No Surprises: Training Robust Lung Nodule Detection for Low-Dose CT Scans by Augmenting with Adversarial Attacks

Siqi Liu, Arnaud Arindra Adiyoso Setio, Florin C. Ghesu, Eli Gibson,
Sasa Grbic, Bogdan Georgescu, Dorin Comaniciu

***Abstract*—Detecting malignant pulmonary nodules at an early stage can allow medical interventions which increases the survival rate of lung cancer patients. Using computer vision techniques to detect nodules can improve the sensitivity and the speed of interpreting chest CT for lung cancer screening. Many studies have used CNNs to detect nodule candidates. Though such approaches have been shown to outperform the conventional image processing based methods regarding the detection accuracy, CNNs are also known to be limited to generalize on under-represented samples in the training set and prone to imperceptible noise perturbations. Such limitations can not be easily addressed by scaling up the dataset or the models. In this work, we propose to add adversarial synthetic nodules and adversarial attack samples to the training data to improve the generalization and the robustness of the lung nodule detection systems. In order to generate hard examples of nodules from a differentiable nodule synthesizer, we use projected gradient descent (PGD) to search the latent code within a bounded neighbourhood that would generate nodules to decrease the detector response. To make the network more robust to unanticipated noise perturbations, we use PGD to search for noise patterns that can trigger the network to give over-confident mistakes. By evaluating on two different benchmark datasets containing consensus annotations from three radiologists, we show that the proposed techniques can improve the detection performance on real CT data. To understand the limitations of both the conventional networks and the proposed augmented networks, we also perform stress-tests on the false positive reduction networks by feeding different types of artificially produced patches. We show that the augmented networks are more robust to both under-represented nodules as well as resistant to noise perturbations.**

I. Introduction

LUNG cancer is the leading cause of all cancer deaths [31]. Detecting malignant pulmonary nodules at an early stage can allow medical interventions which increases the survival rate of lung cancer patients. Early-stage cancer generally manifests in the form of pulmonary nodules which are defined as rounded opacity, well or poorly defined, measuring up to 30mm in diameter [11]. Based on the findings of the National Lung Screening Trial (NLST), the U.S. Centers for Medicare and Medicaid Services (CMS) approved screening for lung cancer of high-risk subjects to be fully reimbursed by insurance companies. The NELSON trial also reported reduced 10 year lung-cancer mortality with CT screening with a randomized trial involving 15789 patients [2]. However, given the sizeable eligible screening population (8.6 million in the US) and the time cost of interpreting 3D chest CT, it substantially increases the efforts for radiologists.

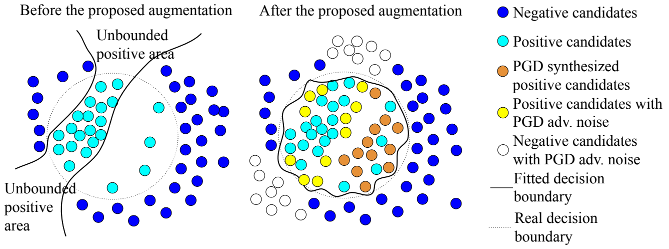

Before the proposed augmentation
Unbounded positive area
Unbounded positive area
After the proposed augmentation
● Negative candidates
● Positive candidates
● PGD synthesized positive candidates
● Positive candidates with PGD adv. noise
○ Negative candidates with PGD adv. noise
Fitted decision boundary
Real decision boundary

Fig. 1: A conceptual illustration of the motivation of the proposed training scheme. Pulmonary nodules in chest CTs follow a long-tail distribution typically with rare and hard nodules under-represented. ReLU networks tend to form open decision boundaries which leave the risk for the network to be activated by arbitrary noise [13]. In this work, we propose adversarial augmentation methods to efficiently search for both hard synthetic nodules and adversarial samples that can improve the robustness of the network.

Motivated by the LUNA16 challenge [30], many studies have attempted to automate the detection of pulmonary nodules using machine learning, in particular deep convolutional neural networks (CNN) in order to assist the radiologists in the lung screening workflow [29], [26], [4], [43], [35], [3]. Following the coarse-to-fine strategy, the majority of the deep learning-based nodule detection methods are implemented as a two-stage system: (1) a candidate generation network with a large field of view is first trained to output initial detection results with a high sensitivity at the cost of low specificity; (2) a false positive reduction (FPR) network is then trained to re-evaluate the confidence of each candidate.

Though many show CNNs can improve both the sensitivity and the specificity comparing to the previous image processing based CAD systems, CNNs can suffer from a few challenges, which we argue cannot be addressed by simply adding more training data or hyper-parameter tuning. First, the observer variability among radiologists is known to be high. For example, only 928 out of 2669 suspected findings

Siqi Liu, Florin C. Ghesu, Eli Gibson, Sasa Grbic, Bogdan Georgescu and Dorin Comaniciu are with Digital Technology & Innovation, Siemens Healthineers, Princeton, NJ, USA. (siqi.liu@siemens-healthineers.com)

Arnaud Arindra Adiyoso Setio is with Digital Technology & Innovation, Siemens Healthineers, Erlangen, Germany.

from the LIDC-IDRI study are agreed as nodules ($\geq$ 3mm) by all the four radiologists [1]. Such variability can be caused by factors such as the vague definition of pulmonary nodules, the imbalanced level of expertise among radiologists or the insufficient information provided by chest CT, etc. Second, the detection networks tend to miss nodules that are under-represented in the training set, such as the small ground-glass nodules, irregular shaped nodules or nodules appearing in under-represented contexts. Because only 3.6% of the screening population have biopsy-proven malignant nodules [33], such malignant nodules can also be under-represented in the training data. Third, neural networks are known to be prone to unexpected image distortions [12]. Such distortions can happen in the real-world low-dose CT imaging though they are rare in both the training and the benchmark datasets. As we show later in this paper, even simple noise patterns can determine an under-augmented nodule detector to giving positive responses. Under- or over-detecting nodules caused by such unanticipated distortions can pose the potential risk of distracting and biasing the radiologists. Therefore, besides achieving overall high sensitivity and a low number of false positives on clean benchmark datasets, a nodule detection system is also expected to (1) be capable of detecting under-represented nodules that are rare in both the training and benchmark datasets (2) be robust to unanticipated noise and distortions in the real-world images.

Motivated by the reasons above, we propose to augment the training set of lung nodule detection by adversarially attacking a pre-trained false positive reduction network with both hard synthetic nodules as well as noise image perturbations. The concept is illustrated in Fig. 1. First, we propose to use projected gradient descent (PGD) [19] to search for the adversarial samples that can determine a trained false positive reduction network into outputting over-confident wrong predictions. These searched patches are then added to the training patches to augment the detector to be more robust to both under-represented nodules and unanticipated image distortions. PGD is used for searching for three types of adversarial augmentation patches: (1) latent codes to sample hard synthetic nodules that the detector fails to detect; (2) perturbation noise that can make the nodule detector fail to detect; (3) noise patterns that can easily determine the nodule detector to giving false-positive findings; To evaluate the proposed methods, we train a baseline nodule detector following the general 2-stage framework using a large-scale training dataset. The adversarial patches are then generated by attacking the baseline false positive reduction (FPR) network and are used for augmenting the FPR network. By evaluating on two different benchmark datasets, we show the proposed techniques can improve the detection performance on clean benchmark data. Using the same techniques, we also generate adversarial samples to stress-test the trained false positive reduction networks. We show that the augmented networks are more robust to both hard nodules and noise perturbations.

