

# Effective Modeling in Medical Imaging with Constrained Data

by

Tzu-Ming Harry Hsu

B.S., National Taiwan University (2016)

B.S.E., National Taiwan University (2016)

S.M. Massachusetts Institute of Technology (2020)

Submitted to the Department of Electrical Engineering and Computer Science  
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2022

© 2022 Tzu-Ming Harry Hsu. All rights reserved.

The author hereby grants to MIT permission to reproduce and to distribute publicly  
paper and electronic copies of this thesis document in whole or in part in any  
medium now known or hereafter created.

Author .....  
Department of Electrical Engineering and Computer Science  
August 26, 2022

Certified by .....  
Peter Szolovits  
Professor of Electrical Engineering and Computer Science  
Thesis Supervisor

Accepted by .....  
Leslie A. Kolodziejcki  
Professor of Electrical Engineering and Computer Science  
Chair, Department Committee on Graduate Students



# Effective Modeling in Medical Imaging with Constrained Data

by

Tzu-Ming Harry Hsu

Submitted to the Department of Electrical Engineering and Computer Science  
on August 26, 2022, in partial fulfillment of the  
requirements for the degree of  
Doctor of Philosophy

## Abstract

Data for modern medical imaging modeling is constrained by their high physical density, complex structure, insufficient annotation, heterogeneity across sites, long-tailed distribution of findings/conditions/diseases, and sparsely presented information. In this dissertation, to utilize the constrained data effectively, we employ various computationally driven and clinically driven techniques, including cross-modal learning, deep reinforcement learning, transfer learning, federated learning, surrogate endpoint modeling, and clinical knowledge infusion. The techniques are demonstrated in a variety of applications, such as risk stratification for pancreatic cancer patients, COVID-19 severity risk assessment, cross-modal X-ray image and report retrieval, X-ray finding report generation from an image, orthopantomogram finding summarization and real-world federated learning benchmarking.

In disease risk stratification applications, we develop an end-to-end body composition assessment system that quantifies fat and muscle amounts from 3-dimensional imaging studies with a two-step approach. The resulting body composition ratios for various tissues are then used to stratify risks in pancreatic cancer or COVID-19 patients. In the pancreatic cancer cohort, muscle loss is shown to be a good indicator of mortality risk; and in COVID-19 patients, visceral fat is more correlated with severity than body mass index is, despite the latter being the current go-to indicator.

Following clinical applications related to body composition analysis, we take advantage of large-scale chest X-ray/report datasets to investigate how the association of the textual modality and the imaging modality can assist modeling. We explore the task of retrieval across radiographs and medical reports by learning a joint embedding space, and find that the retrieval performance can benefit from even a small amount of supervision. On the task of medical report generation, we attempt to describe clinical findings in a chest X-ray as radiologists do. While past works only consider language fluency but not clinical efficacy, we include both in our modeling process. The resulting models turn out to be, unsurprisingly, better at describing diseases and findings, which we identify to be a key trait for an AI system that aims to augment clinicians in their workflows.

We then look at finding summarization from orthopantomogram, or, panoramic dental X-ray. The goal of the summarization is to localize teeth in the permanent dentition and tag them with labels of the six potential findings. To combine the modeling process with existing dental knowledge, we propose a new form of annotation that is quick to provide – a set of 32 binary labels indicating the existence of each tooth. This annotation is used in a novel objective function for the system to optimize and is shown to improve finding summarization accuracy despite its simplicity compared to the pixel-wise supervision typically used in this task.

Finally, we turn to inspect federated learning, which is a learning paradigm for medical institutions to collaboratively learn an AI model without exposing private patient data. As a precursor to medical imaging, we gather two large natural visual classification datasets on real-world scales, aiming to describe the impact of data heterogeneity on the performance of existing federated learning algorithms. Our results show that extreme data heterogeneity can greatly impact algorithms in their ability to classify visual patterns in federated learning setups, and the two novel solutions we bring to the table can somewhat alleviate the performance drop. We believe the conclusions can be extendable to medical imaging problems.

