PLOS ONE

RESEARCH ARTICLE

Calibrated bagging deep learning for image semantic segmentation: A case study on COVID-19 chest X-ray image

Lucy Nwosu[1], Xiangfang Li[1,2], Lijun Qian[1,2], Seungchan Kim[1], Xishuang Dong[1,2]*

1 Center of Computational Systems Biology (CCSB), Department of Electrical and Computer Engineering, Prairie View A&M University, Texas A&M University System, Prairie View, Texas, United States of America, **2** Center of Excellence in Research and Education for Big Military Data Intelligence (CREDIT), Department of Electrical and Computer Engineering, Prairie View A&M University, Texas A&M University System, Prairie View, Texas, United States of America

* xidong@pvamu.edu

Citation: Nwosu L, Li X, Qian L, Kim S, Dong X (2022) Calibrated bagging deep learning for image semantic segmentation: A case study on COVID-19 chest X-ray image. PLoS ONE 17(11): e0276250. https://doi.org/10.1371/journal.pone.0276250

Editor: Jude Hemanth, Karunya Institute of Technology and Sciences, INDIA

Received: May 29, 2022

Accepted: October 3, 2022

Published: November 16, 2022

Data Availability Statement: https://github.com/v7labs/covid-19-xray-dataset.

Funding: This research work is supported in part by the Texas A&M Chancellor's Research Initiative (CRI), the U.S. National Science Foundation (NSF) award 1736196, and by the U.S. Office of the Under Secretary of Defense for Research and Engineering (OUSD(R&E)) under agreement number FA8750-15-2-0119. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and

Abstract

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) causes coronavirus disease 2019 (COVID-19). Imaging tests such as chest X-ray (CXR) and computed tomography (CT) can provide useful information to clinical staff for facilitating a diagnosis of COVID-19 in a more efficient and comprehensive manner. As a breakthrough of artificial intelligence (AI), deep learning has been applied to perform COVID-19 infection region segmentation and disease classification by analyzing CXR and CT data. However, prediction uncertainty of deep learning models for these tasks, which is very important to safety-critical applications like medical image processing, has not been comprehensively investigated. In this work, we propose a novel ensemble deep learning model through integrating bagging deep learning and model calibration to not only enhance segmentation performance, but also reduce prediction uncertainty. The proposed method has been validated on a large dataset that is associated with CXR image segmentation. Experimental results demonstrate that the proposed method can improve the segmentation performance, as well as decrease prediction uncertainty.

1 Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) causes coronavirus disease 2019 (COVID-19) which was first identified in 2019 in Wuhan, Central China [1]. It is spreading globally, resulting in more than 458 million confirmed infections and 6 million deaths, and causing huge economic loss. Although global economics seems to be recovered gradually, early and accurate tests of this disease such as reverse transcription-polymerase chain reaction (RT-PCR), antigen tests, and medical imaging tests must be improved to be ready for future pademics [2, 3]. Compared to RT-PCR tests, medical imaging tests such as chest X-ray (CXR) and computed tomography (CT) are more effective and efficient [4, 5], especially for severe patients, which is of great help to physicians. For instance, in Italy, the United States, and

conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U.S. National Science Foundation (NSF) or the U.S. Office of the Under Secretary of Defense for Research and Engineering (OUSD(R&E)) or the U. S. Government.

Competing interests: NO authors have competing interests

China, the majority of serious COVID-19 cases have been identified through the manifestation characteristics in CT images [6]. Therefore, effective extraction of COVID-related information on medical images will play an important role to fight against a new round of pandemic caused by COVID mutated variant [7].

Deep learning (DL) played an important role in promoting COVID-related information extraction by COVID-19 infection region segmentation and disease classification through analyzing CXR and CT data [8, 9]. Compared with CT images, CXR images are easier to obtain in radiological inspections. Currently, most of DL models, especially convolutional neural networks (CNN), were employed to classify entire CXR images to detect COVID-19 cases [10, 11]. For example, Hemdan *et al.* proposed COVIDX-Net to assist radiologists to diagnose COVID-19 based on CXR features [12]. It integrated various deep convolutional neural networks (DCNNs) models with different structures, such as DenseNet201 [13], Xception [14], and MobileNetV2 [15]. Sethy *et al.* integrated different DCNNs models with a support vector machine (SVM) classifier to recognize COVID-19 [16]. In addition, to address the shortcomings of training data, Castiglioni *et al.* employed transfer deep learning techniques for COVID-19 classification, where the pretrained models were built based on ResNet on ImageNet datasets [17]. Ioannis *et al.* comprehensively evaluated transfer learning based COVID-19 classification by investigating 5 DCNN models, including VGG19, MobileNetV2, Inception, Xception, and InceptionResNetV2 [18]. Similarly, Narin *et al.* applied 3 typical pretrained DCNN models (i.e., ResNet50, InceptionV3, and InceptionResNetV2) to classify COVID-19 on a small-scale CXR dataset [19]. Irfan *et al.* implemented a hybrid COVID-19 classification model by using a multi-model (CNN + LSTM) [20]. Almalki *et al.* used Inception-ResNet block with an extra number of layers branches that consisted of the convolutional layer for COVID-19 classification [21]. Moreover, Lucy *et al.* [22] developed two-path semi-supervised deep learning model to implement COVID-19 classification by using huge amounts of unlabeled data.

Compared with CXR classification, CXR semantic segmentation is a more challenging task that is to classify each pixel into predefined classes [23] to recognize region of interests (ROIs) on CXR images, where a few previous work explored this task [24–26]. However, prediction uncertainty of DL models for this task has not been comprehensively investigated since most of DL models focus on performance improvement on this task such as increasing detection accuracy. For safety-critical applications like medical image processing, the prediction uncertainty of DL models is a key evaluation metic on reliability of model predictions since high prediction uncertainty means low prediction reliability. For example, for COVID-19 applications, applying uncertain predictions to clinical processes would result in disastrous consequences such as missing severe COVID cases or delayed treatments.

