Reducing Risk and Uncertainty of Deep Neural Networks on Diagnosing COVID-19 Infection

[1] **Krishanu Sarker,** [2] **Sharbani Pandit,** [3] **Anupam Sarker,** [1] **Saeid Belkasim,** [1] **Shihao Ji**

[1] Georgia State University
[2] Georgia Institute of Technology
[3] Institute of Epidemiology, Disease Control and Research

Abstract

Effective and reliable screening of patients via Computer-Aided Diagnosis can play a crucial part in the battle against COVID-19. Most of the existing works focus on developing sophisticated methods yielding high detection performance, yet not addressing the issue of predictive uncertainty. In this work, we introduce uncertainty estimation to detect confusing cases for expert referral to address the unreliability of state-of-the-art (SOTA) DNNs on COVID-19 detection. To the best of our knowledge, we are the first to address this issue on the COVID-19 detection problem. In this work, we investigate a number of SOTA uncertainty estimation methods on publicly available COVID dataset and present our experimental findings. In collaboration with medical professionals, we further validate the results to ensure the viability of the best performing method in clinical practice.

Introduction

The incredible success has inspired the use of deep learning in the medical imaging field (Erickson et al. 2017), e.g., Computer-Aided Diagnosis (CAD) (Doi 2004), medical image analysis, etc. However, such systems are still not being utilized in clinical practice (Yanase and Triantaphyllou 2019). One of the major reasons behind is the lack of reliability of existing CAD systems. Even though CAD has been studied widely, the uncertainty estimation of DNNs in medical imaging is remarkably understudied (Laves, Ihler, and Ortmaier 2019; Poduval, Loya, and Sethi 2020). Hence, in this paper we aim to conduct a comprehensive study on mitigating uncertainty of DNNs on COVID-19 detection.

COVID patients often develop lung infection which can be visible through chest X-ray (CXR) and CT scan (Cleverley, Piper, and Jones 2020). A DNN enabled CAD system can potentially be utilized to diagnose COVID through CXR analysis and provide noninvasive detection solutions that would reduce pressure on critical resources. A number of high performing models have been proposed (Chen et al. 2020) to detect COVID from CXR. COVID-Net (Wang and Wong 2020) is one such model that achieves high positive predictive value (PPV) for detecting COVID positive CXR samples. Even though these methods achieve very high accuracy, the predictive uncertainty problem still persists.

The predictive uncertainty of DNNs has received a lot of attention in the literature (Liu et al. 2019; Cordella et al. 1995; Geifman and El-Yaniv 2019; Sarker et al. 2020). However, most of them are evaluated only with benchmark datasets. These empirical performances often translate poorly into real-world scenario. Most of these methods often require extensive modifications to underlying DNNs which makes them unsuitable for wide deployment.

To bridge the gap, we investigate efficacy of three SOTA uncertainty estimation methods on reducing unreliability of DNNs on the COVID-19 detection. Our experiments reveal that abstention framework proposed in (Sarker et al. 2020) outperforms other SOTA methods on COVID detection, while requiring minimum effort to incorporate with existing DNNs. Through visualization we further demonstrate the effectiveness of the best performing method on the COVIDx dataset (Wang and Wong 2020). We also propose a statistical testing based feature selection method to improve the abstention framework to achieve higher Positive Predictive Value (PPV) for COVID-19 positive cases. Please refer to supplementary materials for details.

Our contributions are summarized in the following.

- Investigation of uncertainty estimation methods to detect confusing cases on COVID diagnosis. To the best of our knowledge, we are the first to comprehensively study the uncertainty of CAD systems on COVID diagnoses.

- Validation of the abstained samples by the best performing framework with medical professionals. Expert opinion on confusing samples abstained by the framework further validates the usability of the framework on screening COVID patients.

- Through extensive experimentation and performance analysis, we provide proof of efficacy of the SOTA uncertainty estimation methods on COVID-19 diagnosis.

Related Works

A number of extraordinary research work have been done on COVID-19 Diagnosis from CXR images (Chen et al. 2020). Even though CXR is not very reliable for COVID detection, majority of these works show promising performance. However, most of these SOTA methods (Sethy and Behera 2020; Castiglioni et al. 2020; Wang and Wong 2020; Zhang

et al. 2020) do not consider the issue of predictive uncertainty, which drastically reduces the reliability of such high performing DNN based CAD systems on COVID detection.