To conclude the dissertation, we provide remarks on other important aspects that researchers in medical AI must consider before landing their applications in clinics, as well as some exciting yet under-explored research tracks in medical imaging. While the objective of this dissertation is to provide an extensive coverage of various methods that more effectively model medical imaging tasks when the available data are constrained, our explorations are not exhaustive. We hope the several research topics showcased in this dissertation inspire further research and can fuel explorations down the line, ultimately benefiting humanity on a civilization scale.

Thesis Supervisor: Peter Szolovits

Title: Professor of Electrical Engineering and Computer Science

## Acknowledgments

“*Yay I’m out! Thanks y’all!*” – I hope the acknowledgment can be that short to make this dissertation legendary, but legend always comes with a price, and I am not prepared to pay.

Jokes aside, to be completely honest, I am more than grateful to acknowledge that I am in a privileged position to have received love and help from a train of people. They help me get through what is supposed to be a *lonely journey* as many predecessors in Ph.D. would say. Hence, in this acknowledgment I will try my best to cover everyone I can recall, but please do also accept my generic thank you if you are supposed to be here and are not – I simply have a goldfish memory.

Back in my master’s thesis, I said (Hsu, 2020):

*“Pete guides me around the ridges and trenches that would have caused me trouble. I would love to write up another thesis about how Pete magically finds ways to deal with our requests, questions, and problems, but I am still collecting more data.”*

I do want to follow up on this claim about my advisor **Peter Szolovits**: it still holds 100% true, and the secret to being as magical as he is, as I have figured out finally, is to always care more for students. Pete does spend a great amount of effort not only educating us about ways of research but also getting us through unavoidable *bureaucratic* and administrative processes that slow us down. He is always open to new ideas and that is why I am lucky to be a part of his research lab. Now that I am finishing my Ph.D. with Pete, it is honored for me to be a member of the *Pete’s Students Club*.

Without my thesis committee members, **Alexander Goehler** and **Dina Katabi**, I would not be able to put together work like this. Alex has been more than an academic collaborator – I appreciate his energy and enthusiasm toward AI research for medicine, and how our joint work can make positive impacts. I was lucky to work with Dina from multiple angles – first I learned about her research in my first year at MIT, then I took Dina’s class on computer networks before working with her as a teaching assistant. I cannot go without also mentioning my RQE committee members,

**Piotr Indyk** and **Randall Davis**. Thank you for your feedback and support that makes my work better.

Our research group – Clinical Decision-Making Group (MEDG) used to be a vibrant group of active and curious students until COVID hits in 2020. I miss the organic conversations with lab members which always inspired research ideas, debugged my long-standing coding issues, and fueled multi-disciplinary collaborations. I thank the following members from MEDG for sharing good times with me in the lab: **Wei-Hung Weng, Geeticka Chauhan, Matthew McDermott, Willie Boag, Marzyeh Ghassemi, Elena Sergeeva, Heather Berlin, Emily Alsentzer, Eric Lehman, Di Jin, Sam Finlayson, and Tristan Naumann.**

Along the way, there are many collaborators, in hospitals, in corporations, or in schools, who taught me quite a lot regarding their specialties and I enjoy publishing great articles with them: **Hang Qi** and **Matthew Brown** from Google; **Ronilda Lacson** and **Jennifer Manne-Goehler** from clinical research facilities in Boston; **Shankeeth Vinayahalingam** and **Joachim Krois** from Europe; **Guanxiong Liu** from University of Toronto; **Shunyu Yao, Jun-Yan Zhu, Jiajun Wu** from MIT CSAIL; **Yunfei Ma, Zhihong Luo, and Nicholas Selby** from MIT Media Lab. I hope the collaborations do not stop here and we will be able to work on exciting projects in the future.

