^{-39} and 10^{-19} . As both values are much smaller than 0.05, it can be concluded that there are statistically significant differences in accuracy and AUC of these two models.

Results shown in Fig. 2 indicate that the proposed pipeline does a promising job for COVID-19 detection from medical images. The model AUC and accuracy are 0.95 ± 0.02 and $93.94\% \pm 1.73\%$ which are in par or

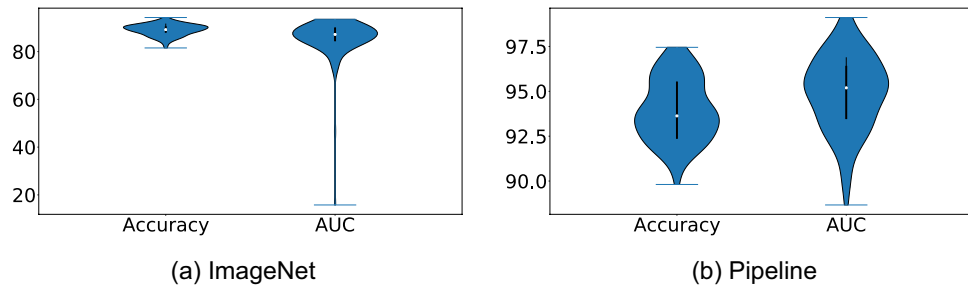


Figure 2. Accuracy and AUC (shown as a percentage) values for DenseNet121 pretrained using ImageNet and CXR datasets. The violin plot is obtained by training and measuring the network performance for 100 times.

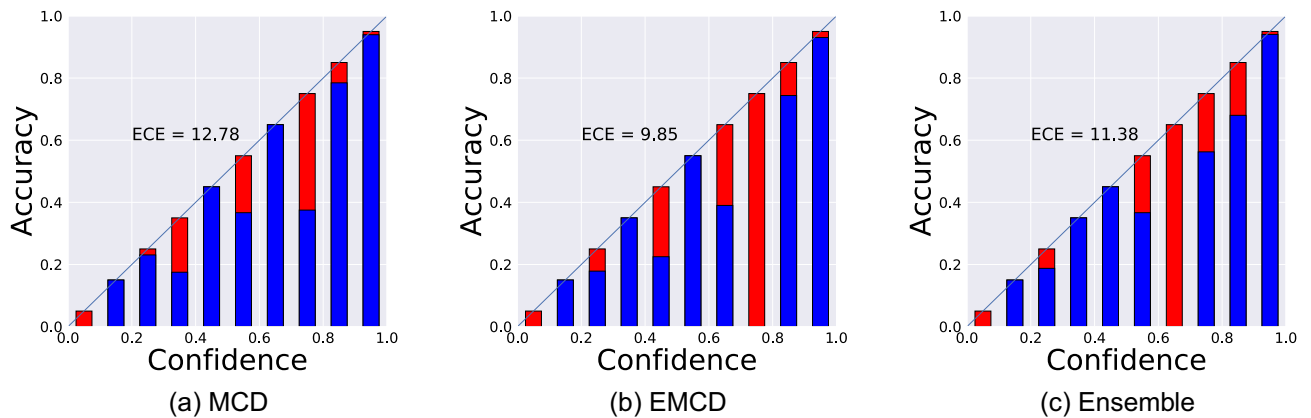


Figure 3. The reliability diagrams (ECE plots) of the trained DNNs. MCD, EMCD, and Ensemble models all have a high ECE indicating miscalibration of generated probabilities.

better than results reported in similar studies⁴⁵. Using CXR dataset (pipeline) for pretraining networks yields higher accuracy compared to using ImageNet dataset. Therefore, we choose networks that are trained on CXR through the rest of the paper. Before starting to analyze predictive uncertainty estimation results, we first check the calibration of predictions generated by DNNs. Figure 3 shows the expected calibration error (ECE) which is a plot of sample accuracy as a function of confidence⁴⁶. To calculate ECE, predictions are grouped in different bins (here M bins) according to their confidence (the value of the max softmax output). The calibration error of each bin measures the difference between the fraction of correctly classified predictions (accuracy) and the mean of the probabilities (confidence). ECE is a weighted average of this error across all bins⁴⁶:

$$ECE = \sum_{m=1}^M \frac{|B_m|}{n} |acc(B_m) - conf(B_m)| \tag{11}$$

where $acc(B_m)$ and $conf(B_m)$ are the accuracy and confidence in the m -th bin:

$$acc(B_m) = \sum \frac{1}{|B_m|} \mathbf{1}(\hat{y}_i = y_i) \tag{12}$$

$$conf(B_m) = \sum \frac{1}{|B_m|} p_i \tag{13}$$

where $\mathbf{1}(\cdot)$ is the indicator function.

