Calibrated Bagging Deep Learning for Image Semantic Segmentation: A Case Study on COVID-19 Chest X-ray Image

Lucy Nwosu, Xiangfang Li, Lijun Qian, Seungchan Kim, Xishuang Dong

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 uncertainties.

Index Terms—COVID-19 Semantic Segmentation, Model Calibration, Ensemble Deep Learning, Majority Voting

I. 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 pandemics [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 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]. Moreover, Lucy et al. [20] 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 [21] to recognize region of interests (ROIs) on CXR images, where a few previous work explored this task [22], [23], [24]. 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 metric 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 [25] and model calibration [26] to enhance performance of semantic segmentation, as well as reduce prediction uncertainty. It includes three stages: 1) training multiple state-of-the-art DL models such as fully convolutional networks (FCN) [27], FCN combined with ResNet [28], FCN combined with MobileNet [27], PSPNet [29], and UNet [30] 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) [26] 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.

The contributions in this study are below.

- We systematically compared performance of various state-of-the-art DL models on semantic segmentation on COVID-19 CXR data with different evaluation metrics. Moreover, the prediction uncertainty 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 demonstrates its effectiveness on improving performance and prediction certainty for semantic segmentation.

II. METHODOLOGY

The proposed method is built based on calibration error [31], [32], [33] and bagging deep learning [25] to enhance image segmentation with higher prediction certainty.

A. Calibration Error

The expected calibration error (ECE) and the maximum calibration error (MCE) are proposed to measure the quality of uncertainty for machine learning models in terms of prediction

Fig. 1. Flow of building and testing calibrated bagging deep learning based on calibration error (CE). SR denotes segmentation result generated by individual deep learning model.
accuracy [35], which is critical for high risk applications such as medical diagnosis [32], [33] and self-driving [34].

- **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

\[
ECE = \frac{1}{N} \sum_{k=1}^{K} \frac{n_k}{N} |acc(k) - conf(k)|
\]

where \(n_k\) is the number of predictions in bin \(k\), \(N\) is the total number of samples predicted, and \(acc(k)\) and \(conf(k)\) denote 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.

\[
MCE = \max_{k=1}^{K} |acc(k) - conf(k)|
\]

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.

**B. 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 [23]. Bagging (or bootstrap aggregating) generates a series of independent subsets from training data to build multiple individual predictors to build an ensemble model [36]. 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.

**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. return DL models \(DL = \{DL_1, DL_2, ..., DL_M\}\) and corresponding \(CE = \{CE_1, CE_2, ..., CE_M\}\)

**III. EXPERIMENT**

**A. Dataset**

We employed COVID-19 chest X-ray dataset [37] 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

https://github.com/v7labs/covid-19-xray-dataset
ground truth with two “Lung” segmentation masks (rendered as polygons, 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).

(a) original image  (b) ground truth

Fig. 3. An example of CXR image and corresponding ground truth.

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

B. Experimental settings

We employed five state-of-the-art individual models as baselines to evaluate performance of semantic segmentation, namely, UNet [30], PSPNet [29], FCN32 [37] (FCN with 32×upsampling), FCN32_ResNet50 (FCN32 combined with ResNet50 [28]), FCN32_MobileNet(FCN32 combined with MobileNet [27]), 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 I.

| Model           | Learning Rate | Batch Size | Epoch |
|-----------------|---------------|------------|-------|
| UNet            | 1e-3          | 2          | 50    |
| PSPNet          | 1e-3          | 2          | 70    |
| FCN32           | 1e-3          | 2          | 50    |
| FCN32_ResNet50  | 1e-3          | 2          | 50    |
| FCN32_MobileNet | 1e-3          | 2          | 50    |

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

C. Evaluation metric

Various evaluation metrics are employed to evaluate the performance of our proposed model, which includes accuracy, F1 score, sensitivity, and specificity. Accuracy is calculated by dividing the number of pixels identified correctly over the total number of pixels in chest X-ray images.

\[
\text{Accuracy} = \frac{N_{\text{correct}}}{N_{\text{total}}}. \tag{3}
\]

\[
F_{\text{score}} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{4}
\]

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

\[
\text{Precision} = \frac{TP}{TP + FP}, \tag{5}
\]

\[
\text{Recall} = \frac{TP}{TP + FN}. \tag{6}
\]

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 [38].