Only a handful of works address the uncertainty of COVID detection methods (Mallick et al. 2020; Ghoshal and Tucker 2020). Mallick et. al. propose a neighborhood components analysis over latent space to estimate uncertainty. However, this approach requires modification to the existing DNN model, which reduces the system's applicability. Ghoshal et. al. evaluates the usefulness of estimating uncertainty approximating Bayesian Convolutional Neural Networks (BCNN). They have eloquently presented how BCNNs help identify uncertain predictions. However, Bayesian networks are known to be intractable, and approximating the solution leads to sub-optimal solutions. Moreover, none of the works present comparative study with SOTA methods.

A large number of research works exist in the literature aiming at the issue of uncertainty estimation of DNNs (Liu et al. 2019; De Stefano, Sansone, and Vento 2000; Bartlett and Wegkamp 2008; Cordella et al. 1995; Geifman and El-Yaniv 2019). However, most of these works only experimented with benchmark datasets ignoring the stochasticity associated with real-world scenarios. Though greatly understudied, some research works on uncertainty estimation, are curated for real-world healthcare datasets (Leibig et al. 2017; Ayhan et al. 2020). Despite that, no work till now addresses the issue of uncertainty on COVID detection.

Methodology

In this work we choose three most recent SOTA uncertainty estimation methods to tackle the issue of predictive uncertainty of DNNs on detecting COVID positive samples. In this section, we will briefly discuss each of these methods and present their pros and cons. For further details on each of these methods, please refer to the corresponding papers.

Test Time Augmentation framework (TTAUG) (Ayhan et al. 2020) proposes an intuitive framework based on test-time augmentation for quantifying the diagnostic uncertainty of Bayesian CNNs. TTAUG is relatively simple to incorporate with any DNNs, as it only augments input samples. Authors propose to use best practice data augmentations to estimate the probabilistic uncertainty with temperature scaling. However, this method require domain knowledge to design appropriate augmentation for the task. Also, augmenting test samples is resource intensive. Moreover, even though Bayesian statistics provide simpler ways to estimate the uncertainty, these type of methods are intractable in most of the real-world scenarios.

SelectiveNet (Geifman and El-Yaniv 2019) proposes a user-defined coverage constraint to learn to abstain samples with high classification loss. By minimizing overall loss, the model learns to abstain from test samples that are difficult to predict. Authors define selective model as a pair (f, g), where f is a prediction function, and $g : X \rightarrow \{0, 1\}$ is a selection function, which is a binary qualifier for f.

$$(f, g)(x) \triangleq \begin{cases} f(x), & \text{if } g(x) = 1; \\ \text{don't know}, & \text{if } g(x) = 0. \end{cases}$$

SelectiveNet can achieve compelling performance when trained with desired abstention rate. However, this method is more complex and requires extensive repetitive training. On top of that, SelectiveNet requires extensive modification to the DNN, which makes it difficult to deploy in practice.

Density-based Filtering Framework (DbFF) (Sarker et al. 2020) proposes a plug-and-play framework that utilizes the underlying data density of training data to differentiate between confusing and certain predictions of data samples. Authors propose to first identify core data distributions for each class using $DBSCAN$ (Ester et al. 1996) clustering algorithm. Based on these core distributions they propose to calculate centroid as the identifier of these clusters.

$$c_j = median([v_{x_i^j}]_{i=0}^{m_{core}}), \quad \forall x_i^j \in l_j. \tag{1}$$

Where, $v_{x_i^j}$ is the feature vector of data sample x_i^j extracted by DNN and m_{core} is the number of core clusters identified by $DBSCAN$.

By calculating distance between a in-the-wild sample, s and the centroids, authors propose a framework to identify samples that are far from the training data distribution. Authors deem s to be confusing if,

$$|d_s^a - d_s^b| < \eta,$$

where a and b are the two nearest clusters from sample s. d_s^a and d_s^b are the distances between sample s and centroids c_a and c_b, respectively, and η is a tolerance parameter that is set empirically. And, the distance is calculated as follows.

$$d_s^j = euclid(v_s, c_j)$$

Through extensive experimentation, authors provide evidence that their proposed method outperform the existing selective prediction methods in most cases with benchmark datasets. As this framework is plug-and-play, it also requires minimal effort to incorporate with any off-the-shelf DNNs. However, this framework has only tested on benchmark datasets.