II. Related Work

1) Deep learning based nodule detection: As one of the most popular applications of computer-aided diagnosis systems, many studies have been dedicated to use image processing and machine learning algorithms to detect lung nodules [41]. The majority of the nodule detection framework generate candidates first either with an image processing pipeline or a fully convolutional neural network. Then a separate classifier is trained to reduce false positives based on the input 3D CT patches centered at the candidate locations. Most of the recent works were developed based on the LUNA challenge [30] which acquired its data from the LIDC-IDRI dataset [1]. Though the annotation process of the LIDC-IDRI dataset has been well documented and is considered reliable, the quantity and diversity of the LIDC-IDRI dataset are highly limited. Besides the LUNA challenge, there have been no benchmarks reported with known statistics. Though the metrics computed from the FROC curves are suitable for reporting the detection performance on a given benchmark dataset, it is not often thoroughly investigated that how robust such detection systems would perform on the rare cases as well as noise perturbations. Our work shows that the conventional CNNs trained without adversarial augmentation would generally fail to recognize rare nodules as well as prone to image noise. For a more comprehensive review of the deep learning based lung nodule detection systems, we would refer our readers to [25], [41].

2) Data synthesis based augmentation in medical image analysis: Inspired by the recent advances in generative models, there have been increasing interests in synthesizing objects in medical images in order to augment the existing training set for better diversity [40]. Many recent studies proposed to use generative networks to synthesize lung nodules in order to improve the performance of diverse lung nodule related applications [39], [17], [14], [36], [37], [7], [10], [34]. Most learning based nodule synthesis methods start with training a generative network to map low dimensional latent codes to realistic lung nodules in chest CT using either variational auto-encoder (VAE) or Generative adversarial networks (GAN). Latent codes are sampled from a predefined prior distribution randomly to synthesize nodules resembling the real ones. These synthetic nodules are blended into the original image contexts by either formulating the training task as either image impainting [14] or using an extra context-blending network [17]. In [17], authors use both the discriminator error and the classification error to select only the hard synthetic cases to be added to the augmented dataset. We show that such sampling strategies can be inefficient. The majority of the synthetic samples would add little values since they can be successfully recognized by a network that is trained on a large-scale dataset. However, hard samples can be drawn from a synthesizer without exhaustive search if the latent codes are optimized to increase the training loss of a trained network.

3) Over-confident neural networks and adversarial training: To build robust computer-aided diagnosis systems that are robust to out of distribution (OOD) samples, one can train the network to estimate the decision uncertainty and reject the samples when the estimated uncertainty is high [8], [28]. Though we also use the beta distribution in our work for uncertainty estimation [8], we show that the uncertainty estimation techniques alone would be insufficient to make the network robust to avoid over-confident decisions on OOD samples. In [19], it is argued that ReLU activated neural networks would always have open decision boundaries which leave the risk of high responses for unseen OOD samples. In another paper, it is argued that batch normalization is also a cause of the adversarial vulnerability [6]. Such network vulnerability are hard to be reflected by the clean medical image benchmark datasets. However, this poses potential risks for deploying the computer-aided diagnosis systems in real clinics as investigated by some recent studies [23], [5], [18], [16], [24], [38]. In [20], [13], it is proposed to use PGD [19] to search for the adversarial augmentation cases from uniform noise or permuted input patches to augment the clean training dataset. We use similar techniques to adversarially sample both hard positive and hard negative nodule samples to enhance the adversarial robustness of the nodule detection networks. Though it was suggested that the adversarially trained networks can generalize slightly worse on clean data [32], we believe such robustness is still vital for real-world medical AI applications.

III. Methods

A. Baseline Detection Architectures

Similar to many new deep learning based nodule detection framework, our baseline framework consists of a candidate generation (CG) module and a false positive reduction (FPR) module as shown in Fig. 3. The candidate generation module is trained to achieve high sensitivity via over-detecting nodule candidates. We use three identical 3D ResUNets [42] as the CG backbone networks without weight sharing. The first CG network is firstly trained to output 3D heatmaps with the nodule centers represented by 3D Gaussian blobs with the same sizes (3D Blob All Nodules). We then fine-tune the first CG with only the ground glass candidates and part-solid candidates since they are under-represented in the training set (3D Blob Ground Glass Nodules). The candidates are derived with non-maximum suppression (NMS) on the fusion heatmap obtained by taking the element-wise maximal of the two network output heatmaps. We found that the blob output CG networks tend to have high sensitivity on small nodules while missing the relatively larger nodules. We thus also finetune the first CG network by adding a 3D region proposal network (RPN) head [27] to outputting 3D bounding boxes (3D RPN Head). We found the 3D RPN network tends to have higher sensitivity on larger nodules. The final candidates are obtained by taking the union of the blob candidates and the bounding box candidates.

The false positive reduction module is then trained to re-evaluate the candidates and prune the false positive findings based on the classification confidence. It is built with a DenseUNet network pre-trained with nodule segmentation. We add shallow classifier layers on top of it to derive the FPR confidence scores. The network is trained using 64^3 patches with a resolution of 0.625^3mm. We train all the CG and FPR networks using the Adam optimizer [15] with the initial learning rate 0.001.

We trained the CG framework first and froze it before performing the analysis presented in this work. For the brevity of this paper, we demonstrate the proposed techniques only to improve the FPR while assuming the CG networks are trained and frozen. However, the same techniques can also be used for improving CG networks.

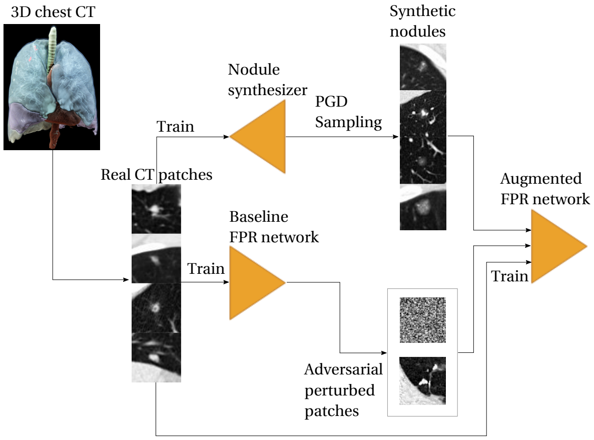


Fig. 2: The data-flow illustration of the proposed adversarial augmentation framework for enhancing the false positive reduction (FPR) network in a nodule detection pipeline.

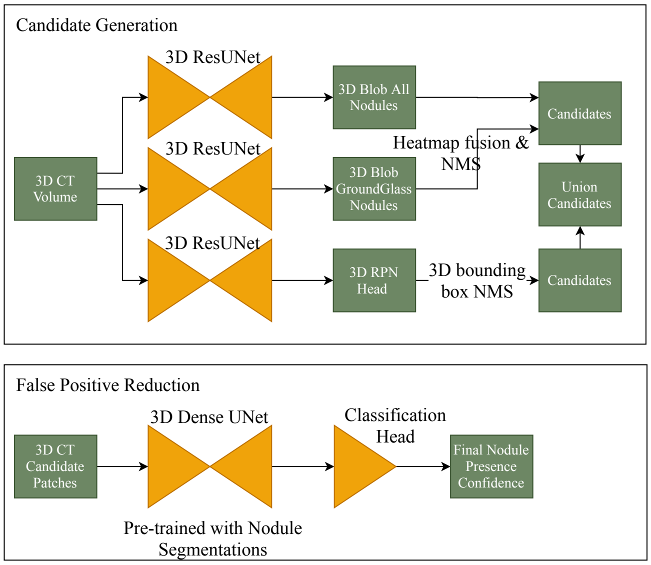


Fig. 3: The baseline two stage nodule detection framework used in this work.