Friends play a significant role in the life of a foreigner like me in the U.S., and I appreciate the emotional support coming from my friends in the Boston area: **Li-Wen Wang, Schrasing Tong, and Hsiang Hsu**; and in Taiwan: **Peggy Tsai** and **Claire Liu.**

As with many other things in life, studying in MIT is not free. My Ph.D. journey was supported and sponsored by the following funding agencies/programs: MIT Media Lab, Taiwan Ministry of Education, Center for Evidence-Based Imaging, Brigham and Women’s Hospital, MIT-Takeda Program, and MIT-IBM Watson AI Lab. I appreciate the generous funding support.

Finally, I would like to thank my family: my mother who raised us almost single-handedly, who has been more than supportive of any big decisions in my life even

though they have absolutely no understanding of what I am pursuing, and who always cares for my physical and mental health.

If you have been reading on until this point, I would like to point out that the girl I thanked in my master's thesis has now become my wife. COVID was indeed huge devastation for people around the world but we took it as an unfortunate blessing and spent time aligning our future plans together while we were in quarantine. We became life partners, and even become partners in academia<sup>1</sup>! Thank you, Chelsea, and I would love to have you in more chapters in my book of life.

---

<sup>1</sup>Chapter 7 is adapted from a publication of Chelsea and myself.





# Contents

<b>List of Figures</b>	<b>15</b>
<b>List of Tables</b>	<b>21</b>
<b>Glossary</b>	<b>25</b>
<b>1 Introduction</b>	<b>29</b>
1.1 Modeling in Medical Imaging . . . . .	29
1.1.1 Computer Aided Detection (CAdE) and Diagnosis (CAdx) . . . . .	30
1.1.2 Quantification and Staging . . . . .	31
1.1.3 Decision Support and Treatment Planning . . . . .	31
1.2 Challenges in Constrained Medical Imaging . . . . .	31
1.2.1 Contributions & Organization . . . . .	33
1.2.2 Publications . . . . .	36
<b>2 Related Works</b>	<b>39</b>
2.1 Clinical Motivations . . . . .	39
2.1.1 Surrogate Modeling Endpoints for Data Re-utilization . . . . .	39
2.1.2 Clinical Knowledge-Infused Modeling for Data Efficiency . . . . .	41
2.2 Computational Motivations . . . . .	42
2.2.1 Annotation Reduction Approaches for Data Hunger Relief . . . . .	42
2.2.2 Federated Learning for Data Collaboration . . . . .	44
<b>3 Transfer Learning for Cancer Mortality Assessment on Restricted</b>	

<b>Institutional Data</b>	<b>47</b>
3.1 Overview . . . . .	48
3.2 Background . . . . .	49
3.3 Materials and Methods . . . . .	50
3.3.1 Data . . . . .	50
3.3.2 Body Composition Measurement . . . . .	51
3.3.3 Algorithm . . . . .	51
3.3.4 Analysis . . . . .	53
3.4 Results . . . . .	55
3.4.1 Model Performance and Generalizability . . . . .	55
3.4.2 Muscle Mass and Visceral Fat – Their Clinical Discriminatory Value . . . . .	57
3.4.3 Clinical Implementation . . . . .	61
3.5 Discussion . . . . .	62
3.6 Summary . . . . .	64
<b>4 Surrogate Endpoint from Limited Imaging Data for COVID-19 Sever-</b>	
<b>ity Prediction</b>	<b>65</b>
4.1 Overview . . . . .	66
4.2 Background . . . . .	67
4.3 Methods . . . . .	68
4.3.1 Data Source . . . . .	68
4.3.2 Ascertaining VAT and SAT Using an AI-Based Body Composi- tion Detector . . . . .	69
4.3.3 Exposures, Outcomes, and Statistical Analysis . . . . .	70
4.4 Results . . . . .	71
4.5 Discussion . . . . .	80
4.6 Summary . . . . .	84
<b>5 Cross-Modal Representation Learning under Sparse Supervision</b>	<b>87</b>
5.1 Overview . . . . .	87