Experimental Analysis

In this section, we present and analyze the experimental study, which demonstrates the efficacy of each of the uncertainty estimation methods described previously. We conducted our experiments on the COVIDx dataset (Wang and Wong 2020), which is the largest publicly available COVID-19 dataset. COVIDx is comprised of a total of 13,917 CXR images (Normal:7966, Pneumonia: 5462, COVID:489) for training and 1578 CXR images (Normal:885, Pneumonia: 593, COVID:100) as test-set.

Experimental Setup

For comparative studies of existing uncertainty estimation methods, we choose VGGNet with 16 layers as our baseline DNN. We implemented SelectiveNet (Geifman and El-Yaniv 2019), TTAUG (Ayhan et al. 2020) and DbFF (Sarker et al. 2020) on the same baseline DNN to ensure fair comparison. Among these three methods, SelectiveNet requires most modification. TTAUG is comparatively simpler to deploy, however, it requires training with the same augmentations that to be used on the test data. On the other hand, DbFF does not require modification either to the DNN or to the loss function. We also incorporated the best performing method from the aforementioned experiment on COVID-Net (Wang and Wong 2020).
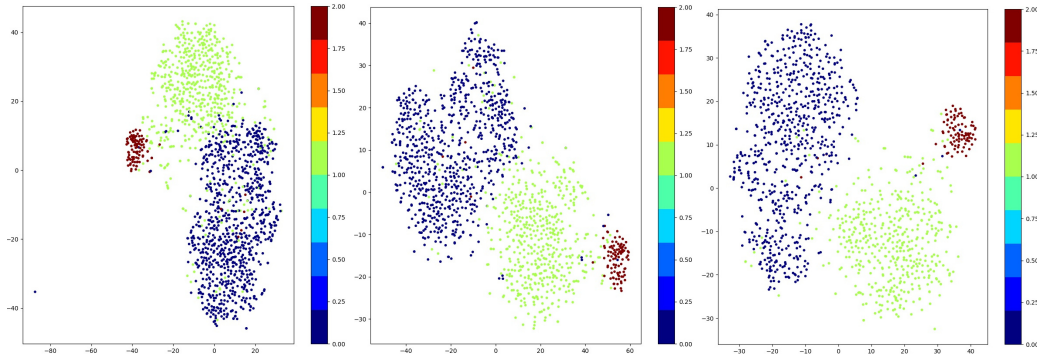
Figure 1: t-SNE visualization of COVIDx test-set in the feature space. Test-set features with 0%, 10% and 20% abstention rate (from left to right).

Abstention	Model		
Rate	TTAUG	SelectiveNet	DbFF
5%	93.07%	93.58%	**94.13%**
10%	94.51%	94.55%	**95.29%**
15%	95.75%	95.18%	**96.13%**
20%	**96.75%**	96.04%	96.70%
25%	97.21%	96.75%	**97.38%**
30%	97.73%	97.14%	**98.10%**

Table 1: Comparative results on COVIDx with varying abstention rates of TTAUG (Ayhan et al. 2020), SelectiveNet (Geifman and El-Yaniv 2019) and DbFF(Sarker et al. 2020). We highlight the best performances with bold-face. Even though TTAUG gains slightly better performance than DbFF with 20% abstention rate, DbFF shows consistency and it is easy to deploy in real-world use-cases.

Comparative Study of Existing Methods

First, we present the experimental studies on the effectiveness of existing methods on the COVIDx dataset (Wang and Wong 2020). We compare state-of-the-art methods, SelectiveNet (Geifman and El-Yaniv 2019), TTAUG (Ayhan et al. 2020) and DbFF (Sarker et al. 2020), and report the results in Table 1. Please note, for fair comparison, we report SelectiveNet results when trained their model with 100% coverage and then calibrated to the desired abstention rate. It can be observed from Table 1 that the DbFF outperforms or achieves similar performance compared with other two methods. Moreover, as mentioned before, SelectiveNet utilizes a specialized loss function which requires modification to the existing DNN, whereas Density-based framework can be utilized with any DNNs in a plug-and-play manner. TTAUG require domain knowledge to design effective augmentation, which hinders the deployment of the method. This establishes the superiority of Density-based Filtering Framework over the existing state-of-the-art.

