

**ADVERSARIAL MACHINE LEARNING FOR
ADVANCED MEDICAL IMAGING SYSTEMS**

by

Xin Li

DISSERTATION

Submitted to the Graduate School,

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

2022

MAJOR: COMPUTER SCIENCE

Approved By:

Advisor

Date

DEDICATION

To my parents and wife for their love.

ACKNOWLEDGEMENTS

I would like to thank to my PhD advisor, Dr. Dongxiao Zhu. This dissertation would not be possible without the support from him. He has also provided insightful discussions about the research. I am very grateful to his patience, knowledge and spirit that guide me to this field. I also have to thank the members of my PhD committee, Dr. Ming Dong, Alexander Kotov, Suzan Arslanturk, and Indrin Chetty for their helpful career advice and suggestions in general.

I would also like to thank my friends, Xiangrui Li, Deng pan, Yao Qiang and Chengyin Li, for their company and wisdom.

TABLE OF CONTENTS

| | |
|---|-----|
| Dedication | ii |
| Acknowledgements | iii |
| List of Tables | vii |
| List of Figures | ix |
| Chapter 1 Introduction | 1 |
| 1.1 Background and Related Work | 1 |
| 1.2 Research Topics | 6 |
| 1.2.1 Advanced Medical Imaging Systems | 6 |
| 1.2.2 Deploying Advanced Medical Imaging System in the Real-World | 8 |
| 1.2.3 Improving Adversarial Robustness via New Loss Functions | 9 |
| 1.2.4 Learning Compact Features via In-Training Representation Alignment | 10 |
| 1.3 Itemized Summary of Contributions and Code Repositories | 10 |
| Chapter 2 Advanced Medical Imaging Systems | 12 |
| 2.1 Introduction | 12 |
| 2.2 Visual Perception and Interpretation (Vispi) of Chest X-rays | 14 |
| 2.2.1 Background and Related Work | 14 |
| 2.2.2 Methods | 17 |
| 2.2.3 Experiments | 19 |
| 2.3 On-Device COVID-19 Patient Triage and Follow-up | 24 |
| 2.3.1 Background and Related Work | 24 |
| 2.3.2 Methods | 31 |
| 2.3.3 Experiments | 34 |

| | | |
|-----------|--|----|
| 2.3.4 | Performance Evaluation on Mobile Devices | 41 |
| 2.4 | Uncertainty-aware Segmentation with Automatic Contour Outlier Mitigation | 43 |
| 2.4.1 | Background and Related Work | 43 |
| 2.4.2 | Methods | 45 |
| 2.4.3 | Results | 52 |
| 2.4.4 | Discussion | 58 |
| 2.5 | Conclusions | 63 |
| Chapter 3 | Deploying Advanced Medical Imaging System in the Real-World . . . | 65 |
| 3.1 | Introduction and Related Work | 65 |
| 3.2 | Unsupervised Detection | 68 |
| 3.2.1 | Motivation | 68 |
| 3.2.2 | Methods | 70 |
| 3.2.3 | Experiments | 73 |
| 3.3 | Semi-supervised Adversarial Training with Adversarial Risk Assessment . . | 78 |
| 3.3.1 | Motivation | 78 |
| 3.3.2 | Method | 79 |
| 3.3.3 | Experiments | 83 |
| 3.4 | Conclusions | 86 |
| Chapter 4 | Improving Adversarial Robustness of DNNs via Probabilistically Compact Loss with Logit Constraints | 88 |
| 4.1 | Introduction and Related Work | 88 |
| 4.2 | Methods | 91 |
| 4.2.1 | Motivation: Predictive Behavior of CNN on Adversarial Samples . . | 92 |

| | |
|--|-----|
| 4.2.2 Probabilistically Compact Loss | 95 |
| 4.2.3 The Logit Constraints | 96 |
| 4.3 Experiments | 99 |
| 4.3.1 Results | 101 |
| 4.3.2 Identifying Gradient Masking | 105 |
| 4.4 Conclusion | 106 |
| Chapter 5 Summary and Future Work | 107 |
| 5.1 Summary | 107 |
| 5.2 Future Work | 107 |
| Chapter 6 Appendix | 110 |
| References | 113 |
| Abstract | 133 |
| Autobiographical Statement | 135 |

LIST OF TABLES

| | | |
|----------|--|----|
| Table 1 | Automatic evaluations on IU dataset. * results from [68]. + results from [150]. | 20 |
| Table 2 | Comparison of disease classification performance using AUROC. * results are from [142]. | 22 |
| Table 3 | Comparison of full and compact DNN model complexity. | 33 |
| Table 4 | Classification performance of MS networks, The values in ./ indicate MobileNetV2 vs. SqueezeNet. | 37 |
| Table 5 | Performance comparison of two feature aggregation schemes (Difference vs. Concatenation) with four different classifiers using two MS Networks (MobileNetV2 and SqueezeNet) as the feature extractor. Values in parentheses indicate the upper bound of accuracy yielded by RF Network (DenseNet-121). | 40 |
| Table 6 | Comparison of resource consumption of the two on-device MS networks deployed to the six Android based mobile devices. | 42 |
| Table 7 | Comparison of U-Net VAE/OM-VAE with different training and test datasets. Values are shown for DSC, HD (mm), NCC and COM distance (mm) computed for each model iteration against the consensus, human-contoured dataset. | 56 |
| Table 8 | Comparison of U-Net, U-Net/VAE and U-Net/OM-VAE models. The U-Net/VAE or U-Net/OM-VAE considers the average of the highest accuracy (based on DSC) contour among 15 contours randomly outputted. The U-Net outputs only one contour on each image slice. DSC and HD (mm) were computed against the consensus, human-contoured dataset. | 57 |
| Table 9 | Comparison of U-Net/VAE with and without data augmentation. Data was augmented by a factor of 100. Data splitting in the ratio of 6:1:3 (training/validation/test) was done prior to augmentation. Values are shown for DSC, HD (mm) and NCC computed for an average of 15 contours against the consensus, human-contoured dataset. | 58 |
| Table 10 | F1 scores are shown for comparing detection performance and AUROC values weighted average over 14 different classes with standard deviation are shown for comparing classification performance of each attack-defense combination. | 75 |

| | | |
|----------|--|-----|
| Table 11 | UAD performance comparison using AUPRC under PGD attack with a perturbation $\epsilon = 0.005$. The last column shows the number of successful adversarial samples. | 85 |
| Table 12 | Systems risk under PGD attack with a perturbation $\epsilon = 0.005$. SSTA* is the risk under a stronger PGD attack ($\epsilon = 0.01$). | 85 |
| Table 13 | Accuracy (%) on K/F/MNIST, CIFAR-10 and SVHN under white-box setting. For CW, the parameter is the confidence. | 97 |
| Table 14 | Accuracy (%) between GCE and our method on MNIST and CIFAR10 under white-box setting. *Results are directly from [13]. | 100 |
| Table 15 | Accuracy (%) on K/F/MNIST, CIFAR-10 and SVHN under black-box setting. | 101 |
| Table 16 | Accuracy (%) on CIFAR-100 and Tiny ImageNet between CE loss, GCE and our new PC loss. | 101 |
| Table 17 | Accuracy (%) on K/F/MNIST and CIFAR-10 with adversarial training under both white- and black-box attacks. | 103 |

LIST OF FIGURES

| | | |
|-----------|--|----|
| Figure 1 | Illustration of an existing medical report generation system (e.g. [58, 150]) (a) and the proposed medical image interpretation system (b). The former uses a coarse grid of image regions as visual features to generate report directly whereas the latter first predicts and localizes disease as semantic features then followed by report generation. | 15 |
| Figure 2 | An automatic workflow of the X-ray interpretation system. | 17 |
| Figure 3 | Comparison of disease classification performance using ROC curves. . | 22 |
| Figure 4 | Illustration of two cases of example outputs of our system. | 24 |
| Figure 5 | The analytical workflow of COVID-19 CXR interpretation systems. . . | 27 |
| Figure 6 | Overview of the three-player KTD training architecture demonstrating the knowledge transfer from AP to RF and the knowledge distillation from RF to MS. The blue and purple arrows demonstrate the training for two tasks: patient triage and follow-up respectively. | 32 |
| Figure 7 | An example of data preparation for a series of longitudinal CXR images with radiological trajectory labels. The patient is in critical condition on t_3 then recovered afterward. | 35 |
| Figure 8 | The upper panel shows the performance of the large-scale RF network and two compact MS networks of discriminating (a) COVID-19 vs. Normal cases; (b) COVID-19 vs. Pneumonia cases and (c) COVID-19 vs. Normal + Pneumonia cases, while the lower panel shows the performance of discriminating (d) "Worse" vs. "Improved" cases; (e) "Worse" vs. "Stable" cases and (f) "Worse" vs. "Improved" + "Stable" cases. | 39 |
| Figure 9 | Overview of on-device deployment of the COVID-MobileXpert. | 42 |
| Figure 10 | Overview of our training architecture demonstrating the transfer learning and automatic contour outlier mitigation. | 50 |

| | | |
|-----------|---|----|
| Figure 11 | Comparison of U-Net/VAE, U-Net/OM-VAE and radiation oncologist-generated contours on three example segmentations for high (upper panel) average (middle panel) and low (lower panel) accuracy cases, respectively. The thick white lines represent the consensus, ground-truth contours while other lines are the generated contours by either radiation oncologist or the model. DSC scores in the columns 2-4 were computed based on 15 randomly selected contours, while the only the best 5 contours are plotted in the figure to ease viewability. For the first column, 5 physician contours are shown along with the consensus contour. DSC was computed based on an average of the 5 physician Dice scores relative to the consensus contour. | 53 |
| Figure 12 | An adversarial attack against a medical image classifier with perturbations generated using FGSM [43]. | 66 |
| Figure 13 | (a) Visualization of input images and feature maps from the first block of a DenseNet-121 [54]. (b) Visualization of feature distributions from the final fully connected layer of clean X-ray images (green) versus adversarial X-ray images (red). | 69 |
| Figure 14 | The proposed defense framework for a chest X-ray disease classification system equipped with our MGM detection module. | 71 |
| Figure 15 | T-SNE visualization of penultimate layer activations of the model trained on the OCT dataset [60]. The clean images are represented by solid circles with each color represents a true class. The adversarial samples (triangles) are crafted by PGD with a perturbation budget $\epsilon = 0.005$ where each color represents a predicted class. For each class, UAD is capable of filtering out the majority of adversarial samples (center) and SSAT enables the model to correctly predict the rest of adversarial samples (close to clean images). | 78 |
| Figure 16 | The proposed robust OCT imaging classification system equipped with SSAT and UAD modules. | 79 |
| Figure 17 | An illustration of assessing systems adversarial risk. Note the system with UAD on the right exhibits a much lower risk represented by smaller red zones. | 82 |
| Figure 18 | The supervised prediction accuracy of the four trained models on 1000 adversarial examples crafted by FGSM, PGD, C&W with an increasing budget and constant c | 84 |
| Figure 19 | T-SNE visualization of the penultimate layer of ResNet-56 trained with CE loss (left) and PC loss (right) on CIFAR-10. | 89 |

- | | | |
|-----------|--|-----|
| Figure 20 | Empirical investigation on the predictive behavior of CNN on adversarial samples from CIFAR-10 and CIFAR-100. The line (black, right y-axis) represents the number of increased successful attacks when ϵ is increased from its previous grid value. Each bar (left y-axis) represents the percentage of misclassification for the increased successful attacks, measuring number of adversarial samples are misclassified into the 2nd, 3rd, 4th and 5th most probable classes. FGSM and MIM are attack methods. | 92 |
| Figure 21 | T-SNE visualization of the penultimate layer of the model trained by CE loss (a,b) and our PC loss (c,d) on MNIST dataset. (a,c) display only clean images whereas (b,d) also include successful attacks generated with FGSM ($\epsilon = 0.3$). | 104 |
| Figure 22 | Examples to illustrate the shortcut features (top left) and non-relevant features (top right). Good features are highlighted with high salience in the second rows, overlapping with radiologists' annotations. The heatmap based DNN interpretations are generated by FullGrad [128]. 108 | |

CHAPTER 1 INTRODUCTION

1.1 Background and Related Work

While the Deep neural networks (DNNs) have achieved state-of-the-art performance on various tasks and become more ubiquitous in our daily life, there still exist fundamental properties of these systems which are still not completely understood. Recent studies [130, 43] demonstrate that these models are vulnerable to carefully crafted adversarial samples which are created by adding only imperceptibly small perturbations to clean images. These DNN based models suffer from a significant performance drop when predicting adversarial samples. This phenomenon has raised substantial safety concerns on the deployment of security-critical applications of DNNs, such as autonomous driving [2], surveillance [127], and medical imaging systems [25]. When deploying security-critical AI systems, an open-world learning framework is desirable that needs to deal with both benign and adversarial inputs from both independent and identically distributed (IID) and out-of-distribution (OOD) test samples.

IID and OOD test samples In machine learning, it's typical to randomly divide the available data into a training and test dataset, with the former being used to teach the model to perform a particular task and the latter being used to check the model's performance. One common assumption is that those two datasets are drawn from the same distribution. In relation to the training dataset, this test dataset is then referred as IID data [40]. Aside from the IID data, recent studies [118, 151] evaluate the performance of AI systems on OOD data, which are systematically different from the IID data with a significant distribution shift. For example, in the medical domain, a test dataset acquired from different

hospitals from the training dataset can be treated as OOD data [27]. The real-world applications need to be resilient to OOD test set in both benign and adversarial environments [50].

Adversarial Samples Generation Adversarial attacks are designed to fool the classifier to misclassify the adversarial samples from the true class to false classes, either targeted or not. One type of attacks are gradient based such as Fast Gradient Sign Method (FGSM) [43] and its variant Basic Iterative Method (BIM), which efficiently generates adversarial samples by perturbing pixel values according to gradient weights of the loss function [65]. Projected Gradient Descent (PGD) [87] introduce a random starting point at each iteration in FGSM within a specified l_∞ norm-ball, i.e., perturbation budget, to enhance attack effects. Momentum iterative method (MIM) [32] use momentum to help iterative gradient-based methods to avoid sticking into local maximum thus further boost attacking performance. Another type of attacks are optimization-based, for example, Carlini and Wagner (C&W) [12] use binary search mechanism to find the minimal perturbation for a successful attack. SPSA [136] is a gradientfree method which approximates gradient to generate attacks and defeats many defenses. It outperforms gradient-based attacks when the loss surface is hard to optimize. Recently, Croce et al. [24] propose an ensemble of parameter-free attacks named AutoAttack, which integrated a novel budget-aware step size-free variant of PGD, to automatically evaluate adversarial robustness without any hyper-parameter tuning. We refer the reader to RobustBench library [23] for overview and comparison of more adversarial sample generation methods.

Adversarial IID Defense Recent research has largely focused on mitigating adversarial attacks on a supervised classification system using multiclass natural image and natural language as IID datasets. To evade adversarial attacks and enhance model robustness, various defensive techniques have been proposed. One line of approaches [65, 126, 154, 120] are based on adversarial training [43] and achieve effective robustness against different adversarial attacks, where the training dataset is augmented with adversarial samples. However, these methods have trade-offs between accuracy and adversarial robustness [134] and are computationally expensive in adversarial samples generation [154]. To reduce the computational burden, Shafahi et al. [120] propose a training algorithm, which improves the efficiency of adversarial training by updating both model parameters and image perturbation in one backward pass. Another line of defending strategy against adversaries, other than augmenting the training dataset, is to learn feature representations with adversarial robustness by using model ensembles or altering network architectures [131, 90, 133, 102]. For example, [131] augment CNNs with the radial basis function kernel to further transform features via kernel trick to improve the class separability in feature space and reduce the effect of perturbation. [90] propose a prototype objective function, together with multi-level deep supervision. Their method ensures the separation in feature space between classes and shows significant improvement of robustness. [102] develop a strong ensemble defense strategy by introducing a new regularizer to encourage diversity among models within the ensemble system, which encourage the feature representation from the same class to be close. Although these approaches avoid high computational cost of adversarial training, they have to modify the network architecture or require extra training process, limiting the flexibility in adapting to different tasks.

More efficient approaches are designing new loss functions to improve model adversarial robustness. By explicitly imposing regularization on latent features, CNNs are encouraged to learn feature representation with more inter-class separability and intra-class compactness [101, 90, 14]. For example, [101] propose a Max-Mahalanobis center (MMC) loss to learn discriminative features. They first calculate Max-Mahalanobis [100] centers for each class and then encourage the features to gather around the centers using Center Loss [144]. However, the assumption of geometrical compactness for latent features (in terms of Euclidean distance or L_2 -norm) may not hold due to inherent intra-class variations in the data and usually requires suitable assumptions on distribution of the latent features. [14] encourages the predicted probabilities of false classes to be equally distributed, whereas our Probabilistically Compact (PC) Loss with Logit Constraints [72] directly enlarges the gap of probabilities between true class and the first several most probable false classes to increase model’s robustness. However, when dataset become complex with more classes, this gap is smaller due to generally lower output probability for the true class, resulting a limited robustness improvement.

Adversarial OOD detection When trained on IID samples, DNNs are known to fail against test inputs that lie far away from training distribution, commonly referred to as OOD samples [50]. Recent ODD detection work considers a multi-class dataset as IID (e.g., CIFAR-10) and use samples from another multi-class dataset as OOD (CIFAR-100) [50, 67, 77, 143]. Existing research either train an OOD detector and a classifier sequentially [118] or simultaneously [15]. For example, [118] employ adversarial training on IID data as well as OOD samples that are close to IID samples to improve learning ro-

bust features. These approaches work well for the so-called closed-world detection where OOD samples are either with simpler data modalities (e.g., medical images with large shared background) or closer to IID samples (CIFAR-10 versus CIFAR-100). Different from IID detection tasks where robust discriminative features are learned from labeled training data, OOD detection needs to learn high-level, task-agnostic and semantic features from the IID dataset to detect diverse OOD inputs at the test time. More recent OOD detection approaches are self-supervised representation learning using only unlabeled training data, which involves two key steps: 1) learning a good (e.g., compact and semantic) feature representation, and 2) modeling features of IID data without requiring class labels. For example, [145] used contrastive training techniques SimCLR [16] to extract semantic features and proposed confusion log probability to determine whether a test example is a near or far OOD example. Using experiments they show their approach is scalable to high-dimensional multimodal OOD samples. [15] also use contrastive loss based label-free training for self-supervised feature learning followed by OOD detection using Mahalanobis distance. Another line of label-free feature learning approaches for OOD detection use flow-based generative models (e.g., VAEs, PixelCNNs, and Glow [62], allowing for the exact formulation of the marginal likelihood, to learn task-agnostic and semantic features to address the OOD detection problem. However, even sophisticated neural generative models trained to estimate feature density distribution (e.g. on CIFAR-10 images) can perform poorly on OOD detection, often assigning higher probabilities to OOD test samples than to IID test samples [92]. Most recent research attempt to learn task-agnostic and semantic features for both IID and OOD images [15, 93, 122, 155], yet text-based detection has unique challenges in learning task-agnostic and semantic representations. Moreover

the critical issue of error control in security-critical applications have not been sufficiently addressed.

In this dissertation, I will mainly focus on tackling the problem of adversarial robustness with novel defense methods for both medical imaging AI systems and general DNN based natural image classification systems. Specifically, Chapter 2 will first introduce advanced medical imaging AI systems that are designed for real-world deployment. To protect these high-stakes and security-critical medical applications, Chapter 3 follows up with novel defense techniques which tackle the unique challenges of defending against adversarial samples on medical images. In chapter 4, We offer an unique insight into the predictive behavior of DNNs on adversarial samples that the former tends to misclassify the latter into the first several most probable classes and propose a new loss function to enhance the adversarial robustness of general DNN based models. Finally, Chapter 5 provides a brief summary of the our current works and description of a future work that tackle the generalization gap via adversarial machine learning.

1.2 Research Topics

1.2.1 Advanced Medical Imaging Systems

Novel DNN based medical image interpretation systems are developed for automatic medical image analysis. We first propose an automatic Visual Perception and Interpretation (Vispi) system of Chest X-ray (CXR) designed to greatly reduce the workload of radiologists. To our knowledge, this is among the first attempts to exploit disease localization for X-ray image report generation with visual supports. Different from the existing medical report generation systems which directly generate reports from a CXR, Vispi first

annotates a CXR via disease classification and localization then followed by a report generation from an attentive LSTM model. The design of Vispi enjoys following advantages: (1) the disease classification and localization module can be pre-trained with a large CXR classification dataset thus avoid the lack of annotated image-report pairs problem; (2) with localized semantic features generated by the first step, the report generation module can produce more detail description of the annotated disease. As a result, Vispi achieves superior performance both in disease classification, localization, and report generation.

With the increasing number of smart devices and improved hardware, there is a growing interest to deploy DNN based models on device to minimize latency and maximize the protection of privacy. Especially during the COVID-19 pandemic, there has been an emerging need for on-device COVID-19 patient triage and follow-up system. In view of this need, we present COVID-MobileXpert: a lightweight DNN based mobile app that can use CXR for COVID-19 case screening and radiological trajectory prediction. In the related task of on-device natural image classification, the large-scale teacher network is pre-trained with ImageNet and distill the knowledge to a lightweight student network. This two-player framework, although seemingly successful, can be problematic for on-device medical imaging based COVID-19 case screening and radiological trajectory prediction task. The large gap between natural images and the medical images of a specific disease such as COVID-19 makes the knowledge distillation less effective than it is supposed to be. The small number of labeled COVID-19 images for training further aggravates the situation. To tackle these challenges, We design and implement a novel three-player knowledge transfer and distillation (KTD) framework including a pre-trained attending physician (AP) network that extracts CXR imaging features from a large scale of lung disease CXR images, a fine-tuned

resident fellow (RF) network that learns the essential CXR imaging features to discriminate COVID-19 from pneumonia and/or normal cases with a small amount of COVID-19 cases, and a trained lightweight medical student (MS) network to perform on-device COVID-19 patient triage and follow-up. The three-player framework provides a more effective way to train the compact on-device model using a smaller dataset while preserving performance.

In addition to the classification task, we also propose a segmentation framework named as U-Net/OM-VAE for task automation in radiation oncology. This Variational autoencoder (VAE) combined with a hierarchical U-Net and an outlier mitigation strategy (UNet/OM-VAE) demonstrates promise towards capturing inter-observer variability and produces accurate prostate auto-contours for radiotherapy planning. The availability of multiple contours for each CT slice enables clinicians to determine trade-offs in selecting the “best fitting” contour on each CT slice. The detection and mitigation of outlier contours in the training dataset improves prediction accuracy but one must be wary of reduction in variability in the training dataset. Such an automated approach is likely to improve efficiency and consistency, in prostate contour delineation and is robust to inter-observer variability.

1.2.2 Deploying Advanced Medical Imaging System in the Real-World

Although DNN based systems trained on medical images have shown state-of-the-art performance in many clinical prediction tasks, recent studies demonstrate that these systems can be fooled by carefully crafted adversarial samples. It has raised concerns on the practical deployment of DNN based medical image classification systems. Although an array of defense techniques have been developed and proved to be effective in computer vision, defending against adversarial attacks on medical images remains largely an uncharted territory due to their unique challenges: (1) crafted adversarial noises added to

a highly standardized medical image can make it a hard sample for the model to predict; (2) label scarcity limits adversarial generalizability. To tackle these challenges, we proposed two novel defense methods that are tailor-designed for medical image AI systems. We first propose an unsupervised learning approach to detect and reject those hard samples (adversarial samples) in an abnormal detection manner. This approach is capable of detecting a wide range of adversarial samples without knowing the attackers nor sacrificing the classification performance. To address the label scarcity problem. In addition to unsupervised abnormal detection, we include the Semi-Supervised Adversarial Training to further enhance the adversarial robustness utilizing both label and unlabeled data. Both approaches can be easily embedded into any deep learning-based medical imaging system as a module to improve the system’s robustness.

1.2.3 Improving Adversarial Robustness via New Loss Functions

DNNs have achieved state-of-the-art performance on various tasks in computer vision. However, recent studies demonstrate that these models are vulnerable to carefully crafted adversarial samples and suffer from a significant performance drop when predicting them. Many methods have been proposed to improve adversarial robustness (e.g., adversarial training and new loss functions to learn adversarially robust feature representations). Here we offer a unique insight into the predictive behavior of CNNs that they tend to misclassify adversarial samples into the most probable false classes. This inspires us to propose a new Probabilistically Compact (PC) loss with logit constraints which can be used as a drop-in replacement for cross-entropy (CE) loss to improve CNN’s adversarial robustness. Specifically, PC loss enlarges the probability gaps between true class and false classes meanwhile the logit constraints prevent the gaps from being melted by a small perturbation.

1.2.4 Learning Compact Features via In-Training Representation Alignment

We further inspect the standard training strategy of DNNs. During the training, in each epoch, the true gradient of the loss function is estimated using a mini-batch sampled from the training set and model parameters are then updated with the mini-batch gradients. Although the latter provides an unbiased estimation of the former, they are subject to substantial variances derived from the size and number of sampled mini-batches, leading to noisy and jumpy updates. To stabilize such undesirable variance in estimating the true gradients, we propose In-Training Representation Alignment (ITRA) that explicitly aligns feature distributions of two different mini-batches with a matching loss in the training process. We also provide a rigorous analysis of the desirable effects of the matching loss on feature representation learning: (1) extracting compact feature representation; (2) reducing over-adaption on mini-batches via an adaptively weighting mechanism; and (3) accommodating to multi-modalities.

1.3 Itemized Summary of Contributions and Code Repositories

- Chapter 2 presents three novel DNN based medical imaging AI systems.
 - Section 2.2 describes the Vispi: a system for automatic visual perception and interpretation of CXR images, which first annotates an image via classifying and localizing common thoracic diseases with visual support and then followed by report generation [70].
 - Section 2.3 introduces the COVID-MobileXpert: a lightweight DNN based mobile app that can use CXR for COVID-19 case screening and radiological trajectory prediction [71]. Code for this Section is available from the following url:

<https://github.com/xinli0928/COVID-Xray>.

- Section 2.4 provides the U-Net/OM-VAE framework for automatic contouring of the prostate gland incorporating inter-observer variation and a novel outlier mitigation technique [69].
- Chapter 3 presents two novel defense techniques against adversarial attacks on medical imaging AI systems.
 - Section 3.2 introduces an unsupervised learning approach to detect adversarial attacks on medical images [76]. Code for this Section is available from the following url: <https://github.com/xinli0928/MGM>
 - Section 3.3 presents a novel robust medical imaging AI framework based on Semi- Supervised Adversarial Training and Unsupervised Adversarial Detection, followed by a new measure for assessing systems adversarial risk [74]. Code for this Section is available from the following url: <https://github.com/xinli0928/Defending>.
- Chapter 4 presents a new loss function named Probabilistically Compact Loss with Logit Constraints to improve the adversarial robustness of general DNN based classification models [72]. Code for this Section is available from the following url: <https://github.com/xinli0928/PC-LC>.
- Chapter 5 describes our future research. We will improve the feature representation learning and generalization of the medical imaging AI system via a novel training strategy named Saliency Guided Adversarial Training.

CHAPTER 2 ADVANCED MEDICAL IMAGING SYSTEMS

2.1 Introduction

Medical imaging contains the essential information for rendering diagnostic and treatment decisions. Inspecting and interpreting image are tedious clinical routines for a radiologist where automation is expected to greatly reduce the workload. Especially during the COVID-19 pandemic, there is increasing demand for interpreting millions of CXR cases. To reduce the workload of radiologists, accurate and rapid medical image analysis systems are urgently needed. In this chapter, we propose two medical image interpretation systems named Vispi and COVID-MobileXpert to meet different clinical needs in the real-world. Despite the rapid development of natural image captioning, computer-aided medical image visual perception and interpretation to generate a report remain a challenging task, largely due to the lack of high-quality annotated image-report pairs and tailor-made generative models for sufficient extraction and exploitation of localized semantic features, particularly those associated with abnormalities. To tackle these challenges, we present Vispi, an automatic medical image interpretation system, which first annotates an image via classifying and localizing common thoracic diseases with visual support and then followed by report generation from an attentive LSTM model. Analyzing an open Indiana University (IU) X-ray dataset, we demonstrate a superior performance of Vispi in disease classification, localization, and report generation using automatic performance evaluation metrics ROUGE [114] and CIDEr [3].

During the COVID-19 pandemic, in addition to the demand for an accurate medical image interpretation system, there has been an emerging need for rapid, dedicated, and

point-of-care COVID-19 patient disposition techniques to optimize resource utilization and clinical workflow. In view of this need, we present COVID-MobileXpert: a lightweight deep neural network (DNN) based mobile app that can use CXR for COVID-19 case screening and radiological trajectory prediction. We design and implement a novel three-player knowledge transfer and distillation (KTD) framework including a pre-trained attending physician (AP) network that extracts CXR imaging features from a large scale of lung disease CXR images, a fine-tuned resident fellow (RF) network that learns the essential CXR imaging features to discriminate COVID-19 from pneumonia and/or normal cases with a small amount of COVID-19 cases, and a trained lightweight medical student (MS) network to perform on-device COVID-19 patient triage and follow-up. To tackle the challenge of vastly similar and dominant fore- and background in medical images, we employ novel loss functions and training schemes for the MS network to learn the robust features. We demonstrate the significant potential of COVID-MobileXpert for rapid deployment via extensive experiments with diverse MS architecture and tuning parameter settings.

In addition to the classification task for medical image analysis, the task automation is essential for efficient and consistent image segmentation in radiation oncology. Here we propose a Variational autoencoder (VAE) combined with a hierarchical U-net and an outlier mitigation strategy (U-Net/OM-VAE) demonstrates a strong promise towards capturing inter-observer variability and producing accurate prostate contours for radiotherapy planning. The availability of multiple contours for each CT slice using the VAE enables physicians to determine clinical tradeoffs in selecting the most appropriate contour on each 2D CT image. Such an automated approach is likely to improve efficiency and consistency, in prostate contour delineation and is robust against inter-observer variability.

2.2 Visual Perception and Interpretation (Vispi) of Chest X-rays

2.2.1 Background and Related Work

X-ray is a widely used medical imaging technique in clinics for diagnosis and treatment of thoracic diseases. Medical image interpretation, including both disease annotation and report writing, is a laborious routine for radiologists. Moreover, the quality of interpretation is often quite diverse due to the differential levels of experience, expertise and workload of the radiologists. To release radiologists from their excessive workload and to better control quality of the written reports, it is desirable to implement a medical image interpretation system that automates the visual perception and cognition process and generates draft reports for radiologists to review, revise and finalize.

Despite the rapid and significant development, the existing natural image captioning models, e.g. [64, 149], fail to perform satisfactorily on medical report generation. The major challenge lies in the limited number of image-report pairs and relative scarcity of abnormal pairs for model training, which are essential for quality radiology report generation. Additional challenge is the lack of appropriate performance evaluation metrics; the n -gram based BLEU scores widely used in natural language processing (NLP) are not suitable for assessing the quality of generated reports.