5.2	Introduction . . . . .	88
5.3	Methodology . . . . .	90
5.3.1	Data . . . . .	90
5.3.2	Methods . . . . .	90
5.3.3	Evaluation . . . . .	93
5.4	Results . . . . .	93
5.5	Summary . . . . .	96
<b>6</b>	<b>Knowledge-Infused Learning for Medical Report Generation from Radiograph</b>	<b>99</b>
6.1	Overview . . . . .	100
6.2	Background . . . . .	100
6.3	Related Works . . . . .	102
6.3.1	Radiology . . . . .	102
6.3.2	Language Generation . . . . .	105
6.3.3	Radiology Report Generation . . . . .	107
6.4	Methods . . . . .	107
6.4.1	Hierarchical Generation via CNN-RNN-RNN . . . . .	109
6.4.2	Reinforcement Learning for Readability . . . . .	111
6.4.3	Novel Reward for Clinically Accurate Reinforcement Learning	111
6.4.4	Implementation Details . . . . .	114
6.4.5	TieNet Re-implementation . . . . .	115
6.5	Experiments . . . . .	117
6.5.1	Datasets . . . . .	117
6.5.2	Evaluation Metrics . . . . .	118
6.5.3	Models . . . . .	118
6.6	Results . . . . .	120
6.6.1	Quantitative Results . . . . .	120
6.6.2	Qualitative Results . . . . .	123
6.7	Limitations & Future Work . . . . .	126

6.8	Reflections on Trends in the Field . . . . .	127
6.8.1	System Generalizability . . . . .	127
6.8.2	Be Careful What You Wish For . . . . .	127
6.9	Summary . . . . .	128
<b>7</b>	<b>Reinforcement Learning for Weakly Supervised Dental Imaging Data</b>	<b>131</b>
7.1	Overview . . . . .	132
7.2	Background . . . . .	132
7.3	Methods . . . . .	134
7.3.1	Model Architecture . . . . .	134
7.3.2	Improved Tooth Localization with Weakly Supervised Reinforcement Learning . . . . .	138
7.3.3	Training Details . . . . .	140
7.4	Experiments . . . . .	140
7.4.1	Dataset . . . . .	141
7.4.2	Overall Evaluation of DeepOPG for Findings Summarization . . . . .	142
7.4.3	Functional Segmentation . . . . .	144
7.4.4	Tooth Localization with Dental Coherence . . . . .	145
7.4.5	Comparing Existing Works . . . . .	147
7.5	Summary . . . . .	149
<b>8</b>	<b>Federated Learning for Heterogeneous Visual Classification</b>	<b>151</b>
8.1	Overview . . . . .	152
8.2	Background . . . . .	153
8.3	Related Work . . . . .	154
8.3.1	Synthetic Client Data . . . . .	154
8.3.2	Realistic Datasets . . . . .	155
8.4	Federated Visual Classification Problems . . . . .	155
8.4.1	Natural Species Classification . . . . .	156
8.4.2	Landmark Recognition . . . . .	156
8.5	Datasets . . . . .	157

8.5.1	iNaturalist-User-120k and iNaturalist-Geo Splits . . . . .	157
8.5.2	Landmarks-User-160k . . . . .	159
8.5.3	CIFAR-10/100 . . . . .	160
8.6	Methods . . . . .	161
8.6.1	Federated Averaging and Server Momentum . . . . .	161
8.6.2	Importance Reweighted Client Objectives . . . . .	164
8.6.3	Splitting Imbalanced Clients with Virtual Clients . . . . .	165
8.6.4	Implementation Details . . . . .	166
8.7	Experiments . . . . .	167
8.7.1	Classification Accuracy vs Distribution Non-Identicalness . . . . .	169
8.7.2	Importance Reweighting . . . . .	170
8.7.3	Federated Virtual Clients . . . . .	170
8.7.4	Federated Visual Classification Benchmarks . . . . .	173
8.7.5	Hyperparameter Sensitivity . . . . .	174
8.7.6	The Effect of Pretraining . . . . .	176
8.7.7	Experiment Run Time . . . . .	178
8.8	Summary . . . . .	180
<b>9</b>	<b>Conclusions</b>	<b>181</b>
9.1	Chapter Review . . . . .	182
9.2	Areas to Be Explored Before Landing AI . . . . .	183
9.3	Novel Research Directions . . . . .	184
9.4	Conclusions . . . . .	186
	<b>Bibliography</b>	<b>187</b>