This paper proposed a novel ensemble deep learning model that integrates bagging deep learning [27] and model calibration [28] to enhance performance of semantic segmentation, as well as reduce prediction uncertainty. It includes three stages shown in Fig 1: 1) training multiple state-of-the-art DL models such as fully convolutional networks (FCN) [29], FCN combined with ResNet [30], FCN combined with MobileNet [29], PSPNet [31], and UNet [32] on training CXR datasets; 2) Calculating calibration errors to measure prediction uncertainties of these DL models on validation CXR datasets, where expected calibration error (ECE) and maximum calibration error (MCE) [28] are employed to measure the prediction uncertainties; 3) Implementing calibrated bagging deep learning with weighted voting, where the weight of each DL model is inversely proportional to the calibration error. The proposed model is validated on a large-scale CXR dataset to examine its effectiveness. Experimental results demonstrate that the proposed method not only enhances the performance of semantic segmentation, but also improves the prediction certainty on CXR data.

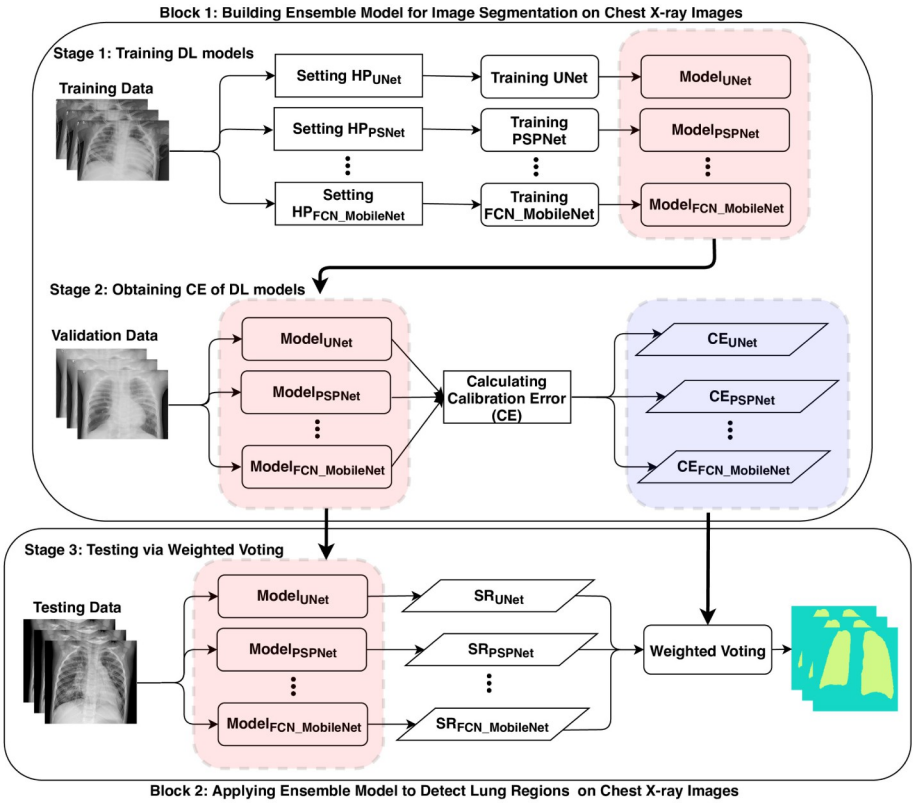

Fig 1. Diagram of building and testing calibrated bagging deep learning based on calibration error (CE). *SR* denotes segmentation result generated by individual deep learning model.

https://doi.org/10.1371/journal.pone.0276250.g001

The contributions in this study are below.

- We systematically compared performance of various state-of-the-art DL models for semantic segmentation on COVID-19 CXR data with different evaluation metrics. Moreover, the prediction uncertainties of these DL models were investigated by measuring expected calibration error (ECE) and maximum calibration error (MCE).
- We implemented a novel ensemble deep learning model based on model calibration and bagging deep learning, which is to calibrate bagging deep learning models through weighted summation of predictions generated by individual models. The proposed approach is easily implemented and scalable to various tasks.
- We validate the proposed method with semantic segmentation on a large COVID-19 CXR dataset based on different evaluation metrics. Experimental results demonstrate its effectiveness on improving performance and prediction certainty, simultaneously.

2 Methodology

The proposed method is built based on calibration error [33–35] and bagging deep learning [27] to enhance image segmentation with higher prediction certainty.

2.1 Calibration error

In the processing of model calibration, the expected calibration error (ECE) and the maximum calibration error (MCE) can be employed to measure the quality of uncertainty for machine learning models in terms of prediction accuracy [36], which is critical for high risk applications such as medical diagnosis [34, 35] and self-driving [37].

- Expected Calibration Error (ECE). It estimates the calibration error in expectation values with three steps: 1) Discretizing the prediction probability region into a fixed number of bins; 2) Assigning each predicted probability to one of these bins; 3) Calculating the difference between the fraction of predictions in the bin that are correct (accuracy) and the mean of the probabilities in the bin (confidence) by

$$ECE = \sum_{k=1}^{K} \frac{n_k}{N} |acc(k) \quad -conf(k)| \tag{1}$$

  where n_k is the number of predictions in bin k, N is the total number of samples predicted, and $acc(k)$ and $conf(k)$ denonte the accuracy and confidence in the bin k, respectively. It is a weighted average of differences of accuracy vs confidence in these bins.

- Maximum Calibration Error (MCE). It measures an upper bound of ECE that is the maximum difference between accuracy and confidence over all predictions across all bins.

$$MCE = \max_{k=1}^{K} |acc(k) \quad -conf(k)| \tag{2}$$

In summary, MCE measures the largest calibration gap across all bins, whereas ECE measures a weighted average of all gaps. Both MCE and ECE equal 0 if the model is perfectly calibrated.

2.2 Bagging learning

Ensemble deep learning combines several individual deep models to improve generalization performance through various ensemble strategies such as bagging and boosting, which integrates the advantages of both deep learning and ensemble learning [27]. Bagging (or bootstrap aggregating) generates a series of independent subsets from training data to build multiple individual predictors to build an ensemble model [38]. In detail, it generates the bagging samples and passes each bag of samples to base models to build multiple predictors. Then, it is to combine predictions of these multiple predictors with specific strategies such as majority voting. Fig 2 presents a diagram for building and testing bagging deep learning with majority voting, where multiple training sets can be generated by sampling with or without replacement.