Figure 4 displays the predictive posterior distribution for two non-COVID-19 (normal) CXR images obtained using MCD method (200 MC iterations). The class with the bigger softmax output for the distribution mean is reported as the predicted outcome. The uncertainty estimate associated with this outcome is also calculated using Eq. (2). The wider the predictive posterior distribution, the less confident the model. For the top case shown in Fig. 4, the predicted output is normal and the model is confident about its correct decision (a low predictive uncertainty estimate of 0.17). In contrast, the model prediction is wrong for the below CXR image in Fig. 4. However, the model generates a wide posterior distribution resulting in a high predictive uncertainty estimate (0.45). This way, it communicates its lack of confidence in this specific prediction and says *I do not know*. As the model is not confident about its prediction, the image could be sent to a medical expert for a *second opinion*¹⁰. In Fig. 3, the more the blue part deviation from the corresponding red parts, the less calibrated the neural network model. A perfectly calibrated model will generate the identity line in this chart. The current plot clearly shows

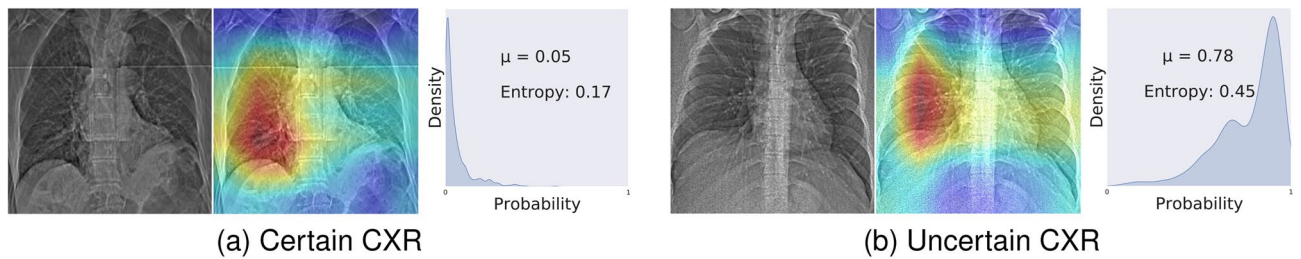


Figure 4. Two healthy (normal) images and their approximate predictive posterior distributions. These distributions, $p(\text{covid}|\text{image})$, are estimated by the MCD algorithm. (a) Correct and certain prediction, (b) incorrect and uncertain prediction. The middle plot in each row shows where DNNs look for making the decision.

UQ Method	UAcc (%)	USen	USpe	UPre
MCD	71.2	0.833	0.704	0.119
EMCD	76.3	0.777	0.762	0.194
Ensemble	77.8	0.833	0.776	0.151

Table 1. Uncertainty performance metrics for the specific threshold of 0.3 for three uncertainty quantification techniques.