Nevertheless several approaches have been developed to generate reports automatically for CXR images using the CNN-RNN architecture developed in natural image captioning research [58, 68, 142, 150] (Fig. 1a). Since the medical report typically consists of a sequences of sentences, [58] use a hierarchical LSTM [64] to generate paragraphs and achieve impressive results on IU X-ray dataset [28]. Instead of only using visual features

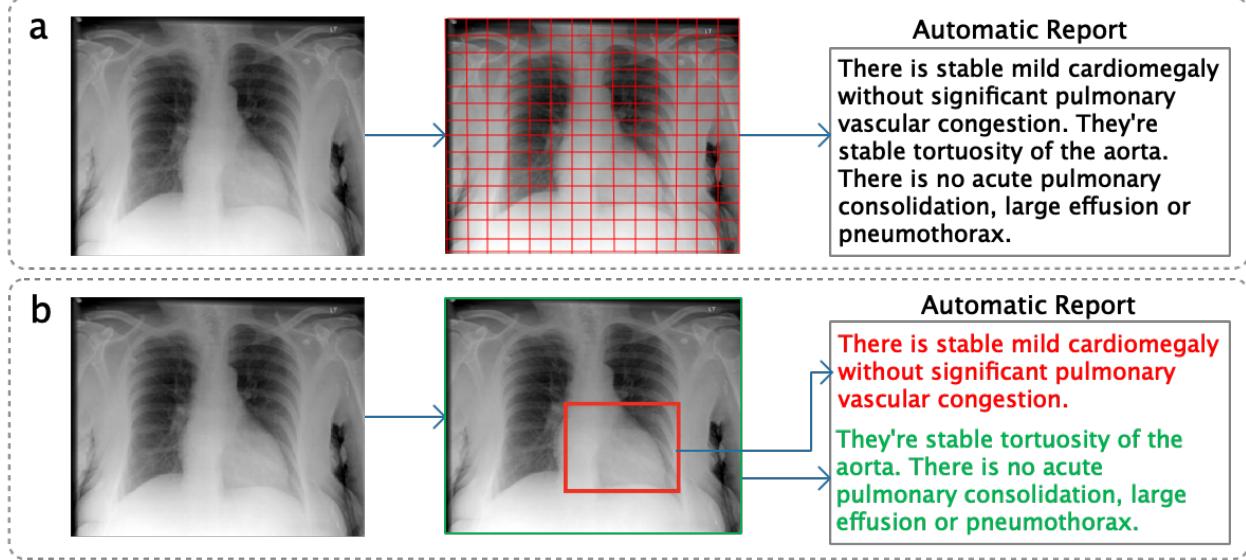


Figure 1: Illustration of an existing medical report generation system (e.g. [58, 150]) (a) and the proposed medical image interpretation system (b). The former uses a coarse grid of image regions as visual features to generate report directly whereas the latter first predicts and localizes disease as semantic features then followed by report generation.

extracted from image, they first predict the Medical Text Indexer (MTI) annotated tags, and then combine semantic features from the tags with visual features from the images for report generation. Similarly, [150] use both visual and semantic features but generate ‘impression’ and ‘findings’ of the report separately. The former one-sentence summary is generated from a CNN encoder whereas the latter paragraph is generated using visual and semantic features. Different from CoAtt, the semantic feature is extracted by embedding the last generated sentence as opposed to the annotated tags. [68] use a hierarchical decision-making procedure to determine whether to retrieve a template sentence from an existing template corpus or to invoke the lower-level decision to generate a new sentence from scratch. The decision priority is updated via reinforcement learning based on sentence-level and word-level rewards or punishments. However, none of these methods demonstrate a satisfactory performance in disease localization and classification, which is

a central issue in medical image interpretation.

TieNet [142] address both disease classification and medical image report generation problems in the same model. They introduce a novel Text-Image Embedding network (TieNet), which integrates self-attention LSTM using textual report data and visual attention CNN using image data. TieNet is capable of extracting an informative embedding to represent the paired medical image and report, which significantly improves the disease classification performance compared to [141]. However, TieNet’s performance on medical report generation improves only marginally over the baseline approach [149], trading the medical report generation performance for the disease classification performance. Moreover, TieNet does not provide a visual support for radiologists to review and revise the automatically generated report.

We present an automatic medical image interpretation system with *in situ* visual support striving for a better performance in both image annotation and report generation (Fig. 1b). To our knowledge this is among the first attempts to exploit disease localization for X-ray image report generation with visual supports. Our contributions are in four-fold: (1) we describe an integrated image interpretation framework for disease annotation and medical report generation, (2) we transfer knowledge from large image data sets (ImageNet and ChestX-ray8) [141] to enhance medical image interpretation using a small number of reports for training (IU X-ray) [28], (3) we evaluate suitability of the NLP evaluation metrics for medical report generation, and (4) we demonstrate the functionality of localizing the key finding in an X-ray with a heatmap.

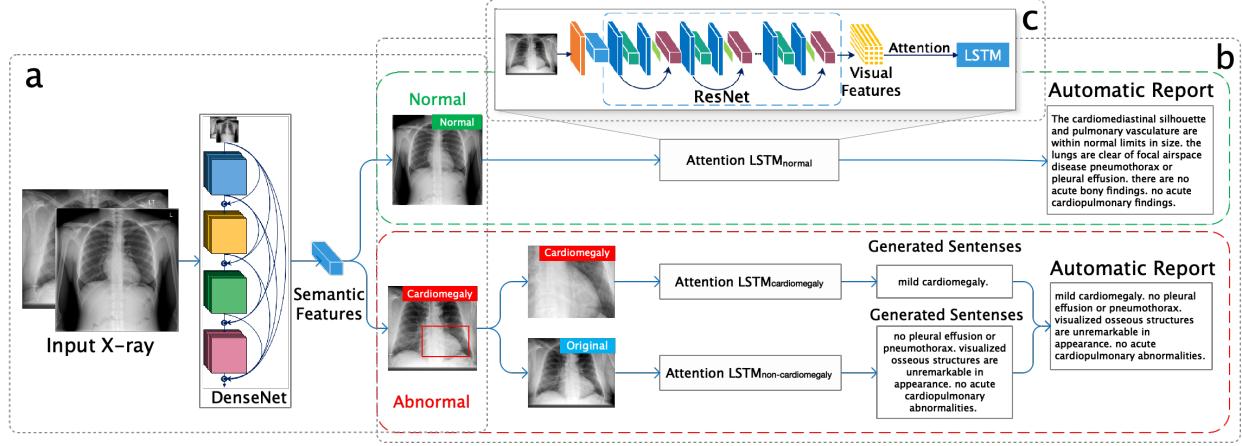


Figure 2: An automatic workflow of the X-ray interpretation system.

2.2.2 Methods

Our workflow (Fig. 2) first annotates an X-ray image by classifying and localizing thoracic diseases (Fig. 2a) and then generates the corresponding sentences to build up the entire report (Fig. 2b). Fig. 2c displays the structure of attentive LSTM used to generate reports.

Disease Classification and Localization Fig. 2a shows our classification module built on a 121-layer Dense Convolutional Network (DenseNet) [54]. Similar to [108], we replace the last fully-connected layer with a new layer of dimension M , where M is the number of diseases. This is a multiple binary classification problem that input is a frontal view X-ray image x and output is a binary vector $y = [y_1, \dots, y_m, \dots, y_M]$, i.e., $y_m \in \{0, 1\}$, indicating absence or presence of a disease m . The binary cross-entropy loss function is defined as:

$$L(x, y) = - \sum_{m=1}^M [y_m(\log g_m(y)) + (1 - y_m) \log(1 - g_m(y))],$$

where $g_m(y)$ is the probability for a target disease m . If $g_m(y) > 0.8$, an X-ray is annotated with disease m for the next level modeling. Otherwise, it is considered as “Normal”. It is worth mentioning that a vast

majority of X-rays are considered as “Normal”, therefore, other choices of thresholds also work well with our system.

We apply Grad-GAMs [119] to localize disease with a heatmap. Gard-CAMs uses the gradient information and flows it back to the final convolutional layer to decipher the importance of each neuron in classifying an image to disease m . Formally, let \mathbf{A}_k be the k th feature maps and weight w_{mk} represents importance of the feature map k for the disease m . We first calculate the gradient of the score for class m , z_m (before the sigmoid), with respect to a feature map \mathbf{A}_k , i.e., $\frac{\partial z_m}{\partial \mathbf{A}_k}$. Thus w_{mk} are calculated by: $w_{mk} = \frac{1}{N} \sum_i \sum_j \frac{\partial z_m}{\partial \mathbf{A}_k}$. (i, j) represents the coordinates of a pixel, and N is the total number of pixels. We then generate a heatmap for disease m by applying weighted average of \mathbf{A}_k , followed by a ReLU activation: $\mathbf{H}_m = \text{ReLU}(\sum_k w_{mk} \mathbf{A}_k)$. The localized semantic features to predict disease m are identified and visualized with the heatmap \mathbf{H}_m . Similar to [141], we apply a thresholding based bounding box (B-Box) generation method. The B-Box bounds pixels whose heatmap intensity is above 90% of the maximum intensity. The resulting region of interest is then cropped for next level modeling.

Attention-based Report Generation Fig. 2b illustrates the process of report generation. If there is no active thoracic disease found in an X-ray, a report will be directly generated by an attentive LSTM based on the original X-ray as shown in the green dashed box. Otherwise (as shown in the red dashed box), the cropped subimage with localized disease from the classification module (Fig. 2a) is used to generate description of abnormalities whereas the original X-ray is used to generate description of normalities in the report.

As shown in the Fig. 2c, the attentive LSTM is based on an encoder-decoder structure

[149], which takes either the original X-ray image or the cropped subimage corresponding to abnormal region as the input and generates a sequence of sentences for the entire report. Our encoder is built on a pre-trained ResNet-101 [48], which extracts the visual features matrix $\mathbf{F} \in \mathbb{R}^{2048 \times 196}$ (reshaped from $2048 \times 14 \times 14$) from the last convolutional layer followed by an adaptive average pooling layer. Each vector $\mathbf{F}_k \in \mathbb{R}^{2048}$ of \mathbf{F} represents one regional feature vector, where $k = \{1, \dots, 196\}$.

The LSTM decoder takes \mathbf{F} as input and generates sentences by producing a word \mathbf{w}_t at each time t . To utilize the spatial visual attention information, we define the weights α_{tk} , which can be interpreted as the relative importance of region feature \mathbf{F}_k at time t . The weights α_{tk} is computed by a multilayer perceptron f : $e_{tk} = f(\mathbf{F}_k, \mathbf{h}_{t-1})$ and $\alpha_{tk} = \text{Softmax}(e_{tk})$, and hence the attentive visual feature vector \mathbf{V}_t is computed by $\mathbf{V}_t = \sum_{k=1}^{196} \alpha_{tk} \mathbf{F}_k$. In addition to the weighted visual feature \mathbf{V}_t and last hidden layer \mathbf{h}_{t-1} , the RNN also accepts the last output word \mathbf{w}_t at each time step as an input. We concatenate the embedding of last output word and visual feature as context vector \mathbf{c}_t . Thus the transition to the current hidden layer \mathbf{h}_t can be calculated as: $\mathbf{h}_t = \text{LSTM}(\mathbf{c}_t, \mathbf{h}_{t-1})$. After model training, a report is generated by sampling words $\mathbf{w}_t \sim p(\mathbf{w}_t | \mathbf{h}_t)$ and updating the hidden layer until hitting the stop token.

2.2.3 Experiments

Datasets. We use the IU X-ray dataset [28], an open image dataset with 3955 radiology reports paired with CXR images (one study per patient) for our experimental evaluation. Each report contains three sections: impression, findings and Medical Subject Headings (MeSH) terms. Similar to [58, 150], we generate sentences in ‘impression’ and ‘findings’ together. The MeSH terms are used as labels for disease classification [142] as well as the

| Model | CIDEr | ROUGE | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 |
|------------------------|--------------|--------------|--------------|--------------|--------------|--------------|
| CNN-RNN [138]* | 0.294 | 0.306 | 0.216 | 0.124 | 0.087 | 0.066 |
| LRCN [31]* | 0.284 | 0.305 | 0.223 | 0.128 | 0.089 | 0.067 |
| AdaAtt [83]* | 0.295 | 0.308 | 0.220 | 0.127 | 0.089 | 0.068 |
| Att2in [111]* | 0.297 | 0.308 | 0.224 | 0.129 | 0.089 | 0.068 |
| CoAtt [58]* | 0.277 | 0.369 | 0.455 | 0.288 | 0.205 | 0.154 |
| HRGR [68]* | 0.343 | 0.322 | 0.438 | 0.298 | 0.208 | 0.151 |
| MRA [150] ⁺ | N\A | 0.366 | 0.464 | 0.358 | 0.270 | 0.195 |
| Vispi | 0.553 | 0.371 | 0.419 | 0.280 | 0.201 | 0.150 |

Table 1: Automatic evaluations on IU dataset. * results from [68]. ⁺ results from [150].

follow-up report generation with abnormality and normality descriptions. We convert all the words to lower-case, remove all non-alphanumeric tokens, replace single-occurrence tokens with a special token and use another special token to separate sentences. We filter out images and reports that are non-relevant to the eight common thoracic diseases included in both ChestX-ray8 [141] and IU X-ray datasets [28], resulting in a dataset with 2225 pairs of X-ray image and report. Finally, we split all the image-report pairs into training, validation and test dataset by ratio 7 : 1 : 2 without patient overlap.

Implementation Details. We implement our model using PyTorch. The dimension of all hidden layers and word embeddings are set to 512. The network is trained with Adam optimizer with a mini-batch size of 16. The training stops when the performance on validation dataset does not increase for 20 epochs. We do not fine-tune the DenseNet pretrained with ChestX-ray8 [141] and ResNet pretrained with ImageNet due to the small sample size of IU X-ray dataset [28]. For each disease class, a specific pair of LSTMs are trained to ensure consistency between the predicted disease annotation(s) and the generated report. For the disease classes with less than 50 samples, we train a shared attentive LSTM across the classes to generate normality description of the report.

Evaluation of Automatic Medical Image Reports. We use the metrics for NLP tasks such as BLEU [104], ROUGE [78], and CIDEr [1] for automatic performance evaluation. As shown in Table 1, our model outperforms all baseline models [31, 83, 111, 138] and demonstrates the best CIDEr and ROUGE scores among all the advanced methods specifically designed for medical report generation [58, 68, 150], despite the fact that we only use a single frontal view X-ray. While BLEU scores measure the percentage of consistency between the automatic report and the manual report in light of the automatic report (precision), it is not illuminative in assessing the amount of information captured in the automatic report in light of the manual report (recall). In real-world clinical applications, both recall and precision are critical in evaluating the quality of an automatic report.

For example, automatic reports often miss description of abnormalities that contained in manual reports written by human radiologists [68, 150], which may decreases recall but does not affect precision. Thus, the automatic report missing the key disease information can still achieve high BLEU scores nevertheless it provides limited insight for medical image interpretation. Therefore, ROUGE is more suitable than BLEU for evaluating the quality of automatic reports since it measures both precision and recall. Further, CIDEr is more suitable for our purpose than ROUGE and BLEU since it captures the notions of grammaticality, saliency, importance and accuracy [1]. Additionally, CIDEr uses TF-IDF to filter out unimportant common words and weight more on disease keywords. As a result, higher ROUGE and CIDEr scores demonstrate a superior performance of our medical image interpretation system.

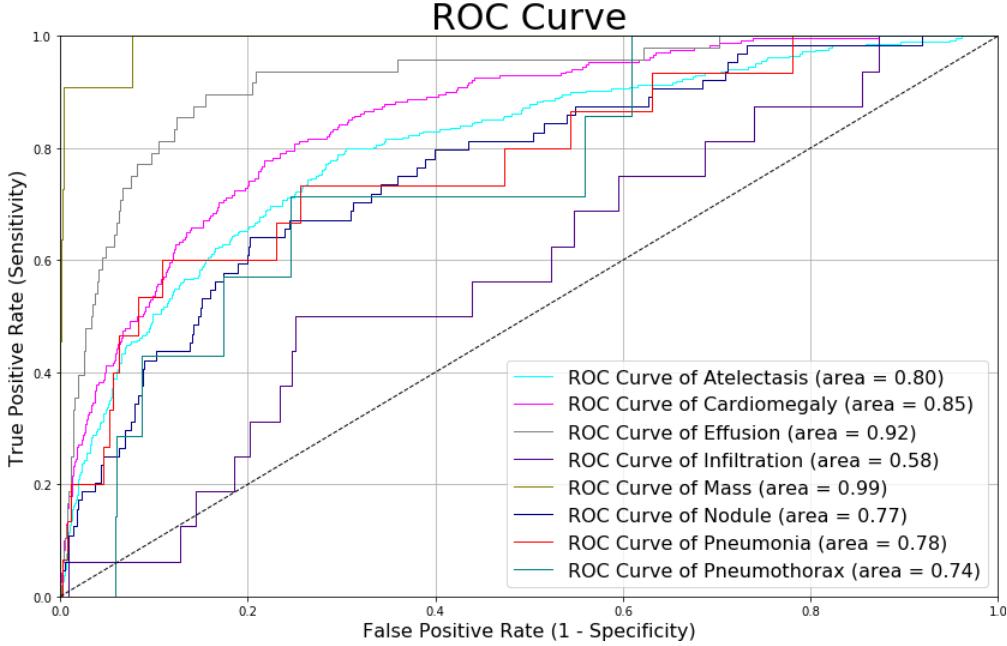


Figure 3: Comparison of disease classification performance using ROC curves.

| Disease | Ate | Cardio | Effusion | Infil | Mass | Nodule | Pneum | Pneumox | Average |
|---------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| TieNet* | 0.744 | 0.847 | 0.899 | 0.718 | 0.823 | 0.658 | 0.731 | 0.709 | 0.757 |
| Vispi | 0.806 | 0.856 | 0.919 | 0.610 | 0.984 | 0.758 | 0.764 | 0.733 | 0.804 |

Table 2: Comparison of disease classification performance using AUROC. * results are from [142].

Evaluation of Disease Classification. Although ROUGE and CIDEr scores are effective in evaluating the consistency of an automatic report to a manual report, none of them, however, are designed for assessing the correctness of medical report annotation in terms of common thoracic diseases. The latter is another key output of a useful image interpretation system. For example, the automatically generated sentence: “no focal airspace consolidation, pleural effusion or pneumothorax” is considered as similar to the manually written sentence: “persistent pneumothorax with small amount of pleural effusion” using both ROUGE and CIDEr scores despite the completely opposite annotations. Therefore, we assess the accuracy in medical report annotation by comparing with TieNet [142] in disease classification using Area Under the ROC (AUROC) as the metric. Our result outper-

forms TieNet’s classification module in 7 out of 8 diseases (Table 2, Fig. 3), even though TieNet is trained on the enhanced version of ChestX-ray8 with 3172 more X-rays and 6 more labeled diseases.

We note that many X-ray based disease classification tasks are multi-label multi-class classification problem. Different from multi-class classification problem where classes are one-hot coded thus mutually exclusive, here we attempt to solve multi-label multi-class classification problem where tasks are inherently related. The classification task of each class is learned simultaneously and synergistically with others using a shared feature representation. As such, the performance of a multi-label classification can benefit considerably from more training samples and classes. Clinically speaking, comorbidity does exist in lung diseases, e.g., Infiltration coexists with Effusion and Atelectasis [141]. Consequently, using additional training samples and extra related disease classes (e.g. 14 classes in [142]) can indeed improve the classification performance. Nevertheless, our approach outperforms TieNet with less number of training samples and classes (8 classes in this study). It is likely that TieNet trades image classification performance for report generation performance whereas our model exploits the former to enhance the latter via a bi-level attention mechanism.

Example System Outputs. Fig. 4 shows two example outputs each with a generated report and image annotation. The first row presents an annotated “Normal” case whereas the second row presents an annotated “Cardiomegaly” case with the disease localized in a red bounding box on the heatmap generated from our classification and localization module. The results show that our medical interpretation system is capable of diagnosing

| Sample Case | Annotation | Manual Report | Automatic Report |
|---|--------------|---|--|
|  | Normal | the heart size and cardiomedastinal silhouette are within normal limits. pulmonary vasculature appears normal. There is no focal air space consolidation.no pleural effusion or pneumothorax. | the cardiomedastinal silhouette and pulmonary vasculature are within normal limits in size. the lungs are clear of focal airspace disease pneumothorax or pleural effusion. there are no acute bony findings. no acute cardiopulmonary findings. |
|  | Cardiomegaly | mild cardiomegaly. mild unfolding of the thoracic aorta. no focal air space opacity. no pleural effusion or pneumothorax. visualized osseous structures are unremarkable in appearance. otherwise no acute cardiopulmonary abnormalities. | mild cardiomegaly. there is no focal consolidation. no pleural effusion or pneumothorax. there is no focal air space consolidation. no pleural effusion or pneumothorax. degenerative changes of the thoracic spine. no acute cardiopulmonary abnormality. |

Figure 4: Illustration of two cases of example outputs of our system.

thoracic diseases, highlighting the key findings in X-rays with heatmaps and generating well-structured reports.

2.3 On-Device COVID-19 Patient Triage and Follow-up

2.3.1 Background and Related Work

Due to its flu-like symptoms and potentially serious outcomes, a dramatic increase of suspected COVID-19 cases are expected to overwhelm the healthcare system during the flu season. Health systems still largely allocate facilities and resources such as Emergency Department (ED) and Intensive Care Unit (ICU) on a reactive manner facing significant labor and economic restrictions. To optimize resource utilization and clinical workflow, a rapid, automated, and point-of-care COVID-19 patient management technology that can triage (COVID-19 case screening) and follow up (radiological trajectory prediction) patients is urgently needed.

CXR, though less accurate than a PCR diagnostic, chest Computed Tomography (CT) or serological test, became an attractive option for patient management due to its impressive portability, availability and scalability [146]. At present, the bottleneck lies in the

shortage of board certified radiologists who are capable of identifying massive COVID-19 positive cases to reduce wait time at ED and determining the radiological trajectory of the COVID-19 patients after admission. The intensive development of deep neural network (DNN) powered CXR image analysis has seen the unprecedented success in automatic classification and segmentation of lung diseases [141]. Using the cloud solutions such as Google Cloud Platform or on-premise computing clusters to train a sophisticated DNN (e.g., DenseNet-121 [55]) with dozens of millions of parameters and hundreds of layers via billions of operations for both training and inference, these large scale Artificial Intelligence (AI) models achieve amazing performance that even outperforms board certified radiologists in some well-defined tasks [108].

With the increasing number of smart devices and improved hardware, there is a growing interest to deploy machine learning models on the device to minimize latency and maximize the protection of privacy. However, up to date on-device medical imaging applications are very limited to basic functions, such as the DICOM image view. In the COVID-19 environment, a mobile AI approach is expected not only to protect patient privacy, but also to provide a rapid, effective and efficient assessment of COVID-19 patients without the immediate need for an on-site radiologist. However, a major challenge that prevents wide adoption of the mobile AI approach is lack of lightweight yet accurate and robust neural networks.

Adequate knowledge has been accumulated from training the large scale DNN systems to accurately discern the subtle difference among the different lung diseases by learning the discriminative CXR imaging features [108]. Leveraging these results, we design and implement a novel three-player knowledge transfer and distillation (KTD) framework

composed of an Attending Physician (AP) network, a Resident Fellow (RF) network, and a Medical Student (MS) network for on-device COVID-19 patient triage and follow-up. In a nutshell, we pre-train a full AP network using a large scale of lung disease CXR images [141, 108], followed by fine-tuning a RF network via knowledge transfer using labeled COVID-19, pneumonia and normal CXR images, then we train a lightweight MS network for on-device COVID-19 patient triage and follow-up via knowledge distillation. After the KTD framework, the lightweight MS network is able to produce expressive features to identify COVID-19 cases as well as predict the radiological trajectory. The unique features of the KTD framework are knowledge transfer from large-scale existing lung disease images to enhance expressiveness of learned representation and novel loss functions to increase robustness of knowledge distillation to the MS network.

To the best of our knowledge, currently, there is no mobile AI system for on-device COVID-19 patient triage and follow-up using CXR images. In this work, we present an AI-powered system, COVID-MobileXpert, to triage and follow up COVID-19 patients using portable X-rays at the patient’s location. At the ED, COVID-MobileXpert calculates COVID-19 probabilistic risk to assist automated triage of COVID-19 patients. At the ICU or general ward (GW), it uses a series of longitudinal CXR images to determine whether there is an impending deterioration in the health condition of COVID-19 patients. Therefore, COVID-MobileXpert is essential to fully realize the potential of CXR to exert both immediate and long-term positive impacts on US healthcare systems. It enjoys the following advantages: 1) accurately detecting positive COVID-19 cases particularly from closely related pneumonia cases; 2) continuously following up admitted patients via radiological trajectory prediction.

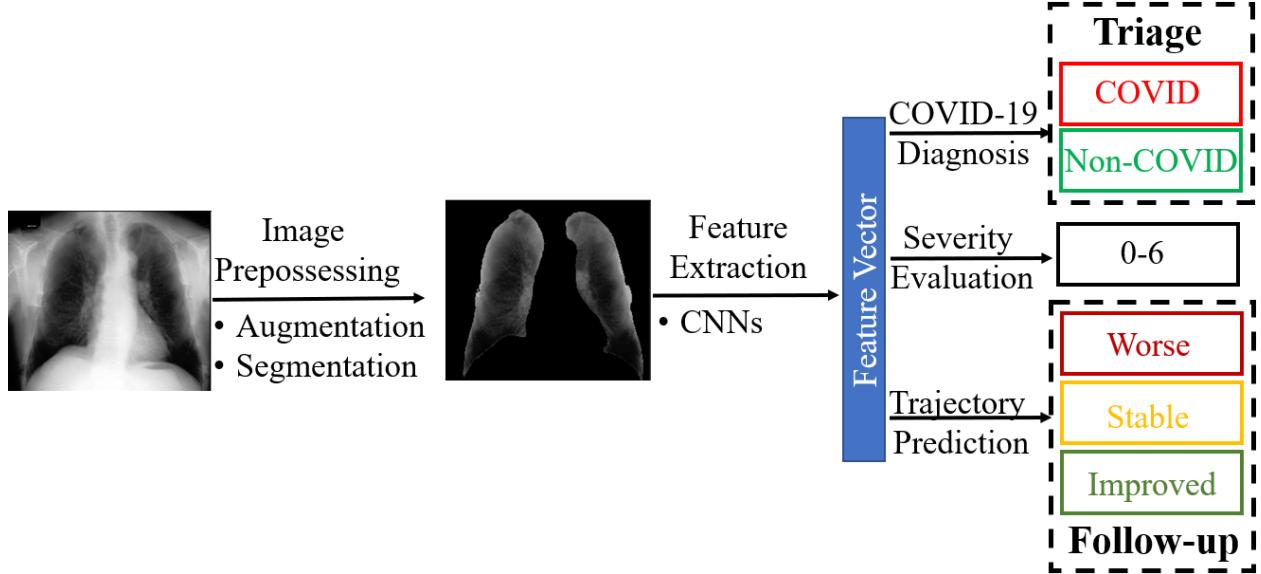


Figure 5: The analytical workflow of COVID-19 CXR interpretation systems.

COVID-19 CXR Interpretation During the COVID-19 pandemic in the past few months, convolutional neural networks (CNNs) have been successfully employed to assist with COVID-19 CXR interpretation. Since mid-February 2020, we collected over 40 works (most of them in a pre-print format) focusing on this subject. Although they may focus on different specific tasks, as shown in Fig. 5, they follow a similar pipeline: image prepossessing, feature extraction, and interpretation.

Data augmentation and segmentation are widely used as part of these approaches to avoid overfitting. In addition to the basic transformation-based method which includes rotating, flipping, scaling used in [19], Khalifa et al. [61] applied Generative Adversarial Network to generate virtual samples for data augmentation. To reduce the interference caused by unrelated area, Yeh et al.[152], Lv et al.[84] and Signoroni et al. [124] applied a U-net based method to perform a fast lung segmentation and preserved the region of interest (ROI) only. After image prepossessing, discriminative patterns for COVID-19 such

as ground-glass opacification-opacity (GGO) are then extracted by CNNs. Most current studies have directly borrowed or adopted well-known architectures such as ResNet [19, 44], InceptionV3 [19], DenseNet [152, 84, 20], and VGG [44, 159].

After feature extraction, three major tasks have been performed: diagnosis, severity evaluation, and trajectory prediction. COVID-19 diagnosis is usually considered to be a classification problem. The most straightforward way of detecting COVID-19 is to train a classifier with cross-entropy loss which is applied within most of these approaches. Other than these straightforward approaches, Zhang et al. [156] employed an unsupervised anomaly detection approach that detects COVID-19 cases as outliers. Moreover, Hassanien et al. [46] developed a classifier based on a support vector machine.

As for severity evaluation, Cohen et al. [20] and Zhu et al. [159] directly predicted lung disease severity scores using a linear regression model based on extracted features. In order to associate each score with a confidence value, Signoroni et al. [124] treated this task as a joint multi-class classification and regression problem using a compound loss function. Based on the severity assessment, the trajectory prediction can be achieved by calculating the difference in severity score between two adjacent CXR images. Other than basic score level interpretation, Duchesne et al. [34] built their trajectory prediction model based on feature level. When the feature from a single CXR was extracted by DenseNet-121, they used logistic regression to classify the trajectory into one of three categories: “Worse”, “Stable”, or “Improved”. However, the feature from a single CXR may not be sufficient to predict radiological trajectory. In order to tackle these challenges, our model forecasts radiological trajectory using feature extracted from a series of longitudinal CXR images of a single patient. By incorporating longitudinal CXR images into our model, novel

imaging features of progressive disease, including subtle changes of radiological features that are invisible to the human eye, can be detected.

On-device AI Model Currently, most AI models trained for COVID-19 interpretation are full DNNs that are not suitable to deploy on resource-constrained mobile devices. As there is no existing on-device medical image interpretation research, the vast majority of the existing work focuses on comparing the performance of different lightweight neural networks such as MobileNetV2 [116], SqueezeNet [56], Condense-Net [53], ShuffleNetV2 [157], MnasNet [132] and MobileNetV3 [52] using small benchmark natural image datasets such as CIFAR 10/100. MnasNet and MobileNetV3 are representative models generated via automatic neural architecture search (NAS) whereas all other networks are manually designed [35]. Due to the practical hardware resource constraint of mobile devices, natural image classification and segmentation performance have been compared based on accuracy, energy consumption, runtime, and memory complexity that no single network has demonstrated superior performance in all tasks [30]. Besides tailor-made network architectures for mobile devices, compression of the full DNN at the different stages of training also stands as a promising alternative. For in-training model compression, for example, Chen et al. [18] designed a novel convolution operation via factorizing the mixed feature maps by their frequencies to store and process feature maps that vary spatially slower at a lower spatial resolution to reduce both memory and computation cost of the image classification. Post-training or fine-tuning model compression techniques such as quantization [109] and/or pruning techniques [45] are often used to reduce the model size at the expense of reduced prediction accuracy. Wang et al. [140] demonstrated using 8-

bit floating-point numbers for representing weight parameters without compromising the model’s accuracy. Lou et al. [82] automatically searched a suitable precision for each weight kernel and chose another precision for each activation layer and demonstrated a reduced inference latency and energy consumption while achieving the same inference accuracy. Tung and Mori [135] combined network pruning and weight quantization in a single learning framework to compress several DNNs without sacrificing accuracy.

In order to improve the performance of the lightweight on-device models, knowledge distillation [51] is also used where a full teacher model is trained on the cloud or an on-premise GPU cluster, and a student model is trained at the mobile device with the ‘knowledge’ distilled via the soft labels from the teacher model. Thus the student model is trained to mimic the outputs of the teacher model as well as to minimize the cross-entropy loss between the true labels and predictive probabilities (soft labels). Knowledge distillation yields compact student models that outperform the compact models trained from scratch without a teacher model [107]. Goldblum et al. [41] attempted to encourage the student network to output correct labels using the training cases crafted with a moderate adversarial attack budget to demonstrate the robustness of knowledge distillation methods. Unlike the natural images, on-device classification of medical images remain largely an uncharted territory due to the following unique challenges: 1) label scarcity in medical images significantly limits the generalizability of the machine learning system; 2) vastly similar and dominant fore- and background in medical images make it hard samples for learning the discriminating features between different disease classes. To tackle these unique challenges we propose a novel three-player framework for training a lightweight network towards accurate and hardware friendly on-device COVID-19 patient triage and

follow-up.

2.3.2 Methods

Model Architecture We employ DenseNet-121 architecture as the template to pre-train and fine-tune the AP and RF networks. In addition, among well-studied lightweight CNNs [30], we select the most well-applied network MobileNetV2, and the most lightweighted network SqueezeNet as the candidate MS networks for on-device COVID-19 case screening and radiological trajectory prediction. Table 3 summarizes the key model complexity parameters [30]. Fig. 6 illustrates the three-player KTD training framework where the knowledge of abnormal CXR images is transferred from AP network to RF network and knowledge of discriminating COVID-19, non-COVID-19, and pneumonia is distilled from the RF network to the MS network.