2.3 Proposed model

We proposed a calibrated bagging deep learning model to enhance generalization performance as well as reduce prediction uncertainty for COVID-19 semantic segmentation that is to recognize lung region of CXR images. Fig 1 presents the flow for building the proposed approach. It includes three stages: 1) training various state-of-the-art deep learning models such as UNet [32], PSPNet [31], and MobileNet [29], on an identical training data for COVID-19 image segmentation models, which differs from the standard strategy for bagging learning that is to generate a bag of training sets on original training data; 2) Estimating calibration error (CE) for these different models. First, it is to complete COVID-19 semantic segmentation on validation

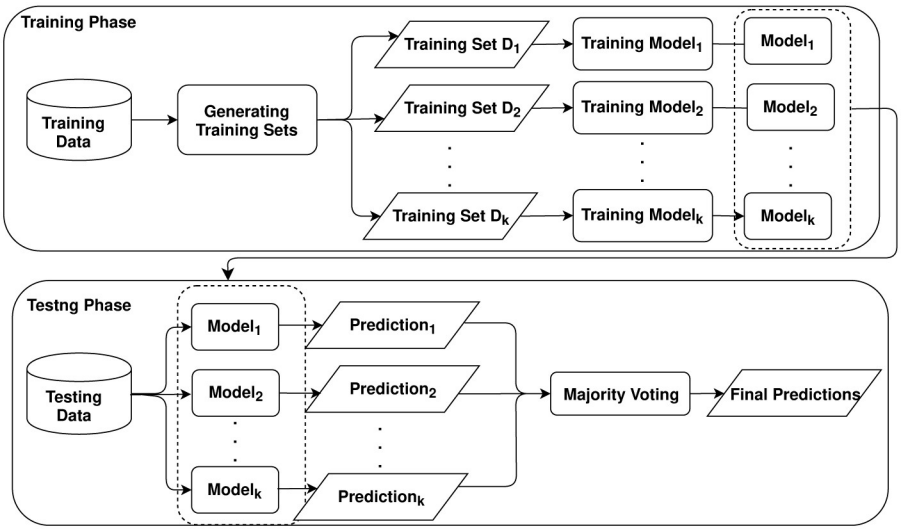

Fig 2. Diagram for building a bagging deep learning model. The model can be different deep learning models such as convolutional neural networks (CNN) and recurrent neural networks RNN) for different applications.

https://doi.org/10.1371/journal.pone.0276250.g002

data by running these DL models to obtain prediction probabilities and accuracy. Then, it calculates CE including ECE and MCE to evaluate uncertainties of these DL models; 3) Testing via weighted voting bagging deep learning. We perform calibrated bagging prediction on testing data through implementing weighted voting, where the weights are built with CE of these DL models. It assumes that lower CE of DL models means higher certainty of these DL models. Moreover, DL models with the higher certainty are assigned with more weights. Therefore, we define the weight of *i*th model as $\frac{1}{CE_i}$, where CE_i is the calibration error for *i*th model. For COVID-19 semantic segmentation, it is to classify each pixel into either Lung or NonLung. If $\sum^{Lung} \frac{1}{CE_i} > \sum^{NonLung} \frac{1}{CE_j}$ for one pixel in a CXR image, this pixel is classified as Lung, otherwise, Non-Lung.

More details on building calibrated bagging deep learning is illustrated in Algorithm 1, where M denotes the number of the-state-of-art deep learning models involved.

Algorithm 1 Building calibrated bagging deep learning

Require: Training set $D_{training}$ and validation set D_{val}
Ensure: Calibrated bagging deep learning
1: **for** $m \leftarrow 1$ to M **do**
2: Setting hyper-parameter (HP) for DL_m
3: Training DL_m on $D_{training}$
4: Calculating CE_m of DL_m on D_{val}
5: **end for**
6: **return** DL models $DL = \{DL_1, DL_2, ..., DL_M\}$ and corresponding $CE = \{CE_1, CE_2, ..., CE_M\}$

3 Experiment

3.1 Dataset

We employed COVID-19 chest X-ray dataset (https://github.com/v7labs/covid-19-xray-dataset) to validate the effectiveness of the proposed method. It includes 6, 402 images of AP/PA chest x-rays/CT scan with pixel-level polygonal lung segmentations. Each image has a corresponding ground truth with two "Lung" segmentation masks (rendered as polygons,

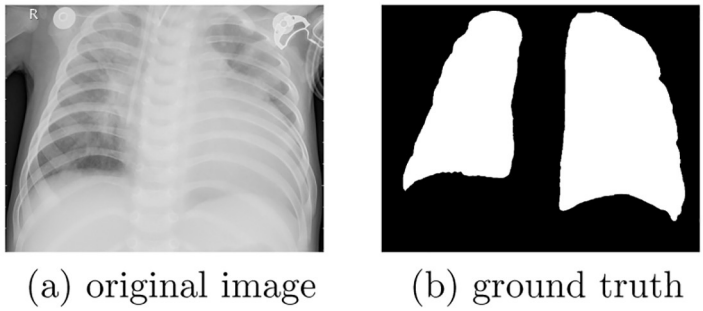

Fig 3. An example of CXR image and corresponding ground truth. (a) original image. (b) ground truth.

https://doi.org/10.1371/journal.pone.0276250.g003

including the posterior region behind the heart), where the masks include most of the heart, revealing lung opacities behind the heart which may be relevant for assessing the severity of viral infection. Fig 3 shows one example of CXR image and corresponding ground truth. In terms of the example, semantic segmentation on CXR images is to classify pixels in the original image into two classes: Lung (white region in ground truth) and NonLung (black region in ground truth).

We split the dataset into training (70% data), validation (10% data), and testing (20% data) datasets.

3.2 Experimental settings

We employed five state-of-the-art individual models as baselines to evaluate performance of semantic segmentation, namely, UNet [32], PSPNet [31], FCN32 [39] (FCN with 32×upsampling), FCN32_ResNet50 (FCN32 combined with ResNet50 [30]), FCN32_MobileNet(FCN32 combined with MobileNet [29]), and an ensemble baseline built based on majority voting, where the ensemble baseline is built based on bagging learning with these results generated by these five baselines (UNet, PSPNet, FCN32, FCN32_ResNet50, and FCN32_MobileNet). Moreover, key hyper-parameters of these individual models are shown in Table 1.