B. Hard-Sample Synthesis with PGD Sampling

We train a nodule synthesizer $f_{generator}$ that can be controlled by the latent code sampled from a prior distribution. We implement the $f_{generator}$ with a 3D convolutional variational encoder. We extract the nodules out of the CT context with the manually annotated nodule segmentation. The boundary of the nodule segmentation is blurred with a distance transform. As shown in Fig. 4, we firstly map the cropped 3D nodules to an encoding space using the encoder network $f_{encoder}$, then the variational encoding is reconstructed back to the nodules in chest CT. We jointly train a WGAN-GP discriminator [9]

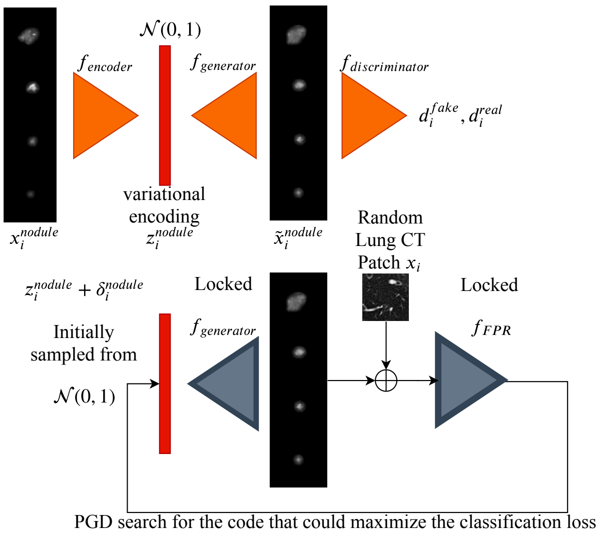


Fig. 4: The illustration of the nodule synthesis framework.

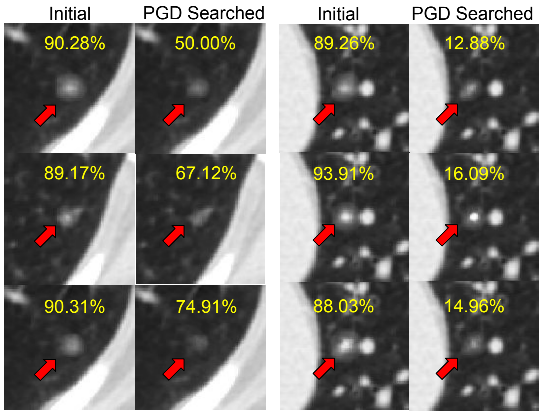


Fig. 5: The demonstrations of the synthetic nodules before and after PGD searching. With slight perturbation in the nodule appearance, the nodule detector trained with conventional strategy would output significantly lower confidence score.

with spectral normalization [21] to enforce the generator to add high frequency details to mimic the real nodules in CT. The data flow can be summarized as

$$\mu_i, \sigma_i = f_{encoder}(x_i^{nodule}) \tag{1}$$

$$z_i^{nodule} \sim \mathcal{N}(\mu_i, \sigma_i^2) \tag{2}$$

$$\tilde{x}_i^{nodule} = f_{generator}(z_i^{nodule}) \tag{3}$$

$$d_i^{fake}, d_i^{real} = f_{discriminator}(\tilde{x}_i^{nodule}, x_i^{nodule}) \tag{4}$$

Here d_i^{fake} and d_i^{real} the discriminator output for the fake and real samples. The training objective of the nodule synthesizer can be summarized as

$$L_{discriminator} = L_{WGAN-GP}(d_i^{fake}, d_i^{real}) \tag{5}$$

$$\begin{aligned} L_{encoder} + L_{generator} = |\tilde{x}_i^{nodule} - x_i^{nodule}| + \\ \lambda_1 D_{KL}(\mathcal{N}(\mu_i, \sigma_i^2)||\mathcal{N}(0,1)) - \\ \lambda_2 L_{WGAN-GP}(d_i^{fake}, d_i^{real}) \end{aligned} \tag{6}$$

where $D_{KL}(\mathcal{N}(\mu_i, \sigma_i^2)||\mathcal{N}(0,1))$ optimizes the probability distribution parameters μ and θ to closely resemble that of $\mathcal{N}(0,1)$. $\lambda_2 L_{WGAN-GP}(d_i^{fake}, d_i^{real})$ is the wasserstein GAN discriminator loss regularized by the gradient penalty defined in [9].

Once the synthesizer is trained, we discard both the encoder network and the discriminator. Only the generator network is kept for sampling synthetic nodules. Random nodules can be sampled by feeding a code to the trained generator $f_{generator}(z_i^{nodule} \sim \mathcal{N}(\mu_i, \sigma_i^2))$. The synthesized nodule can be fused to a random background chest CT patch x_i and then fed to a trained FPR classifier f_{FPR}. Though it is feasible to add another training stage as described in [17] to further blend the generated nodule into its context, we found it non-critical for the sake of improving the nodule detection in practice.

It is inefficient to draw hard-cases directly by randomly sampling from the prior because most of the cases close to the mean have already been learned by the nodule false positive reduction network f_{FPR}. So instead of randomly sampling the encoding of nodules, we use the projected gradient descent (PGD) as originally used for generating adversarial attacks [19] to sample hard nodules. For each sampling, we initialize the encoding from the standard normal distribution $z_i^{nodule} \sim \mathcal{N}(0,1)$ and randomly initialize a perturbation vector δ_i^{nodule} to explore the neighbourhood $\mathcal{S}$ of z_i^{nodule} within a bounded radius. δ_i^{nodule} is updated by PGD to maximize the L_{FPR} as

$$e_i = f_{FPR}(f_{generator}(z_i^{nodule} + \delta_i^{nodule}) \oplus x_i) \tag{7}$$

$$arg \max_{||\delta|| \leq \epsilon} L_{FPR}(e_i, 1) \tag{8}$$

Here, $\oplus$ is the fusion operator that blends the synthetic nodule into the CT context patch x_i. We define $\oplus$ simply as masked image summation. We use the beta distribution as in [8] to measure the classification uncertainty instead of using the sigmoid activation and the binary cross entropy. The FPR network f_{FPR} outputs the classification evidences e_i for positive and negative labels. $L_{FPR}(e_i, 1)$ is the classification loss defined with the beta distribution distance [8]. The perturbation vector δ can be updated as

$$\delta := \mathcal{P}(\delta + \alpha \nabla_\delta L_{FPR}(.)) \tag{9}$$

where $\mathcal{P}$ denotes the projection onto the ball of interest defined by ϵ; α is the step size. In Fig. 5, we show initial synthetic nodules together with the synthetic nodules searched with PGD. Though visually similar, the tiny differences in the nodule appearance can result in large difference in the F_{FPR} responses.