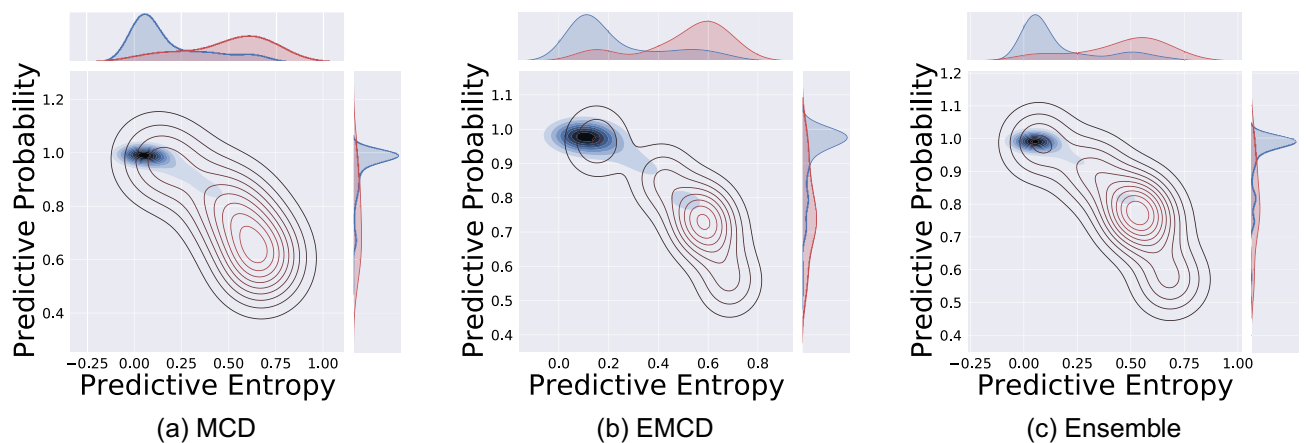


Figure 5. Contour plots of predicted probabilities and uncertainty estimates (entropy) for three uncertainty quantification methods. Distribution of estimated predictive uncertainty estimates grouped by correctly classified and misclassified are shown on the top of plots. The best separation between two groups is obtained by ensemble methods. The red lines (contours) are related to miss-classified predictions with high uncertainty and the blue ones are for correct-classified predictions with low uncertainty.

that probabilities generated by three DNN families investigated in this paper are not calibrated. ECE values reported in Fig. 3 indicate that all models overconfidently classify samples resulting in misleading outcomes. EMCD has the smallest ECE value of 9.85.

We then investigate the overall model ability in gauging and reporting its lack of confidence in its prediction. Figure 5 displays the contour plots of predicted probabilities and uncertainty estimates for three uncertainty quantification methods. The estimated marginal distributions of probabilities and uncertainty estimates are also shown on the sides of plots grouped by correctly classified and misclassified predictions. The plots clearly show that the centers of two groups are well apart from each other resulting in high accuracy. The estimated distribution for predicted probabilities of the correct group is much more compact compared to incorrect group. At the same time, the estimated uncertainties are higher for misclassified images. The visual inspection of three subplots indicates the uncertainty estimate mean for erroneous predictions is on the right of the uncertainty estimate mean for correct predictions. This qualitative investigation highlights the model capability in gauging and communicating its confidence (or lack of confidence) in generated predictions. This finding is of paramount practical importance as reliable uncertainty estimates provide additional valuable information to predicted probabilities. These could be used to flag uncertain predictions and request a second opinion by a medical expert.

We also comprehensively evaluate predictive uncertainty estimates using performance metrics introduced in section “Predictive uncertainty evaluation”. Figure 6 displays uncertainty accuracy (UAcc), uncertainty sensitivity

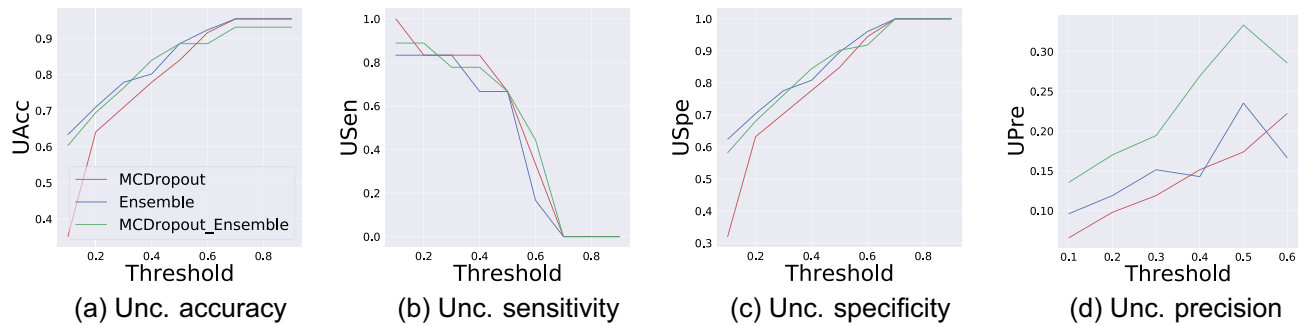


Figure 6. Quantitative evaluation of three uncertainty quantification techniques using performance metrics introduced in section “Predictive uncertainty evaluation.” Uncertainty accuracy, sensitivity, specificity, and prediction are calculated for threshold values between 0.1 and 0.9.