Moreover, we employed sensitivity and specificity to evaluate performance of semantic segmentation [39], 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.

\[
\text{Specificity} = \frac{TN}{TN + FP}. \tag{7}
\]

\[
\text{Sensitivity} = \frac{TP}{TP + FN}. \tag{8}
\]

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 to measure the calibration errors [26] 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.

D. Experimental results

We validate the proposed method from two perspectives: comprehensive performance comparison between the baselines and the proposed method, and hyper-parameter examination.
TABLE II
COMPARING PERFORMANCE BETWEEN THE BASELINES AND THE PROPOSED METHOD BASED ON VARIOUS EVALUATION METRICS AND CORRESPONDING STANDARD DEVIATIONS.

| DL                              | 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        | 94.3±2.5        | 2.5±2.1     | 37.6±19.1 |
| FCN32_ResNet50 (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 |
| FCN32_MobileNet (F32_M)         | 95.2±2.3    | 91.0±4.7     | 90.1±5.6        | 93.3±3.3        | 4.1±1.6     | 38.2±19.9 |
| Ensemble (Majority Voting (MV)) | 98.8±0.6    | 94.1±3.0     | 92.9±3.6        | 96.6±1.7        | 2.4±1.2     | 28.1±14.1 |
| Ensemble (Weighted Voting (ECE))| 99.1±0.5    | 95.4±2.9     | 94.3±2.9        | 97.1±1.5        | 2.3±1.2     | 24.7±12.4 |
| Ensemble (Weighted Voting (MCE))| 98.7±0.7    | 93.9±3.7     | 92.6±3.7        | 96.3±1.9        | 2.4±1.2     | 28.9±14.6 |
| Ensemble (Majority Voting + ECE + MCE (MVEM)) | 99.2±0.4    | 97.7±2.3     | 95.4±2.3        | 98.4±0.8        | 2.1±1.1     | 20.1±10.1 |

Fig. 4. An example of prediction visualization on semantic segmentation generated by the baselines and proposed models. F32_R50 and F32_M denotes FCN32_ResNet50 and FCN32_MobileNet while MV, ECE, and MVEM denotes Ensemble (Majority Voting (MV)), Ensemble (Weighted Voting (ECE)), and Ensemble (Majority Voting + ECE + MCE (MVEM)). 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.

TABLE III
COMPARING PERFORMANCE OF THE PROPOSED METHODS BUILT WITH DIFFERENT NUMBER OF INDIVIDUAL MODELS.

| DL                             | DL          | Accuracy (%) | Sensitivity (%) | Specificity (%) | F1score (%) | ECE (%) | MCE (%) |
|--------------------------------|-------------|--------------|-----------------|-----------------|-------------|---------|---------|
| Ensemble 2 (FCN32_RESNET50 + FCN32) | 95.8±2.1    | 92.3±3.9     | 91.0±4.5        | 94.0±2.0        | 2.5±1.3     | 37.6±18.8 |
| Ensemble (Majority Voting (MV)) | 99.0±0.5    | 95.3±2.4     | 94.4±2.8        | 96.9±1.6        | 2.3±1.2     | 22.3±11.3 |
| Ensemble (Weighted Voting (MCE))| 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) | 98.7±0.7    | 93.9±3.0     | 94.1±3.0        | 96.1±2.0        | 2.7±1.4     | 31.1±15.6 |
| Ensemble (Majority Voting (MV)) | 98.4±0.8    | 95.5±2.3     | 94.9±2.6        | 96.9±1.6        | 2.8±1.4     | 26.1±13.1 |
| Ensemble (Weighted Voting (ECE))| 98.3±0.9    | 93.1±3.5     | 93.0±3.5        | 96.0±2.0        | 2.8±1.4     | 31.2±15.6 |
| Ensemble (Weighted Voting (ECE))| 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 + FCN32_MOBILENET) | 98.7±0.7    | 95.0±2.5     | 94.6±2.7        | 96.5±1.8        | 2.7±1.4     | 24.3±13.5 |
| Ensemble (Majority Voting (MV)) | 99.1±1.1    | 94.1±3.0     | 93.8±3.1        | 96.1±2.0        | 3.0±1.5     | 32.3±15.0 |
| Ensemble (Weighted Voting (MCE))| 98.7±0.7    | 93.9±3.7     | 92.6±3.7        | 96.3±1.9        | 2.4±1.2     | 28.9±14.6 |
| Ensemble (Majority Voting + ECE + MCE (MVEM)) | 99.2±0.4    | 97.7±2.3     | 95.4±2.3        | 98.4±0.8        | 2.1±1.1     | 20.1±10.1 |

1) Performance Comparison: Table II 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. For these individual models, FCN32_ResNet50 outperforms other individual models with higher certainty. In addition, as one baseline,
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. 5. 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 MV, ECE, and MVEM perform better in that regard of predictions since yellow regions in their predictions are smaller, where the proposed method including ECE and MVEM 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.

