A Deep Bayesian Ensembling Framework for COVID-19 Detection using Chest CT Images

AUTHOR(S)

P Tabarisaadi, Abbas Khosravi, Saeid Nahavandi

PUBLICATION DATE

01-01-2020

HANDLE

10536/DRO/DU:30147495

2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)
October 11-14, 2020. Toronto, Canada

A Deep Bayesian Ensembling Framework for COVID-19 Detection using Chest CT Images

Pegah Tabarisaadi, Abbas Khosravi, Saeid Nahavandi
Institute for Intelligent Systems Research and Innovation (IISRI)
Deakin University
VIC, Australia
{abbas.khosravi, saeid.nahavandi}@deakin.edu.au

***Abstract*—The chest computed tomography (CT) images have been used for COVID-19 detection. Automating the process of analyzing can save great amount of time and energy. In this paper a deep bayesian ensembling framework is proposed for automatic detection of COVID-19 cases using the chest CT scans. Data augmentation is applied to increase the size and quality of training data available. Transfer learning is utilized to extract informative features. The extracted features are used to train the three different bayesian classifiers. The uncertainty of the neural network predictions is estimated by anchored, unconstrained and regularized bayesian ensembling methods. The reliability of predictions is then delineated. The epistemic and aleatoric uncertainties are estimated and different bayesian classifiers are compared from different perspectives. We use a small dataset containing only 275 CT images of positive COVID-19 cases. The results sounds promising and they can be improved in the future, as the performance of deep neural networks is reliant to big datasets. Prediction accuracy and predictive uncertainty estimates for unseen chest CT images indicate that the deep bayesian ensembling is a promising framework for COVID-19 detection.**

***Index Terms*—Bayesian Ensembling, Uncertainty quantification, Deep Neural Networks, COVID-19, CT images.**

I. Introduction

Several cases of pneumonia with unknown origins were reported by Wuhan Municipal Health Commission in 31 December 2019 [1]. This was the first officially reported of a novel contagious disease called COVID-19 that is labeled as pandemic by the World Health Organization. As of 9 May 2020, 3855812 confirmed cases of COVID-19 and the total death of 265862 reported globally.

Diagnosing the new cases can be considered as one of the great challenges that the world face in this crisis. COVID-19 tests are used extensively to detect new cases but unfortunately not enough number of tests are available and they can not be produced fast enough. On the other hand it is reported that in some cases fake COVID-19 test kits were found.

The limitation of COVID-19 testing kits, caused to look for other alternative ways to diagnose it. As the corona virus involves respiratory tract including the lungs at the early stages, chest CT is prescribed in suspicious cases [2].

Analyzing the X-ray images needs radiology experts and it will take relatively long time. Whereas there are increasing number of patients around the world, novel solutions are required to remove the burden from the healthcare system.

Recently many researches have focused on automatic diagnosing COVID-19 using chest CT images [3]–[14]. Different dataset are gathered and many deep neural networks are proposed and different algorithms are applied for automatically diagnosing if a patient is affected by corona virus or not. In [15] a dataset containing 275 chest CT images of COVID-19 positive cases is prepared and a deep neural network (NN) is trained on the available images to be able to diagnose whether a patient infected by corona virus or not by analyzing his/her chest CT image. Comparing the predicted NNs results with the ground truth, whenever it is applicable define the accuracy of the NN prediction.

While many studies focuses on improving the accuracy and performance of NNs predictions, the uncertainty quantification (UQ) of the predictions has been largely overlooked [16]–[18]. Quantifying uncertainties associated with NN predictions is of vital importance in safety critical applications such as healthcare. In particular, the epistemic uncertainty must be properly quantified as it represents how much the end user can trust network predictions on new samples [19]. The epistemic uncertainty is mainly related the data deficiency and can be reduced through collection of more quality data. The UQ should also include the aleatoric uncertainty. This represents the inherent uncertainty in data and mainly relates to the measurement noise or image quality.

Bayesian neural networks (BNNs) were proposed to model the uncertainty in NNs, where the parameters of NNs are modeled with probability distributions [20]. While BNNs works ideally in small scale problems, they face infeasibility for larger ones.