C. Over-confident Perturbation with PGD Sampling

Besides searching the latent codes for the nodule synthesizer, PGD can also be used for perturbing the real patches x_i as

$$e_i = f_{FPR}(x_i + \delta_i^{patch}) \tag{10}$$

$$arg \max_{||\delta_i^{patch}|| \leq \epsilon} L_{FPR}(e_i, g_i) \tag{11}$$

where g_i is the groundtruth label for patch x_i. As shown in the first row of Fig. 6, we found for most of the positive nodules

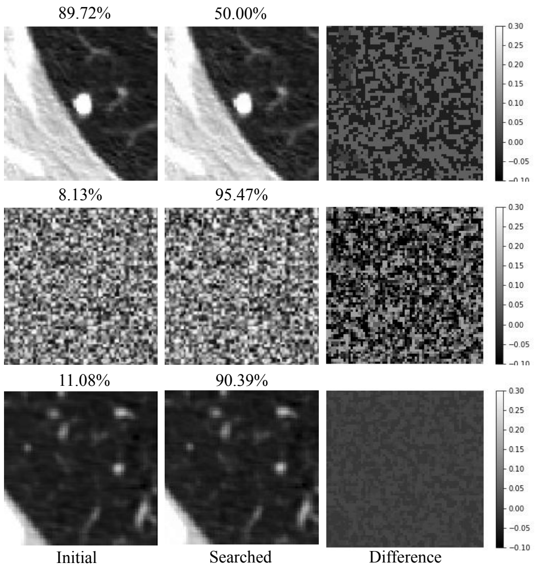


Fig. 6: The upper row demonstrates the noise perturbation on nodule patches. Arbitrary noise can determine a trained nodule detector to ignore a well-defined nodule. The middle and bottom row demonstrates that specific noise patterns can activate a trained nodule detector to output high confidence scores from either pure adversarial noise or the negative CT patches distorted by adversarial noise. The difference patch are shown with the window $[0, 0.3]$ to make the perturbation visible while the image patches are shown with the window $[0, 1]$.

patches, it is easy to find a δ_i^{patch} with a small magnitude to perturb x_i so that f_{FPR} no longer recognizes the nodule resides in it. Such perturbations can disturb the model from recognizing the nodules when the images contain unexpected abnormalities, strong imaging artefacts or malicious noise injections.

We also found that even for noise patches $x_i^{uniform}$ drawn from a uniform distribution, PGD can search for a neighbouring patch and excites the FPR network to output a positive decision, though the searched patch does not contain any interpretable patterns as shown in the second row of Fig. 6. The intersection between the chest CT distribution and the uniform distribution is expected to have close to zero probability mass. As explained in [13], ReLU networks decompose the observation space into a finite set of polytopes in which outer polytopes extend to infinity. Adding the adversarial patches searched by Eq.(11) to augment the FPR network can make it robust to such image perturbations by closing the decision boundary.

In practice, we train a baseline FPR network first by randomly sampling real positive and negative candidate patches with 50% chance each until reaching convergence. Then we finetune the baseline model by also sampling from the augmentation patches generated by attacking the baseline model. For positive sampling, we draw 50% from the real positive patches, and 25% from synthetic nodules and 25% the adversarial positive patches. For negative sampling, we draw 50% from both real negative patches and 50% from the adversarial negative patches.

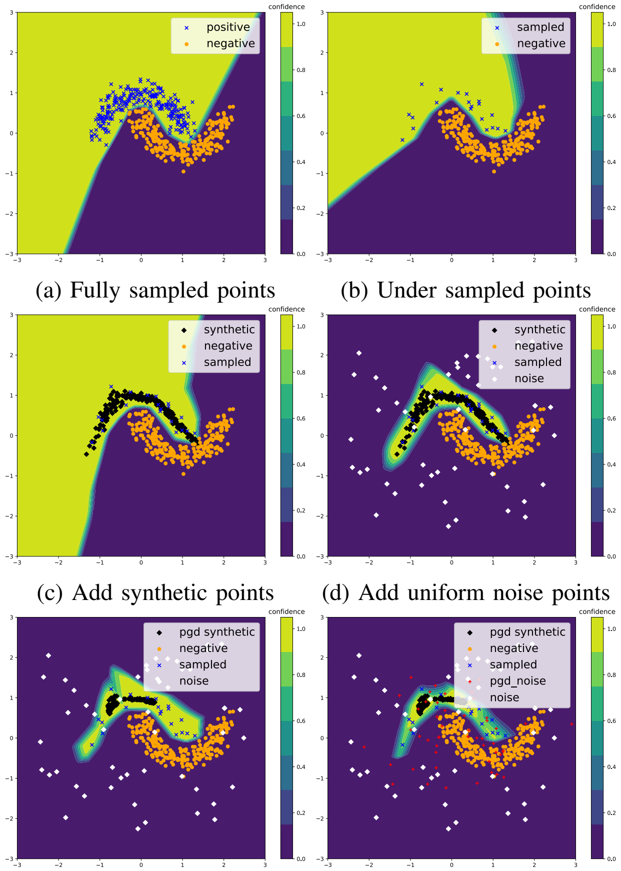
IV. Data

6488 3D chest CT scans were collected for training. The training images were collected from multiple sources, including the LUNA challenge [30], the NLST cohort [33] and an in-house data collection. Each training image contains at least one radiologist confirmed nodule. We annotated the nodule locations and diameters in the training images from our in-house dataset and the NLST subset. Our annotators firstly detected all the potential nodule candidates. Then two radiologists went through all the candidates to confirm the presence of a nodule. 10% of the training images were randomly sampled as the validation set for parameter searching and early stopping. To evaluate the performance, we constructed two benchmark datasets, as summarized in Table I. The In-house Benchmark was built based on a private data collection with 174 challenging images. Besides lung nodules, many patients in the In-house Benchmark also had other types of pulmonary abnormalities which constitute a significant source of false positives for both human and the networks. The NLST Benchmark consists of randomly sampled 272 baseline scans from the NLST cohort. The patients were sampled following the real-world screening distribution [33] (1% with cancer, 25.8% with cancer negative nodules and 73.2% healthy) while ensuring (1) the slice thicknesses are lower than $1.5mm$ (2) there is no gap in the DICOM series (3) each image contains the entire lung. We had three on-board radiologists read the images in both benchmark datasets independently. In the first round, each radiologist marked the nodule candidates individually. All the candidate nodules spotted in the first round were merged and presented to each radiologist to confirm in case there were under-attended nodule candidates. We took the nodules that are the consensus among all three radiologists as the positive locations while the rest as irrelevant findings which were not involved in the metrics computing. We only considered the nodules with the diameters larger than $6mm$ for benchmarking. However, we do not claim this is a critical choice since the size threshold can be adjusted according to the different application scenarios.

All the augmentation patches, including the synthetic nodules, perturbed positive nodule patches and the perturbation noises, were pre-computed and randomly sampled during the FPR model training by attacking the baseline network (baseline-beta-finetune). We generated synthetic nodule patches on 10 random background patches from each training image. The locations of the background patches were constrained within the lungs using the lung segmentation masks predicted by a previously trained network. We also ensured that the background patches do not contain a real nodule inside. For each background patch, we sampled the synthesizer six times with random sampling and the PGD sampling, respectively. It resulted in 389,280 synthetic nodule patches for both sampling strategies. We generated one adversarially perturbed patch for each positive nodule candidate in our training data (22,169 relevant nodules) similarly to the upper row of Fig. 6. We also generated 100,000 pure adversarial

TABLE I: The summary of the two chest CT benchmark datasets.

	In-house Benchmark	NLST Benchmark
Images	174	272
Images w/ Nodules	97	83
Solid Nodules	94	103
Fully Calcified Nodules	7	19
Part-Solid Nodules	13	3
Ground Glass Nodules	36	6
Total Nodules (>=6mm)	150	131



(a) Fully sampled points
(b) Under sampled points
(c) Add synthetic points
(d) Add uniform noise points
(e) Add PGD synthetic points
(f) Add PGD noise points

Fig. 7: A toy experiment to depict the concept of the proposed augmentation methods.

noise patches similarly to the lower row of Fig. 6. To stress-test the robustness of network at random pulmonary locations, we sampled 10 random patches centered in the lungs as the negative stress-test samples from each benchmark CT volume, while avoiding annotated nodules. We add adversarial noise to these negative samples by attacking the baseline network (baseline-beta-finetune).