Effect on COVID-Net

To further explore the effectiveness of DbFF method, we incorporated it with state-of-the-art COVID-Net (Wang and Wong 2020). Please note, though we choose COVID-Net as our base model here, DbFF can easily be extended to any other COVID detection NNs because of its plug-and-play nature. We present the results of this experiment in

Table 2. As can be observed, DbFF can effectively reduce the error rate of COVID-Net with the abstention of confusing samples. It can identify 49.4% of the mistaken samples as confusing by only abstaining 10% of the data for referral. DbFF method also improves PPV and sensitivity of COVID-Net with higher abstention rate. In order to demonstrate the effectiveness of DbFF method in detecting confusing samples, we visualize the feature spaces of the trained COVID-Net model on COVIDx (Wang and Wong 2020) test-set using T-distributed Stochastic Neighbor Embedding (t-SNE) (Maaten and Hinton 2008) in Fig. 1. From the visualization, it is visible how DbFF can identify the confusing samples lying on the boarder of class distributions and with higher abstention rate well defined distributions emerge. However, if closely observed, we can find few samples fall into wrong distributions. We argue that these samples may as well be the result of label noise in the dataset.

Expert Analysis

To understand the true efficacy of DbFF framework on COVIDx dataset, we collaborated with medical professionals, including an Epidemiologist closely working with the COVID-19 outbreak. Our goal here is to understand whether SOTA method could correctly identify the confusing samples, or it is abstaining randomly. To that end, we set up an experiment as follows.

- Randomly sample CXRs from each class on the COVIDx test-set that are predicted correctly by the DNN.

- Sample CXRs from abstained samples, while only 10% of data are abstained. The rationale is to sample the most confusing samples that are abstained for expert referral.

- Sample more CXRs from pool of wrong predictions, yet not abstained by DbFF while abstaining 25% data.

We shared these samples (15 samples per set) with medical professionals without disclosing the sampling criteria to prevent bias. Their analysis of each set is as follows.

- First set were straight forward to diagnose (Fig 2 a-c).

- Samples from the second set were confusing, and medical professionals recommended lateral view CXR or CT scan for further investigation. They mentioned that for some images, the CXR quality was poor (Fig. 2(d)) as a reason for confusion. For some samples, the CXR was not clear

Abstention Rate	Accuracy	Sensitivity			Positive Predictive Value		
		Normal	Pneumonia	COVID	Normal	Pneumonia	COVID
0%	94.82%	94.80%	94.90%	94.00%	96.30%	92.80%	94.00%
10%	97.16%	97.80%	96.60%	94.80%	97.30%	97.10%	95.70%
20%	98.81%	99.60%	98.30%	95.60%	98.60%	99.60%	96.60%
30%	99.18%	99.70%	99.00%	96.60%	99.00%	99.70%	97.70%

Table 2: Experimental results on COVIDx with varying abstention rate of DbFF (Sarker et al. 2020) framework with COVID-Net (Wang and Wong 2020) as the baseline DNN. Note that, the results presented here with 0% abstention rate represent the COVID-Net performance.

(a) COVID(COVID)   (b) Normal(Normal)   (c)   Pneumonia(Pneumonia)