On the other hand, ensembling used as an alternative for determining the uncertainty. Various NNs with different initial conditions are trained and the variance of the predictions is

978-1-7281-8526-2/20/$31.00 ©2020 IEEE

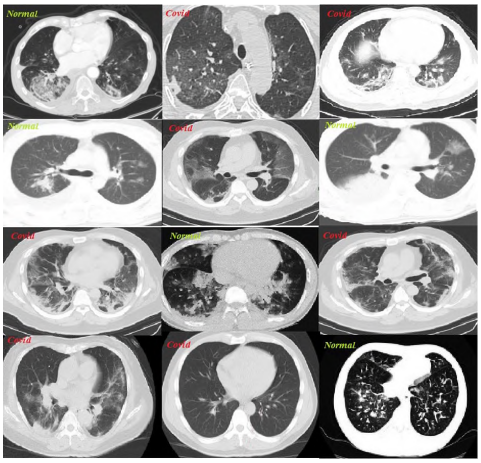

Figure 1. A few examples of the chest computed tomography (CT) images that is used in this study [15]. The positive and negative cases for COVID-19 are marked with Covid and Normal respectively.

regarded as the uncertainty. Also ensembling was successful in many applications [21]–[23] it was criticized for not being bayesian [24].

The so called Bayesian ensembling proposed in [25], bridges the BNNs with ensembling method to take advantage of both strengths.

In this paper we use a small chest X-ray image dataset [15]. A pretrained network that was previously trained on a big and general dataset is applied to address this issue (transfer learning). Data augmentation is applied to prepare more diversity to the available train dataset. Then three bayesian ensembling algorithms are applied to analyze the CT images of suspicious cases to detect if they are infected by COVID-19 or not. On the other hand uncertainties of the predictions are estimated to determine how dependable the predictions are. The three bayesian ensembling methods are compared from different perspective and the results are reported. This can be considered as the main contribution of this study. To the best of the authors knowledge, the UQ of automatic detection of COVID-19 has not been referenced in the literature yet. Collecting big datasets may be a little time consuming but it will be surely fulfilled in near future. So the promising results achieved can also be improved as well.

II. Background

A. Bayesian Neural Networks

Deep NNs achieved great success in recent years. However in real world application the reliability level and robustness of the NNs predictions was always an issue. Bayesian NNs, by joining the probability theory and NNs tries to address this concern.

Consider a supervised learning problem where $D = \{x_i, y_i\}_{i=1}^{N}$ is the set of training examples, considering:

$$P(\theta|D) \propto P(D|\theta)P(\theta) \tag{1}$$

where $P(\theta)$ is the prior distribution of the parameters. It illustrates our prior knowledge about θ.

$$P(D|\theta) = \prod_{(x_i,y_i)\in D} P(y_i|x_i,\theta) \tag{2}$$

$P(D|\theta)$ is known as the likelihood and it is defined as the probability of observing D, given the parameter θ. Given a test input x, the posterior predictive distribution of y is as follows:

$$P(y|x,D) = \int P(y|x,\theta)P(\theta|D)d\theta \tag{3}$$

As equation (3) is intractable, different bayesian approaches have been developed to estimate it [26]. In variational inference (VI), the posterior is approximated through a simple form and the parameters are obtained by optimizing the evidence lower bound [27], [28]. In stochastic gradient Langevian dynamics, the gradient noise is added during training to approximate samples from the posterior [29]. Dropout [30], batch normalization [31] and ensembling [32] can also be interpreted as bayesian approaches.

It is worth mentioning that maximum likelihood estimation (MLE) and maximum a posterior (MAP) estimate the parameter θ as follows:

$$\hat{\theta}_{MLE} = argmax_\theta\{P(D|\theta)\} \tag{4}$$

$$\begin{aligned}\hat{\theta}_{MAP} = argmax_\theta\{P(\theta|D)\} &= argmax_\theta\{\frac{P(D|\theta)P(\theta)}{P(D)}\} \\ &= argmax_\theta\{P(D|\theta)P(\theta)\}\end{aligned} \tag{5}$$

From equations (4) and (5), one can conclude that MLE and MAP are both point estimators, whilst bayesian inference methods returns probability density function.