V. Results

A. Toy Example

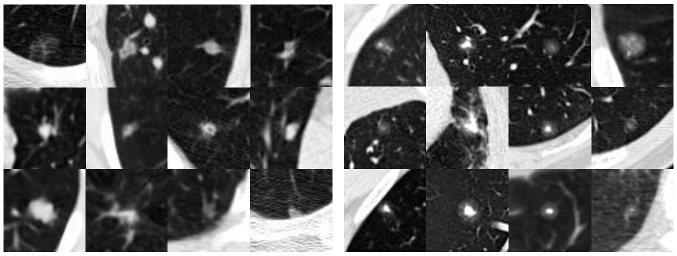
In Fig. 7, we firstly show a toy experiment built with the simple two-moon dataset to demonstrate the presented concept. 500 spots are sampled from both the positive and the negative cluster by adding the Gaussian noise with the standard deviation of 0.15. In our context, they represent the positive and negative candidates used for training the FPR classifier. We train a ReLU activated multi-layer perceptron to mimic the FPR classifier based on the sampled spots to plot the decision boundary. We then sub-sample only 20 positive candidates following a long tail distribution to simulate the real-world training set distribution as Fig. 7b. We trained a small VAE on the 20 positive spots and generated synthetic samples by drawing the latent code from a standard normal distribution. The added synthetic spots help filling the hole in the decision boundary as in Fig. 7c. However, a sizeable out-of-distribution area is also predicted as confident positive as anticipated in [13]. We then sampled another 20 spots that are randomly drawn from a uniform distribution and added them to the negative cluster. In Fig. 7d, it is shown that such noise samples can bound the decision boundary tightly to the positive cluster. Though there is a small chance that the noise spots can also reside in the positive cluster, such cases are extremely rare in the real world 3D inputs. Though we use uniform sampling in this toy example, it is notable that in a high-dimensional input space, the random sampling can be highly in-efficient for both synthesizing real nodules and generating adversarial noise samples. We use PGD to search for the latent code from the trained VAE. As in Fig. 7e, the PGD searched synthetic spots only reside in the under-sampled region. In addition to the uniform spots, we show the PGD searched negative spots which are closer to the positive cluster in Fig. 7f. Such supporting negative spots can be more efficient for refining the decision boundary when the input dimension is higher as in 3D chest CT patches.

B. Benchmark on clean data

Before we analyze the FPR networks, the frozen CG framework achieved 100% sensitivity on the In-house Benchmark and 97.71% sensitivity on the NLST Benchmark when having 100 average candidates per scan. We summarize the FROC curves for benchmarking the nodule detection FPR models trained with different strategies in Table II. The classification head with beta distribution (baseline-beta) produced similar CPM scores as the sigmoid head trained with binary cross entropy (baseline-ce). However, we also show that the classifier would generate slightly higher CPM scores if the network is firstly trained with cross-entropy and then finetuned with the beta-distribution loss (baseline-beta-finetune). In the experiments beta-syn (random) and beta+syn, we respectively added synthetic nodules randomly sampled from the standard normal, and the ones searched using the proposed PGD sampling. Though both types of synthetic nodules can improve the overall network generalization, the nodules searched with PGD consistently outperforms its counterpart, especially at the region of the lower number of false positives. We show that adding the noise perturbation augmentation patches (beta+perturb and beta+perturb+syn) can also slightly improve the overall CPM scores comparing to the conventional training baseline (baseline-beta-finetune). However, they do not show better performance than only using only PGD searched nodules (beta+syn). We show such perturbation augmented

TABLE II: The table summarizes the FROC metrics obtained from the compared training strategies. The CPM [22] score averages the sensitivities sampled at 7 log-scale operating points indicating differnt numbers of false positives (0.125, 0.25, 0.5, 1, 2, 4, 8).

In-house Benchmark											
	PERTURB	SYN	LOSS	**CPM**	FP=0.125	FP=0.25	FP=0.5	FP=1	FP=2	FP=4	FP=8
baseline-ce	✗	✗	CE	88.46%	73.09%	73.09%	89.84%	92.02%	92.28%	94.65%	96.64%
baseline-beta	✗	✗	BETA	89.11%	75.43%	75.43%	88.78%	90.80%	91.77%	94.26%	96.64%
baseline-beta-finetune	✗	✗	BETA	88.90%	75.58%	75.58%	90.79%	91.37%	92.11%	94.24%	96.70%
beta+syn (random)	✗	✓	BETA	90.76%	79.89%	79.89%	92.05%	93.34%	93.35%	94.26%	96.67%
beta+syn	✗	✓	BETA	**91.22%**	**81.09%**	**81.09%**	**92.66%**	93.33%	93.33%	94.17%	96.99%
beta+perturb	✓	✗	BETA	90.07%	76.35%	76.35%	89.91%	**93.42%**	93.61%	**95.40%**	**97.92%**
beta+perturb+syn	✓	✓	BETA	90.47%	77.52%	77.52%	89.90%	92.75%	**93.97%**	94.97%	97.45%
NLST Benchmark											
	PERTURB	SYN	LOSS	**CPM**	FP=0.125	FP=0.25	FP=0.5	FP=1	FP=2	FP=4	FP=8
baseline-ce	✗	✗	CE	82.56%	52.18%	52.18%	84.99%	89.68%	91.68%	93.14%	95.63%
baseline-beta	✗	✗	BETA	80.62%	44.74%	44.74%	83.35%	88.56%	91.38%	93.18%	94.40%
baseline-beta-finetune	✗	✗	BETA	83.60%	53.69%	53.69%	85.30%	91.35%	93.01%	**93.99%**	**95.71%**
beta+syn (random)	✗	✓	BETA	85.81%	66.04%	66.04%	86.55%	90.17%	92.05%	93.15%	95.06%
beta+syn	✗	✓	BETA	**87.89%**	**74.44%**	**74.44%**	87.61%	**90.55%**	**93.30%**	93.80%	94.77%
beta+perturb	✓	✗	BETA	85.38%	58.75%	58.75%	**88.21%**	89.58%	92.43%	93.78%	94.60%
beta+perturb+syn	✓	✓	BETA	85.51%	62.30%	62.30%	85.71%	89.44%	92.09%	93.78%	94.61%



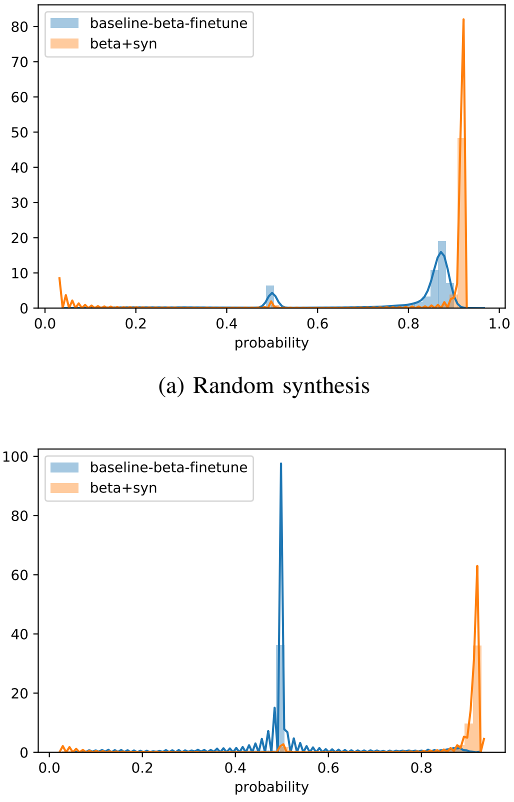
(a) Real nodules (b) PGD synthetic nodules

Fig. 8: The mosaic view to compare the real nodule patches and the synthetic nodules in patches of size 64^3 and 0.625^3mm resolution. Besides being generally smaller, the PGD searched synthetic nodules tend to have round glass component with or without a solid core. Such non-solid or part-solid nodules are relatively rare in the real datasets.

networks are more robust to both uniform and adversarial noise in the next section.