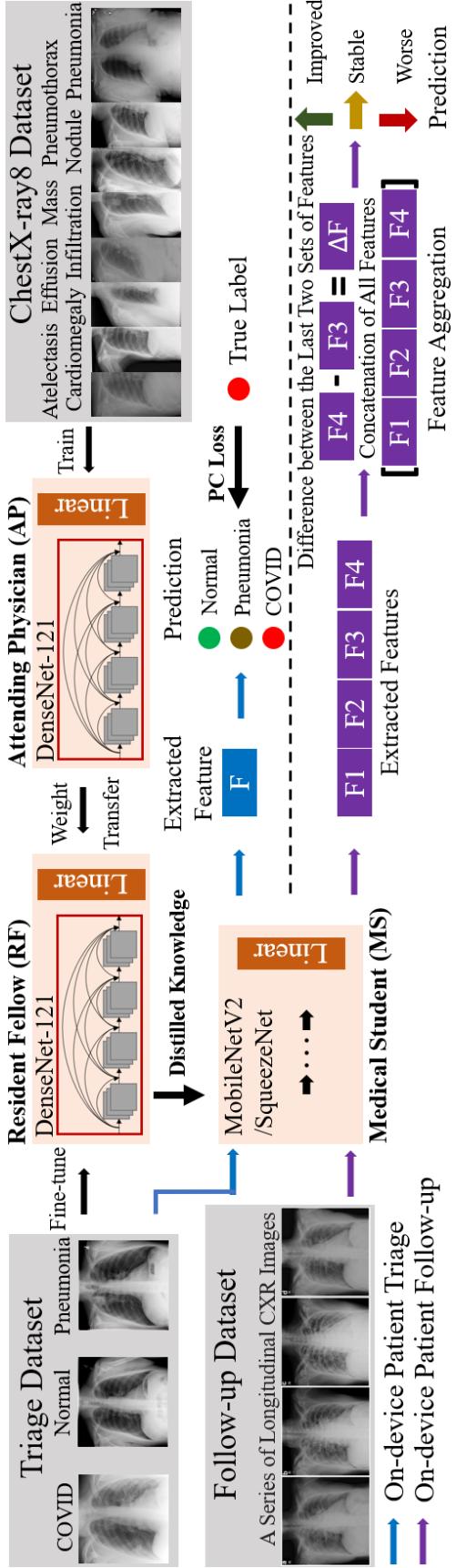


Figure 6: Overview of the three-player KTD training architecture demonstrating the knowledge transfer from AP to RF and the knowledge distillation from RF to MS. The blue and purple arrows demonstrate the training for two tasks: patient triage and follow-up respectively.

| Metric | DenseNet-121 | MobileNetV2 | SqueezeNet |
|------------------|--------------|-------------|------------|
| # of CONV layers | 120 | 20 | 22 |
| Total weights | 7.9M | 3.47M | 0.72M |
| Total MACs | 2900M | 300M | 282M |

Table 3: Comparison of full and compact DNN model complexity.

The KTD Training Scheme We pre-train the AP network as the *source task*, i.e., lung disease classification, and fine-tune, validate and test the RF network as the *destination task*. Different from recent studies [139] that pre-train the models with natural image datasets such as ImageNet, we pre-train the DenseNet-121 based AP network using the more related ChestX-ray8 dataset [141] of 108,948 lung disease cases to extract the CXR imaging features of lung diseases instead of generic natural imaging features. Specifically, beyond the dense block, we employ a shared fully connected layer for extracting the general CXR imaging feature and 8 fully connected disease-specific layers (including pneumonia as one disease layer) to extract disease-specific features (Fig. 6). Following the pre-training using large ChestX-ray8 dataset, the weights defining the general CXR imaging feature and the pneumonia disease feature are transferred to fine-tune the DenseNet-121 based RF network using a smaller compiled dataset of 3 classes of CXR images, i.e., COVID-19, normal and pneumonia. The latter is randomly initialized using two sets of weight parameters corresponding to normal and COVID-19 classes with the initial values of other weight parameters transferred from the pre-trained source model. The RF network is then used to train the lightweight MS network, e.g., MobileNetV2, or SqueezeNet, via knowledge distillation.

As shown in the MS section in Fig. 6, after knowledge distillation, the trained MS network can triage patients by screening COVID-19 cases following the blue arrow. Then

a radiological trajectory prediction model is further developed based on the trained MS network. Following the purple arrow, given a series of longitudinal CXR images from one patient, all images are fed into the pre-trained MS network for extracting disease-specific features. These features are then aggregated using different schemes before prediction. Here we investigate two different schemes: 1) calculating the difference between the last two CXR images' features; 2) chronologically concatenating all features. After feature aggregation, two fully connected layers are randomly initialized and trained with softmax loss function for the trajectory prediction.

Loss Functions As stated above, a unique challenge in medical imaging classification is the so-called “hard sample problem” [73], i.e., a subtle difference on the ROI across the images with a large amount of shared fore- and backgrounds. Motivated by this, we use an in-house developed loss function [72], i.e., PC loss with technical details given in the Chapter 4, for training the MS model and compared with ArcFace [29], the additive angular margin loss for deep face recognition, using the classical softmax loss as the baseline. Both PC and ArcFace losses are designed for improving classification performance on hard samples.

2.3.3 Experiments

In this section, we design and conduct extensive experiments to evaluate the performance of the compact MS network in patient triage and follow-up. In order to gain a holistic view of the model behavior, we investigate the performance concerning multiple choices of loss functions and values of tuning parameters for COVID-19 case screening as well as various choices of feature aggregation schemes and classifiers for radiological

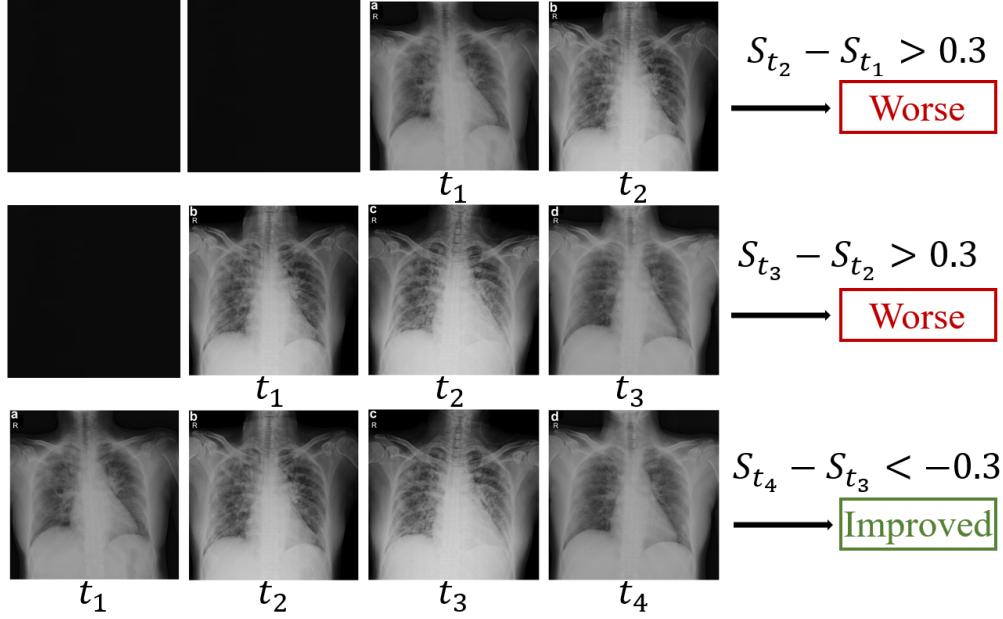


Figure 7: An example of data preparation for a series of longitudinal CXR images with radiological trajectory labels. The patient is in critical condition on t_3 then recovered afterward.

trajectory prediction.

The CXR image dataset for COVID-19 patient triage is composed of 179 CXR images from normal class [96], 179 from pneumonia class [96] and 179 from COVID-19 class containing both PA (posterior anterior) and AP (anterior posterior) positions [22] and we split it into training/validation/test sets with 125/18/36 cases (7:1:2) in each class. Since some patients have multiple CXR images in COVID-19 class, we sample images per patient for each split to avoid images from the same patient being included in both training and test sets.

Datasets For the radiological trajectory dataset, we assign a opacity score S for each COVID-19 positive CXR image in [22] using the scoring system provided by [20]. Fig. 7 shows an example of how we generate CXR image sequences and assign corresponding

radiological trajectory labels (i.e., “Worse”, “Stable”, “Improved”). Given a COVID-19 patient’s CXR images over four time points (the maximum length is set to four time points), we can create three CXR image sequences with zero-padding. For each sequence, we calculate the difference in the score of the last two CXR images. If the difference is larger than 0.3 the sequence is categorized as “Worse”, if the difference is less than -0.3 , it is labeled as “Improved”, otherwise, the category is “Stable”. We collect a total of 159 CXR image sequences from 100 patients in [22] and the dataset contains 76 “Worse” samples, 38 “Stable” samples, and 45 “Improved” samples. Similarly, we split it into training/validation/test sets with 111/16/32 samples (7:1:2).

Implementation Details We implement our model using PyTorch. The network is trained with the Adam optimizer for 50 epochs with a mini-batch size of 32 (triage task) and 10 (follow-up task). The parameter values that give rise to the best performance on the validation dataset are used for test. Similar to [34], when training the radiological trajectory prediction model, we employ the pre-trained MS network as a feature extractor (fixed weights). To overcome the overfitting problem, we also apply a dropout regularization with a rate of 0.5.

Tunning Parameters ξ : in the PC loss formula, a large value encourages the probabilistic intra-class compactness. α : in knowledge distillation framework [51, 41] (Eq. 2.1),

$$\min_{\theta} \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}} [\alpha t^2 \text{KL} (S_{\theta}^t(\mathbf{x}), T^t(\mathbf{x})) + (1 - \alpha)L(S_{\theta}^t(\mathbf{x}), y)], \quad (2.1)$$

| MobileNetV2/SqueezeNet (T=5) | | | | |
|---|---------------------|---------------------|-------------|-------------|
| α | PC($\xi=0.8$) | PC($\xi=0.995$) | ArcFace | SM |
| 0.2 | 0.870/0.798 | 0.833/0.777 | 0.870/0.750 | 0.861/0.777 |
| 0.4 | 0.880 /0.777 | 0.870/0.815 | 0.861/0.796 | 0.833/0.759 |
| 0.6 | 0.851/0.796 | 0.851/0.787 | 0.851/0.805 | 0.861/0.796 |
| 0.8 | 0.880 /0.824 | 0.870/0.796 | 0.851/0.796 | 0.833/0.787 |
| MobileNetV2/SqueezeNet ($\alpha = 0.8$) | | | | |
| T | PC($\xi=0.8$) | PC($\xi=0.995$) | ArcFace | SM |
| 1 | 0.851/0.750 | 0.880 /0.814 | 0.870/0.796 | 0.870/0.796 |
| 5 | 0.880 /0.824 | 0.870/0.796 | 0.851/0.796 | 0.833/0.787 |
| 10 | 0.880 /0.796 | 0.842/0.750 | 0.861/0.787 | 0.870/0.824 |

Table 4: Classification performance of MS networks, The values in ./ indicate MobileNetV2 vs. SqueezeNet.

it regularizes the ‘strength’ of knowledge distillation by specifying the relative contributions of the distillation loss, i.e., $\text{KL}(S_\theta^t(\mathbf{x}), T^t(\mathbf{x}))$, measuring how well the MS model mimics the RF model’s behavior using KL divergence and the classification loss of the MS model, i.e., $L(S_\theta^t(\mathbf{x}), y)$. $S_\theta(\cdot)$ and $T(\cdot)$ represent the RF model and MS model, respectively. The larger value, the stronger knowledge distillation is enforced from the RF model to the MS model.

T : in Eq. 2.1, it represents temperature where $T = 1$ corresponds to the standard softmax loss. As the value of T increases, the probability distribution generated by the softmax loss becomes softer, providing more information regarding which classes the RF model found more similar to the predicted class.

Evaluation of COVID-19 Patient Triage Performance We first report the classification accuracy to select the best MS model under different values of hyperparameters, followed by systematic evaluation of the model’s discriminating power of COVID-19 from non-COVID pneumonia and normal cases using AUROC values. With the knowledge transfer from the AP network pre-trained with a large set of abnormal lung disease cases, the RF

network demonstrates a remarkably high accuracy of 0.935 in the classification of CXR images.

Distilling knowledge from the RF network to the lightweight MS network, we observe an impressive performance that a vast majority of accuracy values are well above 0.850 for CXR image classification. Table 4 shows the classification accuracy results of both MobileNetV2 and SqueezeNet architectures with different loss functions and values of tuning parameters. It is clear that the knowledge distillation is essential to train the lightweight MS network without compromising much accuracy since the MS network alone, without knowledge distillation, achieves a baseline classification accuracy of 0.843 (MobileNetV2) and 0.732 (SqueezeNet), which are lower than those with knowledge distillation shown in Table 4.

Looking at Table 4 in more detail, we note that the performance of MobileNetV2 and SqueezeNet are not sensitive to the choice of temperatures (T) and strengths of distillation (α), however, it is very sensitive to the choice of loss functions. Overall, the PC loss developed in-house that flattens other probable class predictions perform the best across diverse settings of the tuning parameters, indicating the quality of knowledge distilled from the RF network to the MS network plays a pivotal role in training the lightweight MS network to ensure an accurate on-device COVID-19 patient triage.

In order to systematically evaluate the performance of the MS networks under the different decision thresholds, we use the AUROC value to assess how well the model is capable of discriminating COVID-19 cases from normal cases, pneumonia cases as well as normal plus pneumonia cases. In Fig. 8, both compact MS networks, i.e., MobileNetV2 and SqueezeNet, demonstrate a remarkable performance on all discrimination tasks that

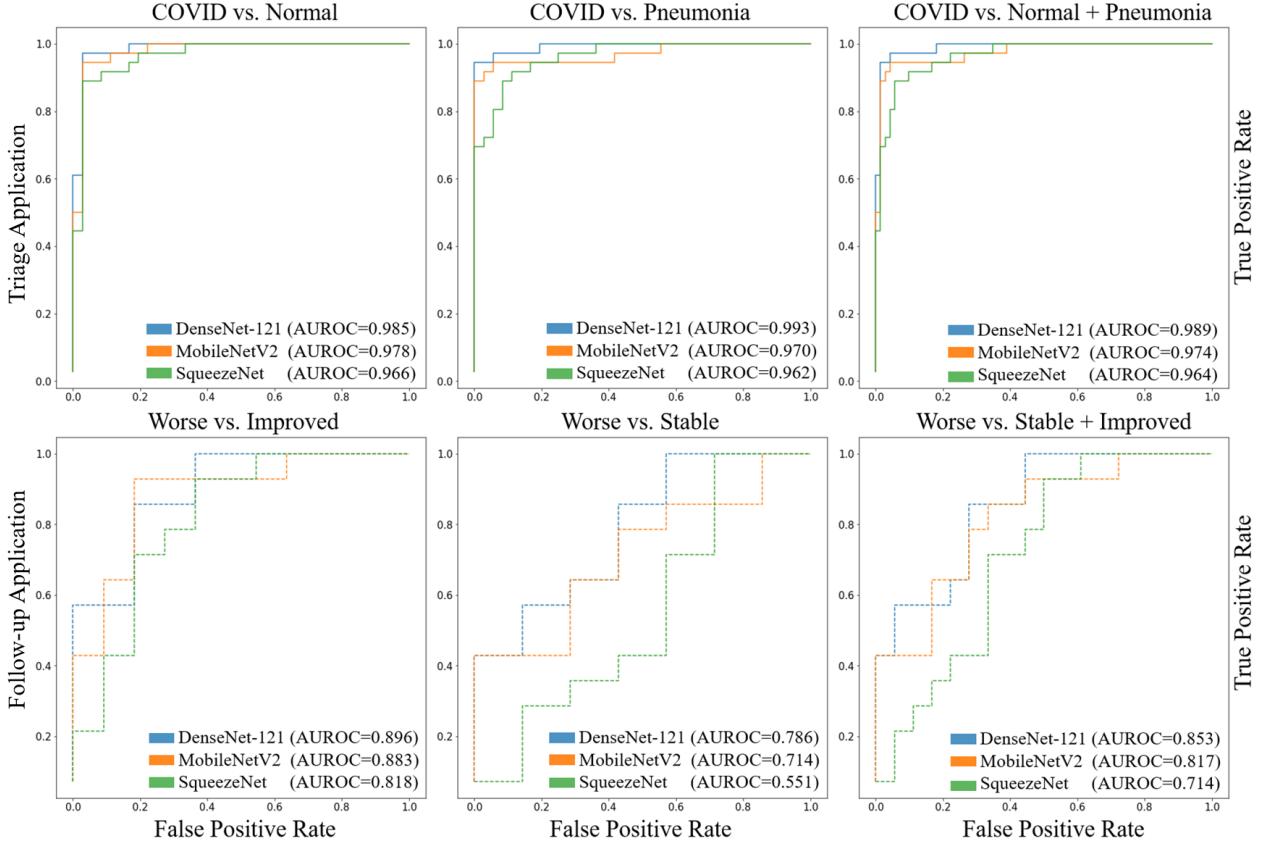


Figure 8: The upper panel shows the performance of the large-scale RF network and two compact MS networks of discriminating (a) COVID-19 vs. Normal cases; (b) COVID-19 vs. Pneumonia cases and (c) COVID-19 vs. Normal + Pneumonia cases, while the lower panel shows the performance of discriminating (d) “Worse” vs. “Improved” cases; (e) “Worse” vs. “Stable” cases and (f) “Worse” vs. “Improved” + “Stable” cases.

are comparable to that of the large scale cloud-based RF network, i.e., DenseNet-121. Importantly, both MobileNetV2 and SqueezeNet achieve high AUROC values of 0.970 and 0.964 when discriminating COVID-19 cases against mixed pneumonia and normal cases demonstrating strong potential for on-device triage using CXR images.

Evaluation of COVID-19 Patient Follow-up Performance Similar to [34], we first report the classification accuracy of discriminating “Worse” versus “Improved” cases to select the best combination of classifiers and feature aggregation schemes for on-device radiolog-

| MobileNetV2/SqueezeNet (DenseNet-121) | | |
|---------------------------------------|---------------------|----------------------------|
| Classifiers | Difference | Concatenation |
| Logistic Regression | 0.560/0.640 (0.720) | 0.760/0.720 (0.800) |
| Gradient Boosting | 0.680/0.640 (0.680) | 0.680/0.680 (0.680) |
| Random Forest | 0.680/0.600 (0.680) | 0.720/0.680 (0.720) |
| Our FC-classifier | 0.720/0.680 (0.720) | 0.800/0.760 (0.800) |

Table 5: Performance comparison of two feature aggregation schemes (Difference vs. Concatenation) with four different classifiers using two MS Networks (MobileNetV2 and SqueezeNet) as the feature extractor. Values in parentheses indicate the upper bound of accuracy yielded by RF Network (DenseNet-121).

ical trajectory prediction, followed by systematic evaluation of the model’s discriminating power of “Worse” cases from “Improved” and “Stable” cases using AUROC values. Based on the features extracted from the MS networks, four classifiers are trained for radiological trajectory prediction: 1) logistic regression; 2) gradient boosting; 3) random forest and 4) MS networks followed by fully connected layers (our FC-classifier).

As shown in Table 5, we observe the classifiers trained based on the feature extracted from both compact MS networks, i.e., MobileNetV2 and SqueezeNet, achieve a very similar level of performance to those trained with the feature from large scale RF network i.e., DenseNet-121. This again demonstrates that KTD training architecture with PC loss performs a high-quality knowledge distillation from RF network to lightweight MS networks.

When doing comparison between the feature aggregation schemes, we can see a significant improvement from using a series of longitudinal features over using only the difference between the last two sets of features. As for classifier selection, compared with the conventional classifiers i.e., logistic regression, gradient boosting and random forest, our FC-classifier is able to learn a series of subtle changes related to radiological features from CXR images, thus achieving a better performance. As a result, the best on-device per-

formance is obtained by our FC-classifier with feature concatenation using MobileNetV2, which attains the upper bound of accuracy (0.800) yielded by DenseNet-121.

Duchesne et al. [34] also report a high accuracy (0.827) of predicting the “Worse” category based on the feature extracted from a single CXR with their highly imbalanced test dataset, which contains over 84.6% samples labeled as “Worse”. However, the reported accuracy is lower than a simple baseline: a dummy classifier that always predicts the most frequent label “Worse” would yield a higher accuracy of 0.846. To make a comparison, we reimplement their model [34] on our more balanced dataset and record a result of 0.600, which implies that using the feature from a single CXR may not be sufficient to predict radiological trajectory. On the other hand, by using feature concatenation from a series of longitudinal CXR images, our model demonstrates better and more reliable performance (0.800).

In order to systematically evaluate the performance of the MS networks under the different decision thresholds, we again use the AUROC value to assess how capable the model is in discriminating “Worse” cases from “Improved” cases and “Stable” cases. As shown in Fig. 8, MobileNetV2 shows a close performance compared to the RF network (DenseNet-121). It is important to note that MobileNetV2 networks can achieve a high AUROC value of 0.883 enabling it to identify “Worse” cases from “Improved” cases and show a significant potential of on-device follow-up using CXR images.

2.3.4 Performance Evaluation on Mobile Devices

For on-device COVID-19 patient triage and follow-up with resource constraints, resource consumption is also an important consideration for performance evaluation in addition to accuracy. In order to systematically assess the performance of our COVID-19

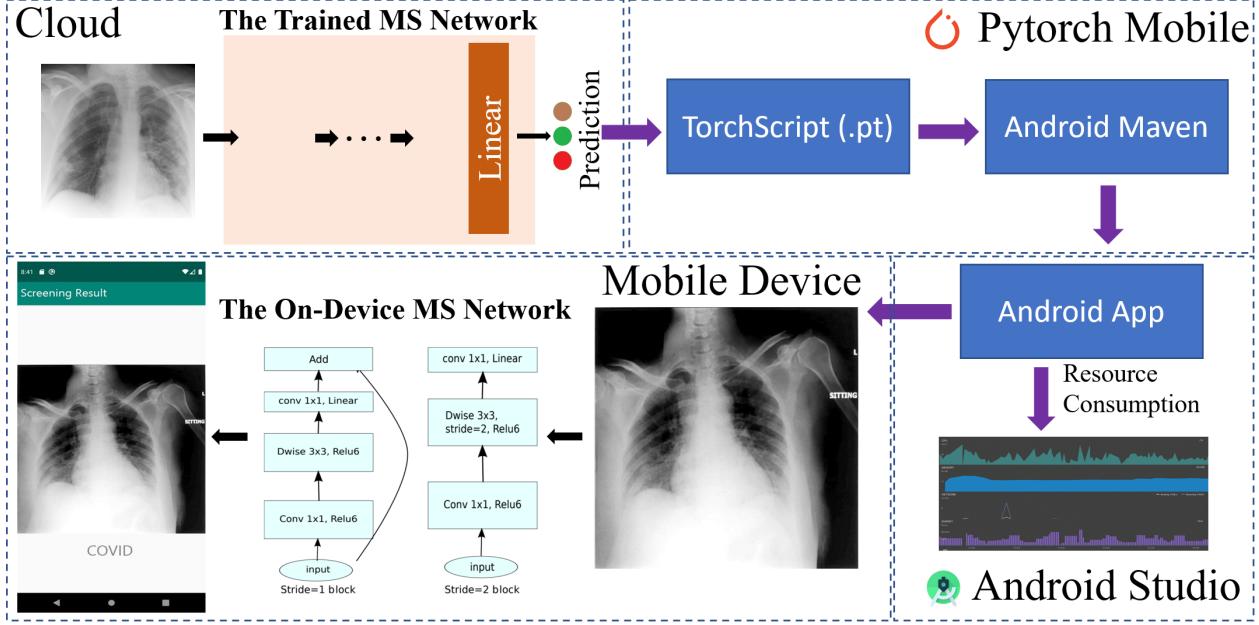


Figure 9: Overview of on-device deployment of the COVID-MobileXpert.

| The MS Network | Mobile Systems | MobileNetV2 | SqueezeNet | Mobile Systems | MobileNetV2 | SqueezeNet |
|----------------|----------------|-------------|------------|----------------|-------------|------------|
| CPU (%) | | 69.3 | 37.7 | | 67.7 | 32.7 |
| Memory (MB) | Nexus One | 69.4 | 67.5 | Nexus S | 88.8 | 64.4 |
| Energy | Heavy | Medium | | Heavy | Medium | |
| CPU (%) | | 67.2 | 29.0 | | 66.2 | 28.8 |
| Memory (MB) | Pixel | 70.5 | 29.0 | Pixel 2 | 69.4 | 70.1 |
| Energy | Heavy | Medium | | Heavy | Medium | |
| CPU (%) | | 68.7 | 26.7 | | 63.6 | 25.8 |
| Memory (MB) | Pixel 2 XL | 72.8 | 68.6 | Pixel 3 XL | 76.5 | 66.1 |
| Energy | Heavy | Medium | | Heavy | Medium | |

Table 6: Comparison of resource consumption of the two on-device MS networks deployed to the six Android based mobile devices.

on-device app, we select six mobile systems released following a chronic order, i.e., Nexus One / Nexus S (low-end); Pixel/ Pixel 2 (mid-range) and Pixel 2 XL/ Pixel 3 XL (high-end). Using the Pytorch Mobile framework, we deploy the three MS networks to the six Android based mobile systems and compare the resource consumption with regard to CPU, memory and energy usages. Fig. 9 describes a workflow to build an Android app based on the MS networks for on-device patient triage and follow-up.

In Table 6, it is clear that the MobileNetV2 based COVID-19 app is resource-hungry,

demonstrated by much higher resource consumption than SqueezeNet. Thus, the high accuracy achieved by MobileNetV2 is at the cost of high resource consumption. Within each app, we observe a downward trend in resource consumption following the chronic order, reflecting a continuous improvement of mobile device hardware. Overall, MobileNetV2 based COVID-19 apps are more suitable for high-performing mobile devices due to the high accuracy achieved with a higher resource consumption. On the other hand, SqueezeNet is more suitable for low-end mobile devices with both lower accuracy and resource consumption.

2.4 Uncertainty-aware Segmentation with Automatic Contour Outlier Mitigation

2.4.1 Background and Related Work

The segmentation of targets and normal tissues using manual approaches for radiotherapy treatment planning is resource-intensive, requiring significant time and effort, and can be prone to uncertainties due to inter- and intra-user variation [10]. To help improve efficiency and improve consistency, automatic segmentation approaches have traditionally focused on atlas-based methods and hybrid approaches [10]. Significant advances in computational technology have enabled implementation of machine learning approaches, such as neural networks and support vector machines [123], which have been implemented for use in automatic medical imaging systems [5]. More recently, applications of deep learning architectures for automatic segmentation of tumors and organs-at-risk (OAR's) in radiation oncology for the pelvis and other anatomic locations have become more commonplace¹ [5]. Many of these studies based on the U-Net deep learning architectures

have been shown to produce accurate results [113]. Liu et al. [79] demonstrating the potential to highly automate target and normal tissue segmentation. U-Net based deep learning methods can improve consistency in segmentation and reduce potential inaccuracies which could otherwise lead to miscomputation of the planning margins and resulting dose distributions to the targets and OAR's [17]. The availability of resource-intensive, on-line adaptive radiotherapy [59] coupled with increased utilization and efficacy of dose-escalated and hypofractionated strategies for prostate radiotherapy [33] further intensifies the need for highly automated auto-segmentation techniques to improve segmentation efficiency and accuracy. In this regard, deep learning algorithms offer a promising potential path forward.

Models incorporating uncertainty within the deep learning framework have been proposed [8, 63, 7, 6] and in the context of segmentation have centered on methods such as modeling the conditional probability distribution of the segmentations for a given input image [8]. Kohl et al. [63] proposed a generative model for lung cancer segmentation based on a combination of a U-Net with a conditional variational autoencoder (VAE). The VAE is able to generate an unlimited number of possible segments to aid in the most optimal choice of contour considering the different clinical tradeoffs [63]. In this work, we modified the general U-Net/VAE method of Kohl et al. [63] by developing an outlier mitigation strategy for prostate tumor auto-segmentation. The VAE enables output of multiple contour outputs on each CT slice, which facilitates flexibility and accuracy in the choice of the optimal contour by the user, and the mitigation of outliers is hypothesized to improve the network accuracy. Moreover, we trained the network using an image dataset, which includes inter-observer variability to generate contours robust to this uncertainty. Therefore,

the main contributions of this study are as follows: (a) We developed an outlier detection and mitigation (OM) technique in which outliers are detected, removed, and replaced in the training dataset; (b) We incorporated a training dataset consisting of prostate CT images contoured by 5 different physicians to account for inter-observer variation in the automatic segmentation process.

To our knowledge, this is the first application of uncertainty-aware deep learning architectures incorporating outlier mitigation and inter-observer variation in the training for prostate gland auto-segmentation in radiotherapy planning.

2.4.2 Methods

Image acquisition and contour For the primary source dataset (source_prim), image and contour data from 300 prostate cancer patients (consisting of over 19,000 2D-CT slices, with approximately 64 slices per dataset) were used for model training in this IRB-approved, retrospective study (HFH IRB# 12934). The source_prim dataset was contoured by a single radiation oncologist. Planning CT image datasets were acquired using a Philips Brilliance Big Bore CT scanner (120 kVp, 300 mAs, 3-mm slice thickness) (Philips Health-Care, Andover, MA). For the source_sec dataset, 10 prostate cancer patient datasets were used (consisting of 640 2D-CT slices, with approximately 64 slices per dataset), on which 5 radiation oncologist experts were previously asked to independently contour the prostate to assess segmentation variability between the observers. Other details about this dataset can be found in Gardner et al.[39].

Image pre-processing Images were re-sampled to a spatial resolution of $1.0 \times 1.0 \times 1.5 mm^2$. A $128 \times 128 \times 64$ voxel patch at the center of each image was cropped for training. A pixel-wise linear transformation was applied to assign HU values to intensity levels

between 0 and 255. Each 2D-CT slice was treated individually for training and test. To tackle the small sample size limitation, we applied data augmentation utilizing random rotation (<5 degrees), cropping and horizontal flipping to each image dataset to increase the sample size by factor of 100. Augmentation was applied to both the source_prim and source_sec training datasets.

Deep Learning Model Architecture A Deep Neural Network (DNN) architecture, comprised of a U-Net and a variational autoencoder (VAE) for automatic contouring of the prostate gland, was implemented for prostate gland segmentation. The Probabilistic Hierarchical U-Net with VAE first described by Baumgartner et al. [8] was used to model the variability on each resolution level of the U-Net to increase the diversity of the generated segmentation. Manual delineation of target and normal tissue segments on treatment planning images can be laborious and can result in some contours being inaccurate due to lack of consistency. Training a DNN with inaccurate contours can be challenging since outliers can be overfit with high model capacity [153]. Moreover, without appropriate intervention during the training, bias can be reinforced within the model. To tackle this challenge, we modified the VAE to incorporate an automatic contour outlier mitigation technique, termed outlier-mitigation VAE (OM-VAE). After the first round of training, low quality contours (as assessed using the Dice similarity coefficient) in the training dataset are replaced by the most accurate contours generated by the OM-VAE. The OM-VAE is then re-trained using the revised training dataset to improve the model accuracy.

Quantitative contour comparison To generate a ground truth contour, the 5 physician-generated contours on the source_sec dataset were previously combined into a consensus contour using a 3-dimensional statistical method3. Thus, the auto-segmentation contours

were compared to the consensus contour using several metrics to quantitatively compare the ground truth consensus segmentations with auto-segmentations using the deep learning model. We built in-house software to compute the evaluation metrics. These metrics included the Dice similarity coefficient (DSC), Hausdorff distance (HD) and Normalized Cross-Correlation (NCC). The DSC is used to assess the amount of overlap between 2 contours (s and y):

$$DSC = \frac{2 \times |s \cap y|}{|s| + |y|}, \quad (2.2)$$

where s is the independent sample form the predicted distribution p_s and y is the ground truth, where p_s is a binary prediction using 0.5 as the threshold. The HD is a measure of the gross error between the auto-segmentation and consensus contour, and is defined as the maximum distance between one point on a contour to the nearest point on another contour:

$$HD(s, y) = \max_{s \in s} \min_{y \in y} \|s - y\|_2. \quad (2.3)$$

To quantify the pixel-wise variability between auto-segmentation and ground truths, the average normalized cross correlation (NCC) of the cross entropy between the mean of the ground truth labels and that of the generated segments was utilized:

$$\mathcal{S}_{NCC}(p_{gt}, p_s) = \mathbb{E}_{y \sim p_{gt}} [\text{NCC}(\mathbb{E}_{s \sim p_s}[\text{CE}(s, s)], \mathbb{E}_{s \sim p_s}[\text{CE}(y, s)])]. \quad (2.4)$$

NCC is a commonly used metric for comparison of correspondence between image datasets. Additionally, we computed the center of mass (COM) distance as a measure of the displacement between auto-segments and manual contours. This metric has been used by others

as a comparative measure in the deep learning setting for auto-segmentation [79].