B. Randomized MAP Sampling(RMS)

In recent years randomized MAP sampling is proposed as a bayesian inference method. In this approach a regularization term is added to the standard MLE loss function and a MAP parameter estimate is returned. In addition adding noise to the this loss and sampling frequently, build a distribution of MAP, that is the estimated of the true parameter posterior distribution [25], [33]–[36] .

C. Transfer Learning

Obtaining large and comprehensive dataset in medical imaging domain is considered as a challenge. When sufficient data are not available one popular approach is transfer learning.

Transfer learning is a well known approach that is used in computer vision tasks where we do not have access to a large enough dataset. In this method a convolutional neural network (CNN) model that is already trained on a large and general dataset (pretrained model) is used for a task even completely different from its origin one.

In medical imaging domain, as sufficient data is always an issue, transfer learning was applied extensively. Typically, the

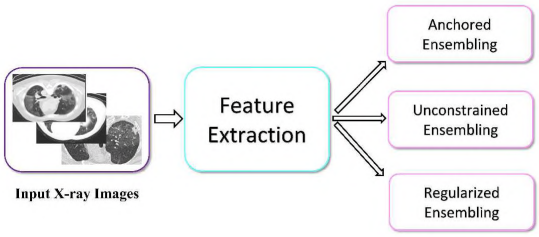

Figure 2. The bayesian ensembling framework for detecting COVID-19 form chest X-ray images. The input chest CT images are fed to the feature extraction and then to three bayesian ensembling classifiers.

pretrained CNN is chosen and applied to the image data, the output are features and they are used to train a separate classifier [37]. For instance [38], [39] used pretrained CNNs as feature extraction for chest pathology and in [40] , [41] the extracted features are used for pulmonary nodule detection.

In this paper we use pretrained network that was trained on ImageNet (where the classes are mostly every day objects such as animals, plants and etc.) to diagnose COVID-19 in chest X-ray images.

III. Methods

We develop a bayesian ensembling framework for diagnosing the Covid-19 from chest X-ray images. We also quantify the uncertainty of the proposed NN predictions. It is worth mentioning that also the dataset we use is the largest currently available dataset it is actually very small for training a deep NN. The main issue with such a small dataset is overfitting. To address this we use two main techniques called transfer learning and data augmentation.

Finally a bayesian ensembling classifier is applied to classify the images and determine the level of uncertainty of our predictions.

A. Extract Features Using VGG16

We apply VGG16 [42] (also known as the OxfordNet). This deep convolutional neural network was already trained on a big dataset (ImageNet). ImageNet contains over 15 millions high quality images belonged to 22000 categories. Also our dataset is completely different in nature from ImageNet, it is very helpful. It was due to the fact that the main duty of VGG16 in our codes is to extract the features of the images.

Our feature extraction that is VGG16 followed by average-pooling and flatten layers has 14,714,688 non trainable parameters. The extracted features that has 512 features are used as inputs to train the bayesian classifier.

B. Data Augmentation

The deep learning success in recent years was partially attributed to availability of large and diverse dataset. In addition, having too few data to learn from leads to overfitting. Infinite dataset helps the model to learn every aspect of the data and the model never overfit. Data augmentation is a technique to artificially generate more training data by modifying the existing training samples [43]. In other words data augmentation provide more diversity in dataset without the need to collect new data.

Cropping, padding and horizontal flipping and other minor alterations are popular techniques for applying data augmentation. The NN considers the augmented data as distinct images.

C. Loss Function

Considering equation (5), one has:

$$\hat{\theta}_{MAP} = argmax_{\theta}\{P(D|\theta)P(\theta)\} \tag{6}$$

Assuming normal distribution for the prior, $P(\theta) = N(\mu, \Sigma)$:

$$\hat{\theta}_{MAP} = argmax_{\theta} log(P(D|\theta)) - \frac{1}{2}(\theta-\mu)^T\Sigma^{-1}(\theta-\mu)\{P(D|\theta)P(\theta)\} \tag{7}$$

Choosing prior covariance as diagonal and $\mu = 0$, one has:

$$\hat{\theta}_{MAP} = argmax_{\theta} log(P(D|\theta)) - \frac{1}{2}||\Sigma^{\frac{-1}{2}}.\theta||_2^2 \tag{8}$$