# List of Figures

3-1	Overview of 2-stage, end-to-end algorithm. . . . .	52
3-2	Overview of experimental set-up. . . . .	53
3-3	Performance of the slice localization. The minimum localization error among all identified slices is plotted against the number of nominated slices. . . . .	56
3-4	Kaplan Meier curve stratified by presence/absence of sarcopenia and low versus high visceral fat as derived by the AI algorithm. . . . .	60
4-1	Visceral fat distribution, overall and by BMI category. **** $p < 0.0001$ . . . . .	72
4-2	Kaplan-Meier curves for intubation or death within 28 days. Abbreviation: VAT, visceral adipose tissue. . . . .	75
4-3	Exemplary visceral fat body compositions by BMI status and gender. . . . .	76
4-4	Scatterplot of VAT by BMI, stratified by sex. . . . .	76
4-5	Cox proportional hazards model for the relationship between quintile of VAT and death or intubation within 30 days (adjusted for age, gender, and diabetes status). . . . .	77
4-6	Adjusted hazard ratio for outcome of death or intubation within 30 days, overall and by BMI group (adjusted for age, gender, and diabetes status). . . . .	79
5-1	The overall experimental pipeline. EA: embedding alignment; Adv: adversarial training. . . . .	91

5-2	Performance measures of retrieval tasks at $k$ retrieved items as a function of the supervision fraction. Higher is better. Note the $x$ -axis is in log scale. Unsupervised is on the left, increasingly supervised to the right. Dashed lines indicate the performance by chance. Vertical bars indicate the 95% confidence interval, and some are too narrow to be visible. . . . .	96
5-3	Different metrics for retrieval on either the <i>impression</i> or <i>findings</i> section using four types of features. 95% confidence intervals are indicated on the bars. . . . .	96
6-1	A chest X-ray and its associated report written by a radiologist. . . . .	102
6-2	<b>The model for our proposed <i>Clinically Coherent Reward</i>.</b> Images are first encoded into image embedding maps, and a sentence decoder takes the pooled embedding to recurrently generate topics for sentences. The word decoder then generates the sequence from the topic with attention on the original images. NLG reward, clinically coherent reward , or combined, can then be applied as the reward for reinforcement policy learning. . . .	108
6-3	<b>Visualization of the generated report and image attention maps.</b> Different words are underlined with its corresponding attention map shown in the same color. Best viewed in color. . . . .	125
7-1	<b>System Overview of DeepOPG.</b> There are three modules in DeepOPG working together to obtain orthopantomogram finding summary on each tooth. (a) Functional Segmentation Module performs per-pixel classification, (b) Tooth localization Module localizes each tooth, and (c) Dental Coherence Module ensures that output is clinically reasonable. Ultimately, we combine the segmentation maps and teeth identity maps to produce findings with explainable predictive values. . . . .	135



7-2	<b>Receiver Operating Characteristic (ROC) Curves.</b> Six finding types are considered. 95% confidence interval is annotated in shades and parentheses, and the operating points with the highest F1, along with the thresholds on percentage area, are labeled. . . . .	143
7-3	<b>Evaluation of the Functional Segmentation Module.</b> The IoU between the predicted segmentation and the ground truth segmentation are shown per class. Standard deviation is also labeled. . . . .	144
7-4	<b>Illustrations of Tooth localization Results.</b> (a) The FDI tooth numbering system, (b) input OPG with ground truth localization, (c) functional segmentation map, (d) localization for DeepOPG (full), (e) localization w/o reinforcement learning, and (f) localization w/o segmentation input. . . . .	146
8-1	<b>iNaturalist Species Distribution.</b> Visualized here are the distributions of Douglas-Fir and Red Maple in the continental US within iNaturalist. In a federated learning context, visual categories vary with location, and users in different locations will have very different training data distributions. . . . .	156
8-2	<b>iNaturalist Distribution.</b> In (a) we show the re-balancing of the original iNaturalist-2017 dataset. In (b) and (c) we show class and example counts vs clients for our 5 iNaturalist partitionings with varying levels of class distribution shift and size imbalance. The client count is different in each partitioning. . . . .	158
8-3	<b>Landmarks-User-160k Distribution.</b> Images are partitioned according to the authorship attribute from the GLD-v2 dataset. Filtering is applied to mitigate long tail in the train split. . . . .	158