C. Stress test

1) Synthetic nodules: The central slices of the randomly selected real nodules and the hard nodules sampled by PGD are shown in Fig. 8. Besides being generally smaller, the PGD searched synthetic nodules tend to have round glass component with or without a solid core. Such non-solid or part-solid nodules are relatively rare in the real datasets. Though one can still visually distinguish a subset of the synthetic nodules from the real nodules, they can be a valuable source to stress-test the FPR network as most of such cases reside at the original decision boundaries. We synthesized 10000 nodules with both random Gaussian sampling and PGD searching respectively. They were fed to the FPR networks trained with (beta+syn) and without (baseline-beta-finetune) synthetic nodules. We ensured that all the synthetic nodules



(a) Random synthesis

(b) PGD searched synthesis

Fig. 9: Confidence histogram obtained by using synthetic nodule to stress-test the nodule detector (false positive reduction). Without augmentation, the conventional detector head tends to output most of the PGD searched nodules as unknown (0.5) while the augmented detector can detect the majority of them with high confidence.

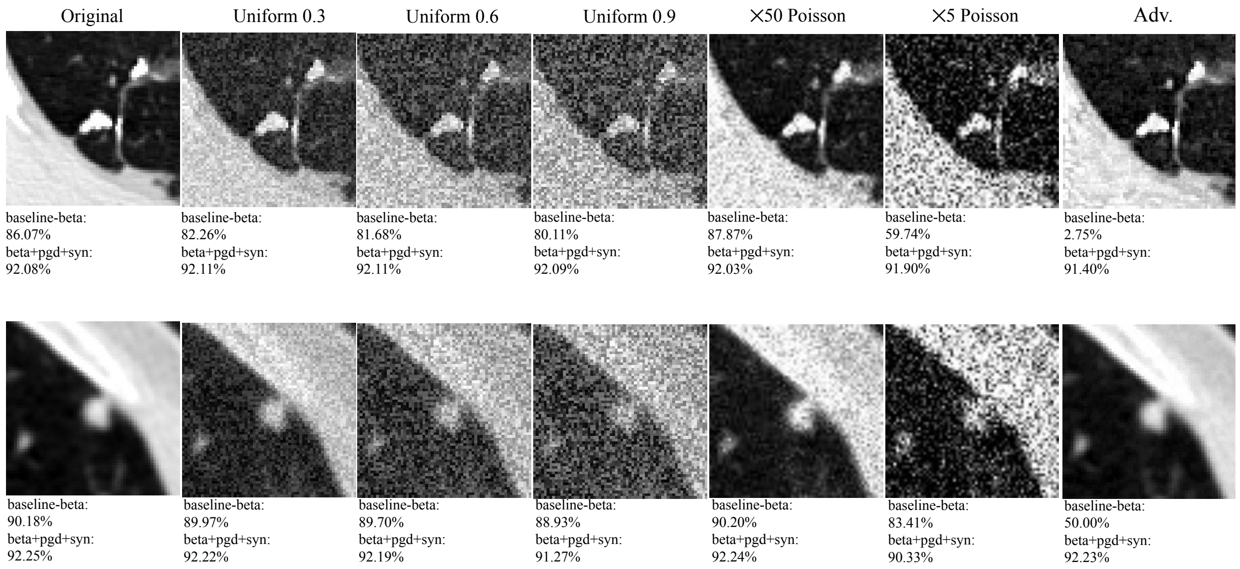
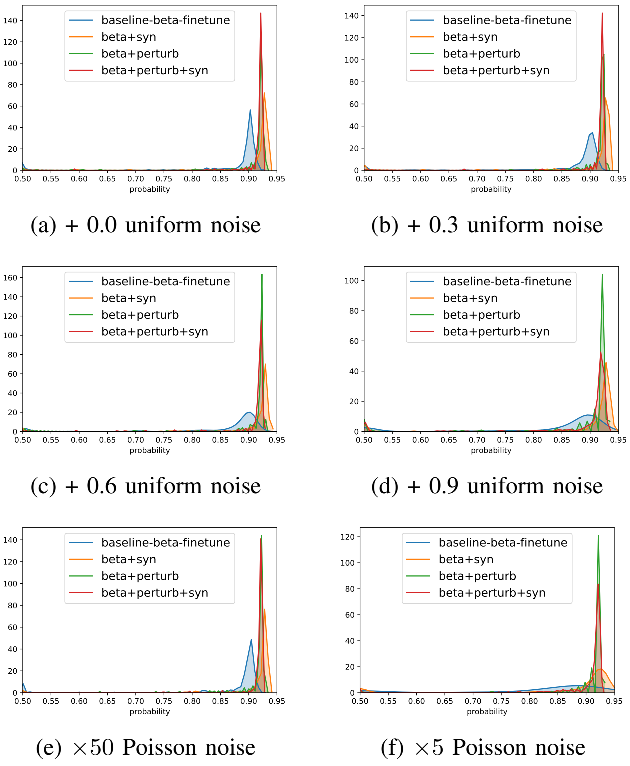

Original
Uniform 0.3
Uniform 0.6
Uniform 0.9
×50 Poisson
×5 Poisson
Adv.
baseline-beta: 86.07% beta+pgd+syn: 92.08%
baseline-beta: 82.26% beta+pgd+syn: 92.11%
baseline-beta: 81.68% beta+pgd+syn: 92.11%
baseline-beta: 80.11% beta+pgd+syn: 92.09%
baseline-beta: 87.87% beta+pgd+syn: 92.03%
baseline-beta: 59.74% beta+pgd+syn: 91.90%
baseline-beta: 2.75% beta+pgd+syn: 91.40%
baseline-beta: 90.18% beta+pgd+syn: 92.25%
baseline-beta: 89.97% beta+pgd+syn: 92.22%
baseline-beta: 89.70% beta+pgd+syn: 92.19%
baseline-beta: 88.93% beta+pgd+syn: 91.27%
baseline-beta: 90.20% beta+pgd+syn: 92.24%
baseline-beta: 83.41% beta+pgd+syn: 90.33%
baseline-beta: 50.00% beta+pgd+syn: 92.23%

Fig. 10: Examples to show different levels of uniform noise and adversarial noise on two nodules randomly drawn from the stress-test.

in this test have diameters at least 6mm. The normalized histograms of the network responses are shown in Fig. 9. Though the conventional network achieved 88.90% CPM, it failed to recognize many sampled nodules even with random sampling. The conventional network predicts the majority of the PGD searched nodules around 50%, which is defined as out-of-distribution samples. The network augmented with PGD synthetic nodules can successfully recognize most of the PGD synthetic nodules with high-confidence.