We implemented two versions of the proposed approach including Ensemble (Weighted Voting (ECE), EECE) and Ensemble (Weighted Voting (MCE), EMCE). EECE is a weighted bagging learning method, where the weights are obtained by calculating expected calibration error (ECE). Similarly, EMCE is a weighted bagging learning method, where the weights are obtained by calculating maximum calibration error (MCE). Moreover, we combine the predictions of Ensemble (Majority Voting (MV), EMV), EECE, and EMCE by majority voting to build Ensemble (Majority Voting + ECE + MCE (MVEM), EMVEM).

3.3 Evaluation metric

Various evaluation metrics are employed to evaluate the performance of our proposed model, which includes accuracy, F1score, sensitivity, and specificity. Accuracy is calculated by

Table 1. Hyper-parameters of baselines for COVID-19 image segmentation.

Model	Learning Rate	Batch Size	Epoch
UNet	1e-3	2	50
PSPNet	1e-3	2	70
FCN32	1e-3	2	50
FCN32_ResNet50 (F32_R50)	1e-3	2	50
FCN32_MobileNet (F32_M)	1e-3	2	50

https://doi.org/10.1371/journal.pone.0276250.t001

dividing the number of pixels identified correctly over the total number of pixels in chest X-ray images.

$$Accuracy = \frac{N_{correct}}{N_{total}}. \tag{3}$$

$$Fscore = \frac{2 \times Precision \times Recall}{Precision + Recall}. \tag{4}$$

where *Precision* defines the capability of a model to represent only correct pixels and *Recall* computes the aptness to refer all corresponding correct pixels.

$$Precision = \frac{TP}{TP + FP}. \tag{5}$$

$$Recall = \frac{TP}{TP + FN}. \tag{6}$$

whereas *TP* (True Positive) counts the total number of pixels that matches the annotated pixels of RIOs. *FP* (False Positive) measures the number of pixels that don't belong to RIOs, but are recognized as pixels of RIOs. *FN* (False Negative) counts the number of pixels of RIOs are recognized as those don't belong to RIOs. The main goal for binary classification is to improve the recall without hurting the precision. However, recall and precision goals are often conflicting, since when increasing the true positive (TP) for the minority class (True), the number of false positives (FP) can also be increased; this will reduce the precision [40].

Moreover, we employed sensitivity and specificity to evaluate performance of semantic segmentation [41], where the sensitivity measures how good a test is at detecting the RIOs while the specificity refers to how good a test is at avoiding false alarms.

$$Specificity = \frac{TN}{TN + FP}. \tag{7}$$

$$Sensitivity = \frac{TP}{TP + FN}. \tag{8}$$

whereas *TN* (True Negative) counts total number of pixels that don't belong to RIOs and are recognized as those don't belong to RIOs.

Finally, we employ expected calibration error (ECE) and MCE (https://www.tensorflow.org/probability/api_docs/python/tfp/stats/expected_calibration_error) to measurethe calibration errors [28] for evaluating the prediction uncertainty, where ECE and MCE are defined as equations (1) and (2), respectively. The lower ECE and MCE are, the higher prediction certainty is.

3.4 Experimental results

We validate the proposed method from two perspectives: comprehensive performance comparison between the baselines and the proposed method, and hyper-parameter examination.

3.4.1 Performance comparison. Table 2 presents the performance comparison between the state-of-the-art individual models and the proposed method in terms of various evaluation metrics and corresponding standard deviations. We can observe that these individual models can perform well on COVID-19 image segmentation regarding F1scores and Accuracy. Moreover, prediction uncertainties of most of them are promising with respect to ECE and MCE.

Table 2. Comparing performance between the baselines and the proposed method based on various evaluation metrics and corresponding standard deviations. F32_R50 and F32_M denotes FCN32_ResNet50 and FCN32_MobileNet while EMV, EECE, EMCE), and EMVEM denotes Ensemble (Majority Voting (MV), EMV), Ensemble (Weighted Voting (ECE), EECE), Ensemble (Weighted Voting (MCE), EMCE), and Ensemble (Majority Voting + ECE + MCE (MVEM), EMVEM).

DL	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1score (%)	ECE (%)	MCE (%)
UNet	95.4±2.5	90.7±3.9	88.9±4.5	93.4±3.0	3.2±1.3	39.7±18.8
PSPNet	95.0±2.0	89.1±3.9	88.2±4.3	92.5±2.9	4.6±1.2	40.6±14.9
FCN32	95.8±2.4	92.3±4.5	91.0±5.0	94.0±3.5	2.5±2.1	37.6±19.1
F32_R50	96.0±2.5	92.3±5.5	91.4±5.9	94.3±3.8	2.3±2.3	29.8±20.3
F32_M	95.2±2.3	91.0±4.7	90.1±5.6	93.1±3.3	4.1±1.6	38.2±19.9
EMV	98.8±0.6	94.1±3.0	92.9±3.6	96.6±1.7	2.4±1.2	28.1±14.1
EECE	99.1±0.5	95.4±2.9	94.3±2.9	97.1±1.5	2.3±1.2	24.7±12.4
EMCE	98.7±0.7	93.9±3.7	92.6±3.7	96.3±1.9	2.4±1.2	28.9±14.6
EMVEM	99.2±0.4	97.7±2.3	95.4±2.3	98.4±0.8	2.1±1.1	20.1±10.1

https://doi.org/10.1371/journal.pone.0276250.t002

For these individual models, FCN32_ResNet50 outperforms other individual models with higher certainty. In addition, as one baseline, EMV performs better than other individual methods with highest prediction certainty by comparing ECE and MCE. It means that combining predictions of these individual models can effectively improve performance and prediction certainty in regard of F1score and ECE.

For the proposed method, EECE can perform better than the baselines including these individual models and EMV by comparing accuracy, recall, and F1score. Moreover, EECE is able to improve the prediction certainty. It means that using appropriate calibration errors as weights to implement weighted bagging deep learning can effectively improve prediction certainty as well as performance. In other words, it is an effective method to calibrate models by using appropriate calibration errors as weights to combine predictions. Furthermore, EMVEM obtains the optimal performance with highest prediction certainty. It indicates that ensemble strategy such as majority voting is effective to combine predictions to further improve performance and prediction certainty. Moreover, EMVEM performed more stable since the standard deviations of performance and calibration errors are lower than those of baselines.