8-4	<b>Synthetic populations with non-identical clients.</b> Distribution among classes is represented with different colors, each standing for a class. (a) 10 clients generated from the sort-and-partition scheme, each assigned with 2 classes. (b–e) populations generated from Dirichlet distribution with different concentration parameters $\alpha$ respectively, 30 random clients each. . . . .	162
8-5	<b>CIFAR-10/100 Distribution.</b> Each curve represents the class counts of clients within a data partitioning synthesized using a Dirichlet concentration parameter $\alpha$ . . . . .	162
8-6	<b>Relative Accuracy vs. Non-identicalness.</b> Federated learning experiments are performed on (a) CIFAR-10 and (b) CIFAR-100 using local epoch $E = 1$ . The top row demonstrates the distributions of EMD of clients with different data partitionings. Total client counts are annotated to the right, and the weighted average of all client EMD is marked. Data is increasingly non-identical to the right. The dashed line indicates the centralized learning performance. The best accuracies over a grid of hyperparameters are reported (see Section 8.7.5). . . . .	168
8-7	<b>Comparing Base Methods with and without FedIR.</b> Accuracy shown at 2.5k communication rounds. Centralized learning accuracy marked with dashed lines. . . . .	169
8-8	<b>Learning with Federated Virtual Clients.</b> Curves on the left are learned on the iNaturalist geo-partitioning Geo-3k and user split User-120k each with 135 clients and 9275 clients. Experiments on multiple iNaturalist partitionings are shown on the right, plotting relative accuracy at 2.5k communication rounds to mean EMD. Centralized learning achieves a 57.9% accuracy. . . . .	172
8-9	<b>Landmarks-User-160k Learning Curves.</b> Only the last two layers of the network are fine-tuned. FedIR is also shown due to its ability to address skewed training distribution as presented in this dataset. . . .	174

8-10	<b>Relative Accuracy of FedAvgM on CIFAR Datasets.</b> Darker shades denote regions of higher relative accuracy. $\eta_{\text{eff}} = \eta / (1 - \beta)$ is the effective learning rate, and $K$ is the reporting goal out of 100 clients. Note that data split is increasingly non-identical to the right.	177
8-11	<b>Learning Curves from ImageNet Pretraining and from Scratch.</b> On the left vertical axis is the relative accuracy while on the right is the absolute accuracy. Two plots are rescaled to have the full span of 100% relative accuracy. . . . .	178



# List of Tables

3.1	Performance of the overall system in quantifying the 3 different tissue components (compared to the manually annotated ground truth). 95% CI labeled in brackets. . . . .	58
3.2	Baseline characteristics of pancreas cohort stratified by sarcopenia status. . . . .	59
3.3	Cox regression model, death . . . . .	61
4.1	<b>Differences in Demographic and Health Characteristics.</b> Comparison done between those with and without imaging data in the MGH COVID-19 Registry. . . . .	73
4.2	<b>Differences in Demographic and Health Characteristics.</b> Comparison done on 378 COVID-19 registry participants, overall and by VAT group. . . . .	74
4.3	Multivariate Adjusted Hazard Ratio for Death or Intubation Within 28 Days From Hospitalization. . . . .	78
4.4	Cox proportional hazards model for death or intubation, stratified for individuals with and without imaging. . . . .	82
4.5	Cox proportional hazards models for death or intubation within 30 days, overall and stratified by imaging during or prior to admission. . . . .	83