Equation (8) is standard $L2$ regularization. For classification, multinomial distribution is common choice for the likelihood.

$$P(D|\theta) \propto \prod_{n=1}^{N}\prod_{c=1}^{C} \hat{y}_{n,c}^{y_{n,c}} \tag{9}$$

where C is the number of classes, N is data points and $\hat{y} \in \{0,1\}$ is the predicted probability and true targets are denoted by $y_{n,c} \in \{0,1\}$.

$$\hat{\theta}_{MAP} = argmax_{\theta} \sum_{n-1}^{N}\sum_{c-1}^{C} y_{n,c} log(\hat{y}_{n,c}) - \frac{1}{2}||\Sigma^{\frac{-1}{2}}.\theta||_2^2 \tag{10}$$

Applying RMS, $\mu prior$ is replaced with a random variable θ_{anc}. A typical choice for θ_{anc} is:

$$\theta_{anc} = N(\mu prior, \sum prior) \tag{11}$$

For classification task the following loss function is optimized:

$$Loss_j = \frac{1}{N}\sum_{n-1}^{N}\sum_{c-1}^{C} y_{n,c} log(\hat{y}_{n,c}) + \frac{1}{N}||\Gamma^{\frac{1}{2}}.(\theta - \theta_{anc,j})||_2^2 \tag{12}$$

Here, we choose $diag(\Gamma_i) = \frac{1}{2\sigma^2_{prior_i}}$ [25]. Choosing $\Gamma = 0$ is the so-called unconstrained ensembling, we refer it as free ensembling in our results. In addition considering $\theta_{anc,j} = 0$ will return the regularized ensembling. In anchored ensembling the parameters are regularized around the prior distribution parameters (equation (12)).

D. Validation

Accuracy is not always a good measure to determine the model performance. In diagnosing a disease it is of vital importance to consider other measures as well because over and under diagnosis may cause bad effects. If our model wrongly determine a healthy person as an infected by corona virus, leads to subject the patient to treatment that is unnecessary. It will also cause unneeded anxiety for the patient. On the other hand if our model diagnose an infected by Covid-19 as healthy, it will lead to spread the virus and can impose huge costs to the community. Sensitivity, specificity, precision, recall, etc are considered as popular measures in medical testing that tell us what the model can not tell us. [44]–[46]

Sensitivity:The sensitivity also known as true positive rate is a measure to determine how accurate is the model in diagnosing positive results for people who have the disease.The sensitivity is calculated as follows:

$$Sensitivity = \frac{TP}{TP + FN} \tag{13}$$

where, TP (True Positive), stands for number of patients who labeled as positive and are truly positive and FN (False Negative), stands for patients that was predicted as negative but they are really positive.

Specificity:The specificity also known as true negative rate is a measure to determine how accurate is the model in predicting negative results for healthy people. The specificity is calculated as follows:

$$Specificity = \frac{TN}{TN + FP} \tag{14}$$

where, TN (True Negative), stands for number of patients who labeled as negative and are truly negative and FP (False Positive), stands for patients that was predicted as positive but they are really negative.

Precision: precision determine the ability of the model to return relevant samples. In other words if the focus is on minimizing FP it is a suitable measure.

$$Precision = \frac{TP}{TP + FP} \tag{15}$$

Recall: recall is a suitable measure when the focus is on minimizing FN.

$$Recall = \frac{TP}{TP + FN} \tag{16}$$

F1: F1 is a weighted mean of precision and recall and it is calculated as follows:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{17}$$

ROC (Receiver Operating Characteristic curve): ROC is the plot of sensitivity against specificity at different thresholds. The area under the ROC curve is known as (AUC). AUC is a measure to determine the quality of the classifier.

Notice that all the measures above are ranged between 0 and 1. The best value for them is 1 and the worst is 0. For instance a random classifier has the $AUC = 0.5$ while the for the perfect classifier the $AUC = 1$.

Table I
STATISTICS OF DATA SPLIT. THE IMAGES ARE DIVIDED TO TEST AND TRAIN SETS.

Item	COVID-19	Normal	Total
Train	220	156	376
Test	55	39	94

Table II
DATASET AND PREPROCESSEING INFORMATION.