Probabilistic hierarchical U-Net with VAE To generate the multiple automatic segments for each CT image slice and capture the inter-observer variability from 5 different observers in the training dataset, we implemented a hierarchical neural network, which enhances a standard U-Net with a VAE as proposed by Kohl et al. [63]. As opposed to the original Probabilistic U-Net, which only includes one latent space for VAE and suffers from the limited diversity of output segmentations, this method decomposes the latent space in the VAE component of the deep neural network into several different scales (resolutions) in a hierarchical way. As a result, the variation on each resolution level is governed by separate latent variables, which helps increasing randomness and circumvents limited diversity in the output samples.

The model aims to approximate the posterior distribution $p(\mathbf{z} \mid \mathbf{s}, \mathbf{x})$ using a variational function, where \mathbf{x} is the input image, \mathbf{s} is the segmentation, and \mathbf{z} is the latent representation. The latent variable $\mathbf{z} = \{\mathbf{z}_1 \dots \mathbf{z}_L\}$ is modeled in a hierarchical way. With the different levels of latent variable \mathbf{z} , the posterior distribution of segmentation \mathbf{s} given an image \mathbf{x} can be written for the general case of L latent level as:

$$p(\mathbf{s} \mid \mathbf{x}) = \int_{\mathbf{z}_1, \dots, \mathbf{z}_L} p(\mathbf{s} \mid \mathbf{z}_1, \dots, \mathbf{z}_L) p(\mathbf{z}_1 \mid \mathbf{z}_2, \mathbf{x}) \cdots p(\mathbf{z}_{L-1} \mid \mathbf{z}_L, \mathbf{x}) p(\mathbf{z}_L \mid \mathbf{x}) d\mathbf{z}_1 \dots d\mathbf{z}_L. \quad (2.5)$$

The posterior distribution $p(\mathbf{z} \mid \mathbf{s}, \mathbf{x})$ can be approximated by a variational function $q(\mathbf{z} \mid \mathbf{s}, \mathbf{x})$ using variational inference. Minimizing the Kullback-Leibler (KL) divergence between

$p(\mathbf{z} \mid \mathbf{s}, \mathbf{x})$ and $q(\mathbf{z} \mid \mathbf{s}, \mathbf{x})$ results in the following lower-bound estimate of $\log p(\mathbf{s} \mid \mathbf{x})$):

$$\log p(\mathbf{s} \mid \mathbf{x}) = \mathcal{L}(\mathbf{s} \mid \mathbf{x}) + \text{KL}(q(\mathbf{z} \mid \mathbf{s}, \mathbf{x}) \| p(\mathbf{z} \mid \mathbf{s}, \mathbf{x})), \quad (2.6)$$

where \mathcal{L} is a lower-bound on $\log p(\mathbf{s} \mid \mathbf{x})$ with equality when the approximation q matches the posterior distribution exactly. The lower bound $\mathcal{L}(\mathbf{s} \mid \mathbf{x})$ can be written as:

$$\begin{aligned} \mathcal{L} &= \mathbb{E}_{q(\mathbf{z}_1, \dots, \mathbf{z}_L \mid \mathbf{x}, \mathbf{s})} [\log p(\mathbf{s} \mid \mathbf{z}_1, \dots, \mathbf{z}_L)] - \alpha_L \text{KL}[q(\mathbf{z}_L \mid \mathbf{s}, \mathbf{x}) \| p(\mathbf{z}_L \mid \mathbf{x})] \\ &\quad - \sum_{\ell=1}^{L-1} \alpha_\ell \mathbb{E}_{q(\mathbf{z}_{\ell+1} \mid \mathbf{s}, \mathbf{x})} [\text{KL}[q(\mathbf{z}_\ell \mid \mathbf{z}_{\ell+1}, \mathbf{s}, \mathbf{X}) \| p(\mathbf{z}_\ell \mid \mathbf{z}_{\ell+1}, \mathbf{X})]], \end{aligned} \quad (2.7)$$

where α 's are hyperparameters which we set to be 1 in our experiment. Following standard practice, the prior and posterior distributions are parametrized as axis aligned normal distributions $\mathcal{N}(\mathbf{z} \mid \mu, \sigma)$ as follows:

$$p(\mathbf{z}_\ell \mid \mathbf{z}_{\ell+1}, \mathbf{x}) = \mathcal{N}\left(\mathbf{z} \mid \phi_\ell^{(\mu)}(\mathbf{z}_{\ell+1}, \mathbf{x}), \phi_\ell^{(\sigma)}(\mathbf{z}_{\ell+1}, \mathbf{x})\right), \quad (2.8)$$

$$q(\mathbf{z}_\ell \mid \mathbf{z}_{\ell+1}, \mathbf{x}, \mathbf{s}) = \mathcal{N}\left(\mathbf{z} \mid \theta_\ell^{(\mu)}(\mathbf{z}_{\ell+1}, \mathbf{s}, \mathbf{x}), \theta_\ell^{(\sigma)}(\mathbf{z}_{\ell+1}, \mathbf{s}, \mathbf{x})\right), \quad (2.9)$$

where ϕ and θ are functions parameterized by the neural networks. The architecture is then trained by maximizing the lower bound in Eq. 2.7. Other details can be found in Baumgartner et al. [8].

Transfer learning with VAE The concept of transfer learning is based on the idea that knowledge from a large, labeled training dataset available from a “source” distribution can be transferred to train a much smaller dataset from a different distribution [137, 99]. The

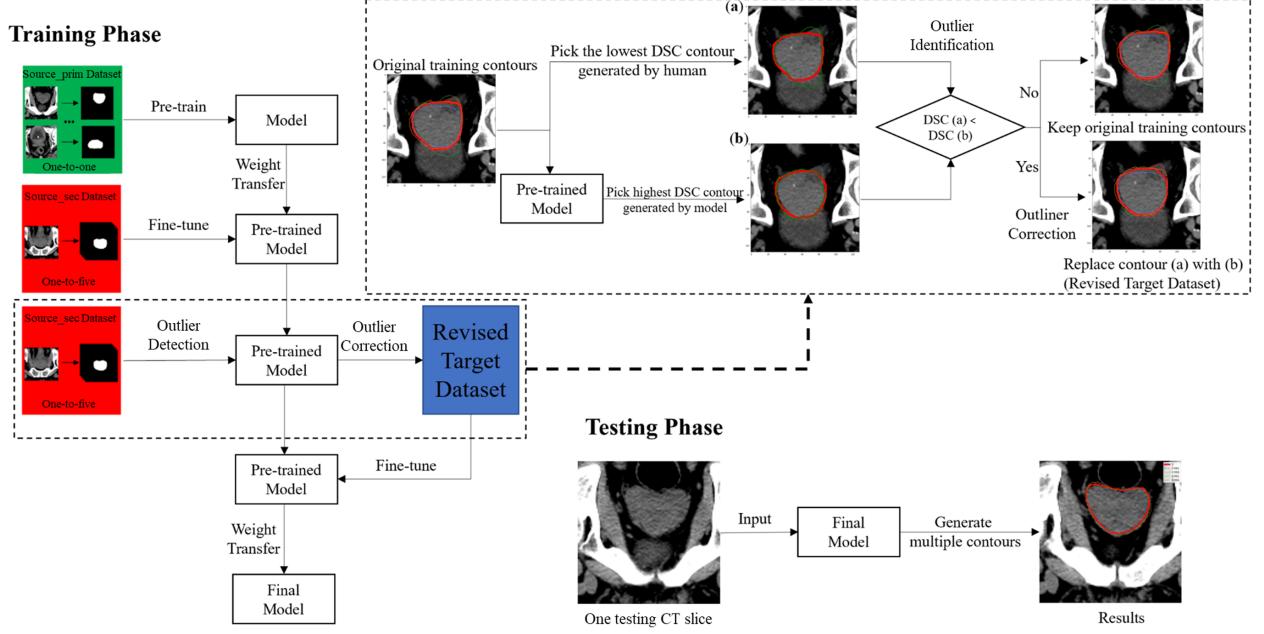


Figure 10: Overview of our training architecture demonstrating the transfer learning and automatic contour outlier mitigation.

goal of transfer learning then is for the smaller dataset to learn a classification method that benefits from already available data originating from one or more sources, corresponding to the much larger dataset. These 2 datasets may be similar but not need not be from the same source, which differentiates transfer learning from supervised learning, where it is assumed that the datasets originate from the same source [137, 99].

Transfer learning (pre-training followed fine tuning) was applied on the `source_sec` dataset following the methodology described in Fig. 10. The VAE model was pre-trained using a primary source (`source_prim`) dataset (consisting of 300 patients and 19,200, 2DCT slices). Weights were then transferred to the smaller, secondary source (`source_sec`) dataset (640 CT slices from 10 patient datasets) incorporating inter-observer variation to finetune the model using the transfer learning concept. Augmentation was applied to both `source_prim` and `source_sec` training datasets.

Automatic contour outlier mitigation We developed an automatic contour outlier detection and mitigation technique (OM-VAE), which identifies low accuracy contours and updates them in the training (source_sec) dataset, as illustrated in Fig. 10. The conceptual framework for the outlier detection and mitigation technique are as follows: The DSC is computed for the OM-VAE contours relative to the consensus contour and the highest value (DSC_{max}) is selected. For the same CT slice, DSC is also computed for 5 observer contours relative to the consensus contour, with the minimum DSC value of the 5 observer contours (DSC_{min}) used for comparison with the OM-VAE maximum DSC value (DSC_{max}). If $DSC_{max} > DSC_{min}$, the lower accuracy contour is replaced with the most accurate contour generated by the OM-VAE in the source_sec training dataset. This process of identifying the low accuracy contours and updating the training dataset was performed only once. Fig. 10 shows the updated training contours after the correction process used for re-training the OM-VAE.

Computational setup and training strategy We set the latent level L set to be 5 and the batch size set to be 24. To train the model, we used the Adam optimizer⁴⁶ with a learning rate of 10^{-5} and weight decay of 10^{-3} . Batch normalization was applied after each convolutional layer on non-output layer to accelerate the training process. The model was pre-trained for 30 epochs using the augmented source_prim dataset. The model parameters were then fine-tuned for 100 epochs using the augmented source_sec dataset. We evaluated the selected model on the test dataset to obtain quantitative results. Computations were performed on the source_sec dataset for 10 iterations, with the dataset randomly splitting into training, validation, and test dataset using a ratio of 6:1:3. More specifically, data splitting for training/validation/test was done first at the patient level before aug-

mentation. For instance, on the source_sec dataset, splitting was done such that 6 patients were randomly assigned for training, 1 for validation, and 3 for test. Augmentation was then applied to the training datasets. Therefore, the images in the training/validation/test datasets are from different patients, which circumvents information leaking. Outputs of the U-Net VAE/OM-VAE consist of 15 contours automatically generated on each CT slice. The value of 15 was selected arbitrarily.

2.4.3 Results

The main goal of Fig. 11 is to provide qualitative comparisons of the different models (U-Net/VAE and U-Net/OM-VAE trained with different source datasets) against ground-truth and consensus contours based on the expert physician observers. Fig. 11 shows examples of prostate segmentation for high, average and low-accuracy cases based on DSC, HD (mm) and NCC. Shown are contours from: 5 radiation oncologists, ground-truth (first column from left); U-Net/VAE trained with source_sec dataset only (second column); U-Net/VAE trained with source_prim followed by source_sec datasets using transfer learning (third column); U-Net/OM-VAE trained with source_prim and source_sec datasets and with outliers replaced (fourth column). For the U-Net/VAE/OM-VAE models (columns 2-4) DSC values are based on an average of 15 contours outputted by each model. Note that we have chosen to just demarcate the 5 contours with the highest accuracy to ease visibility. The consensus, ground-truth contours generated from the 5 radiation oncologists are mapped onto the model predictions (white dashed lines) for comparison. In the first column, DSC values were computed based on the average of the 5 physician contours relative to the consensus contour. For all cases, the DSC values increased for the U-Net/VAE trained using source_sec only versus source_prim followed by source_sec

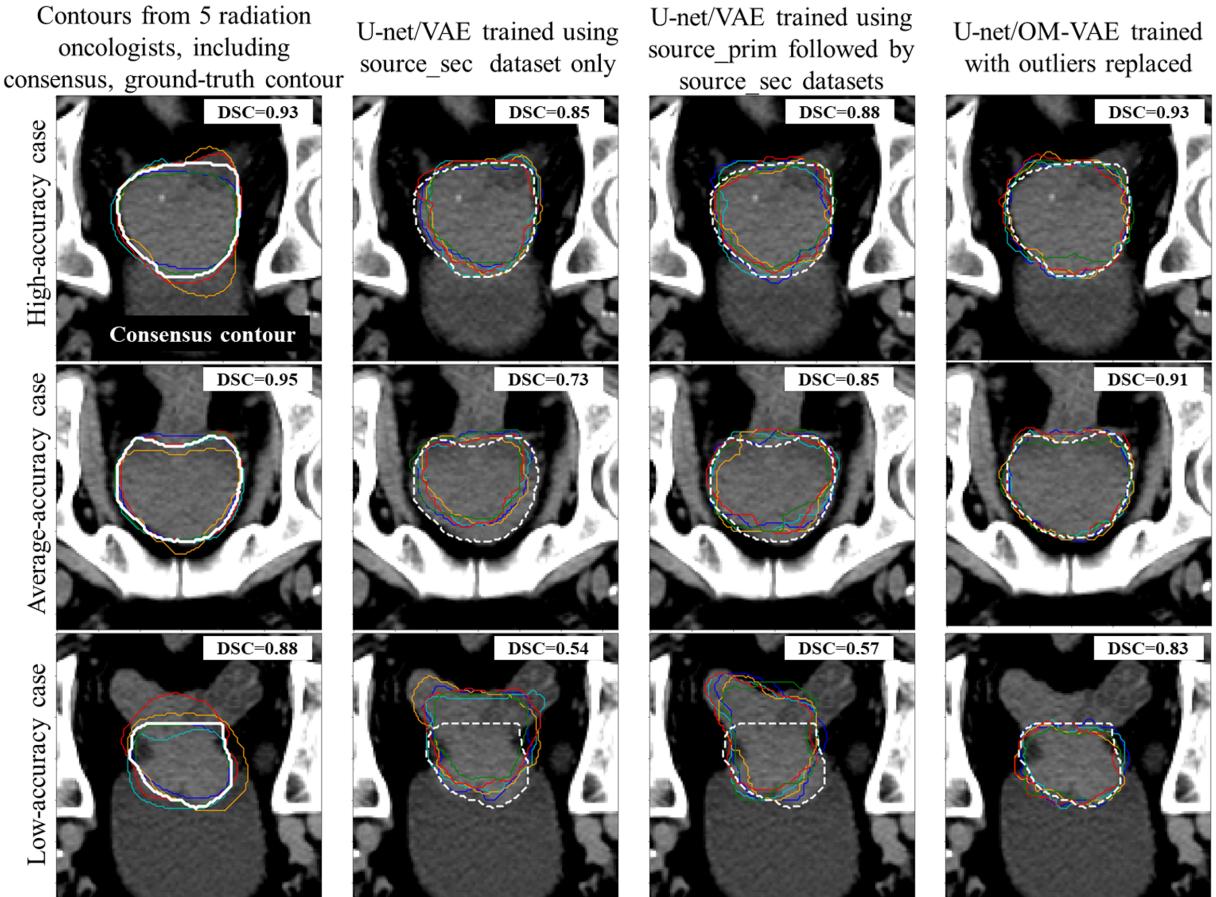


Figure 11: Comparison of U-Net/VAE, U-Net/OM-VAE and radiation oncologist-generated contours on three example segmentations for high (upper panel) average (middle panel) and low (lower panel) accuracy cases, respectively. The thick white lines represent the consensus, ground-truth contours while other lines are the generated contours by either radiation oncologist or the model. DSC scores in the columns 2-4 were computed based on 15 randomly selected contours, while the only the best 5 contours are plotted in the figure to ease viewability. For the first column, 5 physician contours are shown along with the consensus contour. DSC was computed based on an average of the 5 physician Dice scores relative to the consensus contour.

datasets, respectively, and further improvement was noted with the U-Net/VAE-OM, where outliers were replaced. For instance, for an ‘average case’, DSC increased from 0.73 to 0.85 for the U-Net/VAE trained using source_sec only versus source_prim followed by source_sec datasets, to 0.91 for the U-Net/OM-VAE with outliers replaced, respectively. Of note, for a low accuracy case, significant improvement was observed with U-Net/OM-VAE with outliers replaced ($DSC=0.83$) relative to the U-Net/VAE trained with source_prim followed by source_sec datasets ($DSC=0.57$). Note also that from a qualitative perspective, significant improvement is noted in the conformity of the contours between columns 2 (source_sec dataset) and 3 (source_prim followed by source_sec datasets) especially for the ‘average accuracy’ case. For the “low accuracy” case, both models show systematic deviations from the ground-truth though the model trained with source_prim followed by source_sec datasets has lower variability.

The primary goal of Table 7 is to compare results of the U-Net/VAE against those of the U-Net/OM-VAE to demonstrate the utility of the outlier mitigation strategy. Table 1 shows quantitative results of the U-Net VAE/OM-VAE trained with different training and test dataset iterations. For all scenarios, the U-Net/OM-VAE trained with source_sec followed by source_sec datasets and with outliers replaced shows the highest accuracy ($DSC=0.82 \pm 0.01$, $HD=9.18 \pm 1.22$ mm and $COM=3.36 \pm 0.81$ mm) over the average of 15 contours, or $DSC=0.90 \pm 0.02$, $HD=5.14 \pm 0.97$ mm and $COM=1.03 \pm 0.58$ mm if we consider the average of the most accurate contours among 15 contours. Results for the U-Net/OM-VAE with outliers removed over the average of 15 contours ($DSC=0.78 \pm 0.01$, $HD=10.65 \pm 1.95$ mm and $COM=4.17 \pm 0.79$ mm) were generally lower than that of the U-Net/VAE trained with the source_prim and source_sec datasets

(DSC=0.80 ± 0.02, HD=10.18+1.35 mm and COM=4.77 ± 0.96 mm), with the exception of the COM distance which was on average, 0.6 mm better. When comparing the highest accuracy among 15 contours, the U-Net/OM-VAE with outliers removed had DSC=0.88 ± 0.02, HD=7.00 ± 1.17 mm, and COM=1.58 ± 0.63 mm representing improvement over the U-Net/VAE trained with source_prim and source_sec datasets, where DSC=0.85 ± 0.02, HD=7.54 ± 1.36 mm, and COM=1.46 ± 0.68 mm, with the exception of the COM distance which was on average, 0.12 mm worse. For the U-Net/VAE trained with source_sec only versus source_prim and source_sec datasets, differences were statistically significant with p<0.01 (DSC), 0.02 (HD), <0.01 (NCC), and 0.03 (COM). Differences were also statistically significant between the U-Net/VAE and U-Net/OM-VAE with outliers replaced; p=0.02 (DSC), 0.04 (HD), <0.01 (COM) For comparison between U-Net/VAE and U-Net/OM-VAE with outliers removed, p=0.02 (DSC), 0.06 (HD), 0.01 (COM). NCC values show an increase between U-Net/VAE trained with source_sec only versus source_prim and source_sec datasets of 0.52 ± 0.10 vs. 0.62 ± 0.06, respectively with statistically significance reached, p<0.01. NCC decreased to 0.46 ± 0.10 with the U-Net/VAE-OM with outliers removed and increased to 0.59 ± 0.07 with outliers replaced in the training dataset. NCC differences between U-Net/VAE trained with both source_prim and source_sec datasets (0.62 ± 0.06) and U-Net/OM-VAE with outliers replaced (0.59 ± 0.07) were not statistically significant (p=0.12).

| Model | Training dataset | test dataset output | DSC | HD (mm) | NCC | COM (mm) |
|--------------|---|--|-------------|--------------|-------------|-------------|
| U-Net/VAE | source_sec only | average over 15 contours | 0.76 ± 0.03 | 11.48 ± 2.28 | 0.52 ± 0.10 | 5.48 ± 0.88 |
| | source_prim followed by source_sec | average over 15 contours | 0.80 ± 0.02 | 10.18 ± 1.35 | 0.62 ± 0.06 | 4.77 ± 0.96 |
| | source_prim followed by source_sec | average of highest DSC (among 15 contours) | 0.85 ± 0.02 | 7.54 ± 1.36 | N/A | 1.46 ± 0.68 |
| | source_prim followed by source_sec with outliers removed | average over 15 contours | 0.78 ± 0.01 | 10.65 ± 1.95 | 0.46 ± 0.10 | 4.17 ± 0.79 |
| | source_prim followed by source_sec with outliers removed | average of highest DSC (among 15 contours) | 0.88 ± 0.02 | 7.00 ± 1.17 | N/A | 1.58 ± 0.63 |
| | source_prim followed by source_sec with outliers replaced | average over 15 contours | 0.82 ± 0.01 | 9.18 ± 1.22 | 0.59 ± 0.07 | 3.36 ± 0.81 |
| U-Net/OM-VAE | source_prim followed by source_sec with outliers replaced | average of highest DSC (among 15 contours) | 0.90 ± 0.02 | 5.47 ± 0.97 | N/A | 1.03 ± 0.58 |

Table 7: Comparison of U-Net VAE/OM-VAE with different training and test datasets. Values are shown for DSC, HD (mm), NCC and COM distance (mm) computed for each model iteration against the consensus, human-contoured dataset.

| Model | Training dataset | DSC | HD (mm) |
|--------------|---|-----------------|------------------|
| U-Net | source_prim followed by source_sec | 0.72 ± 0.03 | 15.92 ± 2.28 |
| U-Net/VAE | source_prim followed by source_sec | 0.85 ± 0.02 | 7.54 ± 1.36 |
| U-Net/OM-VAE | source_prim followed by source_sec with outliers replaced | 0.90 ± 0.02 | 5.47 ± 0.97 |

Table 8: Comparison of U-Net, U-Net/VAE and U-Net/OM-VAE models. The U-Net/VAE or U-Net/OM-VAE considers the average of the highest accuracy (based on DSC) contour among 15 contours randomly outputted. The U-Net outputs only one contour on each image slice. DSC and HD (mm) were computed against the consensus, human-contoured dataset.

The main goal of Table 8 is to compare results of the U-Net against those of the U-Net/VAE with and without outlier mitigation to demonstrate utility of the VAE relative to the standard U-Net. Table 8 shows the related DSC and HD (mm) values. The U-Net/VAE or U-Net/OM-VAE considers the highest accuracy (based on DSC) contour among 15 contours randomly outputted. The U-Net outputs only one contour on each image slice. For the U-Net/OM-VAE model, outliers were replaced in the training dataset. DSC improves from 0.72 ± 0.03 (U-Net) to 0.85 ± 0.02 (U-Net/VAE) and HD decreases from 15.92 ± 2.28 mm (U-Net) to 7.54 ± 1.36 mm (U-Net/VAE), with $p < 0.01$. Further improvement was noted for the U-Net/OM-VAE with outliers replaced, DSC = 0.90 ± 0.02 and 5.47 ± 0.97 mm.

The primary goal of Table 9 is to demonstrate the utility of data augmentation. DSC, HD (mm) and NCC values (averaged over 15 contours and compared against the manual consensus contour) are shown for the U-Net/VAE with and without data augmentation. Data was augmented by a factor of 100. For the model trained using data augmentation (either source_sec or source_prim followed by source_sec via transfer learning) all metrics improved over those trained without data augmentation (Table 3). For instance, results

| Model | Training dataset | DSC | HD (mm) | NCC |
|------------------------------------|------------------------------------|------|---------|------|
| U-Net/VAE w/o data augmentation | Source_sec only | 0.67 | 19.32 | 0.46 |
| | Source_prim followed by Source_sec | 0.72 | 11.32 | 0.57 |
| U-Net/VAE w/ data augmentation | Source_sec only | 0.78 | 9.75 | 0.62 |
| | Source_prim followed by Source_sec | 0.80 | 8.97 | 0.73 |

Table 9: Comparison of U-Net/VAE with and without data augmentation. Data was augmented by a factor of 100. Data splitting in the ratio of 6:1:3 (training/validation/test) was done prior to augmentation. Values are shown for DSC, HD (mm) and NCC computed for an average of 15 contours against the consensus, human-contoured dataset.

for the model trained using the source_prim followed by source_sec were enhanced with data augmentation by 0.08 (DSC), -2.35 mm (HD) and 0.16 (NCC). Improvement with augmentation is most notable for the model trained with source_sec only (see Table 9, Rows 1 vs. Row 3) since the source_sec dataset suffers from label scarcity relative to training with the source_prim followed by source_sec datasets.

2.4.4 Discussion

Auto-contouring using artificial intelligence (AI) approaches has shown a significant potential to improve efficiency and consistency in the treatment planning setting [81]. The accuracy of the AI model prediction is based on the training datasets and the availability of data with sufficient variability. Using a modified U-Net/VAE we proposed to train the model using a dataset incorporating inter-observer variation such that the model outputs (auto-contours) would be robust to this uncertainty. The multiple contour outputs of the U-NET/VAE afford the ability for clinicians to select the best contour based on tradeoffs between target and normal tissue boundaries. We also developed an outlier mitigation strategy in which outlier contours from the initial output of the U-Net/VAE are replaced in the training dataset with more accurate estimates. We hypothesized that mitigation of

outliers would improve the model segmentation accuracy.

In Fig. 11, it is clear that the auto-contouring accuracy is increased using the outlier mitigation technique (U-Net/OM-VAE) relative to the U-Net/VAE including outliers. This was confirmed in Table 7 where the standard deviations for DSC and HD were reduced using U-Net/OM-VAE with outliers replaced relative to the values for the U-Net/VAE. For U-Net/OM-VAE model with outliers removed the DSC and HD values decreased relative to those of the U-Net/VAE trained with source_prim and source_sec datasets. This is likely related to the data scarcity problem – simply removing contours from the training datasets might significantly reduce variability and opportunity for the model to train on a range of data. Note, however, that for the U-Net/OM-VAE with outliers removed, the highest accuracy contours have better DSC, HD and COM values relative to the U-Net/VAE (Table 7) suggesting that removal of outliers does enhance the “best case” accuracy, despite the reduction in variability. From the low accuracy case (Fig. 11, third row) it is clear that the presence of outliers dominates the accuracy of the U-Net/VAE trained with either the source_sec or source_prim and source_sec datasets (Fig. 11, third row, second and third columns), and that the replacement of the outliers using the U-Net/OM-VAE mitigates this issue (Fig. 11, third row, fourth column).

The NCC values for the U-Net/OM-VAE (with outliers replaced) decreased relative to that of the U-Net/VAE, although not significantly so, $p=0.12$. Unlike the DSC and HD which focus on accuracy of the automated contours, the NCC provides an estimate about the variability among the two distributions. The reduction of NCC for the U-Net/OM-VAE suggests that the variability in the training dataset is reduced when outliers are replaced with the highest accuracy contours after the initial model prediction. Therefore, even

though the DSC and HD values increase significantly when outliers are replaced by highest accuracy contours, variability decreases. For the U-Net/VAE with outliers removed from the training dataset (Table 1), we find that the NCC values are significantly reduced (0.46 ± 0.10) relative to the U-Net/OM-VAE with outliers replaced (0.59 ± 0.07) or the U-Net/VAE without outlier mitigation (0.62 ± 0.06). This implies that the removal of outliers from the training dataset negatively impacts the variability of the training dataset, which can ultimately affect accuracy in particular when test data includes geometry that is highly variable relative to the training dataset. Therefore, caution must be exercised when removing outliers from the training dataset. Replacement of outliers appears to have a smaller impact on variability, however, must be limited. In this work the process of identifying the low accuracy contours and updating the training dataset was performed only once because of the potential to significantly reduce variability in the training dataset when contours are removed for multiple iterations.

The OM strategy compared the max DSC auto-contour with the minimum DSC observer contour, and defined the outlier when $\text{DSC observer_min} < \text{DSC auto-contour_max}$. Using this criterion, we were able to replace outliers in the training dataset and improve the model accuracy. In the outlier mitigation procedure, only 17% of the total contours were detected as outliers. This implies that the best auto-contour could not beat the majority (83%) of observer contours (in terms of DSC). Choosing other criteria, such as comparing the best contour from the observer and model for the outlier detection will result in a very low outlier detection rate since the best observer contour will likely always beat the best auto-contour, considering that the model is trained on these observer, ground-truth contours.

In this work, we implemented a modified version of the probabilistic hierarchical U-Net with VAE described by Baumgartner et al [8]. The VAE is a deep generative, and unsupervised learning technique in which the encoding distribution is regularized via a latent representation for dimensionality reduction to generate good model predictions. Studies [8, 63] have shown that VAE’s provide robust modeling of unstructured, heterogenous data incorporating uncertainty, and thereby provide an advantage over standard U-Net which have been successful with contoured data without uncertainty. Another major advantage of the VAE over the standard U-Net is the ability to produce multiple output contours on a given 2D image, which incorporate uncertainty within the training dataset. We have applied this approach toward modeling of inter-observer segmentation uncertainty for prostate contouring to produce automatic contours which are robust to the inter-observer uncertainty.

The following limitations are worth noting. Currently we only performed one round of outlier mitigation to improve the segmentation accuracy. The reason for this is that after multiple rounds of outlier replacement, we begin to lose the variability in the contours generated by different radiation oncologists, resulting in a lower model variance and resulting potential for overfitting and increasing in the generalization error. A higher variance afforded by greater variability in the training dataset suggests that the model is becoming more sophisticated to fit more patterns of the dataset, which is degraded when training data variability is reduced. Consequently, the overuse of outlier mitigation will negatively affect the model accuracy. We have not investigated the tradeoffs between bias, variance, and generalization error as a function of the number of outlier mitigation iterations, which is a part of future work. Regarding the sample sizes, the source_prim dataset consists of a

total of 300 prostate cancer patient CT image datasets while the source_sec consists of image datasets from 10 patients independently contoured by each of 5 physicians (total of 50, 3D-CT datasets). We then applied data augmentation to increase each of the source_prim and source_sec sizes by a factor of 100. It is possible that there is a tradeoff between sample size and dataset variability. For instance, in comparing a very large dataset with low variability versus a smaller dataset with high variability, it is not clear which of these datasets optimizes trade-offs between model bias and variance, ultimately compromising the model's generalizability. As part of future work, we intend to evaluate the influence of these factors as a function of sample size, variability and level of augmentation and augmentation technique. Despite the increase in flexibility of the U-Net/VAE in generating multiple contour outputs for clinical decision support, one could also argue that it might be time-consuming to review multiple contours for each 2D cut, especially if a large number of segment outputs are selected. In this regard, we intend to evaluate the optimal number of contours needed to optimize the tradeoff between accuracy and efficiency in the clinical workflow.

Our network was applied to segmentation of the prostate gland only primarily because our primary CT dataset was limited to availability of the prostate contours only, and our secondary dataset consisted of multiple clinician contours for the prostate gland only. We are working on acquisition of manual contours for normal organs, which we will evaluate as part of a future study. Additionally, we intend to compare results of our network with that of other newer architectures, such as deeply supervised U-Net, spatial transformer network, mask R-CNN, etc.

The major elements of this work include: (a) development of an outlier mitigation

strategy using initial prediction of a U-Net/VAE to limit the influence of outliers on the model prediction accuracy; (b) application of a combined U-Net/VAE for incorporating inter-observer uncertainty in auto-segmentation of the prostate gland for patients with prostate cancers; (c) application of transfer learning to train a smaller dataset (incorporating inter-observer uncertainty) based on initial training of a large dataset.