In addition to the performance comparison, we show an example of prediction visualization on semantic segmentation generated by the baselines and proposed models in Fig 4.

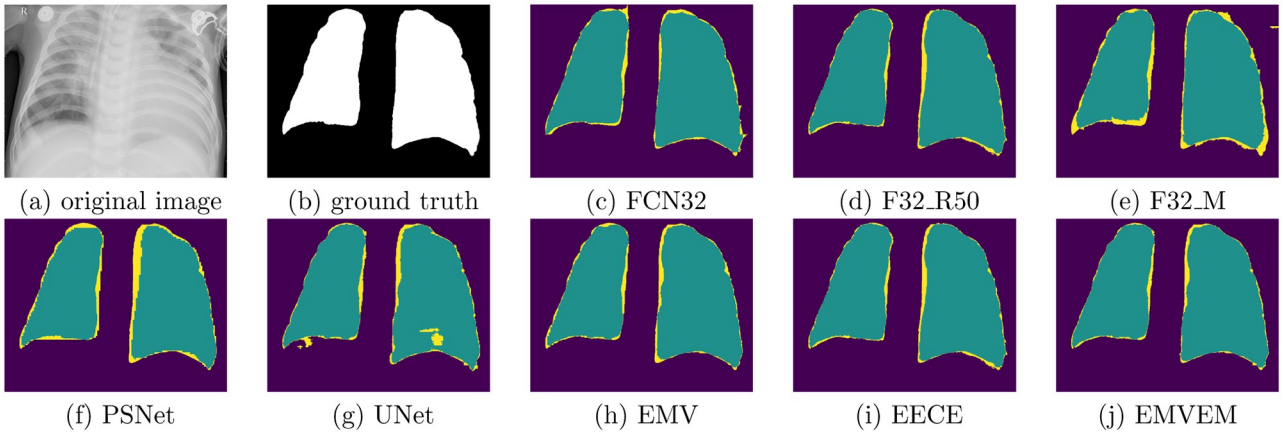

Fig 4. An example of prediction visualization on semantic segmentation generated by the baselines and proposed models. (a) original image, (b) ground truth, (c) FCN32, (d) F32 R50, (e) F32 M, (f) PSNet, (g) UNet, (h) EMV, (i) EECE, (j) EMVEM.

https://doi.org/10.1371/journal.pone.0276250.g004

When we examine the prediction visualization for these individual models, we can observe that they miss some key components (yellow regions) for detecting lung. Taking UNet as an example, through comparing the predictions with ground truth, key components highlighted with yellow color are missed on subfigure (g). On the contrary, ensemble models such as EMV, EECE, and EMVEM perform better in that regard of predictions since yellow regions in their predictions are smaller, where the proposed method including EECE and EMVEM outperform other baselines. It means that the proposed method can effectively improve recall on detecting lung by distributing contributions of prediction based on calibration errors such as ECE and MCE.

3.4.2 Hyper-parameter examination. Fine-tuning hyper-parameter for building deep learning models is an imperative step to obtain optimal performance. The process of building the proposed method involved various hyper-parameters. For example, for each individual DL model, we have to fine-tune learning rate, batch size, and epoch to achieve optimal performance. Specifically, for the proposed bagging deep learning, how many individual models involved is still an open challenge. Here, we examine if the number of individual models will significantly affect the performance of the proposed method.

Table 3 presents the performance comparison for various bagging deep learning models built with different number of individual models. Generally speaking, more individual models will enhance performance and improve prediction certainty regarding F1score and ECE. When we employ five individual models (Ensemble 5 (FCN32_RESNET50 + FCN32 + UNet+ FCN32_MOBILENET + PSPNet)), we obtain the optimal performance and the highest prediction certainty regarding values of accuracy, F1score, and ECE for EECE and EMVEM, where the values of F1score are improved more significantly than other evaluation metrics.

Additionally, Fig 5 shows comparison of prediction visualization produced by the proposed methods built with different number of individual models. It is observed that more individual

Table 3. Comparing performance of the proposed methods built with different number of individual models.

Ensemble 2 (FCN32_RESNET50 + FCN32)						
DL	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**	**F1score (%)**	**ECE (%)**	**MCE (%)**
F32_R50	95.8±2.1	92.3±3.9	91.0±4.5	94.0±3.0	2.5±1.3	37.6±18.8
EECE	99.0±0.5	95.3±2.4	94.4±2.8	96.9±1.6	2.3±1.2	22.3±11.3
EMCE	98.8±0.6	93.8±3.1	93.7±3.2	96.4±1.8	2.5±1.3	25.1±12.6
Ensemble 3 (FCN32_RESNET50 + FCN32 + UNet)						
DL	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**	**F1score (%)**	**ECE (%)**	**MCE (%)**
EMV	98.7±0.7	93.9±3.0	94.1±3.0	96.1±2.0	2.7±1.4	31.1±15.6
EECE	98.4±0.8	95.5±2.3	94.9±2.6	96.9±1.6	2.8±1.4	26.1±13.1
EMCE	98.3±0.9	93.1±3.5	93.0±3.5	96.0±2.0	2.8±1.4	31.2±15.6
EMVEM	98.8±0.6	97.6±1.2	96.4±1.8	98.1±1.0	2.1±1.1	21.1±10.6
Ensemble 4 (FCN32_RESNET50 + FCN32 + UNet + FN32_MOBILENET)						
DL	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**	**F1score (%)**	**ECE (%)**	**MCE (%)**
EECE	98.3±0.9	95.0±2.5	94.6±2.7	96.5±1.8	2.7±1.4	24.3±13.5
EMCE	97.9±1.1	94.1±3.0	93.8±3.1	96.1±2.0	3.0±1.5	32.3±15.0
Ensemble 5 (FCN32_RESNET50 + FCN32 + UNet + FCN32_MOBILENET + PSPNet)						
DL	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**	**F1score (%)**	**ECE (%)**	**MCE (%)**
EMV	98.8±0.6	94.1±3.0	92.9±3.6	96.6±1.7	2.4±1.2	28.1±14.1
EECE	99.1±0.5	95.4±2.9	94.3±2.9	97.1±1.5	2.3±1.2	24.7±12.4
EMCE	98.7±0.7	93.9±3.7	92.6±3.7	96.3±1.9	2.4±1.2	28.9±14.6
EMVEM	99.2±0.4	97.7±2.3	95.4±2.3	98.4±0.8	2.1±1.1	20.1±10.1