Item	Value
Number of images	470
Number of COVID-19 images	275
Number of Normal images	195
Image size	224×224
Augmentation parameter (rg)	15

Table III
THE EPISTEMIC AND ALEATORIC UNCERTAINTIES FOR ANCHORED ENSEMBLING (ANC), REGULARIZED ENSEMBLING (REG) AND UNCONSTRAINED ENSEMBLING (FREE). ANCHORED ENSEMBLING PREDICT THE HIGHEST UNCERTAINTIES IN COMPARISON TO OTHER METHODS.

Method	Total Uncertainty
10 × NNs Anc	1.040 ± 1.38
10 × NNs Reg	0.192 ± 0.515
10 × NNs Free	0.676 ± 1.435

IV. EXPERIMENTS

In this section our finding on detecting if a patient is infected by COVID-19 by analyzing his/her chest CT images is described in details. The dataset we use includes 470 CT scans, where 275 images are COVID-19 positive. It is worth mentioning that some of the positive COVID-19 images are the X-ray images of one patient in different stages of his/her illness. We split the dataset to train and test set, the number of Covid and Normal images in each set is represented in Table I. Table II also provides information about the number of images and images per class. All the images are resized to 224-by-224. The data augmentation is applied by randomly rotating the train images in 15 degree rotation range. The pretrained VGG16 is utilized as the feature extraction model.

We apply three popular approximate and inference methods for classification. The methods are tested for 5, 10 and 15

Table IV
COMPARING THE ACCURACY, SENSITIVITY AND SPECIFICITY REPORTED FOR ANCHORED ENSEMBLING (ANC), REGULARIZED ENSEMBLING (REG) AND UNCONSTRAINED ENSEMBLING (FREE). RESULTS ARE REPORTED FOR 5,10 AND 15 NNS.

Method	Accuracy (%)	Sensitivity	Specificity
5×NNs Anc	79.7	0.86	0.83
10×NNs Anc	82.6	0.85	0.80
15×NNs Anc	81.9	0.85	0.81
5×NNs Reg	82.6	0.86	0.80
10×NNs Reg	80.5	0.81	0.81
15×NNs Reg	82.6	0.823	0.80
5×NNs Free	81.2	0.899	0.77
10×NNs Free	83.2	0.87	0.81
15×NNs Free	83.9	0.886	0.77

Table V
COMPARING THE PRECISION, RECALL AND F1 FOR ANCHORED ENSEMBLING (ANC), REGULARIZED ENSEMBLING (REG) AND UNCONSTRAINED ENSEMBLING (FREE). RESULTS ARE TESTED FOR 5,10 AND 15 NNS.

Method	Precision	Recall	F1
5×NNs Anc	0.85	0.86	0.85
10×NNs Anc	0.83	0.85	0.84
15×NNs Anc	0.83	0.85	0.84
5×NNs Reg	0.83	0.83	0.83
10×NNs Reg	0.84	0.84	0.84
15×NNs Reg	0.84	0.86	0.85
5×NNs Free	0.85	0.85	0.85
10×NNs Free	0.85	0.84	0.84
15×NNs Free	0.83	0.86	0.84

Table VI
COMPARING THE AUC FOR ANCHORED ENSEMBLING (ANC), REGULARIZED ENSEMBLING (REG) AND UNCONSTRAINED ENSEMBLING (FREE). RESULTS ARE REPORTED FOR 5,10 AND 15 NNS.

Method	COVID-19	Normal	Micro-average	Macro-average
5×NNs Anc	0.85	0.85	0.85	0.86
10×NNs Anc	0.88	0.88	0.87	0.88
15×NNs Anc	0.87	0.87	0.87	0.88
5×NNs Reg	0.89	0.89	0.89	0.89
10×NNs Reg	0.88	0.88	0.88	0.88
15×NNs Reg	0.88	0.88	0.88	0.89
5×NNs Free	0.90	0.90	0.90	0.91
10×NNs Free	0.89	0.89	0.89	0.90
15×NNs Free	0.89	0.89	0.90	0.90

NNs. The NNs structure is constructed by 3 fully-connected layers. Anchored ensemble that is trained using equation (12), unconstrained ensembles (where $\Gamma = 0$) and regularized ensembles (where $\theta_{anc,j} = 0$).