2.5 Conclusions

In this chapter, we present three novel DNN based medical imaging AI systems, Vispi, COVID-MobileXpert and U-Net/OM-VAE. Using only a single frontal view CXR, Vispi is capable of accurately annotating X-ray images and generating quality reports. it also provides visual supports to assist radiologists in rendering diagnostic decisions. With more quality training data becomes available in the near future, our medical image interpretation system can be improved by: (1) incorporating both frontal and lateral view of X-rays, (2) predicting more disease classes, and (3) using hand labeled bounding boxes as the target of localization. We will also generalize Vispi by extracting informative features from Electronic Health Record (EHR) data and repeated longitudinal radiology reports to further enhance the performance of our system. Different from the Vispi which interprets CXR on a large scale DNN system, COVID-MobileXpert includes a three-player KTD framework which is designed for on-device medical imaging based COVID-19 case screening and radiological trajectory prediction described herein. In the framework, knowledge transfer from the AP network to the RF network can be viewed as a more effective regularization as they are built on the same network architecture, which in turn, make the knowledge distillation more effective since the RF network and MS network share the same training set. Different

from what has extensively investigated focusing on the impact of distillation strength and temperature, we uncover a pivotal role of employing novel loss functions in refining the quality of knowledge to be distilled. Hence our three-player framework provides a more effective way to train the compact on-device model using a smaller dataset while preserving performance. From a more broad perspective, the three-player KTD framework is generally applicable to train other on-device medical imaging classification and segmentation apps for point-of-care screening of other human diseases such as lung and musculoskeletal abnormalities. In addition to those two classification tasks, U-Net/OM-VAE is a first application of a deep learning architectures incorporating outlier mitigation and inter-observer variation in the training for prostate gland auto-segmentation in radiotherapy planning. It can automatically detect and mitigate the outlier contours in the training dataset and considering the inter-observer variation with the VAE part. In the future, we intend to extend the current method to more organ segmentation tasks.

CHAPTER 3 DEPLOYING ADVANCED MEDICAL IMAGING SYSTEM IN THE REAL-WORLD

3.1 Introduction and Related Work

With the development of deep neural networks (DNNs) and the availability of high quality labeled medical imaging datasets, DNN based medical imaging systems have substantially increased the accuracy and efficiency of the clinical prediction tasks. For example, Daniels et al. [25] extract features from X-rays for lung disease classification, Shaffie et al. [121] detect lung cancer using computed tomography (CT) scans and Reda et al. [110] make an early diagnosis of prostate cancer using magnetic resonance imaging (MRI) scans. Recently, several healthcare start-ups such as Zebra Medical Vision and Aidoc announced U.S. Food & Drug Administration (FDA) clearances for their AI medical imaging systems. These FDA approvals indicate that DNN based medical imaging systems are potentially applicable for clinical diagnosis in the near future.

In parallel to the progress in DNN based medical imaging systems, the so-called adversarial images have exposed vulnerabilities of these systems in different clinical domains [38]. Adversarial images are inputs of deep learning models that are intentionally crafted to fool the trained models. Recent studies [105, 97] have specifically explored the reliability of DNN models in both classification and segmentation tasks of medical imaging. They show that these medical DNN models can be even more vulnerable to adversarial samples compared to natural image models. With human imperceptible perturbations added to clean images, adversarial samples can completely fool the trained DNN model into making incorrect predictions. To generate adversarial samples, various types of methods have been proposed, such as Fast Gradient Sign Method (FGSM) [43] and its variant with stronger

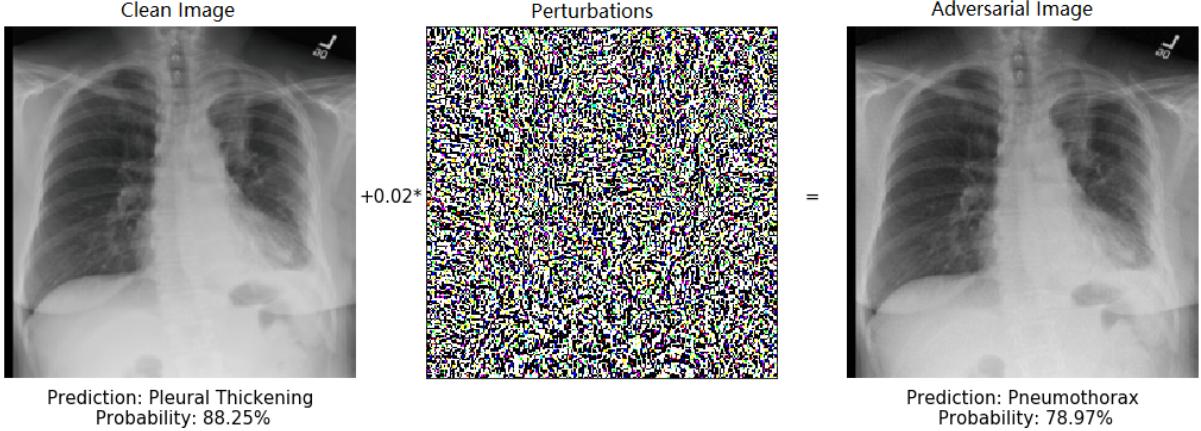


Figure 12: An adversarial attack against a medical image classifier with perturbations generated using FGSM [43].

attacks Projected Gradient Descent (PGD) [87], and optimization-based attack Carlini & Wagner (C&W) [12]. For segmentation tasks, Ozbulak et al. [97] propose an adaptive segmentation mask attack (ASMA), which produces a crafted mask to fool the trained DNN model. Figure 12 shows how a clean image is manipulated to attack a medical image classification system. With only imperceptibly small perturbations added to a clean X-ray image, the system incorrectly classifies “Pleural Thickening” as “Pneumothorax”. Consequently, without proper safeguards, users of such systems can be exposed to unforeseen hazardous situations, such as diagnostic errors, medical reimbursement fraud and so on. Therefore, an effective defense strategy needs to be implemented before these systems can be safely deployed.

To defend against these adversarial attacks, an array of strategies have been developed. One major line of those methods is adversarial training (AT) [87], which improves model’s adversarial robustness by augmenting the training set with adversarial samples. However, AT for DNNs in medical imaging is problematic as they are primarily designed for natural

images and require a large labeled training set [129] whereas medical data sets are usually with a small number of labeled samples. Recently several techniques are proposed to improve the effectiveness of defensive methods for medical images. In segmentation tasks, He et al. [49] find that global contexts and global spatial dependencies are effective against adversarial samples, thus they propose a non-local context encoder in the medical image segmentation system to improve adversarial robustness. In classification tasks, Taghanaki et al. [131] use a radial basis mapping kernel to transform and separate features on a manifold to diminish the influence of adversarial perturbation. Based on features extracted from a trained DNN model, Ma et al. [86] attempt to distinguish adversarial samples from clean ones via density estimation in the subspace of deep features learned by a classification model. Although it achieves impressive performance, the so-called ‘detection’ methods rely on estimating the density of adversarial samples, e.g., via local [85] or Bayesian uncertainty [37] approaches, consequently the effectiveness is limited to the attack methods that are previously seen.

In this chapter, we propose two defending methods to tackle the challenges mentioned above. To defend against diverse unseen attacks, we propose a robust detection strategy for adversarial images that can effectively thwart the adversarial attacks against DNN based medical image classification systems. Inspired by [158], we focus on unsupervised abnormal detection using features extracted from a trained CNN classifier. Our approach does not assume any prior knowledge of attack methods, hence it can robustly defend against diverse unseen attacks, white-box or black-box. For the small data problem, in addition to the unsupervised abnormal detection strategy, we present a hybrid approach that enhances defensive performance using semi-supervised adversarial training (SSAT)

and unsupervised adversarial detection (UAD). Specifically, we utilize both labeled and unlabeled data to generate pseudo-labels for SSAT to improve the robustness of class prediction. Furthermore, both defense strategies can be easily incorporated in any medical imaging system without modifying the architecture nor compromising the performance.

3.2 Unsupervised Detection

In this section, we present the proposed robust detection method. We start from the analysis of the unique properties of adversarial attacks on medical imaging AI systems. Then we describe our detection module equipped with three different unsupervised abnormal detection techniques.

3.2.1 Motivation

The adversarial image is crafted by adding subtle perturbations to the original image; as a result, the perturbations at the pixel level look like noise which do not impede human recognition. However such noise is influential at feature levels of CNN models. We demonstrate these characteristics of adversarial medical images by visualizing the feature maps of a CNN model. In Figure 13a, given one clean X-ray image (top left) and its adversarial counterpart (top right), the corresponding feature maps extracted from the first block of a DenseNet-121 [54] are shown in bottom left and bottom right, respectively. It suggests that adversarial perturbations, albeit are subtle at the pixel level and hard to be detected by human eyes, lead to substantial “noise” at feature levels.

Furthermore, this “noise” can be exacerbated by the convolution-pooling operations implemented in CNN models during forward propagation [147], and finally leads to misclassification. On the other hand, since the magnitude of perturbations increases layer by

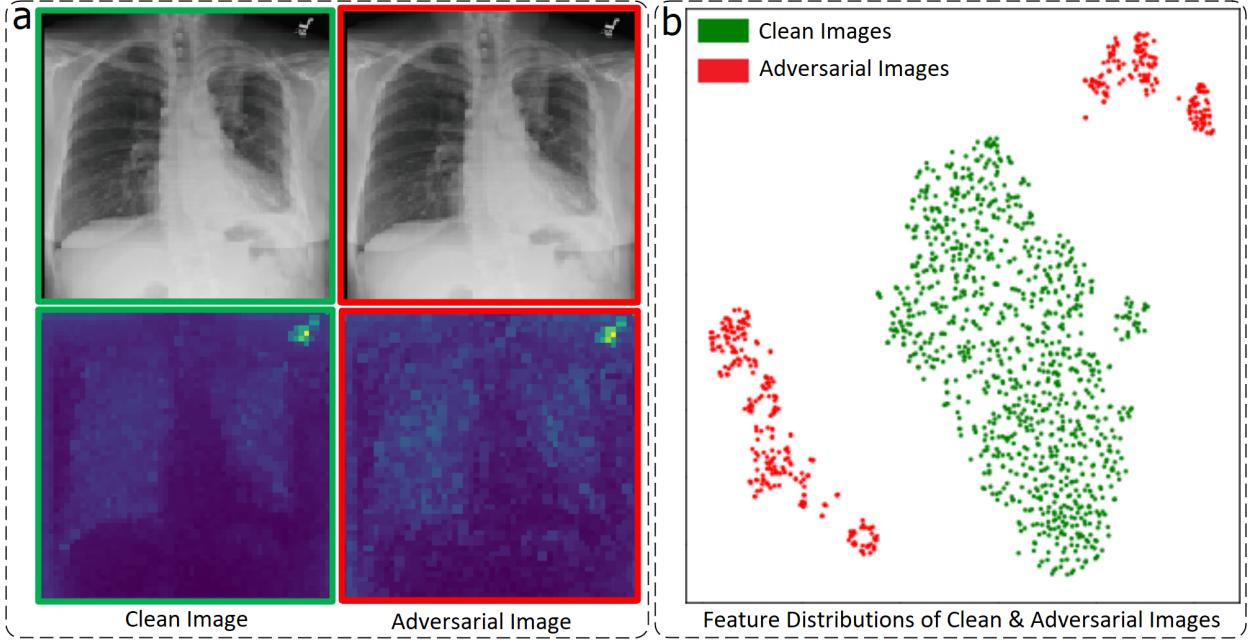


Figure 13: (a) Visualization of input images and feature maps from the first block of a DenseNet-121 [54]. (b) Visualization of feature distributions from the final fully connected layer of clean X-ray images (green) versus adversarial X-ray images (red).

layer, the clean and adversarial images can be easily distinguished based on the high-level features. This assumption is verified from Figure 13b, which visualizes feature distributions of clean and adversarial X-ray images extracted from the final fully connected layer of the DenseNet-121 using t-SNE method. All X-ray images are randomly selected, which cover different types of pathologies. It is obvious that the clean images can be modeled as a unimodal multivariate density (green) whereas adversarial images (red) can be treated as outliers. Different from natural images that may be affected by changes in lighting and position, medical images are highly standardized since they are generally captured with pre-defined and well-established positioning and exposure. Consequently, the trained DNN based imaging system is more sensitive to these crafted perturbations.

3.2.2 Methods

We propose to augment the medical image classification system with an adversarial image detection module. Figure 14 illustrates an example framework of the chest X-ray disease classification system equipped with our detection module. After training the CNN classifier with all clean images to extract the high-level features for learning the detection module, the lower panel illustrates the process of detection and testing. Given a new (clean or adversarial) image, the system extracts features using the trained CNN classifier as the input of detection module. The input image is rejected if detected as an adversarial image, otherwise, it continues to the loss layer to predict classification labels. To accommodate diverse adversarial attacks, we use unsupervised anomaly detection techniques for the detection module. Specifically, we use unimodal multivariate Gaussian model (MGM) as the attacker detection method whereas Isolation Forest (ISO) [80] and One-class SVM (OCSVM) [117] as competing methods.

The high-level feature distribution of clean images can be modeled using MGM: $z \sim \mathcal{N}(\mu, \Sigma)$, where z represents the feature extracted using the final fully connected layer given a clean input image x . The $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d \times d}$ are mean vector and covariance matrix, where d denotes dimension of MGM. Given features extracted from clean training images $Z = \{z_1, \dots, z_n\}$, we estimate

$$\mu = \frac{1}{n} \sum_{i=1}^n z_i, \quad \Sigma = \frac{1}{n} \sum_{i=1}^n (z_i - \mu)(z_i - \mu)^T + \lambda I, \quad (3.1)$$

where λI is the non-negative regularization added to the diagonal of covariance matrix.

After training MGM, for a new (clean or adversarial) image x^* , we compute the prob-

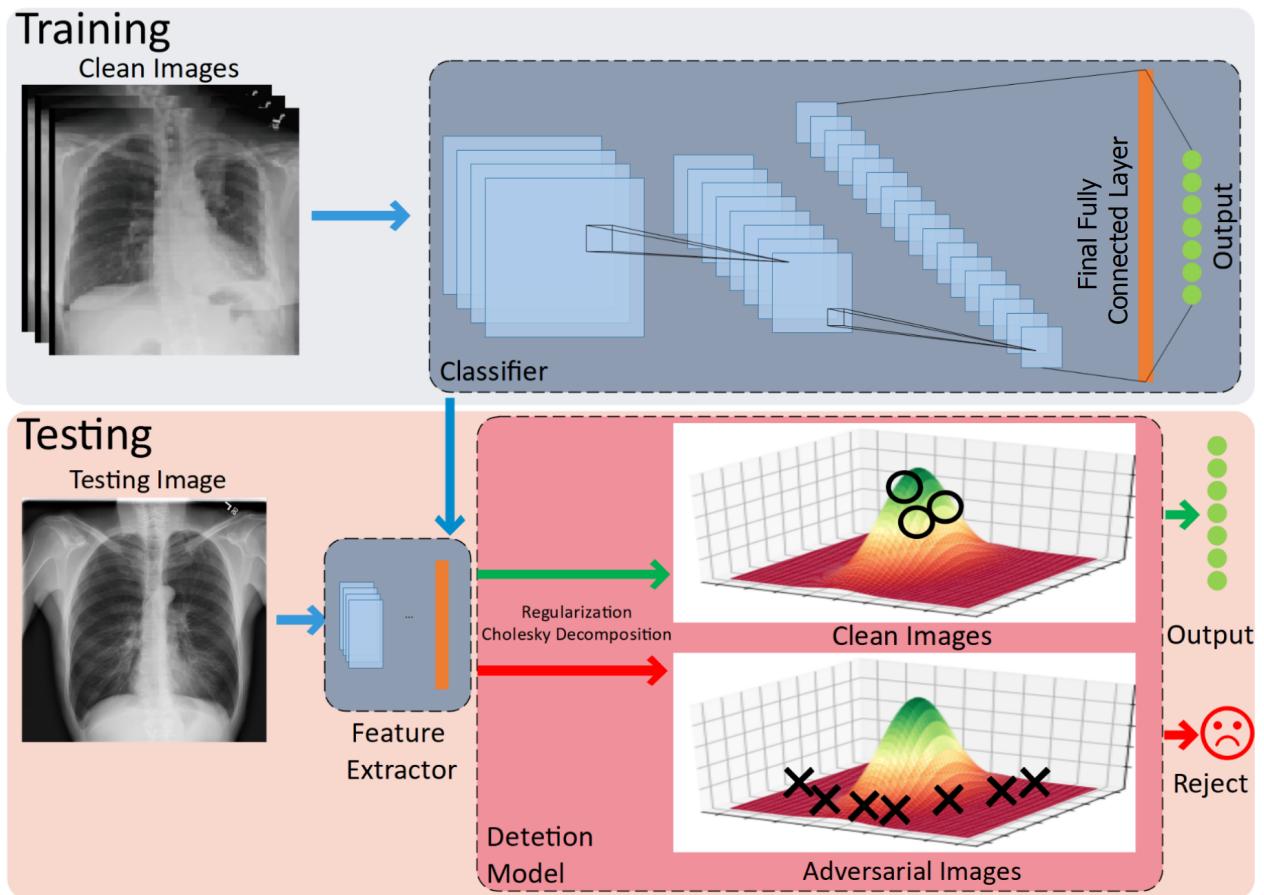


Figure 14: The proposed defense framework for a chest X-ray disease classification system equipped with our MGM detection module.

ability of $\mathbf{z}^* = H(\mathbf{x}^*)$ belonging to the clean image distribution by:

$$p(\mathbf{z}^*) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{z}^* - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{z}^* - \boldsymbol{\mu})\right). \quad (3.2)$$

However, in practice, the high dimension, i.e., $d = 1024$, makes $p(\mathbf{z}^*)$ computational expensive, and the value of $p(\mathbf{z}^*)$ is so close to zero that cause arithmetic underflow. To overcome these technical difficulties, we use Cholesky decomposition to re-parametrize the covariance matrix: $\Sigma = \mathbf{R}\mathbf{R}^T$ and rewrite the probability density function into log form:

$$\log p(\mathbf{z}^*) = -\frac{1}{2}[2 \times \left(\sum_{i=1}^d \mathbf{R}_{ii}\right) + \|\mathbf{R}^{-1}(\mathbf{z}^* - \boldsymbol{\mu})\|^2 + d \log(2\pi)]. \quad (3.3)$$

Finally, as shown in the Figure 14, \mathbf{x}^* will be detected as an adversarial image and rejected if $\log p(\mathbf{z}^*)$ is lower than a threshold. The threshold can be determined by keeping 95% of the training data as clean images.

ISO algorithm builds an isolation tree (itree) by recursively dividing Z with a random feature and a random cut-off value. By creating many itrees, the average path length of unsuccessful search $c(n)$ is used to assign the anomaly score: $s(\mathbf{z}, n) = 2^{(-E(h(\mathbf{z}))/c(n))}$, where $E(h(\mathbf{z}))$ is the average path length of a single input \mathbf{z} . The new (clean or adversarial) image \mathbf{x}^* is rejected if $s(\mathbf{z}^*, n)$ is close to 1. OCSVM is another competitor used in our experiment, it can be summarized as mapping the clean training data \mathbf{z} to a feature space and finding the maximal margin which separates the mapped data from the origin. In our context, let Φ to be the kernel function that transforms \mathbf{z} to another space, and \mathbf{w} and ρ are the parameters to be learned to characterize the maximal margin. After training, given

a new (clean or adversarial) image \mathbf{z}^* , it will be detected as an adversarial image if the decision function

$$f(\mathbf{z}^*) = \text{sgn}((\mathbf{w} \cdot \Phi(\mathbf{z}^*)) - \rho) = -1. \quad (3.4)$$

3.2.3 Experiments

Dataset To verify the performance of our proposed defense approach on medical image classification, experiments are conducted on a large public chest X-ray dataset. The NIH ChestX-ray14 [141] contains 112,120 frontal-view chest X-rays taken from 30,805 patients, where around 46% images are labeled with at least one of 14 pathologies. Following the pre-processing of [141], we split the dataset into training, validation and test datasets by a ratio of 7:1:2 for the image classification system which is DenseNet-121 in our experiment. The features extracted from the entire clean training and validation datasets are used for training and validating the detection module. We then randomly select 1000 clean images from the test dataset for crafting adversarial images using four adversarial attack methods, i.e., fast gradient sign method (FGSM) [43], projected gradient descent (PGD) [87], basic iterative method (BIM) [65], and momentum iterative method (MIM) [32] (the winner of NIPS 2017 adversarial attacks competition). For each attack method, we craft 1000 adversarial images based on the 1000 clean images.

Attacks We evaluate our defense approaches (MGM, ISO and SVM) against the four attack methods mentioned above. Two attack settings are used in the experiment. 1) White-box Attack: attackers know all details of the true CNN classifier (DenseNet-121), and directly use gradients from the model to craft adversarial images. 2) Black-box Attack: attackers know nothing about the true CNN classifier and use an arbitrary substitute

classifier (ResNet-50 [48]) to craft adversarial images. Since the disease classification problem is a multiple binary classification problem and attackers would not know the true label, for each clean image, we use the class with the highest predicted probability to craft the adversarial images. The perturbations are calculated by using the gradient of cross-entropy loss function on the selected class. To ensure the perturbations are subtle enough to remain undetectable from human recognition, the maximum perturbation is limited by 0.05 for black-box setting and 0.02 for white-box setting.

| Attacks | White-box Attack (F1 / AUROC ± STD) | | | | MIM |
|-------------------------------------|--|--|--|--|-----|
| | FGSM | BIM | PGD | | |
| No Defense | 0.500 / 0.702 ± 0.063 | 0.500 / 0.617 ± 0.071 | 0.500 / 0.616 ± 0.071 | 0.500 / 0.591 ± 0.063 | |
| ISO | 0.838 / 0.786 ± 0.077 | 0.874 / 0.810 ± 0.083 | 0.874 / 0.810 ± 0.083 | 0.874 / 0.810 ± 0.083 | |
| SVM | 0.870 / 0.783 ± 0.077 | 0.931 / 0.816 ± 0.083 | 0.931 / 0.816 ± 0.089 | 0.931 / 0.816 ± 0.094 | |
| MGM | 0.936 / 0.801 ± 0.089 | 0.975 / 0.820 ± 0.089 | 0.975 / 0.820 ± 0.089 | 0.975 / 0.820 ± 0.089 | |
| Black-box Attack (F1 / AUROC ± STD) | | | | | |
| Attacks | BIM | | | | MIM |
| | FGSM | BIM | PGD | | |
| No Defense | 0.500 / 0.749 ± 0.077 | 0.500 / 0.737 ± 0.077 | 0.500 / 0.741 ± 0.077 | 0.500 / 0.719 ± 0.077 | |
| ISO | 0.871 / 0.810 ± 0.083 | 0.759 / 0.777 ± 0.089 | 0.735 / 0.776 ± 0.089 | 0.837 / 0.801 ± 0.083 | |
| SVM | 0.903 / 0.812 ± 0.077 | 0.777 / 0.781 ± 0.083 | 0.754 / 0.776 ± 0.083 | 0.859 / 0.792 ± 0.089 | |
| MGM | 0.958 / 0.819 ± 0.083 | 0.924 / 0.809 ± 0.089 | 0.903 / 0.808 ± 0.083 | 0.957 / 0.818 ± 0.083 | |

Table 10: F1 scores are shown for comparing detection performance and AUROC values weighted average over 14 different classes with standard deviation are shown for comparing classification performance of each attack-defense combination.

Metrics We evaluate our defense approach against each attack method based on detection performance and follow-up classification performance. The detection performance is evaluated by F1 score, representing the best trade-off between precision and recall. For comparing performance of the follow-up classification, we use AUROC weighted average from 14 different classes because ROC curve has the advantage of determining the optimal cut off values for classification decisions based on the class probabilities.

Results Table 10 shows the detection performance for each attack-defense combination under both white-box and black-box settings. Since the test dataset consists of 1000 clean images and 1000 adversarial images, the F1 score of the classification system without a detection module (the weak baseline) is always 0.5. All detection methods demonstrate robust performance against these attacks under the white-box setting with MGM has the best performance. We note that the adversarial images crafted using one-step FGSM [43], an earlier adversarial attack method, are more effective compared to others under the white-box setting evident by a lower F-1 score. Similar to the white-box setting, MGM demonstrates the best performance among all detection methods against all attacks under the black-box setting where the architecture of the true CNN classifier is unknown to the attackers. However, the trend is reversed under the black-box setting that adversarial images crafted using one-step FGSM are easier to be detected compared to others. We explain this phenomenon below.

Since detection is based on the features extracted from the true CNN classifier, an adversarial image is easier to be detected if it is contaminated with more “noise” at feature levels. Under the white-box setting, adversarial images crafted from the iterative methods

(e.g., BIM, PGD, MIM) are easier to be detected because they iteratively increase perturbations to maximize the “noise” at feature levels. However, under the black-box setting, adversarial images are crafted using a substitute classifier (ResNet-50), which can be quite different from the true CNN classifier (DenseNet-121). Thus adversarial images crafted by the iterative methods can maximize “noise” for the substitute classifier but not for the true CNN classifier, making it much lower “noise” at feature levels thus harder to be detected.

We also report the follow-up classification performance in Table 10 under both white-box and black-box settings, which is consistent with detection performance. The system equipped with the MGM detection module has the best performance among all detection methods under both settings evident by the highest AUROC values. It is interesting to point out that the proposed framework with a detection module, such as MGM under the white-box setting, can have a better classification performance on mixed clean and adversarial images (0.820) than the true CNN classifier tested only on clean images (0.817), which is possibly due to: (1) the detection module effectively rejects all adversarial images, ensuring the system’s non-compromised classification performance as using a clean dataset, and (2) the detection module can also erroneously reject some clean images as adversarial images. These clean images can be problematic for the CNN classifier since they are at tails of the distribution. Therefore, rejecting these clean images can improve classification performance.

3.3 Semi-supervised Adversarial Training with Adversarial Risk Assessment

In this section, we mainly tackle the label scarcity challenge in the medical images against adversarial attacks. Specifically, we propose a novel robust medical imaging AI framework based on SSAT and UAD, followed by designing a new measure for assessing systems adversarial risk.

3.3.1 Motivation

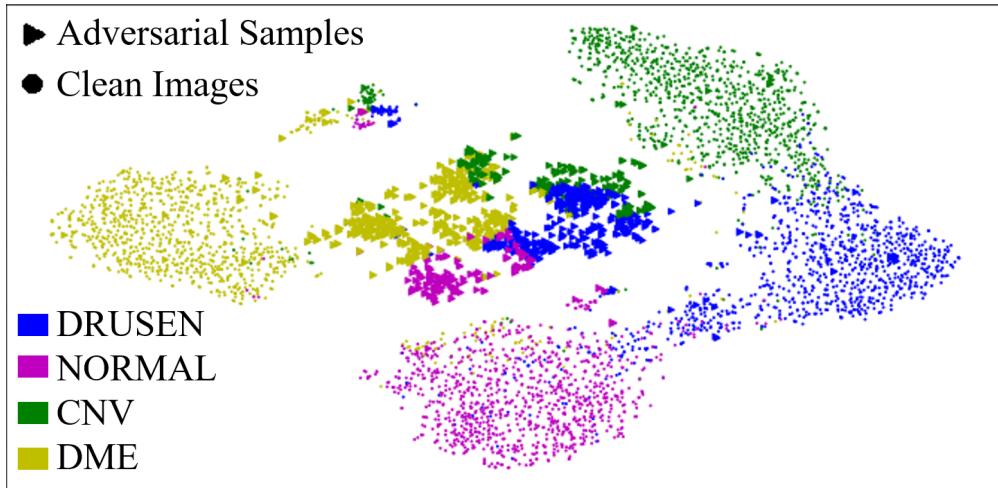


Figure 15: T-SNE visualization of penultimate layer activations of the model trained on the OCT dataset [60]. The clean images are represented by solid circles with each color represents a true class. The adversarial samples (triangles) are crafted by PGD with a perturbation budget $\epsilon = 0.005$ where each color represents a predicted class. For each class, UAD is capable of filtering out the majority of adversarial samples (center) and SSAT enables the model to correctly predict the rest of adversarial samples (close to clean images).

The medical image classification problem is to train a prediction function $f_\theta(\cdot)$ by minimizing the loss in mapping a clean image $x \in \mathcal{X}$ to its true label y . Due to the existence of adversarial samples $x' \in \mathcal{X}'$, it is necessary to have a detection function $g_\phi(\cdot)$ that can distinguish whether an input of f_θ is perturbed by an adversary. Ideally g_ϕ takes inputs from

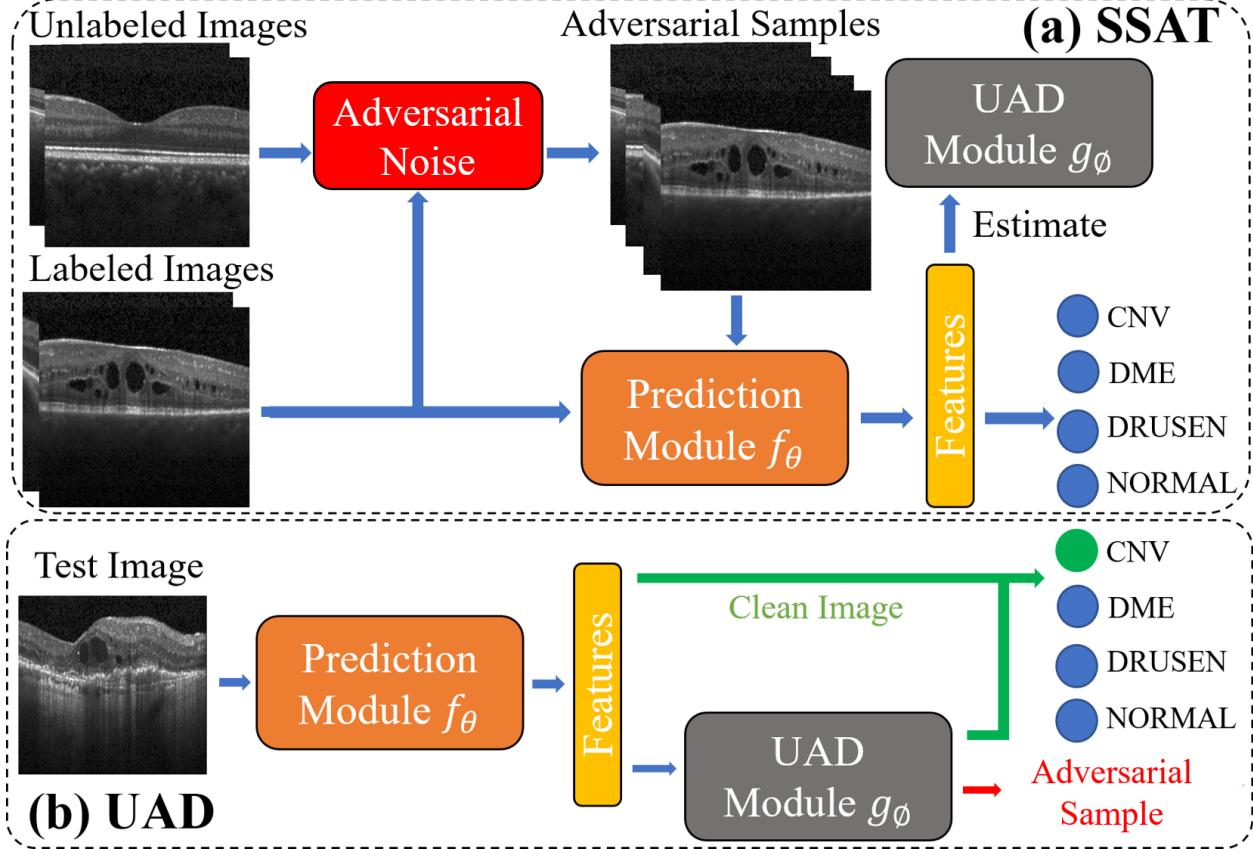


Figure 16: The proposed robust OCT imaging classification system equipped with SSAT and UAD modules.

both \mathcal{X} and \mathcal{X}' , rejects all x' from \mathcal{X}' , then f_θ only takes x from \mathcal{X} to make predictions. A promising solution is to design a UAD function g_ϕ to reject all adversarial samples from \mathcal{X}' . However, it is a challenging task since some of adversarial samples are very close to clean images (Figure 15). As such, a supervised prediction function f_θ that is capable of correctly classifying those adversarial samples using a limited labeled training set is also indispensable for maximizing the defense effectiveness.