(d) Normal(Pneumonia) (e) Normal(Pneumonia) (f) Pneumonia(COVID)   (g) Pneumonia(Normal)   (h) COVID(Normal)
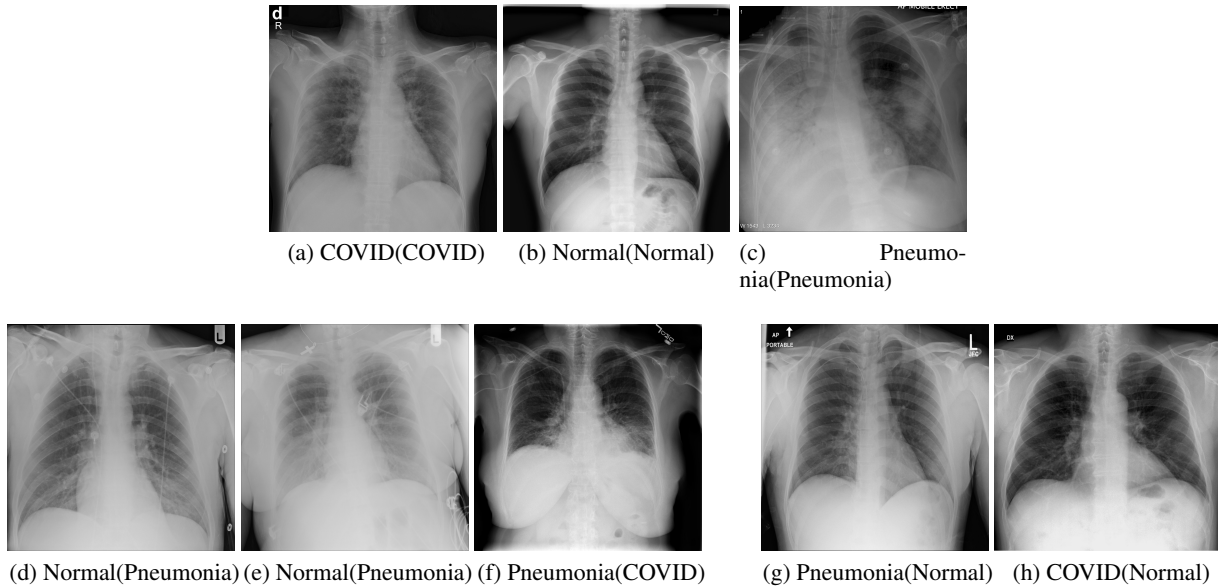
Figure 2: Samples from COVIDx dataset; The texts in and outside the parenthesis represent the predicted label and ground truth respectively. (a-c) samples that were correctly classified by model; (d-f) samples that were deemed as confused when 10% data were abstained; (g-h) samples that were not abstained yet mistaken by model. Sample (g) and (h) were predicted as normal by the model while the ground truths are Pneumonia and COVID positive, respectively.

due to the obesity of the patients (Fig. 2(e)). There were a few samples where DbFF made mistakes, but the experts could diagnose. They pointed out that these CXRs had breast shadows as the patients were female (Fig. 2(f)).

- The last set were mostly identifiable expect a few poor quality samples. However, the experts agreed with the DNN's prediction over the ground truth on two samples (Figs. 2(g-h)). We argue that, these samples may be contaminated by label noise.

Recommendations from the Experts

CXRs are not a very reliable indicator of diagnosis. However, a CT scan or RT PCR test may not be available in remote parts of the world, where CXR can be available. Hence, detecting critical patients via CXR analysis could save their lives. Our collaborating medical professionals suggested using a better quality CXR for training and detection. They also recommended associating metadata with CXR analysis, e.g., sex, BMI index, other clinical features, etc. for more reliable detection performance.

Conclusion

COVID-19 has been causing devastation in every aspects of our life. Detection and intervention are critical for patients who develop COVID-pneumonia. The research community has come together to create a reliable and accurate COVID-19 detection system with deep learning. On this consolidated effort, we intend to add our contribution. To the best of our knowledge, we are the first to address the predictive uncertainty issue of DNNs on COVID-19 detection. Through extensive experimentation, we demonstrated that uncertainty estimation framework, such as DbFF (Sarker et al. 2020), can effectively improve the reliability of existing CAD systems. In collaboration with medical professionals, we further analyzed the samples to gain valuable insights regarding such CAD systems. Lastly, we came across a number of potential areas that require further investigation: collecting high-quality CXR data, handling potential label noise, incorporating clinical metadata with CXR analysis, and handling data bias. We leave these areas open for future research.

References

Ayhan, M. S.; Kuehlewein, L.; Aliyeva, G.; Inhoffen, W.; Ziemssen, F.; and Berens, P. 2020. Expert-validated estimation of diagnostic uncertainty for deep neural networks in diabetic retinopathy detection. *Medical Image Analysis* 101724.

Bartlett, P. L.; and Wegkamp, M. H. 2008. Classification

with a reject option using a hinge loss. *Journal of Machine Learning Research* 9(Aug): 1823–1840.

Castiglioni, I.; Ippolito, D.; Interlenghi, M.; Monti, C. B.; Salvatore, C.; Schiaffino, S.; Polidori, A.; Gandola, D.; Messa, C.; and Sardanelli, F. 2020. Artificial intelligence applied on chest X-ray can aid in the diagnosis of COVID-19 infection: a first experience from Lombardy, Italy. *medRxiv*.

Chen, J.; Li, K.; Zhang, Z.; Li, K.; and Yu, P. S. 2020. A Survey on Applications of Artificial Intelligence in Fighting Against COVID-19. *arXiv preprint arXiv:2007.02202* .