In regularized ensembles all NNs are encouraged to the same solution and the diversity is decreased, while in unconstrained there is no sense of prior and may cause overfitting.

Table III compares the epistemic and aleatoric uncertainties for these ensembling methods. It can be concluded that the anchored ensembling produces the most conservative predictions in comparison to other ensembling methods. Table IV compares the accuracy, sensitivity and specificity of these methods for three-layer NNs with ReLU nonlinearity. Table V illustrate the precision, recall and F1. Table VII and Table IV show the confusion matrices and AUC for all the scenarios. In Fig. 3 the ROC curve for anchored ensembling with 10 NNs is illustrated.

Table VII
THE NORMALIZED CONFUSION MATRIX FOR ANCHORED, UNCONSTRAINED AND REGULARIZED ENSEMBLING. THE RESULTS ARE REPORTED FOR 5, 10 AND 15 NNS. TN, FP, FN AND TP STANDS FOR NORMALIZED TRUE NEGATIVE, FALSE POSITIVE, FALSE NEGATIVE, AND TRUE POSITIVE RESPECTIVELY.

NN	TN	FP	FN	TP
5×NNs Anc	0.83	0.17	0.14	0.86
10×NNs Anc	0.80	0.20	0.15	0.85
15×NNs Anc	0.81	0.19	0.15	0.85
5×NNs Reg	0.80	0.20	0.16	0.84
10×NNs Reg	0.81	0.19	0.16	0.84
15×NNs Reg	0.81	0.19	0.14	0.86
5×NNs Free	0.83	0.17	0.15	0.85
10×NNs Free	0.81	0.19	0.15	0.85
15×NNs Free	0.80	0.20	0.14	0.86

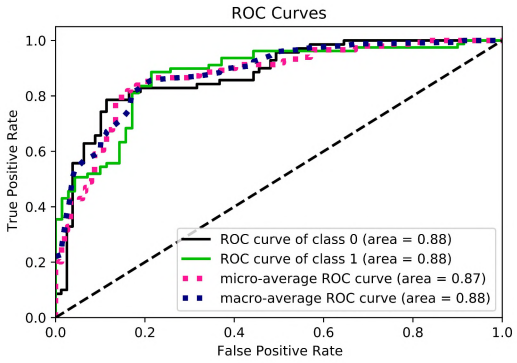

Figure 3. The ROC curve for Anchored ensembling with 10 NNs. The area under the ROC curve (AUC) is an important performance measurement. AUC for random classifier is 0.5 and AUC for a perfect classifier is 1. Class 1 is positive covid-19 and class 0 is negative cases.

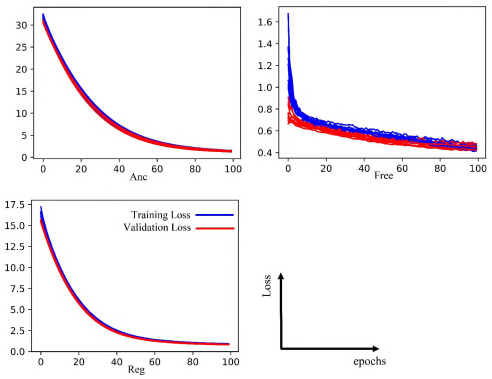

Figure 4. The training and validation loss per epochs for Anchored, Unconstrained and Regularized ensembling.

The training and validation loss for anchored ensembling, regularized ensembling and unconstrained ensembling is illustrated in Fig. 4. It can be concluded that all the methods converge after 60 epochs and regularized ensembling converges best among others.

V. CONCLUSION

This paper aimed to apply bayesian framework for detecting COVID-19 from chest X-ray images. VGG16 with the weights trained on ImageNet is applied as feature extraction. The extracted features are applied to three bayesian ensembling classifiers. Anchored ensembling, regularized ensembling and unconstrained ensembling are tested separately and the results are reported and compared. The reported results show the effectiveness of all in this task however it is concluded that the anchored ensembling is the most conservative one as it predicts the highest uncertainties in comparison to others. Although the proposed framework achieves good results, access to bigger

dataset will definitely lead to more accurate diagnosis with higher confidence.