https://doi.org/10.1371/journal.pone.0276250.t003

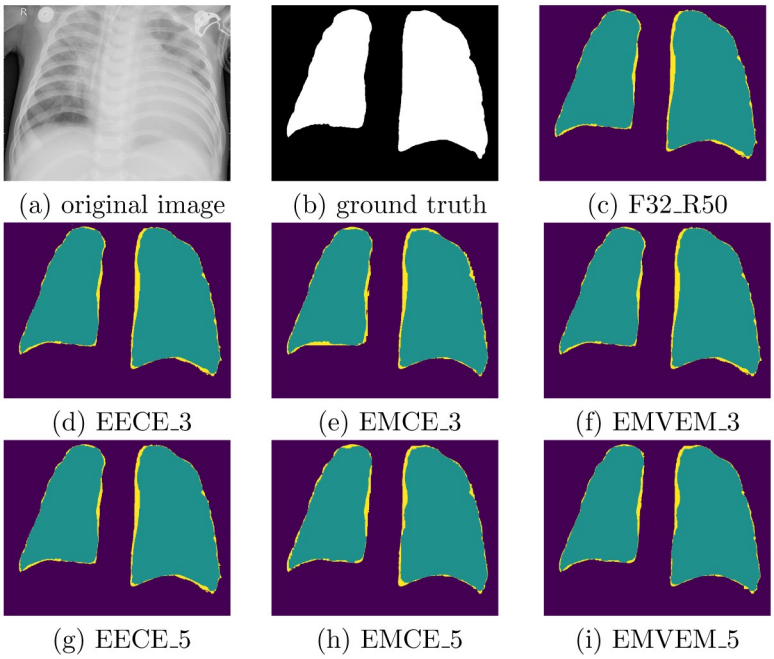

Fig 5. Comparison of prediction visualization produced by the proposed methods built with different number of individual models. For instance, in the second row, it presents the prediction results generated by ensemble models that are built with three individual models, namely, FCN32_RESNET50, FCN32, UNet with voting strategies. In the predictions, purple color, yellow color, and green color denotes background, incorrect prediction, and correct prediction, respectively, where the smaller region of yellow color means higher performance. (a) original image, (b) ground truth, (c) F32 R50, (d) EECE 3, (e) EMCE 3, (f) EMVEM 3, (g) EECE 5, (h) EMCE 5, (i) EMVEM 5.

https://doi.org/10.1371/journal.pone.0276250.g005

models involved in the proposed approach will reduce the size of missing components. Moreover, EMVEM outperforms other ensemble methods, which means that majority voting based on more individual DL models can further enhance the performance of recognition of RIOs.

In summary, in terms of observations mentioned above, the proposed method can effectively improve semantic segmentation, as well as reduce the prediction uncertainty through using the calibration error as weights of DL models to combine their predictions. Moreover, more individual DL models involved in the implementation of the proposed approach can further enhance the performance and prediction certainty, which meets the intuition of majority voting for bagging deep learning. To some extent, it is an effective method to combine advantages of these individual DL models to improve the task performance without complex implementations.

4 Related work

This paper aims to build a novel bagging learning method to implement COVID-19 semantic segmentation through combining bagging deep learning and model calibration. Semantic segmentation has achieved significant successes by developing deep learning models such as U-Net [32] and V-Net [42]. In the biomedical domain, there have been numerous techniques for lung segmentation with different purposes [43, 44]. The U-Net is an effective technique for segmenting both lung regions and lung lesions in COVID applications [45]. The U-Net built with fully convolutional network [32] has a U-shape architecture with two symmetric paths: encoding path and decoding path. The layers at the same level in two paths are connected by the shortcut connections, which is to learn better visual semantics as well as detailed

contexture. Zhou *et al.* [46] proposed the UNet++ that inserts a nested convolutional structure between the encoding and decoding path. In addition, Milletari *et al.* [42] built V-Net using the residual blocks as the basic convolutional block, and optimized the network by a Dice loss. Furthermore, Shan *et al.* [47] built VB-Net for more efficient segmentation by equipping the convolutional blocks with the so-called bottleneck blocks. Moreover, U-Net and its variants have been developed, achieving reasonable segmentation results in COVID-19 diagnosis [48]. In recent years, attention mechanisms can learn the most discriminant part of the features in deep learning models. Oktay *et al.* [49] proposed an Attention U-Net to capture fine structures in medical images, thereby suitable for segmenting lesions and lung nodules in COVID-19 applications.

Safety-critical applications like medical image processing [50], autonomous driving [51], and precipitation forecasting [52] not only require high accuracy, but also need high prediction uncertainty measured by the model calibration. Two categories of methods are proposed to calculate the model calibration, namely, Bayesian-based and Non-Bayesian-based. Bayesian-based methods refer to Bayesian neural networks that estimates prediction/model uncertainty based on Bayesian process. The main concern of such methods is associate with its high computation complex and prior assumption on model weights. To reduce the computation complexity and enhance the scalability of Bayesian neural networks for data analysis on larger datasets, Hernández-Lobato *et al.* [53] proposed probabilistic back-propagation for learning Bayesian neural networks. Non-Bayesian-based methods develop various strategies such as model ensemble [54] and prior assumption on predictions [55] to estimate the prediction uncertainty, which is to reduce the cost of estimating the uncertainty. To reduce computation cost and training difficulty, Lakshminarayanan *et al.* [54] proposed deep ensemble that is simply to implement, trained in a parallel manner, requires less hyper-parameter tuning, and estimates high quality predictive uncertainty. However, it is very tricky to obtain the optimal number of individual models to build deep ensemble for various applications. Moreover, to reduce the cost of the memory usage and inference of Bayesian neural networks and deep ensembles, Liu *et al.* [56] proposed approaches to estimate uncertainty by building only one neural networks with two steps: 1) Measuring the distance between testing samples and training samples; 2) Implementing spectral-normalized neural Gaussian process (SNGP) that is to improve the measurement of the distance by adding a weight normalization step during training and replacing the output layer with a Gaussian process. However, experimental results on dialog intent detection indicated that deep ensemble performed better than the proposed method on many evaluation metrics such as accuracy. Recently, Wilson *et al.* [57] systematically summarized Bayesian deep learning and claimed that deep ensemble can be treated as approximate Bayesian marginalization of model parameters. On the other side, they also claimed that Bayesian methods were not perfect regarding prior assumptions on model weights.