2) Noise: To stress-test the network resistance to different levels of noise, we first add uniform noise with different magnitudes to the nodule patches as depicted by Fig. 10. The uniform noise can significantly reduce the response from the baseline network as shown in Fig. 11(a)-(d). We found that the network augmented with either synthetic nodules (beta+syn) or PGD noises (beta+perturb and beta+perturb+syn) can be more robust to uniform noise. To simulate the Poisson noise in CT, we rescaled the CT patches to [0, 50] and [0, 1] respectively and then sample from them following the Poisson process. In Fig. 11e and Fig. 11f, similarly to the uniform noise, stronger Poisson noise can deactivate the baseline FPR network while affect less on the augmented networks. Though beta+syn is more robust to mild uniform and Poisson noise, it can not resist the perturbation with adversarial noise (adv. noise) while little difference can be observed from the noise augmented networks in Fig. 12. We also tested the FPR network by feeding randomly generated noise patches and PGD adversarial noise patches, as shown in Fig. 6. In Fig. 13, the conventional FPR network would normally not be activated by random uniform noise, meaning most of the responses are below 50%. However, the adversarial noise patches can easily activate it. The networks augmented by the adversarial noise augmentation (beta+perturb and beta+perturb+syn) were mostly robust to both types of noise patterns. In Fig. 14b, we show that the augmented networks are also more robust to the adversarial noise added to the real negative CT patches than the baseline network.



(a) + 0.0 uniform noise

(b) + 0.3 uniform noise

(c) + 0.6 uniform noise

(d) + 0.9 uniform noise

(e) ×50 Poisson noise

(f) ×5 Poisson noise

Fig. 11: Stress-test by perturbing the positive patches with different levels of uniform noise perturbation.

VI. Discussions and Conclusions

In this paper, we propose adversarial augmentation methods to improve both the generalization and the robustness of the nodule detection framework. We first use the beta-distribution to replace the sigmoid output of the false positive reduction network to estimate the observation uncertainty explicitly at the output layer. Then we add both adversarial synthetic nodules and adversarial perturbation noise to the training set that is searched using the project gradient descent (PGD). The overview of the framework is shown in Fig. 2. By evaluating on two benchmark datasets with different statistics, we show that the proposed augmentation methods can improve the detection CPM scores on the clean datasets. We also use the synthetic nodules and the generated perturbations to stress test the trained models and show the augmented networks can be more robust to both hard nodules as well as different types of noise distortions. By using the beta distribution based uncertainty estimation, we also showed that uncertainty estimation alone might not be sufficient to make the network robust to the out-of-distribution inputs, especially when the inputs are adversarially generated.

As one of the early attempts to enhance the robustness of the medical image analysis CNNs, this study has a few limitations that can be targeted in the future works. We use a relatively simple nodule synthesizer network to sample the lung nodules from the latent space. This synthesizier was not capable of synthesizing all types of different nodules, such as nodules with spiculation. It was also not constrained to maintain the size of a synthetic nodule, therefore we had to filter out the synthetic nodules that are smaller than the relevant threshold. We only investigated the network robustness towards three types of image noise. The improved robustness towards other types of image artefacts, such as metal artefacts and motion distortion, etc., remains unknown. As a proof of concept study, the proposed techniques were only applied to the false positive reduction (FPR) of the lung nodule detection pipeline for brevity. However, the same perturbations can also affect the candidate generation networks. We also found that in practice it is hard to generate adversarial noise by attacking the noise augmented networks without showing visually detectable artefacts. However, it is possible to attack the augmented networks with the same techniques. Though we only evaluated the proposed techniques in the context of nodule detection, we believe such techniques can also be helpful for the other deep CNN based medical imaging applications with minor technical adjustments.

Disclaimer: The concepts and information presented in this paper are based on research results that are not commercially available

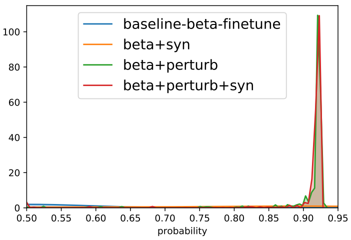
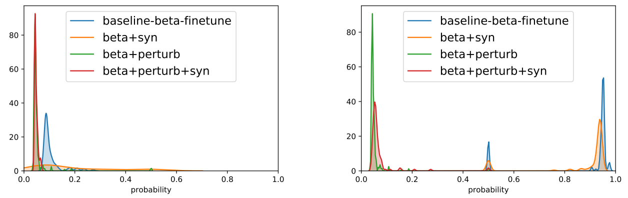


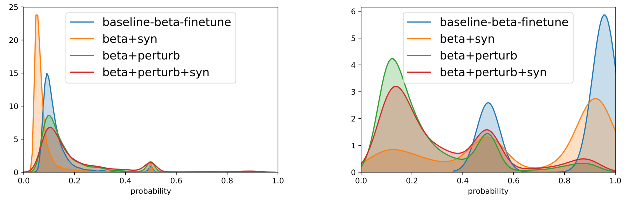
Fig. 12: Stress-test by feeding the PGD perturbed positive patches to different FPR networks.



(a) uniform noise

(b) adversarial noise

Fig. 13: Stress test by feeding noise patches to the network. The baseline model can resist to pure uniform while classifies most of the adversarial noise patches to positive. The model augmented only by the synthetic nodules is also fooled by the adversarial perturbation. Only the two models augmented by adversarial noise can be robust to most of the adversarial noise patterns.



(a) negative samples

(b) adversarial negative samples

Fig. 14: Stress test by (a) feeding negative CT samples to the network and (b) the negative CT samples distorted by PGD adversarial noise.

References

[1] Samuel G. Armato et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A completed reference database of lung nodules on CT scans. *Medical Physics*, 38(2):915–931, 1 2011.

[2] Harry J. de Koning, Carlijn M. van der Aalst, Pim A. de Jong, Ernst T. Scholten, Kristiaan Nackaerts, Marjolein A. Heuvelmans, Jan-Willem J. Lammers, Carla Weenink, Uraujh Yousaf-Khan, Nanda Horeweg, Susan van t Westeinde, Mathias Prokop, Willem P. Mali, Firdaus A.A. Mohamed Hoesein, Peter M.A. van Ooijen, Joachim G.J.V. Aerts, Michael A. den Bakker, Erik Thunnissen, Johny Verschakelen, Rozemarijn Vliegenthart, Joan E. Walter, Kevin ten Haaf, Harry J.M. Groen, and Matthijs Oudkerk. Reduced lung-cancer mortality with volume ct screening in a randomized trial. *New England Journal of Medicine*, 382(6):503–513, 2020. PMID: 31995683.

[3] Jia Ding, Aoxue Li, Zhiqiang Hu, and Liwei Wang. Accurate pulmonary nodule detection in computed tomography images using deep convolutional neural networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 559–567. Springer, 2017.

[4] Qi Dou, Hao Chen, Lequan Yu, Jing Qin, and Pheng-Ann Heng. Multilevel contextual 3-D CNNs for false positive reduction in pulmonary nodule detection. *IEEE Transactions on Biomedical Engineering*, 64(7):1558–1567, 2016.

[5] Samuel G Finlayson, John D Bowers, Joichi Ito, Jonathan L Zittrain, Andrew L Beam, and Isaac S Kohane. Adversarial attacks on medical machine learning. *Science*, 363(6433):1287–1289, 2019.

[6] Angus Galloway, Anna Golubeva, Thomas Tanay, Medhat Moussa, and Graham W Taylor. Batch normalization is a cause of adversarial vulnerability. *arXiv preprint arXiv:1905.02161*, 2019.

[7] Chufan Gao, Stephen Clark, Jacob Furst, and Daniela Raicu. Augmenting LIDC dataset using 3D generative adversarial networks to improve lung nodule detection. In *Medical Imaging 2019: Computer-Aided Diagnosis*, volume 10950, page 109501K. International Society for Optics and Photonics, 2019.

[8] Florin C. Ghesu, Bogdan Georgescu, Eli Gibson, Sebastian Guendel, Mannudeep K. Kalra, Ramandeep Singh, Subba R. Digumarthy, Sasa Grbic, and Dorin Comaniciu. Quantifying and leveraging classification uncertainty for chest radiograph assessment. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 676–684, Cham, 2019. Springer International Publishing.