3.3.2 Method

Figure 16 illustrates our adversarial defense approach. During training (Figure 16a), we learn the robust feature representation via SSAT for both prediction and UAD modules.

During inference (Figure 16b), given an unseen test image, the system extracts the feature as the input for UAD module. The test image is rejected if it is detected as an adversarial sample, otherwise, it continues to the loss layer to predict a class label. We describe the technical details of SSAT and UAD modules in the following subsections.

Semi-supervised Adversarial Training Adversarial training (AT) [43] is a powerful way to improve the adversarial robustness of a prediction module when the labeled training set is abundant. Recently adversarial samples generated from unlabeled data with pseudo labels have been shown to be valuable for improving the adversarial robustness [129]. For training the prediction module f_θ with labeled images, we use the supervised AT, i.e.,

$$\mathcal{L}_{\text{sup}}(\theta) = \mathbb{E}_{\mathbf{x} \in \mathcal{X}} \sup_{\mathbf{x}' \in \mathcal{N}_\epsilon(\mathbf{x})} \text{xent}(y, f_\theta(\mathbf{x}')), \quad (3.5)$$

where xent is the cross-entropy loss, $\mathcal{N}_\epsilon(\mathbf{x})$ denotes the neighborhood of a clean image \mathbf{x} and $\|\mathbf{x} - \mathbf{x}'\|_\infty < \epsilon$. The inner maximization can be approximated by any available attack method, such as PGD and FGSM. For training with unlabeled images, we first find their pseudo labels $\hat{y}(\mathbf{x})$ predicted by f_θ , followed by AT, i.e., minimizing

$$\mathcal{L}_{\text{unsup}}(\theta) = \mathbb{E}_{\mathbf{x} \in \mathcal{X}} \sup_{\mathbf{x}' \in \mathcal{N}_\epsilon(\mathbf{x})} \text{xent}(\hat{y}(\mathbf{x}), f_\theta(\mathbf{x}')). \quad (3.6)$$

We then minimize the loss function to perform SSAT in an effort to enhance model's adversarial robustness:

$$\mathcal{L}_{\text{semi-sup}}(\theta) = \mathcal{L}_{\text{sup}}(\theta) + \lambda \mathcal{L}_{\text{unsup}}(\theta), \quad (3.7)$$

where λ is a hyper-parameter tuned based on relative abundances of labeled and unlabeled data.

Unsupervised Adversarial Detection To filter out adversarial samples x' from being fed into f_θ , we design an UAD module g_ϕ with the goal to exclude the majority of adversarial samples $x' \in \mathcal{X}'$, and simultaneously prevent $x \in \mathcal{X}$ from being erroneously rejected. As shown in Figure 15, the clean images have a different distribution from adversarial samples classified into the same class (color). Inspired by this observation, we estimate a probability density only for clean images as the UAD module and reject images deviating away from this density as adversarial samples. Unlike the detection methods described in [85, 86, 37], our proposed UAD is completely unsupervised that does not need to estimate the adversarial density in whatever way. As a result, it is not limited to detecting the adversarial samples from the known attack types. Specifically, let Z be the latent feature extracted from the penultimate layer of f_θ using x as input and we employ a Gaussian mixture model (GMM) for UAD module g_ϕ . Let $\mu_{ij} \in \mathbb{R}^d$ and $\Sigma_{ij} \in \mathbb{R}^{d \times d}$ represents the mean and covariance matrix of the j th Gaussian component of class i , respectively. For a single class, given all features extracted from clean training samples $Z = \{z_1, \dots, z_n\}$, we can estimate parameters of the GMM using the EM algorithm. The high dimension of Z may cause numerical issues during training. Thus a small non-negative regularization is added to the diagonal of the covariance matrices to alleviate these issues [106].

Adversarial Risk Evaluation We propose a new adversarial risk evaluation measure for comparing systems performance in terms of adversarial defense. We assess the risk derived

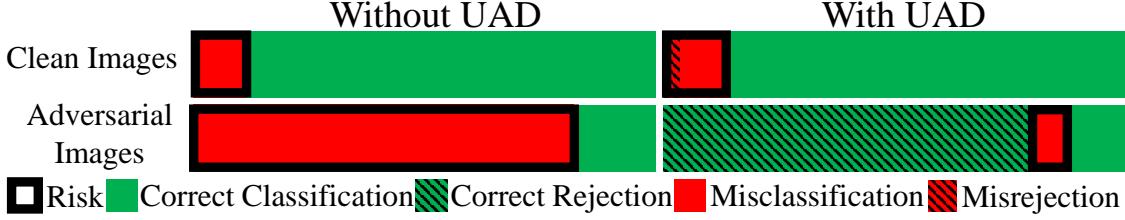


Figure 17: An illustration of assessing systems adversarial risk. Note the system with UAD on the right exhibits a much lower risk represented by smaller red zones.

from clean images based on the following intuition: 1) a clean image incurs no risk if it can be correctly classified; 2) a clean image being rejected by the UAD incurs risk r_{cln}^{uad} ; and 3) a clean image being accepted by the UAD but misclassified by prediction model incurs risk r_{cln}^{prd} . Assume that for clean images, the number of accepted images that incorrectly predicted is $N_{cln}^{inc}(f, g)$, and the number of clean images being rejected is $N_{cln}^{rej}(g)$, the risk derived from misclassifying (first term) and erroneously rejecting (second term) clean images is calculated as $R_{cln}(f, g) = N_{cln}^{inc}(f, g) \cdot r_{cln}^{prd} + N_{cln}^{rej}(g) \cdot r_{cln}^{uad}$. If only f is used to make predictions (without UAD), the second term is zeroed out. Lets denote $N_{cln}^{inc}(f)$ as the number of clean images being misclassified by f , then the risk is calculated as

$$R_{cln}(f) = N_{cln}^{inc}(f) \cdot r_{cln}^{prd}.$$

Similarly for adversarial samples, we have the following intuition: 1) being correctly rejected by UAD or bypassed but correctly classified incurs no risk; and 2) being erroneously accepted by UAD and misclassified incurs a risk r_{adv}^{prd} . Assume the number adversarial samples in 2) is $N_{adv}^{inc}(f, g)$, the risk derived from adversarial samples is calculated as $R_{adv}(f, g) = N_{adv}^{inc}(f, g) \cdot r_{adv}^{prd}$. When only f is used to make predictions (without UAD), since $N_{adv}^{inc}(f, g) = N_{adv}^{inc}(f)$ and $N_{adv}^{inc}(f)$ is the number of misclassified adversarial samples, the risk is calculated as $R_{adv}(f) = N_{adv}^{inc}(f) \cdot r_{adv}^{prd}$. The total risk, incurred by both clean and adversarial samples, thus can be calculated by $R = R_{cln} + R_{adv}$. The value of different risks

$(r_{cln}^{uad}, r_{cln}^{prd}, r_{adv}^{prd})$ are determined empirically, then we have the risk measures for AI systems with UAD as $R(f, g) = N_{cln}^{inc}(f, g) \cdot r_{cln}^{prd} + N_{cln}^{rej}(g) \cdot r_{cln}^{uad} + N_{adv}^{inc}(f, g) \cdot r_{adv}^{prd}$ and without UAD as $R(f) = N_{cln}^{inc}(f) \cdot r_{cln}^{prd} + N_{adv}^{inc}(f) \cdot r_{adv}^{prd}$.

These evaluation measures are illustrated in Figure 17. Using the above equations, we can assess and compare average adversarial risks between UAD based ($r(f, g) = R(f, g)/N$) and not UAD based ($r(f) = R(f)/N$) defense approaches.

3.3.3 Experiments

We use experiments to demonstrate that: 1) The SSAT module can significantly increase model's adversarial robustness without compromising classification performance of clean images. 2) The UAD module can detect and exclude a majority of successful adversarial examples. 3) Our medical imaging AI system (UAD + SSAT) minimizes adversarial risk compared to other existing AI systems.

Dataset and Experiment Settings The experiments are conducted on a public retinal OCT image dataset, originally released in [60]. It contains 84,495 images taken from 4,686 patients with 4 classes: choroidal neovascularization (CNV), diabetic macular edema (DME), drusen, and normal. To demonstrate the advantages of using unlabeled images for semi-supervised training, we randomly sample 4,000 images for training, 1,000 images for test and additional 1,000 images as the unlabeled dataset for SSAT. The 4 classes are balanced in each data set. Following the standard prepossessing [43], all images are center-cropped to 224×224 and all pixels are scaled to $[0,1]$. For AT and SSAT, we augment the data set by generating adversarial samples for each mini-batch using FGSM with a uniformly sampling perturbation from the interval $[0.001, 0.003]$. The number of ad-

versarial and clean images remains 1 : 1 within each mini-batch. We use ResNet-18 [48] pre-trained with ImageNet to learn robust feature representations against adversarial attacks. The networks are trained with the SGD optimizer for 10 epochs with a batch size of 64. We set $\lambda = 5$ for SSAT as in [129].

SSAT Performance We evaluate class prediction performance under the most challenging threat: 'white-box' setting [12]. Compared to the benign 'white-box' setting, the adversary possesses complete knowledge of the target model, including architecture and model parameters. We compare our SSAT with three baseline methods in terms of classification accuracy: natural training (NT) with cross-entropy loss, AT with cross-entropy loss [43] and NT with guided complement entropy (GCE) loss [13]. The 1,000 attacks are crafted by 1-step FGSM, 10-step PGD, and C&W.

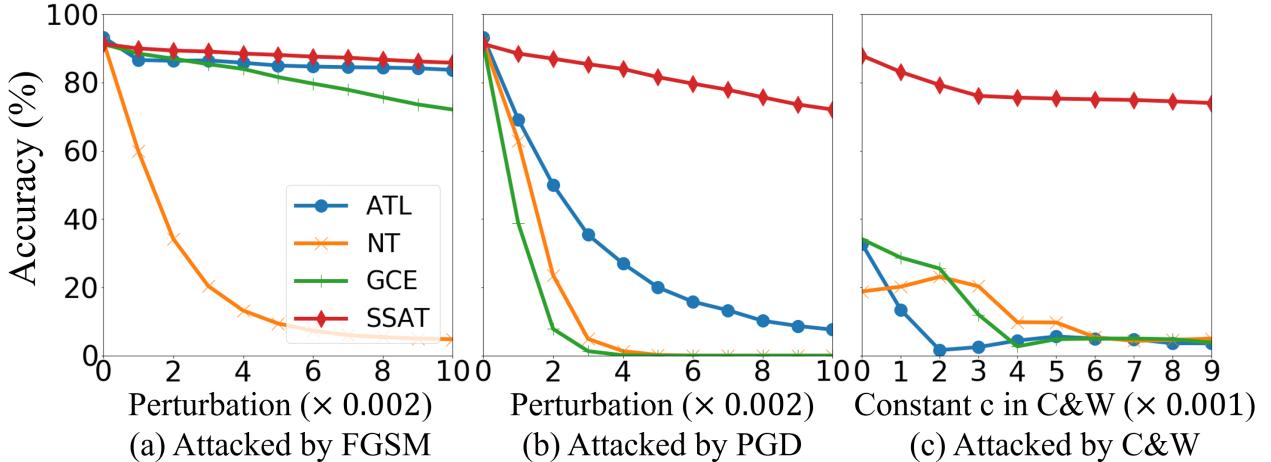


Figure 18: The supervised prediction accuracy of the four trained models on 1000 adversarial examples crafted by FGSM, PGD, C&W with an increasing budget and constant c .

Figure 18 demonstrates that SSAT markedly outperforms other baselines in all white-box attack settings while maintaining a comparable or better performance on the clean image classification (when the perturbation budget is zero). The NT appears very suscep-

tible to easy attacks generated using FGSM with a very small perturbation budget whereas GCE and AT demonstrate a solid performance against easy attacks but fail under strong attacks such as those generated using PGD and C&W. For AT, label scarcity has significantly limited its adversarial generalizability. For GCE, widening the gap in the manifold between different classes may not work well for medical images due to significant overlaps in both the fore- and backgrounds.

| Classes | CNV | DME | DRUSE | NORMAL | Average | # cases |
|---------|--------------|--------------|--------------|--------------|--------------|---------|
| NT | 0.897 | 0.802 | 0.852 | 0.859 | 0.852 | 885 |
| GCE | 0.943 | 0.902 | 0.930 | 0.931 | 0.927 | 970 |
| AT | 0.890 | 0.932 | 0.841 | 0.903 | 0.892 | 580 |
| SSAT | 0.965 | 0.987 | 0.967 | 0.974 | 0.973 | 136 |

Table 11: UAD performance comparison using AUPRC under PGD attack with a perturbation $\epsilon = 0.005$. The last column shows the number of successful adversarial samples.

| Method | NT | GCE | AT | SSAT | SSAT* |
|--------------------------|--------------|--------------|--------------|--------------|--------------|
| Adversarial Risk w/o UAD | 0.483 | 0.529 | 0.324 | 0.112 | 0.456 |
| Adversarial Risk w. UAD | 0.446 | 0.367 | 0.317 | 0.108 | 0.225 |
| Adv Samples Accuracy | 11.5% | 0.3% | 42% | 86.4% | 17.5% |

Table 12: Systems risk under PGD attack with a perturbation $\epsilon = 0.005$. SSTA* is the risk under a stronger PGD attack ($\epsilon = 0.01$).

UAD Performance We use features extracted from 4000 clean images in the training set to estimate mixture model density for UAD. Then the 1000 images from test set and its successful adversarial counterparts are used for assessing performance of UAD. As shown in Table 11, UAD is effective in detecting and excluding adversarial samples evident by high area under the Precision-Recall curve (AUPRC) values among all settings. Furthermore, SSAT is more effective than other training strategies, i.e., NT, AT or GCE. Since the classes of clean images and successful attacks are highly imbalanced (136:1000), AUPRC

is a suitable metric for evaluation [115]. The average AUPRC value of 0.973 shows the proposed UAD can correctly filter out a vast majority of adversarial samples.

Comparison of Adversarial Risks Finally, we demonstrate that UAD complementing with SSAT gives rise to the lowest adversarial risk in terms of the new measure proposed in Section 3.3.2. In Table 12, it is clear that UAD based systems have consistently lower risks compared to those are not, regardless of the training methods used. Note that the reduction of risk is not significant for SSAT against PGD attacks with a smaller budget ($\epsilon = 0.005$). The main reason is that these adversarial samples are relatively weak (highest class prediction accuracy of 86.4% in the last row) that SSAT can successfully predict their labels without the need for UAD. After we double the perturbation budget of PGD attack ($\epsilon = 0.01$), as shown in the last column, the adversarial risk decreases by half (from 0.456 to 0.225) with UAD, highlighting the striking robustness of our system against stronger PGD attacks compared with those without UAD.

3.4 Conclusions

In this chapter, we propose two novel methods to defend adversarial attacks on medical images: the robust detection method and UAD complemented with SSAT. The robust detection method detects the adversarial attacks by modeling the high-level features learned from the clean images using a standard CNN classifier in an unsupervised abnormal detection way. Thus it can defend against diverse unseen attacks. To tackle the label scarcity problem in training a robust classifier, the proposed UAD complemented with SSAT introduces the semi-supervised adversarial training which utilizes both labeled and unlabeled data to improve the adversarial robustness of the class prediction. Through experiments,

our systems demonstrate superior performance in adversarial defense to competing techniques. Furthermore, Both strategies do not need any prior knowledge of attack methods nor modification of the CNN architecture. As a result, these effective strategies can be combined with other defense methods and are sufficiently flexible for many medical imaging applications with diverse image formats. We expect deployment of our approaches would enhance the security of DNN based medical imaging classification systems. For future works, we plan to extend the current method to accommodate more complex datasets that may follow multimodal distributions and more medical image tasks such as segmentation. In addition, we also plan to investigate new dimension reduction approaches to reduce the number of training examples required to estimate the distribution.

CHAPTER 4 IMPROVING ADVERSARIAL ROBUSTNESS OF DNNs VIA PROBABILISTICALLY COMPACT LOSS WITH LOGIT CONSTRAINTS

4.1 Introduction and Related Work

Convolutional neural networks (CNNs) have achieved significant progress for various challenging tasks in computer vision, including image classification [73], semantic segmentation [47], and image generation [42]. Despite their success, CNNs are highly vulnerable to adversarial samples [130]. With imperceptibly small perturbation added to a clean image, adversarial samples can drastically change models' prediction, resulting in a significant drop in CNN's predictive performance. This phenomenon poses a serious threat to security-critical applications of deep learning, such as autonomous driving [2], surveillance system [127], and medical imaging system [25]. Furthermore, studies have shown that adversarial robustness is also a key property to obtain human interpretation in computer vision and other application fields [95, 98]. Therefore, improving models' adversarial robustness is critical to build trustworthy Artificial Intelligence systems to prevent unforeseen hazardous situations.

To improve CNN's adversarial robustness, many methods have been proposed. One strategy is to modify the inputs during inference time via noise removal [89], super-resolution [91] and JPEG compression [26] to diminish the impact of perturbation, but can be easily evaded by strong attacks [4]. Another type of strategy [133, 65, 126, 154, 120] is based on adversarial training [43] that can effectively increase the model's robustness by utilizing crafted adversarial examples as data augmentation. However, it is computationally expensive and compromise model classification performance on clean images [134]. Other than modifying data, some techniques directly enhance model robustness by alter-

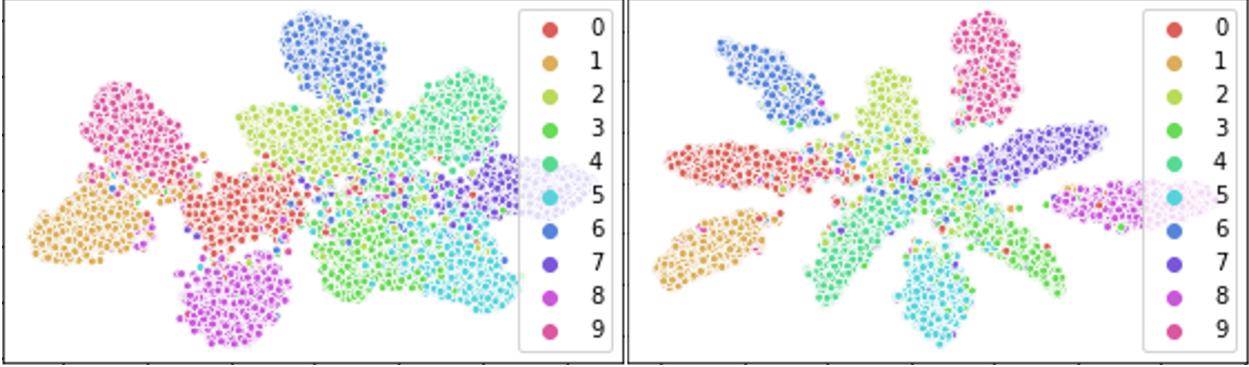


Figure 19: T-SNE visualization of the penultimate layer of ResNet-56 trained with CE loss (left) and PC loss (right) on CIFAR-10.

ing network architectures [131, 90], or constructing ensembles of networks [133, 102]. However, they require additional processes and are not flexible to be adopted to other models.

While previous works have successfully improved CNN robustness against adversarial attacks, the connection of CNN predictions between adversarial and clean samples is not known. In this paper, we investigate this connection in the typical setting when CNNs are trained with the cross-entropy (CE) loss. When an adversarial sample successfully fools the trained CNN with small perturbations, it tends to be misclassified into the first several most probable false classes when predicting the original clean sample, i.e., the classes with larger predicted probabilities. This consistent pattern of CNN’s predictive behaviors is intuitive and potentially implies a deeper connection between the CNN feature learning and its adversarial robustness.

The tendency of misclassifying adversarial samples enlightens us that the adversarial robustness of CNN can be benefited from the training that focuses on the differentiation between sample’s true class and its first several most probable false classes. Hence, in this paper, we propose a novel training objective, termed as Probabilistically Compact (PC) loss

with logit constraints, which can improve adversarial robustness and achieve comparable classification accuracy without extra training procedure and computational burden.

Unlike CE loss which focuses only on maximizing the output probability of the true class, PC loss aims at maximizing probability gaps between the true class and the most probable false classes. Meanwhile, the logit constraints suppress logit which ensures that the gaps is not only large, but also difficult to be crossed. Consequently, this formulation helps to widen the gaps between different classes in feature space, and ensures that it is difficult for an adversary to fool the trained model with small perturbations. We demonstrate the synergistic effect of the gaps at both probability and feature levels using benchmark datasets. For example, the average probability gaps between the true class and the most probable false class of ResNet-56 on CIFAR-100 test data are 0.527 (CE loss) vs. 0.558 (PC loss), respectively. And for Tiny ImageNet test data the gaps become 0.131 (CE loss) vs. 0.231 (PC loss). These results demonstrate that our PC loss can directly enlarge the probability gap of prediction and the effect is more pronounced for more challenging dataset (Tiny ImageNet). As shown in Figure 19, ResNet-56 trained with our PC loss has clear margin boundaries and samples of each classes are evenly distributed around the center with a minimal overlap on CIFAR-10 test data.

Our main contributions are summarized as follows: (1) We offer an unique insight into the predictive behavior of CNN on adversarial samples that the former tends to misclassify the latter into the first several most probable classes. (2) We formulate the problem by proposing a new loss function, i.e., PC loss with logit constraints to improve CNN’s adversarial robustness, where these two components are systematically integrated and simultaneously optimized during training process. (3) Our PC loss can be used as a drop-in

replacement of the CE loss to supervise CNN training without extra procedure nor additional computational burden for improving adversarial robustness. Experimental results show that when trained with our method, CNNs can achieve significantly improved robustness against adversarial samples without compromising performance on predicting clean samples.

4.2 Methods

Notation Let $D = (\mathbf{x}_i, y_i)_{i=1}^N$ be the set of training samples of size N , where $\mathbf{x}_i \in \mathbf{R}^p$ is the p -dimensional feature vector and $y_i = k (k = 1, \dots, K)$ is the true class label, and $S_k = \{(\mathbf{x}_i, y_i) : y_i = k\}$ the subset of D for the k -th class. The bold $\mathbf{y}_i = (y_i^1, \dots, y_i^K)$ is used to represent the one-hot encoding for y_i : $y_i^k = 1$ if $y_i = k$, 0 otherwise.

Cross-entropy (CE) loss Assume that CNN's output layer, after convolutional layers, is a fully connected layer of K neurons with bias terms, then the predicted probability for sample \mathbf{x} being classified into k -th class is calculated using the softmax activation (the k -th logit $a_k = \mathbf{W}_k \mathbf{h}_x + b_k$):

$$f_k(\mathbf{x}) = p(y = k | \mathbf{x}) = \frac{\exp(a_k)}{\sum_{j=1}^K \exp(a_j)} \quad (k = 1, \dots, K), \quad (4.1)$$

where \mathbf{h}_x is the feature representation of \mathbf{x} , \mathbf{W}_k and b_k are parameters of the k -th neuron in the output layer. Then CE loss, which is equivalent to the maximum likelihood approach, is given as follows:

$$L(\boldsymbol{\theta}) = - \sum_{k=1}^K \sum_{i_k \in S_k} \log f_k(\boldsymbol{\theta}; \mathbf{x}_{i_k}), \quad (4.2)$$

where $\boldsymbol{\theta}$ is the vector of trainable model parameters.

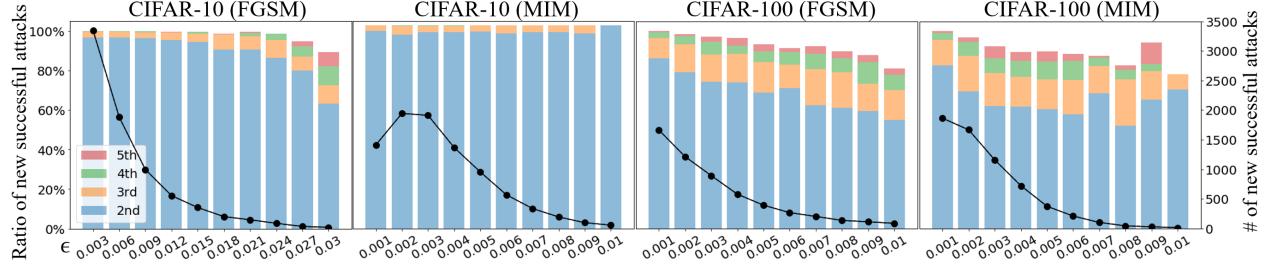


Figure 20: Empirical investigation on the predictive behavior of CNN on adversarial samples from CIFAR-10 and CIFAR-100. The line (black, right y-axis) represents the number of increased successful attacks when ϵ is increased from its previous grid value. Each bar (left y-axis) represents the percentage of misclassification for the increased successful attacks, measuring number of adversarial samples are misclassified into the 2nd, 3rd, 4th and 5th most probable classes. FGSM and MIM are attack methods.

4.2.1 Motivation: Predictive Behavior of CNN on Adversarial Samples

Previous studies have shown that when trained to optimum, *i.e.* $\theta^* = \arg \min_{\theta} L(\theta)$, CNNs can misclassify adversarial samples that are only slightly different from the original clean samples. This vulnerability has recently inspired many methods for generating adversarial samples (attack), defending adversarial attacks and detecting adversarial samples. Here, we take a different perspective on the attacks and empirically investigate if there is a systematic tendency on how CNNs misclassify adversarial samples.

Specifically, for a test (clean) sample (x, y) , the (untargeted) attack seeks a small perturbation ϵ that leads to the misclassification of x when the perturbation is added to x :

$$\min_{\epsilon} \|\epsilon\|_p, \text{s.t. } y' = \arg \max_k f_k(x + \epsilon) \text{ and } y \neq y', \quad (4.3)$$

where $\|\cdot\|_p$ is the norm such as L_1 , L_2 and L_∞ . When Eq. (4.3) is optimized and the attack succeeds, a natural question to ask is “*are there any connections between y' and y for the trained CNN misclassifying $x + \epsilon$ into class y' ?*” Note that $y = \arg \max_k f_k(x)$.

Intuitively, we could expect that y' is likely to be the most probable class except the true class label y , *i.e.*, the class corresponding to the 2nd largest value of CNN predicted probabilities. This conjectures that solving Eq. (4.3) is equivalent to solve

$$\min_{\epsilon} \|\epsilon\|_p, \text{ s.t. } \arg \max_k f_k(\mathbf{x} + \epsilon) = \arg \#_2 \max_k f_k(\mathbf{x}), \quad (4.4)$$

where $\arg \#_2 \max$ represents the operation of taking the 2nd largest value¹.

Here we provide an analysis as our motivation behind this conjecture. Assuming that the CNN is Lipschitz continuous [36], then we have the inequality:

$$\|\mathbf{f}(\mathbf{x} + \epsilon) - \mathbf{f}(\mathbf{x})\|_p \leq l\|(\mathbf{x} + \epsilon) - \mathbf{x}\|_p = l\|\epsilon\|_p, \quad (4.5)$$

where l is the Lipschitz constant and $\mathbf{f}(\cdot) = (f_1(\cdot), \dots, f_K(\cdot))$. The Lipschitz continuity implies that the change of the output is bounded by the change of the input, which is the small perturbation in adversarial attacks. To misclassify $\mathbf{x} + \epsilon$, the possible minimal value of the LHS in Eq. (4.5) is to seek an ϵ such that $f_j(\mathbf{x} + \epsilon) \geq f_y(\mathbf{x} + \epsilon)$, where y is true class and j is 2nd most probable class. To see this, consider the following two cases:

- Case 1. $f_y(\mathbf{x} + \epsilon) \geq f_y(\mathbf{x})$. To misclassify $\mathbf{x} + \epsilon$, the possible minimal value $\|\mathbf{f}(\mathbf{x} + \epsilon) - \mathbf{f}(\mathbf{x})\|_p$ is to reduce $f_{k'}(\mathbf{x})$ ($k' \neq y, j$) to compensate $f_j(\mathbf{x} + \epsilon)$ so that $f_j(\mathbf{x} + \epsilon) \geq f_y(\mathbf{x} + \epsilon)$.
- Case 2. $f_y(\mathbf{x} + \epsilon) < f_y(\mathbf{x})$. The possible minimal value is that $f_y(\mathbf{x} + \epsilon) = f_y(\mathbf{x}) - (\frac{f_y(\mathbf{x}) - f_j(\mathbf{x})}{2})$, $f_j(\mathbf{x} + \epsilon) = f_j(\mathbf{x}) + (\frac{f_y(\mathbf{x}) - f_j(\mathbf{x})}{2})$ and all other $f_{k'}(\mathbf{x})$ ($k' \neq y, j$) remain

¹We may relax it to the first several most probable classes such as 3rd and 4th.

unchanged.

Those two cases may not be achievable in practice, but provide a lower bound on $\|\mathbf{f}(\mathbf{x} + \epsilon) - \mathbf{f}(\mathbf{x})\|_p$. The same analysis can be further relaxed to the 3rd and 4th most probable classes. Observing Eq. (4.5), solving Eq. (4.3) provides an upper-bound for the LHS of Eq. (4.5). With Lipschitz continuity, we hence conjecture that CNN tends to misclassify adversarial samples into classes that have large predicted probabilities when predicting the original clean samples.

To verify our conjecture, we perform an empirical study on CIFAR-10 and CIFAR-100 datasets. FGSM and MIM are used as the adversarial attack algorithms and generate adversarial samples for the standard test data of CIFAR-10 and CIFAR-100. We do not solve Eq. (4.3) for each test samples as it is computationally expensive for the test data of size 10,000. Instead, we take a fine grid of perturbation values and summarize the misclassification of newly successful attacks when the perturbation ϵ is increased from ϵ_m to $\epsilon_{m+1} = \epsilon_m + \Delta$ (Δ is the value of increment). Figure 20 displays the summary of the misclassification results. From the figure, we can see that for CIFAR-10, every time the perturbation is increased, the newly successful attacks are mostly misclassified into the 2nd most probable class for the clean samples. For CIFAR-100, the misclassification follows a similar trend considering it has 100 classes. We also notice that as the perturbation gets larger in FGSM, more newly successful attacks are classified into the 3rd, 4th and 5th most probable classes of predicting clean samples. A possible reason is that the large perturbation results in overshoot in the misclassification as the difference between 2nd most probable class and 3rd most probable class is small when a clean sample needs large

perturbation to be adversarial. Different from FGSM, as perturbation increases, MIM always maintains a high percentage of classifying newly successful adversarial samples into the 2nd most probable class of predicting clean samples for CIFAR-10, due to its iterative procedure in generating adversarial attacks. Overall, Figure 20 empirically agrees with our analysis that motivates our proposed PC loss.

4.2.2 Probabilistically Compact Loss

The predictive behavior of CNN on adversarial samples in the last section inspires us that to improve model robustness to adversaries, CNN needs to focus on the differentiation between the true class and the first several most probable classes. In terms of predicted probability, CNN robustness is benefited from the large gap between true class $f_y(\mathbf{x})$ and false class $f_{y'}(\mathbf{x})$ ($y' \neq y$). Indeed, [94] shows that the gap $f_y(\mathbf{x}) - \max_{y' \neq y} f_{y'}(\mathbf{x})$ can be used to measure the generalizability of deep neural networks.

With the aforementioned motivation, we propose the PC loss to improve CNN's adversarial robustness as follows:

$$L_{pc}(\boldsymbol{\theta}) = \frac{1}{N} \sum_{y' \neq y_i, i \in D} \max(0, f_{y'}(\mathbf{x}_i) + \xi - f_{y_i}(\mathbf{x}_i)), \quad (4.6)$$

where N is the number of training samples, $\xi > 0$ is the probability margin treated as hyperparameter. Here, we include all non-target classes in the formulation and penalize any classes for each training sample that violate the margin requirement for two considerations: (1) if one of the most probable classes satisfies the margin requirement, all less probable classes will automatically satisfy this requirement and hence have no effect in PC loss; (2) since the first several most probable classes are unknown and can change during

the training process, it is necessary to maintain the margin requirement for all classes.