(USen), uncertainty specificity (USpe), and uncertainty precision (UPre) calculated for uncertainty thresholds between 0.1 and 0.9. UAcc, USpe, and UPre are positively correlated with the uncertainty threshold. This correlation is negative for USen. UAcc, USen, and USpe all achieve values close to one for a wide range of thresholds. Achieving a high USen means that all three uncertainty quantification methods are able to flag incorrect predictions with high uncertainty. These methods are quite capable of flagging erroneous predictions for further investigation. None of the uncertainty quantification methods achieves a high UPre close to one. This is because there are many correct predictions that have a high uncertainty (FU). This can be observed in the long tail of the estimated distributions of predictive uncertainties in Fig. 5 (top side plots). It is also important to note that the number of correctly classified images is much greater than the number of misclassified images. This makes the number of TU predictions always much smaller than FU resulting in a low UPre. This is an expected pattern for models with high accuracy.

UAcc, USpe, and UPre achieve their maximum values for thresholds close to one. Thresholds close to zero lead to the highest values for USpe. Selecting the best uncertainty threshold value depends on the users’ preferences, e.g., sensitivity vs. specificity. Setting it to 0.3 results in a good trade off between four uncertainty performance metrics for three uncertainty quantification methods. Table 1 reports uncertainty performance metrics for all methods. UAcc for the MCD method is 71.2% which is much smaller than UAcc for both ensemble methods. The ensemble method achieves the highest UAcc amongst uncertainty quantification methods and its UAcc (77.8%) is slightly superior to that of EMCD (76.3%). The same pattern holds for other performance metrics of three methods.

Conclusion

In this paper, we investigate the competency of deep uncertainty quantification techniques for the task of COVID-19 detection from CXR images. A novel confusion matrix and multiple performance metrics for the evaluation of predictive uncertainty estimates are introduced. Our investigations reveal that deep learning models pretrained using medical imaging datasets outperform models pretrained using natural datasets such as ImageNet. Through comprehensive evaluation, we also find that ensemble methods better capture uncertainties associated with their predictions resulting in more trustworthy diagnosis solutions. The proposed uncertainty confusion matrix also shows that uncertainty quantification methods achieve high uncertainty sensitivity and specificity. However, they often fail at producing uncertainty estimates resulting in high precision.

There are many rooms for improving the proposed uncertainty evaluation metrics and its application for DNN development. For future work, we will include the proposed uncertainty evaluation metrics as the loss function in the process of training DNNs. This will lead to networks that are optimized based on performance metrics of both point predictions and uncertainty estimates.

Received: 9 September 2021; Accepted: 6 January 2022

Published online: 17 January 2022

References

- Zu, Z. Y., Jiang, M. D., Xu, P. P., Chen, W., Ni, Q. Q., Lu, G. M. & Zhang, L. J. Coronavirus disease 2019 (covid-19): A perspective from China. *Radiology*, 200490 (2020).
- Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**(7639), 115–118 (2017).
- Esteva, A. *et al.* A guide to deep learning in healthcare. *Nat. Med.* **25**(1), 24–29 (2019).
- Minaee, S., Kafieh, R., Sonka, M., Yazdani, S. & Soufi, G. J. Deep-covid: Predicting covid-19 from chest x-ray images using deep transfer learning. *arXiv preprint arXiv:2004.09363* (2020).
- Shoebi, A., Khodatars, M., Alizadehsani, R., Ghassemi, N., Jafari, M., Moridian, P., Khadem, A., Sadeghi, D., Hussain, S., Zare, A. *et al.*, Automated detection and forecasting of covid-19 using deep learning techniques: A review. *arXiv preprint arXiv:2007.10785* (2020).
- Lalmuanawma, S., Hussain, J. & Chhakchhuak, L. Applications of machine learning and artificial intelligence for covid-19 (sars-cov-2) pandemic: A review. *Chaos Solitons Fractals*, 110059 (2020).
- Shi, F., Wang, J., Shi, J., Wu, Z., Wang, Q., Tang, Z., He, K., Shi, Y. & Shen, D. Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for covid-19. *IEEE Rev. Biomed. Eng.* **14**, 4–15 (2020).
- Hanumanthu, S. Rekha. Role of intelligent computing in covid-19 prognosis: A state-of-the-art review. *Chaos Solitons Fractals*, 109947 (2020).