[9] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved Training of Wasserstein GANs. *ArXiv e-prints*, page arXiv:1704.00028, 2017.

[10] Changhee Han, Yoshiro Kitamura, Akira Kudo, Akimichi Ichinose, Leonardo Rundo, Yujiro Furukawa, Kazuki Umemoto, Yuanzhong Li, and Hideki Nakayama. Synthesizing diverse lung nodules wherever massively: 3D multi-conditional GAN-based CT image augmentation for object detection. In *2019 International Conference on 3D Vision (3DV)*, pages 729–737. IEEE, 2019.

[11] David M Hansell, Alexander A Bankier, Heber MacMahon, Theresa C McLoud, Nestor L Muller, and Jacques Remy. Fleischner society: glossary of terms for thoracic imaging. *Radiology*, 246(3):697–722, 2008.

[12] Douglas Heaven. Why deep-learning ais are so easy to fool. *Nature*, 574(7777):163, 2019.

[13] Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 41–50, 2019.

[14] Dakai Jin, Ziyue Xu, Youbao Tang, Adam P Harrison, and Daniel J Mollura. CT-realistic lung nodule simulation from 3D conditional generative adversarial networks for robust lung segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 732–740. Springer, 2018.

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[16] Yi Li, Huahong Zhang, Camilo Bermudez, Yifan Chen, Bennett A. Landman, and Yevgeniy Vorobeychik. Anatomical context protects deep learning from adversarial perturbations in medical imaging. *Neurocomputing*, 2019.

[17] Siqi Liu, Eli Gibson, Sasa Grbic, Zhoubing Xu, Arnaud Arindra Adiyoso Setio, Jie Yang, Bogdan Georgescu, and Dorin Comaniciu. Decompose to manipulate: Manipulable object synthesis in 3D medical images with structured image decomposition. *arXiv preprint arXiv:1812.01737*, 2018.

[18] Xingjun Ma, Yuhao Niu, Lin Gu, Yisen Wang, Yitian Zhao, James Bailey, and Feng Lu. Understanding adversarial attacks on deep learning based medical image analysis systems. *arXiv preprint arXiv:1907.10456*, 2019.

[19] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[20] Alexander Meinke and Matthias Hein. Towards neural networks that provably know when they don't know. *arXiv preprint arXiv:1909.12180*, 2019.

[21] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *ArXiv e-prints*, page arXiv:1802.05957, February 2018.

[22] M. Niemeijer, M. Loog, M. D. Abramoff, M. A. Viergever, M. Prokop, and B. van Ginneken. On combining computer-aided detection systems. *IEEE Transactions on Medical Imaging*, 30(2):215–223, Feb 2011.

[23] Magdalini Paschali, Sailesh Conjeti, Fernando Navarro, and Nassir Navab. Generalizability vs. Robustness: Adversarial Examples for Medical Imaging. *arXiv e-prints*, page arXiv:1804.00504, Mar 2018.

[24] Magdalini Paschali, Walter Simson, Abhijit Guha Roy, Muhammad Ferjad Naeem, Rüdiger Göbl, Christian Wachinger, and Nassir Navab. Data Augmentation with Manifold Exploring Geometric Transformations for Increased Performance and Robustness. *arXiv e-prints*, page arXiv:1901.04420, Jan 2019.

[25] Lea Marie Pehrson, Michael Bachmann Nielsen, and Carsten Ammitzbøl Lauridsen. Automatic pulmonary nodule detection applying deep learning or machine learning algorithms to the LIDC-IDRI database: A systematic review. *Diagnostics*, 9(1):29, 2019.

[26] Gustavo Pérez and Pablo Arbeláez. Automated detection of lung nodules with three-dimensional convolutional neural networks. In *13th International Conference on Medical Information Processing and Analysis*, volume 10572, page 1057218. International Society for Optics and Photonics, 2017.

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[28] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 3179–3189. Curran Associates, Inc., 2018.

[29] Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Geert Litjens, Paul Gerke, Colin Jacobs, Sarah J van Riel, Mathilde Marie Winkler Wille, Matiullah Naqibullah, Clara I Sánchez, and Bram van Ginneken. Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks. *IEEE transactions on medical imaging*, 35(5):1160–1169, 2016.

[30] Arnaud Arindra Adiyoso Setio, Alberto Traverso, Thomas De Bel, Moira SN Berens, Cas van den Bogaard, Piergiorgio Cerello, Hao Chen, Qi Dou, Maria Evelina Fantacci, Bram Geurts, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. *Medical image analysis*, 42:1–13, 2017.

[31] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. Cancer statistics, 2019. *CA: a cancer journal for clinicians*, 69(1):7–34, 2019.

[32] David Stutz, Matthias Hein, and Bernt Schiele. Disentangling adversarial robustness and generalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6976–6987, 2019.

[33] National Lung Screening Trial Research Team. The national lung screening trial: overview and study design. *Radiology*, 258(1):243–253, 2011.

[34] Qingfeng Wang, Xuehai Zhou, Chao Wang, Zhiqin Liu, Jun Huang, Ying Zhou, Changlong Li, Hang Zhuang, and Jie-Zhi Cheng. WGAN-based synthetic minority over-sampling technique: improving semantic fine-grained classification for lung nodules in CT images. *IEEE Access*, 7:18450–18463, 2019.

[35] Hongtao Xie, Dongbao Yang, Nannan Sun, Zhineng Chen, and Yongdong Zhang. Automated pulmonary nodule detection in CT images using deep convolutional neural networks. *Pattern Recognition*, 85:109–119, 2019.

[36] Ziyue Xu, Xiaosong Wang, Hoo-Chang Shin, Holger Roth, Dong Yang, Fausto Milletari, Ling Zhang, and Daguang Xu. Tunable CT lung nodule synthesis conditioned on background image and semantic features. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 62–70. Springer, 2019.

[37] Ziyue Xu, Xiaosong Wang, Hoo-Chang Shin, Dong Yang, Holger Roth, Fausto Milletari, Ling Zhang, and Daguang Xu. Correlation via synthesis: end-to-end nodule image generation and radiogenomic map learning based on generative adversarial network. *arXiv preprint arXiv:1907.03728*, 2019.

[38] Fei-Fei Xue, Jin Peng, Ruixuan Wang, Qiong Zhang, and Wei-Shi Zheng. Improving robustness of medical image diagnosis with denoising convolutional neural networks. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 846–854, Cham, 2019. Springer International Publishing.

[39] Jie Yang, Siqi Liu, Sasa Grbic, Arnaud Arindra Adiyoso Setio, Zhoubing Xu, Eli Gibson, Guillaume Chabin, Bogdan Georgescu, Andrew F Laine, and Dorin Comaniciu. Class-aware adversarial lung nodule synthesis in CT images. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1348–1352. IEEE, 2019.

[40] Xin Yi, Ekta Walia, and Paul Babyn. Generative adversarial network in medical imaging: A review. *Medical image analysis*, page 101552, 2019.

[41] Junjie Zhang, Yong Xia, Hengfei Cui, and Yanning Zhang. Pulmonary nodule detection in medical images: a survey. *Biomedical Signal Processing and Control*, 43:138–147, 2018.

[42] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. Road extraction by deep residual U-Net. *IEEE Geoscience and Remote Sensing Letters*, 15(5):749–753, 2018.

[43] Wentao Zhu, Chaochun Liu, Wei Fan, and Xiaohui Xie. DeepLung: Deep 3D dual path nets for automated pulmonary nodule detection and classification. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 673–681. IEEE, 2018.