Compared with previous works [90, 101, 131] that improve adversarial robustness via explicitly learning features with large intra-class compactness, PC loss avoids assumptions on the feature space. Instead, PC loss only encourages the feature learning that leads to probabilistic intra-class compactness by imposing a probability margin ξ .

In training CNN with PC loss, the latter is differentiable and hence can be optimized with stochastic gradient descent. The gradient of PC loss can be calculated as (w.r.t. logit a_y)

$$\frac{\partial(f_{y'} + \xi - f_y)}{\partial a_y} = -f_y(1 - f_y + f_{y'}), \quad (4.7)$$

where the gradient is computed for the softmax function. For other logits ($a_{y'}$), the gradients can be similarly computed.

4.2.3 The Logit Constraints

In last section we introduce PC loss to enhance adversarial robustness by enlarging the probability gaps. In this section, we further propose logit constraints as a complement of PC loss, which suppress logit to ensure that the gaps are not only large, but also difficult to be crossed. Here we explain the necessity of using both parts together. We use $|\cdot|$ to denote $\|\cdot\|_2$ for simplicity. The conclusion can be extended to other L_p norms. Assume there is a clean image x , and a corrupted image $x + \epsilon$ attacked by some adversarial algorithm, where $|\epsilon| < \tau$, with $\tau > 0$ be a small constant.

We take the log probability for simplicity, i.e., PC loss is equivalently enlarging $\log f_y(x) - \log f_j(x)$ between true class y and most probable false class j . Given a perturbation ϵ , the

| Attacks | Param. | MNIST | | KMNIST | | FMNIST | | Param. | CIFAR-10 | | SVHN | |
|---------|--------|-------|-------------|--------|-------------|--------|-------------|--------|----------|-------------|------|-------------|
| | | CE | Ours | CE | Ours | CE | Ours | | CE | Ours | CE | Ours |
| Clean | - | 99.2 | 99.2 | 95.5 | 95.4 | 90.1 | 90.2 | - | 91.6 | 91.2 | 94.9 | 94.7 |
| FGSM | 0.1 | 71.5 | 80.5 | 26.2 | 62.8 | 17.5 | 58.0 | 0.04 | 4.3 | 53.1 | 8.7 | 39.5 |
| | 0.2 | 51.6 | 76.3 | 1.8 | 39.7 | 9.5 | 43.3 | 0.12 | 11.7 | 30.3 | 4.5 | 24.2 |
| | 0.3 | 31.8 | 65.0 | 1.2 | 34.1 | 7.6 | 31.8 | 0.2 | 11.5 | 18.7 | 2.6 | 17.1 |
| BIM | 0.1 | 52.8 | 72.0 | 55.8 | 82.5 | 0.0 | 18.7 | 0.04 | 0.0 | 29.0 | 1.3 | 26.2 |
| | 0.2 | 4.5 | 48.6 | 28.7 | 73.3 | 0.3 | 8.4 | 0.12 | 0.0 | 21.0 | 0.0 | 18.2 |
| | 0.3 | 1.5 | 39.5 | 16.8 | 60.5 | 0.0 | 6.4 | 0.2 | 0.0 | 20.0 | 0.0 | 17.6 |
| PGD | 0.1 | 49.0 | 72.3 | 31.3 | 62.4 | 0.0 | 15.7 | 0.04 | 0.0 | 27.6 | 0.0 | 27.6 |
| | 0.2 | 3.3 | 50.2 | 3.9 | 39.9 | 0.0 | 7.0 | 0.12 | 0.0 | 14.6 | 0.0 | 22.0 |
| | 0.3 | 0.8 | 39.7 | 2.0 | 33.2 | 0.0 | 4.6 | 0.2 | 0.0 | 7.5 | 0.0 | 21.0 |
| MIM | 0.1 | 49.8 | 73.8 | 26.0 | 65.4 | 0.0 | 14.8 | 0.04 | 0.0 | 34.3 | 0.0 | 29.2 |
| | 0.2 | 5.0 | 54.0 | 4.2 | 45.4 | 0.0 | 6.3 | 0.12 | 0.0 | 32.7 | 0.0 | 27.6 |
| | 0.3 | 1.5 | 43.3 | 2.0 | 36.9 | 0.0 | 4.5 | 0.2 | 0.0 | 32.4 | 0.0 | 26.0 |
| CW | 0.0 | 42.2 | 78.0 | 19.5 | 57.3 | 0.2 | 21.8 | 0.0 | 0.0 | 30.2 | 0.0 | 36.2 |

Table 13: Accuracy (%) on K/F/MNIST, CIFAR-10 and SVHN under white-box setting. For CW, the parameter is the confidence.

corresponding log probabilities can be estimated via first order approximation

$$\log f_y(\mathbf{x} + \boldsymbol{\epsilon}) = \log f_y(\mathbf{x}) + \boldsymbol{\epsilon} \cdot \nabla_{\mathbf{x}} \log f_y(\mathbf{x}), \quad (4.8)$$

$\log f_j(\mathbf{x} + \boldsymbol{\epsilon})$ can also be approximated in the same manner. To prevent $\log f_y(\mathbf{x} + \boldsymbol{\epsilon}) - \log f_j(\mathbf{x} + \boldsymbol{\epsilon}) < 0$ (i.e., false prediction with perturbation $\boldsymbol{\epsilon}$), we should solve $\min_{\theta} \boldsymbol{\epsilon} \cdot (\nabla_{\mathbf{x}} \log f_j(\mathbf{x}) - \nabla_{\mathbf{x}} \log f_y(\mathbf{x}))$.

Lets denote vector $\mathbf{b} = \nabla_{\mathbf{x}} \log f_j(\mathbf{x}) - \nabla_{\mathbf{x}} \log f_y(\mathbf{x})$. As the attackers can always choose the worst $\hat{\boldsymbol{\epsilon}}$ that maximizes $\boldsymbol{\epsilon} \cdot \mathbf{b}$ by letting $\hat{\boldsymbol{\epsilon}}$ in the same direction as \mathbf{b} , i.e. $\boldsymbol{\epsilon} \cdot \mathbf{b} \leq |\hat{\boldsymbol{\epsilon}}| \cdot |\mathbf{b}|$. Our goal becomes to minimize the upper bound $|\hat{\boldsymbol{\epsilon}}| \cdot |\mathbf{b}|$

$$\begin{aligned} \min_{\theta} |\hat{\boldsymbol{\epsilon}}| |\mathbf{b}| &\implies \min_{\theta} |\mathbf{b}| \\ &\implies \min_{\theta} |\nabla_{\mathbf{x}} \log f_j(\mathbf{x}) - \nabla_{\mathbf{x}} \log f_y(\mathbf{x})|. \end{aligned} \quad (4.9)$$

Hence to prevent the prediction changes after perturbation $\boldsymbol{\epsilon}$, we should minimize

$|\nabla_{\mathbf{x}}(\log f_y(\mathbf{x}) - \log f_j(\mathbf{x}))|$. Let a_k denote the logit for the k th class softmax output, observe that

$$\begin{aligned}\nabla_{\mathbf{x}} \log f_y - \nabla_{\mathbf{x}} \log f_j &= \frac{\nabla_{\mathbf{x}} f_y}{f_y} - \frac{\nabla_{\mathbf{x}} f_j}{f_j} \\ &= \nabla_{\mathbf{x}} a_y - \nabla_{\mathbf{x}} a_j,\end{aligned}\tag{4.10}$$

because $\nabla_{\mathbf{x}} f_y = -\sum_k f_k f_y \nabla_{\mathbf{x}} a_k + f_y \nabla_{\mathbf{x}} a_y$, and the same holds for $\nabla_{\mathbf{x}} f_j$. We can equivalently change our objective to $\min_{\boldsymbol{\theta}} |\nabla_{\mathbf{x}}(a_y - a_j)|$. We can estimate $|\nabla_{\mathbf{x}}(a_y(\mathbf{x}) - a_j(\mathbf{x}))|$ using

$$|\nabla_{\mathbf{x}}(a_y - a_j)| \approx |(a_y(\mathbf{x}) - a_j(\mathbf{x})) - (a_y(\mathbf{x} + \hat{\boldsymbol{\epsilon}}) - a_j(\mathbf{x} + \hat{\boldsymbol{\epsilon}}))| / |\hat{\boldsymbol{\epsilon}|},\tag{4.11}$$

where we denote $|\boldsymbol{\epsilon}| < |\hat{\boldsymbol{\epsilon}}| = \tau$ that upper bounds $|\boldsymbol{\epsilon}|$. Note that an adversarial attack tends to minimize $a_y(\mathbf{x} + \hat{\boldsymbol{\epsilon}}) - a_j(\mathbf{x} + \hat{\boldsymbol{\epsilon}})$ so that $a_y(\mathbf{x} + \hat{\boldsymbol{\epsilon}}) - a_j(\mathbf{x} + \hat{\boldsymbol{\epsilon}}) < a_y(\mathbf{x}) - a_j(\mathbf{x})$. And a robust model should instead prevent $a_y(\mathbf{x} + \boldsymbol{\epsilon}) - a_j(\mathbf{x} + \boldsymbol{\epsilon}) < 0$ to ensure a correct prediction when attacked, so we have the following inequality for a robust model under attack

$$0 < a_y(\mathbf{x} + \hat{\boldsymbol{\epsilon}}) - a_j(\mathbf{x} + \hat{\boldsymbol{\epsilon}}) < a_y(\mathbf{x}) - a_j(\mathbf{x}).\tag{4.12}$$

Then Eq. (4.11) can be upper bounded by

$$|\nabla_{\mathbf{x}}(a_y - a_j)| < |a_y(\mathbf{x}) - a_j(\mathbf{x})| / |\hat{\boldsymbol{\epsilon}}|.\tag{4.13}$$

Substitute this inequality back to Eq. (4.10), we get a logit constraint condition to ensure

model robustness

$$|\nabla_{\mathbf{x}}(\log f_y - \log f_j)| < |a_y(\mathbf{x}) - a_j(\mathbf{x})|/|\hat{\epsilon}| < C, \quad (4.14)$$

where C is an arbitrary positive constant thresholding robustness, hence we can optimize PC loss subject to the above condition

$$\min_{\boldsymbol{\theta}} L_{pc}(\boldsymbol{\theta}), \text{s.t. } |a_y(\mathbf{x}; \boldsymbol{\theta}) - a_j(\mathbf{x}; \boldsymbol{\theta})| < C' \text{ for } \forall \mathbf{x}, \quad (4.15)$$

where $C' = |\hat{\epsilon}|C$. It is equivalent to write the above minimization problem with a multiplier λ

$$\min_{\boldsymbol{\theta}, \lambda} \left(L_{pc}(\boldsymbol{\theta}) + \frac{\lambda}{N} \sum_{\mathbf{x} \in D} (d_{yj} - C') \right), \quad (4.16)$$

where N is the number of samples, and D is the set of training samples. λ is treated as a hyper-parameter in training, $d_{yj} = \max(0, a_y(\mathbf{x}; \boldsymbol{\theta}) - a_j(\mathbf{x}; \boldsymbol{\theta}))$. As shown in Eq. (5.1), PC loss and logit constraints are systematically integrated and simultaneously optimized during training process to enhance model adversarial robustness.

4.3 Experiments

In this section, we evaluate our proposed PC loss with logit constraints along with analysis that our method does not rely on the ‘gradient masking’ that provides a false sense of security [4].

Datasets and models: We analyze seven benchmark datasets: MNIST, KMNIST, Fashion-MNIST (FMNIST), CIFAR-10, CIFAR-100, Street-View House Numbers (SVHN), and Tiny Imagenet. We scale all pixel values to $[0, 1]$ following the preprocessing procedure in [90,

102]. For gray-scale image datasets (K/F/MNIST), we use a LeNet-5 model [66], and for color image datasets (CIFAR-10, CIFAR-100, SVHN, Tiny Imagenet), we use a VGG-13 model [125]. All these models are trained using Adam optimizer with a initial learning rate of 0.01 and a batch size of 256. For our method, we first warm up the training process for T epochs ($T = 50$ for K/F/MNIST and $T = 150$ for other datasets) using CE loss, and then train the model using our method shown in Eqs. (4.6) and (5.1) ($\xi = 0.995, \lambda = 0.05$) for another T epochs whereas we directly train the baseline using CE loss for $2T$ epochs.

| Attacks | Param. | MNIST | | | Param. | CIFAR-10 | | |
|---------|--------|-------|-------------|-------------|--------|----------|-------------|-------------|
| | | CE | GCE* | Ours | | CE | GCE* | Ours |
| FGSM | 0.1 | 71.5 | 87.7 | 80.5 | 0.04 | 12.7 | 41.2 | 58.4 |
| | 0.2 | 51.6 | 62.7 | 76.3 | 0.12 | 10.3 | 14.8 | 17.3 |
| | 0.3 | 31.8 | 47.2 | 65.0 | 0.2 | 7.0 | 11.8 | 12.0 |
| BIM | 0.1 | 52.8 | 61.9 | 72.0 | 0.04 | 0.0 | 19.6 | 16.6 |
| | 0.2 | 4.5 | 34.5 | 48.6 | 0.12 | 0.0 | 3.0 | 3.4 |
| | 0.3 | 1.5 | 33.5 | 39.5 | 0.2 | 0.0 | 2.0 | 2.6 |
| PGD | 0.1 | 49.0 | 51.9 | 72.3 | 0.04 | 0.0 | 5.9 | 10.2 |
| | 0.2 | 3.3 | 9.6 | 50.2 | 0.12 | 0.0 | 1.9 | 3.5 |
| | 0.3 | 0.8 | 2.2 | 39.7 | 0.2 | 0.0 | 1.6 | 2.7 |
| MIM | 0.1 | 49.8 | 61.2 | 73.8 | 0.04 | 0.0 | 15.4 | 16.0 |
| | 0.2 | 5.0 | 39.8 | 54.0 | 0.12 | 0.0 | 13.1 | 11.6 |
| | 0.3 | 1.5 | 38.8 | 43.3 | 0.2 | 0.0 | 12.7 | 11.2 |
| C&W | 0.0 | 0.0 | 25.6 | 30.1 | 0.0 | 0.0 | 0.8 | 3.3 |

Table 14: Accuracy (%) between GCE and our method on MNIST and CIFAR10 under white-box setting. *Results are directly from [13].

Attack types In the adversarial setting, there are two main threat models: white-box attacks where the adversary possesses complete knowledge of target model, including its architecture, training method and learned parameters, and black-box attacks where the adversary does not have access to the information about trained classifier but is aware of the classification task. We evaluate the robustness of our proposed method against both white-box and black-box attacks.

| Attacks | Param. | MNIST | | KMNIST | | FMNIST | | Param. | CIFAR-10 | | SVHN | |
|---------|--------|-------|-------------|--------|-------------|--------|-------------|--------|----------|-------------|------|-------------|
| | | CE | Ours | CE | Ours | CE | Ours | | CE | Ours | CE | Ours |
| PGD | 0.1 | 96.6 | 98.1 | 90.8 | 92.5 | 65.5 | 74.0 | 0.04 | 19.5 | 42.2 | 43.6 | 48.8 |
| | 0.2 | 84.3 | 92.8 | 76.7 | 85.0 | 50.4 | 57.1 | 0.12 | 13.2 | 38.0 | 19.5 | 28.1 |
| | 0.3 | 61.1 | 85.6 | 59.1 | 77.7 | 47.8 | 53.5 | 0.2 | 16.7 | 35.6 | 13.2 | 24.5 |
| MIM | 0.1 | 96.4 | 98.1 | 90.4 | 92.2 | 62.4 | 72.3 | 0.04 | 17.2 | 42.0 | 40.1 | 44.2 |
| | 0.2 | 84.2 | 94.7 | 74.9 | 83.0 | 43.3 | 54.2 | 0.12 | 1.3 | 16.3 | 13.3 | 21.7 |
| | 0.3 | 56.7 | 81.2 | 48.4 | 63.8 | 30.5 | 36.0 | 0.2 | 0.3 | 11.2 | 10.0 | 16.2 |
| SPSA | 0.3 | 72.9 | 95.7 | 50.2 | 78.0 | 4.3 | 39.8 | 0.3 | 0.0 | 45.3 | 4.0 | 58.0 |

Table 15: Accuracy (%) on K/F/MNIST, CIFAR-10 and SVHN under black-box setting.

| Attacks | Param. | CIFAR-100 | | | Tiny ImageNet | | |
|---------|--------|-----------|------|-------------|---------------|------|-------------|
| | | CE | GCE | Ours | CE | GCE | Ours |
| Clean | - | 40.2 | 64.5 | 67.7 | 38.2 | 32.8 | 37.7 |
| PGD | 0.005 | 11.4 | 24.4 | 56.7 | 11.3 | 8.9 | 24.8 |
| | 0.010 | 2.0 | 14.8 | 54.6 | 2.8 | 2.6 | 19.2 |
| | 0.015 | 0.4 | 9.1 | 52.6 | 0.8 | 1.0 | 15.8 |
| MIM | 0.005 | 8.7 | 21.9 | 55.8 | 7.9 | 7.3 | 23.5 |
| | 0.010 | 1.4 | 12.3 | 52.8 | 1.9 | 2.1 | 18.0 |
| | 0.015 | 0.3 | 7.4 | 48.9 | 0.6 | 1.4 | 14.6 |
| SPSA | 0.015 | 3.9 | 11.5 | 22.1 | 6.3 | 7.3 | 16.2 |

Table 16: Accuracy (%) on CIFAR-100 and Tiny ImageNet between CE loss, GCE and our new PC loss.

4.3.1 Results

Performance on white-box attacks Following the attack settings in [13], we crafted adversarial examples in a non-targeted way with respect to allowed perturbation ϵ for gradient-based attacks, i.e., FGSM, BIM, PGD and MIM. The number of iterations is set to 10 for BIM and 40 for MIM and PGD while perturbation of each step is 0.01. For parameters of optimization-based attack C&W, the maximum iteration steps are set to 100, with a learning rate of 0.001, and the confidence is set to 0.

The results (Table 13) demonstrate that our proposed PC loss with logit constraints outperforms the CE loss under white-box attacks while maintaining the comparable level of performance on the clean image classification. The improvement is even more significant

on stronger attacks.

Besides comparing to the standard CE loss, we also compare our defense approach with a closely related Guided Complement Entropy (GCE) approach [13]. To ensure a fair comparison we use the exactly same models (LeNet-5 for MNIST and ResNet56 for CIFAR-10) and parameters (max iterations of C&W is 1000) as in the GCE paper. In Table 14, it is evident that our method outperforms GCE in the vast majority of settings.

Performance on black-box attacks The performance under black-box setting is critical to substantiate adversarial robustness since it is closer to the real-world scenario where an adversary has no access to the trained classifier. During inference time, black-box adversary uses a substitute model trained on the same dataset to generate adversarial samples to attack the target model. In our cases, we use a 3-layer CNN as the substitute model for LeNet-5 and ResNet-56 for VGG-13 to generate black-box attacks. Similar to [102], we adopt PGD and MIM, the two most commonly used attack methods under the black-box setting. We then further evaluate our defense method using a gradient-free attack approach, i.e., SPSA, as in [11], which performs numerical approximation on the gradients using test data. The learning rate of SPSA is set to 0.01, and the step size is $\delta = 0.01$ [136]. As shown in Table 15, the model trained with our PC loss improves robustness against the black-box attacks.

Larger-scale experiments on CIFAR-100 and Tiny ImageNet We also evaluate our method on larger and more complex CIFAR-100 and Tiny ImageNet datasets under both white-box attacks (PGD, MIM) and black-box attack (SPSA). Similar to [102], we reduce the perturbation budget to the range of [0.005, 0.015] and attack iterations to 10 due to the increased data complexity and scale. As shown in Table 16, our method significantly im-

| Attacks | | BIM | PGD | MIM | C&W | SPSA |
|----------|---------|-------------|-------------|-------------|-------------|-------------|
| | Param. | 0.3 | 0.3 | 0.3 | 0.0 | 0.3 |
| MNIST | CE+AT | 27.0 | 3.2 | 10.8 | 75.5 | 77.3 |
| | GCE+AT | 28.3 | 26.8 | 27.7 | 69.4 | 56.6 |
| | Ours+AT | 86.1 | 72.3 | 78.8 | 96.4 | 97.1 |
| | Ours | 39.5 | 39.7 | 43.3 | 78.0 | 95.7 |
| KMNIST | CE+AT | 58.4 | 37.2 | 22.3 | 47.9 | 72.0 |
| | GCE+AT | 8.1 | 0.5 | 15.7 | 48.2 | 69.9 |
| | Ours+AT | 65.9 | 48.7 | 51.9 | 67.1 | 79.0 |
| | Ours | 60.5 | 33.2 | 36.9 | 57.3 | 78.0 |
| FMNIST | CE+AT | 0.1 | 0.0 | 0.0 | 4.0 | 14.2 |
| | GCE+AT | 4.8 | 2.7 | 1.4 | 21.5 | 30.0 |
| | Ours+AT | 18.6 | 13.4 | 9.1 | 29.8 | 41.6 |
| | Ours | 4.9 | 2.8 | 2.2 | 21.8 | 39.8 |
| Param. | | 0.04 | 0.04 | 0.04 | 0.0 | 0.3 |
| CIFAR-10 | CE+AT | 17.3 | 11.4 | 9.0 | 0.0 | 6.5 |
| | Ours+AT | 38.7 | 33.7 | 33.3 | 33.7 | 39.6 |

Table 17: Accuracy (%) on K/F/MNIST and CIFAR-10 with adversarial training under both white- and black-box attacks.

proves the model’s adversarial robustness compared to the CE loss and GCE while maintaining the comparable level of performance on the clean image classification. Recall our observation that the most probable false classes are more vulnerable to attacks. GCE flattens the probabilities on false classes and thus enlarges the gap between true class and the most probable false class to increase model’s robustness. However, when dataset become complex with more classes, this gap is smaller due to generally lower output probability for the true class, resulting a limited robustness improvement. On the other hand, our method directly maximizes the probability gap and thus is more suitable for large scale datasets.

Combining with adversarial training To demonstrate our method’s compatibility and synergy with other adversarial defense techniques, we investigate the performance of our method in combination with adversarial training. Our goal is not to beat adversarial training, instead we attempt to show our method can be combined with it to further improve

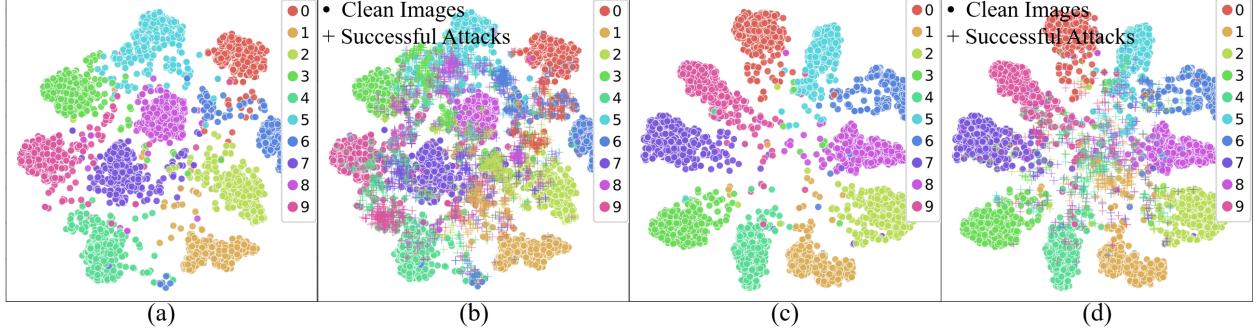


Figure 21: T-SNE visualization of the penultimate layer of the model trained by CE loss (a,b) and our PC loss (c,d) on MNIST dataset. (a,c) display only clean images whereas (b,d) also include successful attacks generated with FGSM ($\epsilon = 0.3$).

adversarial robustness. During training, we augment the dataset with adversarial samples generated using FGSM with perturbation range of $[0.1, 0.3]$ for gray-scale image datasets (K/F/MNIST), and 5-step PGD with perturbation range of $[0.0, 0.1]$ for color image dataset (CIFAR-10). The ratio of adversarial examples and clean images in each training mini-batch remains 1 : 1. For gray-scale image datasets, table 17 shows that integrating our method with adversarial training further improves the model’s adversarial robustness under both white-box (PGD, BIM, MIM, CW) and black-box attack settings (SPSA). Furthermore, our PC loss with adversarial training outperforms GCE with adversarial training, which demonstrates our method has better compatibility with other defense techniques. It is worth mentioning that our method alone outperforms the fast version adversarial training on gray-scale datasets, which generates adversarial training examples by one-step FGSM attack. And for color image dataset (CIFAR-10), we augment the dataset with adversarial examples crafted by more advanced PGD attacks. The result shows the same trend, and the performance gain is more pronounced on this more challenging dataset.

Feature Space Visualization In order to visually dissect the advantages of PC loss over the CE loss, we also inspect the feature space of trained models using t-SNE on MNIST

datasets. As shown in Figure 21a, the model trained with CE loss has a large portion of clean images lay across the boundaries between different classes thus easily to be manipulated to become adversarial samples. On the contrary, for the model trained with our PC loss with logit constraints, in Figure 21c, the samples of each class have clear boundaries and are evenly distributed around the center with a minimal overlap. Note that the samples locate near the center are ‘hard samples’ for a classifier even without attacks.

Looking into the successful attacks (labeled with ‘+’) in Figure 21, we find the predictive behavior of CNN on adversarial samples is consistent to our hypothesis. In Figure 21b, for a model (LeNet-5) trained with CE loss, adversarial samples are mostly located to the nearest classes corresponding to the most probable false classes. For example, many adversarial attacks generated based on class 5 are located within the class 8 of clean images and *vice versa*. While in Figure 21d, for a model trained with PC loss, due to the large margin between classes, the adversarial samples are harder to cross the boundaries with the only exception that the adversarial samples are distributed near the center of the feature space where hard samples are usually located.

4.3.2 Identifying Gradient Masking

Previous defense strategies [9, 148] rely on the effect of gradient masking, which was considered as a false sense of security [4]. Briefly, these defenses deteriorate the gradient information to make gradient-based attack methods hard to generate effective adversarial examples. However, these defenses can be easily defeated by black-box or gradient-free attackers. We show that our method does not rely on gradient masking on the basis of characteristics defined in [4, 11]. (1) Iterative attacks have better performance than one-step attack: Our results in Table 13 indicate that the iteration-based attacks (BIM, MIM, PGD)

are more successful in generating adversarial attacks than single step method (FGSM). (2) Robustness against Black-box attacks is higher than white-box attacks: When model’s gradients information is manipulated by the defender, the attacker can recover the gradient with black-box attacks and perform more successful attacks than using white-box attacks [103]. However, the results in Tables 13 & 15 demonstrate that our method is more effective against black-box attacks and thus does not obfuscate gradients. (3) Increasing perturbation budget will increase attack success: As shown in the Table 13, increase of perturbations monotonically enhances the attacks. With a large budget ($\epsilon = 0.3$), the success rate is close to 100%.

4.4 Conclusion

We propose a novel PC loss with logit constraints inspired by the predictive behavior of CNN on adversarial samples. A CNN trained with our PC loss can achieve impressive robustness against adversarial samples without compromising performance on clean images nor requires additional procedures/computing, making it scalable to large-scale datasets. In addition, our PC loss is flexible and compatible with other defense methods, e.g., as a drop-in replacement of CE loss to supervise adversarial training. In future work, we plan to extensively investigate the connection of predictions between adversarial and clean samples in more general settings.

CHAPTER 5 SUMMARY AND FUTURE WORK

5.1 Summary

In this dissertation, we introduce novel DNN based medical imaging AI systems, defense techniques against adversarial attacks that tailor-made for those systems and a new loss function to improve the adversarial robustness of general DNN based classification models. Beyond those attacks and defenses game, we intend to use the adversarial machine learning technique to tackle more fundamental problems in machine learning such as feature representation learning and generalization gap in the future work.

5.2 Future Work

Learning good feature representation that generalizes well to OOD test sets is a central challenge in machine learning. Recently, DNN has demonstrated impressive performance in classification tasks on IID test sets [72]. Model regularization techniques, e.g., those based on parameter sparsity and loss function smoothing, used in conjunction with adversarial training, have been proven effective on mitigating *robust overfitting* [112] on IID test sets. Nevertheless, the performance degradation on OOD test sets remains a salient problem [122]. One observation is that the current approach introduces a nearly ideal scenario for DNN to learn spurious shortcuts or non-relevant features [40] that do not exist in OOD test sets. In medical imaging systems, the problem becomes even more salient due to the significant distribution shift between imaging data sets acquired from different hospitals, populations, and time periods. As a result, the AI imaging system that is seemingly effective on training sets often does not generalize well to new hospitals or data sets [27].

Recent studies [88, 21] demonstrate that CXR classification systems might depend more

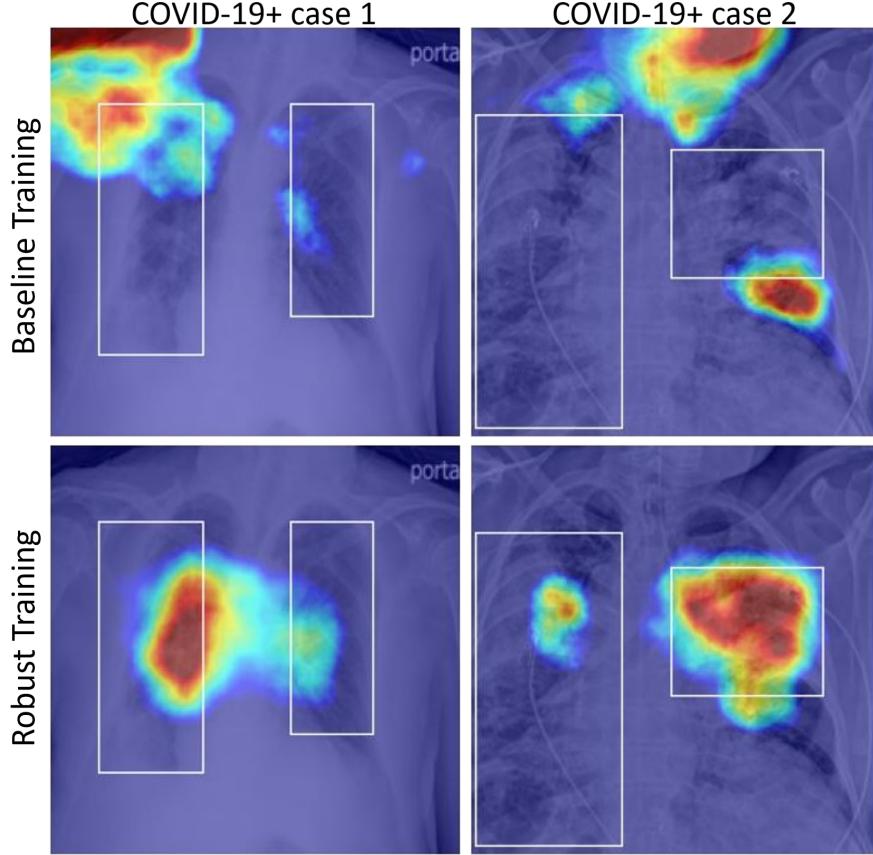


Figure 22: Examples to illustrate the shortcut features (top left) and non-relevant features (top right). Good features are highlighted with high salience in the second rows, overlapping with radiologists’ annotations. The heatmap based DNN interpretations are generated by FullGrad [128].

on nuisance features generated by different medical devices with various manufacturing standards and acquisition parameters. Similar to adversarial perturbation, those nuisance features do not impede human recognition but are obvious to DNN models, particularly when they lay on extremely clean background around the CXR borders [76]. As shown by case 1 (top left in Figure 22), model using those shortcut features would have a poor generalization on OOD test sets. To tackle this challenge, we hypothesize that adversarial training [87] can eliminate those shortcut features since adversarial perturbation are also imperceptible and usually considered the worst case noise. For the non-relevant features

(top right in Figure 22), since a recent study [57] demonstrates that masking noisy gradients can improve model’s interpretability, we argue that it also helps prevent the model from extracting non-relevant features. As a promising future direction, we will leverage model interpretability constraints on adversarial training to learn good features that ensure generalization performance on OOD test sets. To initially test this idea, a preliminary experiment is conducted with the following loss function:

$$\frac{1}{n} \sum_{i=1}^n [\mathcal{L}(f_\theta(X_i), y_i) + \lambda D_{KL}(f_\theta(X_i) || f_\theta(X'_i))], \quad (5.1)$$

where the first term is the standard cross-entropy classification loss and second term is the KL divergence between the output of model on clean input X_i and it corresponding adversarial masked input X'_i . As shown in the Fig. 22, the robustly trained model (bottom row) has a more visually coherent feature focus on the radiologists’ annotations which shows a significant potential to tackle the failure of medical imaging system in new hospitals or on new test datasets. A preliminary version of this work has been accepted for publication at AdvML workshop @ICML-2022 [75].