In terms of previous work on model calibration and semantic segmentation, we proposed the calibrated ensemble model to not only enhance performance on semantic segmentation, but also reduce the prediction uncertainty.

5 Conclusion and future work

In this paper, a novel bagging deep learning model is proposed for COVID-19 image segmentation on chest x-ray images. It combines the model calibration and traditional bagging learning to not only enhance the segmentation performance, but also improve the prediction certainty that is extremely important to high-risk applications in biomedical domain. We validate the proposed method on a large chest x-ray dataset that is associated with COVID-19.

Experimental results demonstrate that the proposed model could recognize the lung region more effectively through comparing with state-of-the-art baselines. For the future work, we plan to extend the proposed model for building an end-to-end model for both COVID-19 image classification and image segmentation.

Author Contributions

Conceptualization: Lijun Qian, Xishuang Dong.

Data curation: Lucy Nwosu.

Methodology: Lucy Nwosu, Lijun Qian, Xishuang Dong.

Project administration: Xishuang Dong.

Supervision: Xishuang Dong.

Validation: Lucy Nwosu, Seungchan Kim, Xishuang Dong.

Visualization: Seungchan Kim.

Writing – original draft: Lucy Nwosu, Xiangfang Li, Lijun Qian, Seungchan Kim, Xishuang Dong.

Writing – review & editing: Xishuang Dong.

References

1. Wang X, Song X, Guan Y, Li B, Han J. Comprehensive named entity recognition on cord-19 with distant or weak supervision. arXiv preprint arXiv:200312218. 2020;.
2. Öztürk Ş, Özkaya U, Barstuğan M. Classification of Coronavirus (COVID-19) from X-ray and CT images using shrunken features. International Journal of Imaging Systems and Technology. 2021; 31(1):5–15. https://doi.org/10.1002/ima.22469 PMID: 32904960
3. Zhang X, Lu S, Wang SH, Yu X, Wang SJ, Yao L, et al. Diagnosis of COVID-19 pneumonia via a novel deep learning architecture. Journal of Computer Science and Technology. 2021; 1.
4. Pan F, Ye T, Sun P, Gui S, Liang B, Li L, et al. Time course of lung changes on chest CT during recovery from 2019 novel coronavirus (COVID-19) pneumonia. Radiology. 2020; p. 200370.
5. Wong HYF, Lam HYS, Fong AHT, Leung ST, Chin TWY, Lo CSY, et al. Frequency and distribution of chest radiographic findings in COVID-19 positive patients. Radiology. 2020; p. 201160.
6. Kumar M, Gupta S, Kumar K, Sachdeva M. Spreading of COVID-19 in India, Italy, Japan, Spain, UK, US: a prediction using ARIMA and LSTM model. Digital Government: Research and Practice. 2020; 1 (4):1–9. https://doi.org/10.1145/3411760
7. Liang T, et al. Handbook of COVID-19 prevention and treatment. The First Affiliated Hospital, Zhejiang University School of Medicine Compiled According to Clinical Experience. 2020; 68.
8. Ghoshal B, Tucker A. Estimating uncertainty and interpretability in deep learning for coronavirus (COVID-19) detection. arXiv preprint arXiv:200310769. 2020;.
9. Gozes O, Frid-Adar M, Greenspan H, Browning PD, Zhang H, Ji W, et al. Rapid ai development cycle for the coronavirus (covid-19) pandemic: Initial results for automated detection & patient monitoring using deep learning ct image analysis. arXiv preprint arXiv:200305037. 2020;.
10. Wang L, Wong A. COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-Ray Images. arXiv preprint arXiv:200309871. 2020;.
11. Wang Y, Hu M, Li Q, Zhang XP, Zhai G, Yao N. Abnormal respiratory patterns classifier may contribute to large-scale screening of people infected with COVID-19 in an accurate and unobtrusive manner. arXiv preprint arXiv:200205534. 2020;.
12. Hemdan EED, Shouman MA, Karar ME. Covidx-net: A framework of deep learning classifiers to diagnose covid-19 in x-ray images. arXiv preprint arXiv:200311055. 2020;.
13. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 4700–4708.
14. Chollet F. Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 1251–1258.

15. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 4510–4520.

16. Sethy PK, Behera SK, Ratha PK, Biswas P. Detection of coronavirus disease (COVID-19) based on deep features and support vector machine. International Journal of Mathematical, Engineering and Management Sciences. 2020; 5(4). https://doi.org/10.33889/IJMEMS.2020.5.4.052

17. Castiglioni I, Ippolito D, Interlenghi M, Monti CB, Salvatore C, Schiaffino S, et al. Artificial intelligence applied on chest X-ray can aid in the diagnosis of COVID-19 infection: a first experience from Lombardy, Italy. MedRxiv. 2020;.

18. Apostolopoulos ID, Mpesiana TA. Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks. Physical and engineering sciences in medicine. 2020; 43 (2):635–640. https://doi.org/10.1007/s13246-020-00865-4 PMID: 32524445

19. Narin A, Kaya C, Pamuk Z. Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks. Pattern Analysis and Applications. 2021; 24(3):1207–1220. https://doi.org/10.1007/s10044-021-00984-y PMID: 33994847

20. Irfan M, Iftikhar MA, Yasin S, Draz U, Ali T, Hussain S, et al. Role of hybrid deep neural networks (HDNNs), computed tomography, and chest X-rays for the detection of COVID-19. International Journal of Environmental Research and Public Health. 2021; 18(6):3056. https://doi.org/10.3390/ijerph18063056 PMID: 33809665

21. Almalki YE, Qayyum A, Irfan M, Haider N, Glowacz A, Alshehri FM, et al. A novel method for COVID-19 diagnosis using artificial intelligence in chest X-ray images. In: Healthcare. vol. 9. Multidisciplinary Digital Publishing Institute; 2021. p. 522.