Cleverley, J.; Piper, J.; and Jones, M. M. 2020. The role of chest radiography in confirming covid-19 pneumonia. *bmj* 370.

Cordella, L. P.; De Stefano, C.; Tortorella, F.; and Vento, M. 1995. A method for improving classification reliability of multilayer perceptrons. *IEEE Transactions on Neural Networks* 6(5): 1140–1147.

De Stefano, C.; Sansone, C.; and Vento, M. 2000. To reject or not to reject: that is the question-an answer in case of neural classifiers. *IEEE Transactions on Systems, Man, and Cybernetics* 30(1): 84–94.

Doi, K. 2004. Overview on research and development of computer-aided diagnostic schemes. In *Seminars in Ultrasound, CT and MRI*, volume 25, 404–410. Elsevier.

Erickson, B. J.; Korfiatis, P.; Akkus, Z.; and Kline, T. L. 2017. Machine learning for medical imaging. *Radiographics* 37(2): 505–515.

Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X.; et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, 226–231.

Geifman, Y.; and El-Yaniv, R. 2019. Selectivenet: A deep neural network with an integrated reject option. *arXiv preprint arXiv:1901.09192* .

Ghoshal, B.; and Tucker, A. 2020. Estimating uncertainty and interpretability in deep learning for coronavirus (COVID-19) detection. *arXiv preprint arXiv:2003.10769* .

Laves, M.-H.; Ihler, S.; and Ortmaier, T. 2019. Uncertainty Quantification in Computer-Aided Diagnosis: Make Your Model say" I don't know" for Ambiguous Cases. *arXiv preprint arXiv:1908.00792* .

Leibig, C.; Allken, V.; Ayhan, M. S.; Berens, P.; and Wahl, S. 2017. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific reports* 7(1): 1–14.

Liu, Z.; Wang, Z.; Liang, P. P.; Salakhutdinov, R. R.; Morency, L.-P.; and Ueda, M. 2019. Deep gamblers: Learning to abstain with portfolio theory. In *NIPS*, 10623–10633.

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9(Nov): 2579–2605.

Mallick, A.; Dwivedi, C.; Kailkhura, B.; Joshi, G.; and Han, T. 2020. Probabilistic neighbourhood component analysis: Sample efficient uncertainty estimation in deep learning. *arXiv preprint arXiv:2007.10800* .

Poduval, P.; Loya, H.; and Sethi, A. 2020. Functional Space Variational Inference for Uncertainty Estimation in Computer Aided Diagnosis. *arXiv preprint arXiv:2005.11797* .

Sarker, K.; Yang, X.; Li, Y.; Belkasim, S.; and Ji, S. 2020. A Unified Plug-and-Play Framework for Effective Data Denoising and Robust Abstention. *arXiv preprint arXiv:2009.12027* .

Sethy, P. K.; and Behera, S. K. 2020. Detection of coronavirus disease (covid-19) based on deep features. *Preprints* 2020030300: 2020.

Wang, L.; and Wong, A. 2020. COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-Ray Images. *arXiv preprint arXiv:2003.09871* .

Yanase, J.; and Triantaphyllou, E. 2019. The seven key challenges for the future of computer-aided diagnosis in medicine. *International journal of medical informatics* 129: 413–422.

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

Supplementary Material: Reducing Risk and Uncertainty of Deep Neural Networks on Diagnosing COVID-19 Infection

[1] **Krishanu Sarker,** [2] **Sharbani Pandit,** [3] **Anupam Sarker,** [1] **Saeid Belkasim,** [1] **Shihao Ji**

[1] Georgia State University
[2] Georgia Institute of Technology
[3] Institute of Epidemiology, Disease Control and Research

Refinement of Feature Vector

Lack of information or misinformation often cause confusion for DNNs. Hence, DNNs often make prediction on samples based on sub-optimal features extracted from noisy samples. To some extent, DNN classifiers (e.g. softmax or sigmoid) are robust to these feature noise due to the supervised feedback process. However, DbFF (Sarker et al. 2020) method heavily rely on the features learned by the pretrained DNN, while unlike other complex methods, it does not require retraining from scratch. Moreover, precise calculation of centroids is prerequisite to the success of the framework, as these centroids are utilized to determine the confusing samples. Hence, we explore ways to minimize the variance of the DNNs by filtering out the noisy features. To address this, we propose to utilize statistical analysis of the features to filter out noisy or constant features rather than using all of them. Specifically, we utilize chi-square test to obtain a scores on each features and based on an empirical threshold we filter-out the features with low statistical scores. χ^2 is defined as,

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Where, O is the observed value and E is the expected value of the distribution. Here, if two features are independent, the observed count would be close to the expected count, which provides evidence of dependency of these two features.