Overall, adversarial machine learning is about more than just attack and defense; it also helps models learn robust features that are semantic and faithful. As a result, it is a an essential component of creating trustworthy applications in the real-world.

CHAPTER 6 APPENDIX

List of Publications and Preprints

Total Citations 261, H-index: 7, I10-index: 7, Google Scholar Link (06/12/2022)

- Li, X., Li, X., Pan, D. and Zhu, D. "Learning Compact Features via In-Training Representation Alignment." Submitted to Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS-2022).
- Li, X., Bagher-Ebadian, H., Gardner, S., Kim, J. Mohamed, E., Movsas, B., Zhu, D. and Chetty, I.J. "An uncertainty-aware deep learning architecture with outlier mitigation for prostate gland segmentation in radiotherapy treatment planning." Submitted to Medical Physics.
- Li, X., Qiang, Y. Li, C., Liu, S. and Zhu, D. "Saliency guided adversarial training for tackling generalization gap with applications to medical imaging classification system." In new frontiers in adversarial machine learning workshop at ICML, 2022
- Pan, D., Li, X. and Zhu, D. 2021. "Explaining Deep Neural Network Models with Adversarial Gradient Integration." In the proceedings of 30th International Joint Conference on Artificial Intelligence (IJCAI-21), Montreal, Canada.
- Li, X., Bagher-Ebadian, H., Mohamed, E., Movsas, B., Zhu, D. and Chetty, I.J., 2021, June. "On the Application of a Variational Autoencoder (VAE) and Transfer Learning to Account for Inter-Observer Uncertainties in Automatic Prostate Gland Segmentation." Medical Physics, 48(6).

- Li, X., Pan, D. and Zhu, D. 2021. “Defending against adversarial attacks on medical imaging AI system, classification or detection?” In the proceedings of IEEE International Symposium on Biomedical Imaging (ISBI-21), virtual conference.
- Li, X., Li, X., Pan, D. and Zhu, D. 2021. “Improving adversarial robustness via probabilistically compact loss with logit constraints.” In the proceedings of Thirty-Five AAAI Conference on Artificial Intelligence (AAAI-21), virtual conference.
- Manwar, R., Li, X., Mahmoodkalayeh, S., Asano, E., Zhu, D. and Avanaki, K. 2020. “Deep learning protocol for improved photoacoustic brain imaging.” Journal of Biophotonics, 13(10), p.e202000212.
- Li, X., Li, C. and Zhu, D. 2020. “COVID-MobileXpert: On-device COVID-19 patient triage and follow-up using chest X-rays.” In 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 1063-1067.
- Qiang, Y., Li, X. and Zhu, D. 2020. “Toward tag-free aspect based sentiment analysis: a multiple attention network approach.” In the proceedings of International Joint Conference on Neural Networks (IJCNN-20), Glasgow, Scotland, UK.
- Pan, D., Li, X., Li, X. and Zhu, D. 2020. “Explainable recommendation via interpretable feature mapping and evaluating explainability.” In the proceedings of 29th International Joint Conference on Artificial Intelligence (IJCAI-20), Yokohama, Japan.
- Li, X., Bagher-Ebadian, H., Li, C., Mohamed, E., Siddiqui, F., Movsas, B., Zhu, D. and Chetty, I. 2020. “Automatic Segmentation of the Prostate on CT Images Using

a Bi-Directional Convolutional LSTM U-Net with Novel Loss Function.” In Medical Physics, 47(6), E584-E584.

- Li, X., Li, X., Pan, D. and Zhu, D. 2020. “On the learning behavior of logistic and softmax losses for deep neural networks.” In the proceedings of Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20), New York, USA.
- Li, X. and Zhu, D. 2020. “Robust detection of adversarial attacks on medical images.” IEEE International Symposium on Biomedical Imaging (ISBI-20), Iowa City, USA.
- Li, X., Pan, D, Li, X. and Zhu, D. 2020. “Improve SGD Training via Aligning Min-batches.” arXiv:2002.09917 [cs.LG].
- Li, X., Cao, R., and Zhu, D. 2020. “Vispi: Automatic visual perception and interpretation of chest X-rays.” In the proceedings of the Medical Imaging with Deep Learning (MIDL-20) conference, Montreal, CA.

REFERENCES

- [1] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Parikh, and D. Batra. Vqa: Visual question answering. *Int. J. Comput. Vision*, 123(1):4–31, May 2017.
- [2] M. Al-Qizwini, I. Barjasteh, H. Al-Qassab, and H. Radha. Deep learning algorithm for autonomous driving using googlenet. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 89–96. IEEE, 2017.
- [3] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.
- [4] A. Athalye, N. Carlini, and D. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018.
- [5] A. Balagopal, S. Kazemifar, D. Nguyen, M.-H. Lin, R. Hannan, A. Owrangi, and S. Jiang. Fully automated organ segmentation in male pelvic ct images. *Physics in Medicine & Biology*, 63(24):245015, 2018.
- [6] A. Balagopal, H. Morgan, M. Dohopolski, R. Timmerman, J. Shan, D. F. Heitjan, W. Liu, D. Nguyen, R. Hannan, A. Garant, et al. Psa-net: Deep learning-based physician style-aware segmentation network for postoperative prostate cancer clinical target volumes. *Artificial Intelligence in Medicine*, 121:102195, 2021.
- [7] A. Balagopal, D. Nguyen, H. Morgan, Y. Weng, M. Dohopolski, M.-H. Lin, A. S. Barkousaraie, Y. Gonzalez, A. Garant, N. Desai, et al. A deep learning-based framework for segmenting invisible clinical target volumes with estimated uncertainties

for post-operative prostate cancer radiotherapy. *Medical image analysis*, 72:102101, 2021.

- [8] C. F. Baumgartner, K. C. Tezcan, K. Chaitanya, A. M. Hötker, U. J. Muehlematter, K. Schawkat, A. S. Becker, O. Donati, and E. Konukoglu. Phiseg: Capturing uncertainty in medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 119–127. Springer, 2019.
- [9] J. Buckman, A. Roy, C. Raffel, and I. Goodfellow. Thermometer encoding: One hot way to resist adversarial examples. In *The International Conference on Learning Representations (ICLR)*, 2018.
- [10] C. E. Cardenas, J. Yang, B. M. Anderson, L. E. Court, and K. B. Brock. Advances in auto-segmentation. In *Seminars in radiation oncology*, volume 29, pages 185–197. Elsevier, 2019.
- [11] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. Goodfellow, A. Madry, and A. Kurakin. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019.
- [12] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017.
- [13] H.-Y. Chen, J.-H. Liang, S.-C. Chang, J.-Y. Pan, Y.-T. Chen, W. Wei, and D.-C. Juan. Improving adversarial robustness via guided complement entropy. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4881–4889, 2019.
- [14] H.-Y. Chen, P.-H. Wang, C.-H. Liu, S.-C. Chang, J.-Y. Pan, Y.-T. Chen, W. Wei, and D.-C. Juan. Complement objective training. *arXiv preprint arXiv:1903.01182*, 2019.

- [15] J. Chen, Y. Li, X. Wu, Y. Liang, and S. Jha. Robust out-of-distribution detection via informative outlier mining. *CoRR*, abs/2006.15207, 2020.
- [16] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.
- [17] X. Chen, S. Sun, N. Bai, K. Han, Q. Liu, S. Yao, H. Tang, C. Zhang, Z. Lu, Q. Huang, et al. A deep learning-based auto-segmentation system for organs-at-risk on whole-body computed tomography images for radiation therapy. *Radiotherapy and Oncology*, 160:175–184, 2021.
- [18] Y. Chen, H. Fan, B. Xu, Z. Yan, Y. Kalantidis, M. Rohrbach, S. Yan, and J. Feng. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3435–3444, 2019.
- [19] M. E. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M. A. Kadir, Z. B. Mahbub, K. R. Islam, M. S. Khan, A. Iqbal, N. Al-Emadi, et al. Can ai help in screening viral and covid-19 pneumonia? *arXiv preprint arXiv:2003.13145*, 2020.
- [20] J. P. Cohen, L. Dao, P. Morrison, K. Roth, Y. Bengio, B. Shen, A. Abbasi, M. Hoshmand-Kochi, M. Ghassemi, H. Li, et al. Predicting covid-19 pneumonia severity on chest x-ray with deep learning. *arXiv preprint arXiv:2005.11856*, 2020.
- [21] J. P. Cohen, M. Hashir, R. Brooks, and H. Bertrand. On the limits of cross-domain generalization in automated x-ray prediction. In *Medical Imaging with Deep Learning*, pages 136–155. PMLR, 2020.

- [22] J. P. Cohen, P. Morrison, and L. Dao. Covid-19 image data collection. *arXiv preprint arXiv:2003.11597*, 2020.
- [23] F. Croce, M. Andriushchenko, V. Sehwag, E. Debenedetti, N. Flammarion, M. Chiang, P. Mittal, and M. Hein. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*, 2020.
- [24] F. Croce and M. Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020.
- [25] Z. A. Daniels and D. N. Metaxas. Exploiting visual and report-based information for chest x-ray analysis by jointly learning visual classifiers and topic models. In *ISBI*, 2019.
- [26] N. Das, M. Shanbhogue, S.-T. Chen, F. Hohman, L. Chen, M. E. Kounavis, and D. H. Chau. Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression. *arXiv preprint arXiv:1705.02900*, 2017.
- [27] A. J. DeGrave, J. D. Janizek, and S.-I. Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021.
- [28] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *JAMIA*, 23(2):304–310, 2015.
- [29] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

- [30] S. Dhar, J. Guo, J. Liu, S. Tripathi, U. Kurup, and M. Shah. On-device machine learning: An algorithms and learning theory perspective. *arXiv preprint arXiv:1911.00623*, 2019.
- [31] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *ICCV*, pages 2625–2634, 2015.
- [32] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018.
- [33] C. Draulans, U. A. van der Heide, K. Haustermans, F. J. Pos, J. v. d. V. van Zyp, H. De Boer, V. H. Groen, E. M. Monninkhof, R. J. Smeenk, M. Kunze-Busch, et al. Primary endpoint analysis of the multicentre phase ii hypo-flame trial for intermediate and high risk prostate cancer. *Radiotherapy and Oncology*, 147:92–98, 2020.
- [34] S. Duchesne, D. Gourdeau, P. Archambault, C. Chartrand-Lefebvre, L. Dieumegarde, R. Forghani, C. Gagne, A. Hains, D. Hornstein, H. Le, et al. Tracking and predicting covid-19 radiological trajectory using deep learning on chest x-rays: Initial accuracy testing. *medRxiv*, 2020.
- [35] T. Elsken, J. H. Metzen, and F. Hutter. Neural architecture search: A survey. *arXiv preprint arXiv:1808.05377*, 2018.
- [36] M. Fazlyab, A. Robey, H. Hassani, M. Morari, and G. Pappas. Efficient and accurate estimation of lipschitz constants for deep neural networks. In *Advances in Neural Information Processing Systems*, pages 11423–11434, 2019.

- [37] R. Feinman, R. R. Curtin, S. Shintre, and A. B. Gardner. Detecting adversarial samples from artifacts. *arXiv preprint arXiv:1703.00410*, 2017.
- [38] S. G. Finlayson et al. Adversarial attacks against medical deep learning systems. *arXiv preprint arXiv:1804.05296*, 2018.
- [39] S. J. Gardner, N. Wen, J. Kim, C. Liu, D. Pradhan, I. Aref, R. Cattaneo, S. Vance, B. Movsas, I. J. Chetty, et al. Contouring variability of human-and deformable-generated contours in radiotherapy for prostate cancer. *Physics in Medicine & Biology*, 60(11):4429, 2015.
- [40] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.
- [41] M. Goldblum, L. Fowl, S. Feizi, and T. Goldstein. Adversarially robust distillation. *arXiv preprint arXiv:1905.09747*, 2019.
- [42] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- [43] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.
- [44] L. O. Hall, R. Paul, D. B. Goldgof, and G. M. Goldgof. Finding covid-19 from chest x-rays using deep learning on a small dataset, 2020.
- [45] S. Han, H. Mao, and W. J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

- [46] A. E. Hassanien, L. N. Mahdy, K. A. Ezzat, H. H. Elmousalami, and H. A. Ella. Automatic x-ray covid-19 lung image classification system based on multi-level thresholding and support vector machine. *medRxiv*, 2020.
- [47] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.
- [48] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.
- [49] X. He et al. Non-local context encoder: Robust biomedical image segmentation against adversarial attacks. In *AAAI*, volume 33, pages 8417–8424, 2019.
- [50] D. Hendrycks and K. Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.
- [51] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
- [52] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1314–1324, 2019.
- [53] G. Huang, S. Liu, L. Van der Maaten, and K. Q. Weinberger. Condensenet: An efficient densenet using learned group convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2752–2761, 2018.
- [54] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *CVPR*, pages 4700–4708, 2017.
- [55] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition*

- tion (CVPR), Jul 2017.*
- [56] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer. SqueezeNet: Alexnet-level accuracy with 50x fewer parameters and < 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.
- [57] A. A. Ismail, H. Corrada Bravo, and S. Feizi. Improving deep learning interpretability by saliency guided training. *Advances in Neural Information Processing Systems*, 34, 2021.
- [58] B. Jing, P. Xie, and E. Xing. On the automatic generation of medical imaging reports. *arXiv preprint arXiv:1711.08195*, 2017.
- [59] S. Kazemifar, A. Balagopal, D. Nguyen, S. McGuire, R. Hannan, S. Jiang, and A. Owranghi. Segmentation of the prostate and organs at risk in male pelvic ct images using deep learning. *Biomedical Physics & Engineering Express*, 4(5):055003, 2018.
- [60] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKown, G. Yang, X. Wu, F. Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018.
- [61] N. E. M. Khalifa, M. H. N. Taha, A. E. Hassanien, and S. Elghamrawy. Detection of coronavirus (covid-19) associated pneumonia based on generative adversarial networks and a fine-tuned deep transfer learning model using chest x-ray dataset. *arXiv preprint arXiv:2004.01184*, 2020.
- [62] A. M. King, S. Danagoulian, M. Lynch, N. Menke, Y. Mu, M. Saul, M. Abesamis, and A. F. Pizon. The effect of a medical toxicology inpatient service in an academic tertiary care referral center. *Journal of Medical Toxicology*, 15(1):12–21, 2019.

- [63] S. Kohl, B. Romera-Paredes, C. Meyer, J. De Fauw, J. R. Ledsam, K. Maier-Hein, S. Eslami, D. Jimenez Rezende, and O. Ronneberger. A probabilistic u-net for segmentation of ambiguous images. *Advances in neural information processing systems*, 31, 2018.
- [64] J. Krause, J. Johnson, R. Krishna, and L. Fei-Fei. A hierarchical approach for generating descriptive image paragraphs. In *CVPR*, pages 3337–3345. IEEE, 2017.
- [65] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.
- [66] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [67] K. Lee, H. Lee, K. Lee, and J. Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*, 2017.
- [68] C. Y. Li, X. Liang, Z. Hu, and E. P. Xing. Hybrid retrieval-generation reinforced agent for medical image report generation. *arXiv preprint arXiv:1805.08298*, 2018.
- [69] X. Li, H. Bagher-Ebadian, S. J. Gardner, M. A. Elshaikh, B. Movsas, D. Zhu, and I. J. Chetty. An uncertainty-aware deep learning architecture with outlier mitigation for prostate gland segmentation in radiotherapy treatment planning. *Submitted to Medical physics*.
- [70] X. Li, R. Cao, and D. Zhu. Vispi: Automatic visual perception and interpretation of chest x-rays. *arXiv preprint arXiv:1906.05190*, 2019.
- [71] X. Li, C. Li, and D. Zhu. Covid-mobileexpert: On-device covid-19 screening using snapshots of chest x-ray. *arXiv preprint arXiv:2004.03042*, 2020.

- [72] X. Li, X. Li, D. Pan, and D. Zhu. Improving adversarial robustness via probabilistically compact loss with logit constraints. *arXiv preprint arXiv:2012.07688*, 2020.
- [73] X. Li, X. Li, D. Pan, and D. Zhu. On the learning property of logistic and softmax losses for deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4739–4746, 2020.
- [74] X. Li, D. Pan, and D. Zhu. Defending against adversarial attacks on medical imaging ai system, classification or detection? *arXiv preprint arXiv:2006.13555*, 2020.
- [75] X. Li, Y. Qiang, C. Li, S. Liu, and D. Zhu. Saliency guided adversarial training for tackling generalization gap with applications to medical imaging classification system. In *new frontiers in adversarial machine learning workshop at ICML*, 2022.
- [76] X. Li and D. Zhu. Robust detection of adversarial attacks on medical images. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1154–1158. IEEE, 2020.
- [77] S. Liang, Y. Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.
- [78] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004.
- [79] C. Liu, S. J. Gardner, N. Wen, M. A. Elshaikh, F. Siddiqui, B. Movsas, and I. J. Chetty. Automatic segmentation of the prostate on ct images using deep neural networks (dnn). *International Journal of Radiation Oncology* Biology* Physics*, 104(4):924–932, 2019.
- [80] F. T. Liu, K. M. Ting, and Z.-H. Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422. IEEE, 2008.

- [81] X. Liu, L. Song, S. Liu, and Y. Zhang. A review of deep-learning-based medical image segmentation methods. *Sustainability*, 13(3):1224, 2021.
- [82] Q. Lou, F. Guo, M. Kim, L. Liu, and L. Jiang. Autoq: Automated kernel-wise neural network quantization. In *International Conference on Learning Representations*, 2020.
- [83] J. Lu, C. Xiong, D. Parikh, and R. Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*, volume 6, page 2, 2017.
- [84] D. Lv, W. Qi, Y. Li, L. Sun, and Y. Wang. A cascade network for detecting covid-19 using chest x-rays, 2020.
- [85] X. Ma, B. Li, Y. Wang, S. M. Erfani, S. Wijewickrema, G. Schoenebeck, D. Song, M. E. Houle, and J. Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. *arXiv preprint arXiv:1801.02613*, 2018.
- [86] X. Ma, Y. Niu, L. Gu, Y. Wang, Y. Zhao, J. Bailey, and F. Lu. Understanding adversarial attacks on deep learning based medical image analysis systems. *arXiv preprint arXiv:1907.10456*, 2019.
- [87] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.
- [88] G. Maguolo and L. Nanni. A critic evaluation of methods for covid-19 automatic detection from x-ray images. *Information Fusion*, 2021.
- [89] D. Meng and H. Chen. Magnet: a two-pronged defense against adversarial examples. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 135–147, 2017.

- [90] A. Mustafa, S. Khan, M. Hayat, R. Goecke, J. Shen, and L. Shao. Adversarial defense by restricting the hidden space of deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3385–3394, 2019.
- [91] A. Mustafa, S. H. Khan, M. Hayat, J. Shen, and L. Shao. Image super-resolution as a defense against adversarial attacks. *IEEE Transactions on Image Processing*, 29:1711–1724, 2019.
- [92] E. Nalisnick, A. Matsukawa, Y. W. Teh, D. Gorur, and B. Lakshminarayanan. Do deep generative models know what they don’t know? *arXiv preprint arXiv:1810.09136*, 2018.
- [93] E. Nalisnick, A. Matsukawa, Y. W. Teh, D. Gorur, and B. Lakshminarayanan. Hybrid models with deep and invertible features. In *International Conference on Machine Learning*, pages 4723–4732. PMLR, 2019.
- [94] B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, pages 5947–5956, 2017.
- [95] A. Noack, I. Ahern, D. Dou, and B. Li. Does interpretability of neural networks imply adversarial robustness? *CoRR*, abs/1912.03430, 2019.
- [96] R. S. of North America. RSNA pneumonia detection challenge. <https://www.kaggle.com/c/rsna-pneumonia-detection-challenge>, 2018.
- [97] U. Ozbulak, A. Van Messem, and W. De Neve. Impact of adversarial examples on deep learning models for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 300–308. Springer, 2019.

- [98] D. Pan, X. Li, X. Li, and D. Zhu. Explainable recommendation via interpretable feature mapping and evaluation of explainability. In C. Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 2690–2696. ijcai.org, 2020.
- [99] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.
- [100] T. Pang, C. Du, and J. Zhu. Max-mahalanobis linear discriminant analysis networks. *arXiv preprint arXiv:1802.09308*, 2018.
- [101] T. Pang, K. Xu, Y. Dong, C. Du, N. Chen, and J. Zhu. Rethinking softmax cross-entropy loss for adversarial robustness. *arXiv preprint arXiv:1905.10626*, 2019.
- [102] T. Pang, K. Xu, C. Du, N. Chen, and J. Zhu. Improving adversarial robustness via promoting ensemble diversity. *arXiv preprint arXiv:1901.08846*, 2019.
- [103] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519, 2017.
- [104] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318. Association for Computational Linguistics, 2002.
- [105] M. Paschali et al. Generalizability vs. robustness: adversarial examples for medical imaging. *arXiv preprint arXiv:1804.00504*, 2018.
- [106] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau,

- M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [107] M. Phuong and C. Lampert. Towards understanding knowledge distillation. In *International Conference on Machine Learning*, pages 5142–5151, 2019.
- [108] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.
- [109] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi. Xnor-net: Imagenet classification using binary convolutional neural networks. In *European conference on computer vision*, pages 525–542. Springer, 2016.
- [110] I. Reda et al. A new cnn-based system for early diagnosis of prostate cancer. In *ISBI*, pages 207–210. IEEE, 2018.
- [111] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. Self-critical sequence training for image captioning. In *CVPR*, pages 7008–7024, 2017.
- [112] L. Rice, E. Wong, and Z. Kolter. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, pages 8093–8104. PMLR, 2020.
- [113] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- [114] L. C. ROUGE. A package for automatic evaluation of summaries. In *Proceedings of Workshop on Text Summarization of ACL, Spain*, 2004.
- [115] T. Saito and M. Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3),

2015.

- [116] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.
- [117] B. Schölkopf, R. C. Williamson, A. J. Smola, J. Shawe-Taylor, and J. C. Platt. Support vector method for novelty detection. In *Advances in neural information processing systems*, pages 582–588, 2000.
- [118] V. Sehwag, A. N. Bhagoji, L. Song, C. Sitawarin, D. Cullina, M. Chiang, and P. Mittal. Analyzing the robustness of open-world machine learning. In *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security*, pages 105–116, 2019.
- [119] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.
- [120] A. Shafahi, M. Najibi, M. A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L. S. Davis, G. Taylor, and T. Goldstein. Adversarial training for free! In *Advances in Neural Information Processing Systems*, pages 3353–3364, 2019.
- [121] A. Shaffie et al. Radiomic-based framework for early diagnosis of lung cancer. In *ISBI*, pages 1293–1297. IEEE, 2019.
- [122] R. Shao, P. Perera, P. C. Yuen, and V. M. Patel. Open-set adversarial defense. *arXiv preprint arXiv:2009.00814*, 2020.
- [123] G. Sharp, K. D. Fritscher, V. Pekar, M. Peroni, N. Shusharina, H. Veeraraghavan, and J. Yang. Vision 20/20: perspectives on automated image segmentation for

- radiotherapy. *Medical physics*, 41(5):050902, 2014.
- [124] A. Signoroni, M. Savardi, S. Benini, N. Adami, R. Leonardi, P. Gibellini, F. Vaccher, M. Ravanelli, A. Borghesi, R. Maroldi, et al. End-to-end learning for semiquantitative rating of covid-19 severity on chest x-rays. *arXiv preprint arXiv:2006.04603*, 2020.
- [125] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [126] A. Sinha, H. Namkoong, and J. Duchi. Certifiable distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2, 2017.
- [127] G. Sreenu and M. S. Durai. Intelligent video surveillance: a review through deep learning techniques for crowd analysis. *Journal of Big Data*, 6(1):48, 2019.
- [128] S. Srinivas and F. Fleuret. Full-gradient representation for neural network visualization. *Advances in neural information processing systems*, 32, 2019.
- [129] R. Stanforth, A. Fawzi, P. Kohli, et al. Are labels required for improving adversarial robustness? *arXiv preprint arXiv:1905.13725*, 2019.
- [130] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.
- [131] S. A. Taghanaki, K. Abhishek, S. Azizi, and G. Hamarneh. A kernelized manifold mapping to diminish the effect of adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11340–11349, 2019.

- [132] M. Tan, B. Chen, R. Pang, V. Vasudevan, M. Sandler, A. Howard, and Q. V. Le. Mnasnet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2820–2828, 2019.
- [133] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.
- [134] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.
- [135] F. Tung and G. Mori. Clip-q: Deep network compression learning by in-parallel pruning-quantization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7873–7882, 2018.
- [136] J. Uesato, B. O’Donoghue, A. v. d. Oord, and P. Kohli. Adversarial risk and the dangers of evaluating against weak attacks. *arXiv preprint arXiv:1802.05666*, 2018.
- [137] A. Van Opbroek, M. A. Ikram, M. W. Vernooij, and M. De Bruijne. Transfer learning improves supervised image segmentation across imaging protocols. *IEEE transactions on medical imaging*, 34(5):1018–1030, 2014.
- [138] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, pages 3156–3164, 2015.
- [139] 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.

- [140] N. Wang, J. Choi, D. Brand, C.-Y. Chen, and K. Gopalakrishnan. Training deep neural networks with 8-bit floating point numbers. In *Advances in neural information processing systems*, pages 7675–7684, 2018.
- [141] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.
- [142] X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9049–9058, 2018.
- [143] X. Wei, H. Wang, B. Scotney, and H. Wan. Minimum margin loss for deep face recognition. *Pattern Recognition*, 97:107012, 2020.
- [144] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016.
- [145] J. Winkens, R. Bunel, A. G. Roy, R. Stanforth, V. Natarajan, J. R. Ledsam, P. MacWilliams, P. Kohli, A. Karthikesalingam, S. Kohl, et al. Contrastive training for improved out-of-distribution detection. *arXiv preprint arXiv:2007.05566*, 2020.
- [146] H. Y. F. Wong, H. Y. S. Lam, A. H.-T. Fong, S. T. Leung, T. W.-Y. Chin, C. S. Y. Lo, M. M.-S. Lui, J. C. Y. Lee, K. W.-H. Chiu, T. Chung, et al. Frequency and distribution of chest radiographic findings in covid-19 positive patients. *Radiology*, page 201160, 2020.

- [147] C. Xie et al. Feature denoising for improving adversarial robustness. In *CVPR*, pages 501–509, 2019.
- [148] C. Xie, J. Wang, Z. Zhang, Z. Ren, and A. Yuille. Mitigating adversarial effects through randomization. *arXiv preprint arXiv:1711.01991*, 2017.
- [149] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pages 2048–2057, 2015.
- [150] Y. Xue, T. Xu, L. R. Long, Z. Xue, S. Antani, G. R. Thoma, and X. Huang. Multi-modal recurrent model with attention for automated radiology report generation. In *MICCAI*, pages 457–466. Springer, 2018.
- [151] J. Yang, K. Zhou, Y. Li, and Z. Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021.
- [152] C.-F. Yeh, H.-T. Cheng, A. Wei, H.-M. Chen, P.-C. Kuo, K.-C. Liu, M.-C. Ko, R.-J. Chen, P.-C. Lee, J.-H. Chuang, C.-M. Chen, Y.-C. Chen, W.-J. Lee, N. Chien, J.-Y. Chen, Y.-S. Huang, Y.-C. Chang, Y.-C. Huang, N.-K. Chou, K.-H. Chao, Y.-C. Tu, Y.-C. Chang, and T.-L. Liu. A cascaded learning strategy for robust covid-19 pneumonia chest x-ray screening, 2020.
- [153] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.
- [154] D. Zhang, T. Zhang, Y. Lu, Z. Zhu, and B. Dong. You only propagate once: Accelerating adversarial training via maximal principle. In *Advances in Neural Information Processing Systems*, pages 227–238, 2019.

- [155] H. Zhang, A. Li, J. Guo, and Y. Guo. Hybrid models for open set recognition. In *European Conference on Computer Vision*, pages 102–117. Springer, 2020.
- [156] J. Zhang, Y. Xie, Z. Liao, G. Pang, J. Verjans, W. Li, Z. Sun, J. He, Y. Li, C. Shen, and Y. Xia. Viral pneumonia screening on chest x-ray images using confidence-aware anomaly detection, 2020.
- [157] X. Zhang, X. Zhou, M. Lin, and J. Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856, 2018.
- [158] Z. Zheng and P. Hong. Robust detection of adversarial attacks by modeling the intrinsic properties of deep neural networks. In *NIPS*, pages 7913–7922, 2018.
- [159] J. Zhu, B. Shen, A. Abbasi, M. Hoshmand-Kochi, H. Li, and T. Q. Duong. Deep transfer learning artificial intelligence accurately stages covid-19 lung disease severity on portable chest radiographs. *PLoS one*, 15(7):e0236621, 2020.

ABSTRACT**ADVERSARIAL MACHINE LEARNING FOR
ADVANCED MEDICAL IMAGING SYSTEMS**

by

Xin Li**June 2022****Advisor:** Dr. Dongxiao Zhu**Major:** Computer Science**Degree:** Doctor of Philosophy

Although deep neural networks (DNNs) have achieved significant advancement in various challenging tasks of computer vision, they are also known to be vulnerable to so-called adversarial attacks. With only imperceptibly small perturbations added to a clean image, adversarial samples can drastically change models' prediction, resulting in a significant drop in DNN's performance. This phenomenon poses a serious threat to security-critical applications of DNNs, such as medical imaging, autonomous driving, and surveillance systems. In this dissertation, we present adversarial machine learning approaches for natural image classification and advanced medical imaging systems.

We start by describing our advanced medical imaging systems to tackle the major challenges of on-device deployment: automation, uncertainty, and resource constraint. It is followed by novel unsupervised and semi-supervised robust training schemes to enhance the adversarial robustness of these medical imaging systems. These methods are designed to tackle the unique challenges of defending against adversarial attacks on medical imaging systems and are sufficiently flexible to generalize to various medical imaging modalities

and problems. We continue on developing novel training scheme to enhance adversarial robustness of the general DNN based natural image classification models. Based on a unique insight into the predictive behavior of DNNs that they tend to misclassify adversarial samples into the most probable false classes, we propose a new loss function as a drop-in replacement for the cross-entropy loss to improve DNN's adversarial robustness. Specifically, it enlarges the probability gaps between true class and false classes and prevents them from being melted by small perturbations. Finally, we conclude the dissertation by summarizing original contributions and discussing our future work that leverages DNN interpretability constraint on adversarial training to tackle the central machine learning problem of generalization gap.

AUTOBIOGRAPHICAL STATEMENT

Xin Li received his B.S. in Electrical and Computer Engineering from Shanghai Jiao Tong University, B.S. and M.S. in Atmospheric and Space Sciences and M.S. in Industrial and Operations Engineering from University of Michigan. Xin Li is currently a Ph.D. candidate at the Department of Computer Science, Wayne State University. His research interests are in trustworthy machine learning and applications with emphasis on adversarial robustness and explainability. He have published extensively in high-impact AI and medical imaging venues (e.g., AAAI, IJCAI, ISBI), and served on program committees of flagship AI conferences (e.g., ICML, NeurIPS, AAAI, IJCAI).