22. Nwosu L, Li X, Qian L, Kim S, Dong X. Semi-supervised learning for COVID-19 image classification via ResNet. arXiv preprint arXiv:210306140. 2021;.

23. Paul HY, Kim TK, Lin CT. Generalizability of deep learning tuberculosis classifier to COVID-19 chest radiographs: new tricks for an old algorithm? Journal of Thoracic Imaging. 2020; 35(4):W102–W104. https://doi.org/10.1097/RTI.0000000000000532

24. Teixeira LO, Pereira RM, Bertolini D, Oliveira LS, Nanni L, Cavalcanti GD, et al. Impact of lung segmentation on the diagnosis and explanation of COVID-19 in chest X-ray images. Sensors. 2021; 21 (21):7116. https://doi.org/10.3390/s21217116 PMID: 34770423

25. Chakraborty S, Saha AK, Nama S, Debnath S. COVID-19 X-ray image segmentation by modified whale optimization algorithm with population reduction. Computers in Biology and Medicine. 2021; 139:104984. https://doi.org/10.1016/j.compbiomed.2021.104984 PMID: 34739972

26. Bhattacharyya A, Bhaik D, Kumar S, Thakur P, Sharma R, Pachori RB. A deep learning based approach for automatic detection of COVID-19 cases using chest X-ray images. Biomedical Signal Processing and Control. 2022; 71:103182. https://doi.org/10.1016/j.bspc.2021.103182 PMID: 34580596

27. Ganaie M, Hu M, et al. Ensemble deep learning: A review. arXiv preprint arXiv:210402395. 2021;.

28. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: International Conference on Machine Learning. PMLR; 2017. p. 1321–1330.

29. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:170404861. 2017;.

30. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770–778.

31. Zhao H, Shi J, Qi X, Wang X, Jia J. Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 2881–2890.

32. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer; 2015. p. 234–241.

33. Murphy AH, Epstein ES. Verification of probabilistic predictions: A brief review. Journal of Applied Meteorology and Climatology. 1967; 6(5):748–755. https://doi.org/10.1175/1520-0450(1967)006%3C0748:VOPPAB%3E2.0.CO;2

34. Crowson CS, Atkinson EJ, Therneau TM. Assessing calibration of prognostic risk scores. Statistical methods in medical research. 2016; 25(4):1692–1706. https://doi.org/10.1177/0962280213497434 PMID: 23907781

35. Jiang X, Osl M, Kim J, Ohno-Machado L. Calibrating predictive model estimates to support personalized medicine. Journal of the American Medical Informatics Association. 2012; 19(2):263–274. https://doi.org/10.1136/amiajnl-2011-000291 PMID: 21984587

36. Naeini MP, Cooper G, Hauskrecht M. Obtaining well calibrated probabilities using bayesian binning. In: Twenty-Ninth AAAI Conference on Artificial Intelligence; 2015.

37. Sayin B, Krivosheev E, Ramírez J, Casati F, Taran E, Malanina V, et al. Crowd-Powered Hybrid Classification Services: Calibration is all you need. In: 2021 IEEE International Conference on Web Services (ICWS). IEEE; 2021. p. 42–50.

38. Breiman L. Bagging predictors. Machine learning. 1996; 24(2):123–140. https://doi.org/10.1007/BF00058655

39. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2015. p. 3431–3440.

40. Chawla NV. Data mining for imbalanced datasets: An overview. In: Data mining and knowledge discovery handbook. Springer; 2009. p. 875–886.

41. Saood A, Hatem I. COVID-19 lung CT image segmentation using deep learning methods: U-Net versus SegNet. BMC Medical Imaging. 2021; 21(1):1–10. https://doi.org/10.1186/s12880-020-00529-5

42. Milletari F, Navab N, Ahmadi SA. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV). IEEE; 2016. p. 565–571.

43. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: International conference on medical image computing and computer-assisted intervention. Springer; 2016. p. 424–432.

44. Isensee F, Petersen J, Klein A, Zimmerer D, Jaeger PF, Kohl S, et al. nnu-net: Self-adapting framework for u-net-based medical image segmentation. arXiv preprint arXiv:180910486. 2018;.

45. Cao Y, Xu Z, Feng J, Jin C, Han X, Wu H, et al. Longitudinal assessment of covid-19 using a deep learning–based quantitative ct pipeline: Illustration of two cases. Radiology: Cardiothoracic Imaging. 2020; 2(2):e200082. https://doi.org/10.1148/ryct.2020200082 PMID: 33778563

46. Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J. Unet++: A nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer; 2018. p. 3–11.

47. Shan F, Gao Y, Wang J, Shi W, Shi N, Han M, et al.. Lung Infection Quantification of COVID-19 in CT Images with Deep Learning; 2020.

48. Chen J, Wu L, Zhang J, Zhang L, Gong D, Zhao Y, et al. Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography: a prospective study. medRxiv. 2020.

49. Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, et al. Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:180403999. 2018;.

50. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. nature. 2017; 542(7639):115–118. https://doi.org/10.1038/nature21056 PMID: 28117445

51. Bojarski M, Del Testa D, Dworakowski D, Firner B, Flepp B, Goyal P, et al. End to end learning for self-driving cars. arXiv preprint arXiv:160407316. 2016;.

52. Sønderby CK, Espeholt L, Heek J, Dehghani M, Oliver A, Salimans T, et al. Metnet: A neural weather model for precipitation forecasting. arXiv preprint arXiv:200312140. 2020;.

53. Hernández-Lobato JM, Adams R. Probabilistic backpropagation for scalable learning of bayesian neural networks. In: International conference on machine learning. PMLR; 2015. p. 1861–1869.

54. Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. arXiv preprint arXiv:161201474. 2016;.

55. Sensoy M, Kaplan L, Kandemir M. Evidential deep learning to quantify classification uncertainty. arXiv preprint arXiv:180601768. 2018;.

56. Liu JZ, Lin Z, Padhy S, Tran D, Bedrax-Weiss T, Lakshminarayanan B. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. arXiv preprint arXiv:200610108. 2020;.

57. Wilson AG. The case for Bayesian deep learning. arXiv preprint arXiv:200110995. 2020;.