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. Specially, 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 III 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 Ensemble (Weighted Voting (ECE)) and Ensemble (Majority Voting + ECE + MCE (MVEM)), where the values of Recall and 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 models involved in the proposed approach will reduce the size of missing components. Moreover, MVEM 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.
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 [30] and V-Net [40]. In the biomedical domain, there have been numerous techniques for lung segmentation with different purposes [41], [42]. The U-Net is an effective technique for segmenting both lung regions and lung lesions in COVID applications [43]. The U-Net built with fully convolutional network [30] 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. [44] proposed the UNet++ that inserts a nested convolutional structure between the encoding and decoding path. In addition, Milletari et al. [40] built V-Net using the residual blocks as the basic convolutional block, and optimized the network by a Dice loss. Furthermore, Shan et al. [45] 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 [46]. In recent years, attention mechanisms can learn the most discriminant part of the features in deep learning models. Oktay et al. [47] 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 [48], autonomous driving [49], and precipitation forecasting [50] 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. [51] proposed probabilistic back-propagation for learning Bayesian neural networks. Non-Bayesian-based methods develop various strategies such as model ensemble [52] and prior assumption on predictions [53] 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. [52] 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 (2020) et al. [54] 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. [55] 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.

V. 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.

Acknowledgment

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 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.

References

[1] X. Wang, X. Song, Y. Guan, B. Li, and J. Han, “Comprehensive named entity recognition on cori-19 with distant or weak supervision,” arXiv preprint arXiv:2003.12218, 2020.
[2] Ş. Öztürk, Ü. Özkaya, and M. Barstuğan, “Classification of coronavirus (covid-19) from x-ray and ct images using shrunked features,” International Journal of Imaging Systems and Technology, vol. 31, no. 1, pp. 5–15, 2021.
[46] J. Chen, L. Wu, J. Zhang, L. Zhang, D. Gong, Y. Zhao, S. Hu, Y. Wang, X. Hu, B. Zheng et al., “Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography: a prospective study,” medRxiv, 2020.

[47] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz et al., “Attention u-net: Learning where to look for the pancreas,” arXiv preprint arXiv:1804.03999, 2018.

[48] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, “Dermatologist-level classification of skin cancer with deep neural networks,” nature, vol. 542, no. 7639, pp. 115–118, 2017.

[49] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang et al., “End to end learning for self-driving cars,” arXiv preprint arXiv:1604.07316, 2016.

[50] C. K. Sønderby, L. Espeholt, J. Heek, M. Dehghani, A. Oliver, T. Salimans, S. Agrawal, J. Hickey, and N. Kalchbrenner, “Metnet: A neural weather model for precipitation forecasting,” arXiv preprint arXiv:2003.12140, 2020.

[51] J. M. Hernández-Lobato and R. Adams, “Probabilistic backpropagation for scalable learning of bayesian neural networks,” in International conference on machine learning. PMLR, 2015, pp. 1861–1869.

[52] B. Lakshminarayanan, A. Pritzel, and C. Blundell, “Simple and scalable predictive uncertainty estimation using deep ensembles,” arXiv preprint arXiv:1612.01474, 2016.

[53] M. Sensoy, L. Kaplan, and M. Kandemir, “Evidential deep learning to quantify classification uncertainty,” arXiv preprint arXiv:1806.01768, 2018.

[54] J. Z. Liu, Z. Lin, S. Padhy, D. Tran, T. Bedrax-Weiss, and B. Lakshminarayanan, “Simple and principled uncertainty estimation with deterministic deep learning via distance awareness,” arXiv preprint arXiv:2006.10108, 2020.

[55] A. G. Wilson, “The case for bayesian deep learning,” arXiv preprint arXiv:2001.10995, 2020.