9. Daniel, T. M. Toman's tuberculosis. case detection, treatment, and monitoring. questions and answers. *Am. J. Trop. Med. Hygiene* 73(1), 229 (2005).
10. Raghu, M., Blumer, K., Sayres, R., Obermeyer, Z., Kleinberg, B., Mullainathan, S., & Kleinberg, J. Direct uncertainty prediction for medical second opinions. In *International Conference on Machine Learning*, pp. 5281–5290 (2019).
11. Mukhoti, J. & Gal, Y. Evaluating bayesian deep learning methods for semantic segmentation. *arXiv preprint arXiv:1811.12709* (2018).
12. Subedar, M., Krishnan, R., Meyer, P. L., Tickoo, O. & Huang, J. Uncertainty-aware audiovisual activity recognition using deep bayesian variational inference. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 6301–6310 (2019).
13. Bernardo, J. M. & Smith, A. F. *Bayesian Theory* Vol. 405 (Wiley, London, 2009).
14. Graves, A. Practical variational inference for neural networks. In *Advances in Neural Information Processing Systems*, pp. 2348–2356 (2011).
15. Blundell, C., Cornebise, J., Kavukcuoglu, K. & Wierstra, D. Weight uncertainty in neural networks. *arXiv preprint arXiv:1505.05424* (2015).
16. Gal, Y. & Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pp. 1050–1059 (2016).
17. Lakshminarayanan, B., Pritzel, A. & Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pp. 6402–6413 (2017).
18. Choi, S., Lee, K., Lim, S. & Oh, S. Uncertainty-aware learning from demonstration using mixture density networks with sampling-free variance modeling. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6915–6922. IEEE (2018).
19. van Amersfoort, J., Smith, L., Teh, Y. W. & Gal, Y. Uncertainty estimation using a single deep deterministic neural network.
20. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. & Bengio, Y. Generative adversarial nets. *Adv. Neural Inform. Process. Syst.* 27 (2014).
21. Singh, K. K. & Singh, A. Diagnosis of covid-19 from chest x-ray images using wavelets-based depthwise convolution network. *Big Data Min. Anal.* 4(2), 84–93 (2021).
22. Pang, J., Huang, Y., Xie, Z., Li, J. & Cai, Z. Collaborative city digital twin for the covid-19 pandemic: A federated learning solution. *Tsinghua Sci. Technol.* 26(5), 759–771 (2021).
23. Kumari, R. *et al.* Analysis and predictions of spread, recovery, and death caused by covid-19 in India. *Big Data Min. Anal.* 4(2), 65–75 (2021).
24. Alberti, T. & Faranda, D. On the uncertainty of real-time predictions of epidemic growths: A covid-19 case study for china and Italy. *Commun. Nonlinear Sci. Numer. Simul.* 90, 105372 (2020).
25. Ghoshal, B., Tucker, A., Sanghera, B. & Wong, W. L. Estimating uncertainty in deep learning for reporting confidence to clinicians when segmenting nuclei image data. In *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 318–324. IEEE (2019).
26. Ghoshal, B. & Tucker, A. Estimating uncertainty and interpretability in deep learning for coronavirus (covid-19) detection. *arXiv preprint arXiv:2003.10769* (2020).
27. Ghoshal, B. & Tucker, A. 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. IEEE (2021).
28. Jokandan, A. S., Asgharnezhad, H., Jokandan, S. S., Khosravi, A., Kebria, P. M., Nahavandi, D., Nahavandi, S. & Srinivasan, D. An uncertainty-aware transfer learning-based framework for covid-19 diagnosis. *arXiv preprint arXiv:2007.14846* (2020).
29. Alizadehsani, R., Sharifrazi, D., Izadi, N. H., Joloudari, J. H., Shoeibi, A., Gorriz, J. M., Hussain, S., Arco, J. E., Sani, Z. A., Khozeimeh, F. *et al.*, Uncertainty-aware semi-supervised method using large unlabelled and limited labeled covid-19 data. *arXiv preprint arXiv:2102.06388* (2021).