Effects of Feature Selection on Uncertainty Estimation

We experimented with the proposed feature selection method on the COVIDx dataset. The results are presented on Table 1. In this experiment we empirically set 1024 features as optimum number of features and a threshold was set accordingly. As we can see, the chi-square test based selection method achieves better performance over vanilla DbFF framework. The effectiveness of the feature selection method can be observed with the t-SNE visualization in Figs. 1(a-b). We can observe that the distributions are much more well defined for the later figure (Fig. 1(b)). We argue that static or noisy features often create issues with the centroid calculation utilized in DbFF method. After filtering these unwanted features, class constrained centroids become more robust to noise, hence improving performance.

References

Sarker, K.; Yang, X.; Li, Y.; Belkasim, S.; and Ji, S. 2020. A Unified Plug-and-Play Framework for Effective Data Denoising and Robust Abstention. *arXiv preprint arXiv:2009.12027* .

Wang, L.; and Wong, A. 2020. COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-Ray Images. *arXiv preprint arXiv:2003.09871* .

Abstention Rate	Positive Predictive Value W/O Feature Selection			Positive Predictive Value W Feature Selection		
	Normal	Pneumonia	COVID	Normal	Pneumonia	COVID
10%	97.30%	97.10%	95.70%	**97.40%**	**97.50%**	**96.20%**
20%	**98.60%**	**99.60%**	96.60%	**98.60%**	99.50%	**98.80%**

Table 1: Experiment demonstrating the effects of feature selection with DbFF (Sarker et al. 2020) method integrated with COVID-Net (Wang and Wong 2020) on COVIDx dataset.
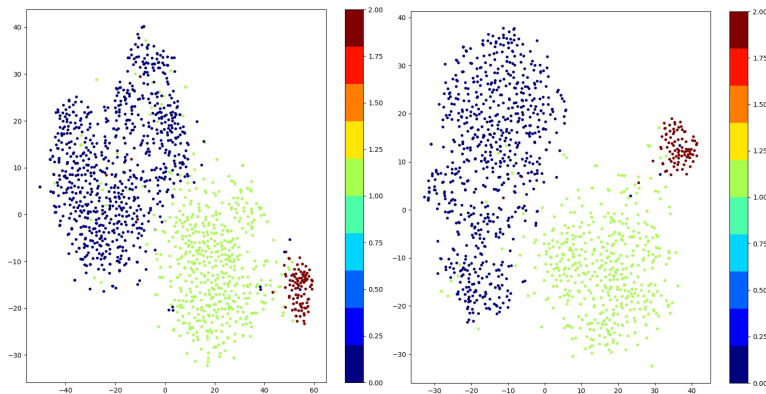
(a) Test-set features with 10% and 20% abstention rate (from left to right).
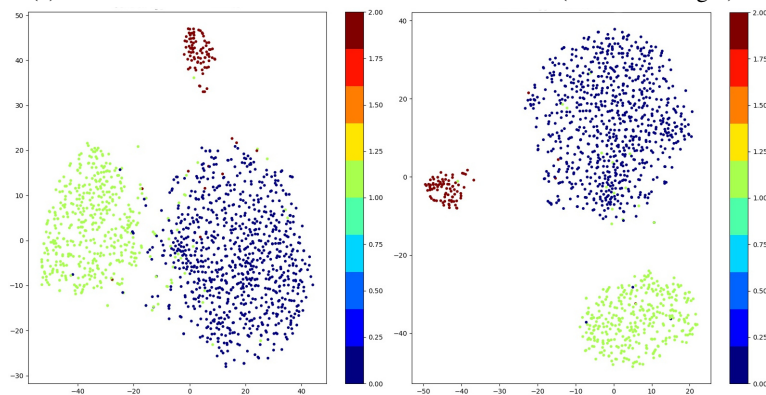
(b) Test-set features with 10% and 20% abstention rate after feature selection (from left to right).

Figure 1: t-SNE visualization of COVIDx test-set in the feature space.