REFERENCES

[1] W.H.O, "Covid-19timeline," 2020. https://www.who.int/newsroom/detail/.

[2] S. Salehi, A. Abedi, S. Balakrishnan, and A. Gholamrezanezhad, "Coronavirus disease 2019 (COVID-19): a systematic review of imaging findings in 919 patients," *am j roentgenol*, pp. 1–7, 2020.

[3] L. Wang and A. Wong, "COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest radiography images," *arXiv preprint arXiv:2003.09871*, 2020.

[4] L. Li, L. Qin, Z. Xu, Y. Yin, X. Wang, B. Kong, J. Bai, Y. Lu, Z. Fang, Q. Song, *et al.*, "Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT," *Radiology*, p. 200905, 2020.

[5] L. Huang, R. Han, T. Ai, P. Yu, H. Kang, Q. Tao, and L. Xia, "Serial quantitative chest CT assessment of COVID-19: deep-learning approach," *Radiology: Cardiothoracic Imaging*, vol. 2, no. 2, p. e200075, 2020.

[6] C. Butt, J. Gill, D. Chun, and B. A. Babu, "Deep learning system to screen coronavirus disease 2019 pneumonia," *Applied Intelligence*, p. 1, 2020.

[7] 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.

[8] I. D. Apostolopoulos and T. A. Mpesiana, "Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks," *Physical and Engineering Sciences in Medicine*, p. 1, 2020.

[9] J. Zhang, Y. Xie, Y. Li, C. Shen, and Y. Xia, "Covid-19 screening on chest x-ray images using deep learning based anomaly detection," *arXiv preprint arXiv:2003.12338*, 2020.

[10] E. E.-D. Hemdan, M. A. Shouman, and M. E. Karar, "Covidx-net: A framework of deep learning classifiers to diagnose covid-19 in x-ray images," *arXiv preprint arXiv:2003.11055*, 2020.

[11] M. Farooq and A. Hafeez, "Covid-resnet: A deep learning framework for screening of covid19 from radiographs," *arXiv preprint arXiv:2003.14395*, 2020.

[12] C. Zheng, X. Deng, Q. Fu, Q. Zhou, J. Feng, H. Ma, W. Liu, and X. Wang, "Deep learning-based detection for COVID-19 from chest CT using weak label," *medRxiv*, 2020.

[13] A. Shoeibi, M. Khodatars, R. Alizadehsani, N. Ghassemi, M. Jafari, P. Moridian, A. Khadem, D. Sadeghi, S. Hussain, A. Zare, *et al.*, "Automated detection and forecasting of covid-19 using deep learning techniques: a review," *arXiv preprint arXiv:2007.10785*, 2020.

[14] A. S. Jokandan, H. Asgharnezhad, S. S. Jokandan, A. Khosravi, P. M. Kebria, D. Nahavandi, S. Nahavandi, and D. Srinivasan, "An uncertainty-aware transfer learning-based framework for covid-19 diagnosis," *arXiv preprint arXiv:2007.14846*, 2020.

[15] J. Zhao, Y. Zhang, X. He, and P. Xie, "COVID-CT-Dataset: a CT scan dataset about COVID-19," *arXiv preprint arXiv:2003.13865*, 2020.

[16] H. M. Kabir, A. Khosravi, A. Kavousi-Fard, S. Nahavandi, and D. Srinivasan, "Optimal uncertainty-guided neural network training," *arXiv preprint arXiv:1912.12761*, 2019.

[17] H. Quan, A. Khosravi, D. Yang, and D. Srinivasan, "A survey of computational intelligence techniques for wind power uncertainty quantification in smart grids," *IEEE Transactions on Neural Networks and Learning Systems*, 2019.

[18] H. M. D. Kabir, A. Khosravi, M. A. Hosen, and S. Nahavandi, "Neural network-based uncertainty quantification: a survey of methodologies and applications," *IEEE access*, vol. 6, pp. 36218–36234, 2018.

[19] J. Postels, F. Ferroni, H. Coskun, N. Navab, and F. Tombari, "Sampling-free epistemic uncertainty estimation using approximated variance propagation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2931–2940, 2019.