30. Calderon-Ramirez, S., Yang, S., Moemeni, A., Coleeavy-Donnelly, S., Elizondo, D. A., Oala, L., Rodríguez-Capitán, J., Jiménez-Navarro, M., López-Rubio, E. & Molina-Cabello, M. A. Improving uncertainty estimation with semi-supervised deep learning for covid-19 detection using chest x-ray images. *IEEE Access* (2021).
31. Hoebel, K., Andrearczyk, V., Beers, A., Patel, J., Chang, K., Depeursinge, A., Müller, H. & Kalpathy-Cramer, J. An exploration of uncertainty information for segmentation quality assessment. In *Medical Imaging 2020: Image Processing*, vol. 11313, p. 113131K, International Society for Optics and Photonics (2020).
32. Mallick, A., Dwivedi, C., Kailkhura, B., Joshi, G. & Han, T. Y.-J. Can your ai differentiate cats from covid-19? sample efficient uncertainty estimation for deep learning safety. *Choice*, 50, 6.
33. Jin, Y., Guo, W. & Zhang, Y. A time-aware dynamic service quality prediction approach for services. *Tsinghua Sci. Technol.* 25(2), 227–238 (2019).
34. van Amersfoort, J., Smith, L., Teh, Y. W. & Gal, Y. Simple and scalable epistemic uncertainty estimation using a single deep deterministic neural network. *arXiv preprint arXiv:2003.02037* (2020).
35. Cohen, J. P., Dao, L., Morrison, P., Roth, K., Bengio, Y., Shen, B., Abbasi, A., Hoshmand-Kochi, M., Ghassemi, M., Li, H. *et al.* Predicting covid-19 pneumonia severity on chest x-ray with deep learning. *arXiv preprint arXiv:2005.11856* (2020).
36. Shih, G. *et al.* Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiol. Artif. Intell.* 1(1), e180041 (2019).
37. Irvin, J. *et al.* Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proc. AAAI Conf. Artif. Intell.* 33, 590–597 (2019).
38. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M. & Summers, R. M. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2097–2106, (2017).
39. Majkowska, A. *et al.* Chest radiograph interpretation with deep learning models: Assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology* 294(2), 421–431 (2020).
40. Johnson, A. E., Pollard, T. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., Peng, Y., Lu, Z., Mark, R. G., Berkowitz, S. J. & Horng, S. Mimi-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042* (2019).
41. Bustos, A., Pertusa, A., Salinas, J.-M. & de la Iglesia-Vayá, M. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Med. Image Anal.* 101797 (2020).
42. Demner-Fushman, D. *et al.* Preparing a collection of radiology examinations for distribution and retrieval. *J. Am. Med. Inform. Assoc.* 23(2), 304–310 (2016).
43. Raghu, M., Zhang, C., Kleinberg, J. & Bengio, S. Transfusion: Understanding transfer learning for medical imaging. In *Advances in Neural Information Processing Systems*, pp. 3347–3357 (2019).
44. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708 (2017).
45. Islam, M., Karray, F., Alhaji, R., Zeng, J. *et al.* A review on deep learning techniques for the diagnosis of novel coronavirus (covid-19). *arXiv preprint arXiv:2008.04815* (2020).
46. Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. On calibration of modern neural networks. *arXiv preprint arXiv:1706.04599* (2017).

Acknowledgements

This research was partially supported by the Australian Research Council's Discovery Projects funding Scheme (Project DP190102181 and DP210101465).

Author contributions

H.A. and A.S. are contributed equally and mostly to this work. R.A. and A.K. are helping partially with writing the text. Z.A.S., S.N., D.S., and S.M.S.I. have supervised us during the writing of the paper with their invaluable comments. At the end, this manuscript has resulted by the collaboration of all authors.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.S.

Reprints and permissions information is available at www.nature.com/reprints.

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



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022