[20] D. J. C. MacKay, "A practical bayesian framework for backpropagation networks," *Neural computation*, vol. 4, no. 3, pp. 448–472, 1992.

[21] R. Tibshirani, "A comparison of some error estimates for neural network models," *Neural Computation*, vol. 8, no. 1, pp. 152–163, 1996.

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

[23] I. Osband, C. Blundell, A. Pritzel, and B. Van Roy, "Deep exploration via bootstrapped DQN," in *Advances in neural information processing systems*, pp. 4026–4034, 2016.

[24] Y. Gal, *Uncertainty in deep learning*. PhD thesis, 2016.

[25] T. Pearce, M. Zaki, A. Brintrup, N. Anastassacos, and A. Neely, "Uncertainty in neural networks: Bayesian ensembling," *arXiv preprint arXiv:1810.05546*, 2018.

[26] P. Myshkov and S. Julier, "Posterior distribution analysis for bayesian inference in neural networks," in *Workshop on Bayesian Deep Learning, NIPS*, 2016.

[27] A. Graves, "Practical variational inference for neural networks," in *Advances in neural information processing systems*, pp. 2348–2356, 2011.

[28] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural networks," *arXiv preprint arXiv:1505.05424*, 2015.

[29] M. Welling and Y. W. Teh, "Bayesian learning via stochastic gradient Langevin dynamics," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 681–688, 2011.

[30] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[31] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[32] D. Duvenaud, D. Maclaurin, and R. Adams, "Early stopping as nonparametric variational inference," in *Artificial Intelligence and Statistics*, pp. 1070–1077, 2016.

[33] Y. Gu, D. S. Oliver, *et al.*, "An iterative ensemble Kalman filter for multiphase fluid flow data assimilation," *Spe Journal*, vol. 12, no. 04, pp. 438–446, 2007.

[34] Y. Chen and D. S. Oliver, "Ensemble randomized maximum likelihood method as an iterative ensemble smoother," *Mathematical Geosciences*, vol. 44, no. 1, pp. 1–26, 2012.

[35] J. M. Bardsley, "MCMC-based image reconstruction with uncertainty quantification," *SIAM Journal on Scientific Computing*, vol. 34, no. 3, pp. A1316–A1332, 2012.

[36] J. M. Bardsley, A. Solonen, H. Haario, and M. Laine, "Randomize-then-optimize: A method for sampling from posterior distributions in nonlinear inverse problems," *SIAM Journal on Scientific Computing*, vol. 36, no. 4, pp. A1895–A1910, 2014.

[37] H. Greenspan, B. Van Ginneken, and R. M. Summers, "Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1153–1159, 2016.

[38] Y. Bar, I. Diamant, L. Wolf, and H. Greenspan, "Deep learning with non-medical training used for chest pathology identification," in *Medical Imaging 2015: Computer-Aided Diagnosis*, vol. 9414, p. 94140V, International Society for Optics and Photonics, 2015.

[39] Y. Bar, I. Diamant, L. Wolf, S. Lieberman, E. Konen, and H. Greenspan, "Chest pathology detection using deep learning with non-medical training," in *2015 IEEE 12th international symposium on biomedical imaging (ISBI)*, pp. 294–297, IEEE, 2015.

[40] B. Van Ginneken, A. A. A. Setio, C. Jacobs, and F. Ciompi, "Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans," in *2015 IEEE 12th International symposium on biomedical imaging (ISBI)*, pp. 286–289, IEEE, 2015.

[41] D. Zhao, D. Zhu, J. Lu, Y. Luo, and G. Zhang, "Synthetic medical images using F&BGAN for improved lung nodules classification by multi-scale VGG16," *Symmetry*, vol. 10, no. 10, p. 519, 2018.

[42] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[43] F. Chollet, *Deep Learning with Python*. Manning publication Co, 2017.

[44] N. Ramachandran, S. C. Hong, M. J. Sime, and G. A. Wilson, "Diabetic retinopathy screening using deep neural network," *Clinical & experimental ophthalmology*, vol. 46, no. 4, pp. 412–416, 2018.

[45] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321–331, 2018.

[46] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature medicine*, vol. 25, no. 1, p. 65, 2019.