5.1	Comparison among supervised (upper) and unsupervised (lower) methods. Subscripts show the half width of 95% confidence intervals. <b>Bold</b> denotes the best performance in each group. <i>Chance</i> is the expected value if we randomly yield retrievals. Higher is better for all metrics.	94
6.1	A description of each available chest X-ray datasets. Open-I (Demner-Fushman et al., 2015), Chest-XRay8 (Wang et al., 2017b) which utilized DNORM (Leaman et al., 2015) and MetaMap (Aronson and Lang, 2010), CheXpert (Irvin et al., 2019), PadChest (Bustos et al., 2019), and MIMIC-CXR (Johnson et al., 2019).	104
6.2	<b>Automatic Evaluation Scores.</b> The table is divided into natural language metrics and clinical finding accuracy scores. BLEU- $n$ counts up $n$ -gram for evaluation, and accuracy is the averaged macro accuracy across all clinical findings. <i>Major class</i> always predicts negative findings.	121
6.3	<b>Clinical Finding Scores.</b> The precision scores for each of the labels are listed and aggregated into the overall precision scores. Recall scores are shown in the last two rows. Macro denotes averaging the numbers in the table directly and micro accounts for class prevalence.	122
6.4	<b>Sample images along with ground truth and generated reports.</b> Note that upper case tokens are results of anonymization.	124
7.1	<b>AUC Comparisons.</b> We compare the AUCROC for six OPG finding types for two settings of DeepOPG. See Section 7.4.4 for method descriptions.	142
7.2	<b>Comparisons of Detection Metrics.</b> We show detection metrics for various settings of DeepOPG. We report the metric values and their standard errors. $AP_x$ denotes $AP@IoU = x$ . See Section 7.4.4 for method descriptions.	145

7.3	<b>Comparison to Prior Works.</b> We compare our full model DeepOPG under similar conditions to prior works on various tasks. Note the evaluations are done on different dataset in each work. We report the metric values and their standard errors. $AP_x$ denotes $AP@IoU = x$ .	148
8.1	<b>Training Dataset Statistics.</b> Note that while CIFAR-10/100 and iNaturalist datasets each have different partitionings with different levels of identicalness, the underlying data pool is unchanged and thus sharing the same centralized learning baselines.	167
8.2	<b>Accuracy of Federated Virtual Client on iNaturalist.</b> $Acc@round$ denotes the accuracy at a FL communication round. $Acc@batch$ denotes the batch count accumulated over the largest clients per round, and is a proxy for a fixed time budget.	171
8.3	<b>iNaturalist-User-120k accuracy.</b> Numbers reported at fixed communication rounds. $K$ denotes the report goal per round.	173
8.4	<b>Landmarks-User-160k Accuracy.</b>	175
8.5	<b>Communication Rounds to Reach Relative Accuracy.</b> Note that models have different centralized learning accuracy (51.4% from scratch and 57.9% from pretrained). The multipliers are calculated row-wise, using $Rounds@10\%$ as the baseline. Experiments that do not reach the target relative accuracy even after $t$ rounds is marked $> t$ .	179





# Glossary

---

<b>Abbreviation</b>	<b>Full Name</b>
ACE-2	Angiotensin-Converting Enzyme 2
aHR	Adjusted Hazard Ratio
AI	Artificial Intelligence
AP	Anteroposterior
AP@IoU	Average Precision at Fixed Intersection Over Union
AUPRC	Area Under the Precision-Recall Curve
AUROC	Area Under the Receiver Operator Characteristic
BLEU	Bilingual Evaluation Understudy Score
BMI	Body Mass Index
CAD	Coronary Artery Disease
CCR	Clinically Coherent Reward
CHF	Congestive Heart Failure
CIDeR	Consensus-Based Image Description Evaluation
CNN	Convolutional Neural Networks
COCO	Common Objects in Context
COPD	Chronic Obstructive Pulmonary Disease
COVID-19	Coronavirus Disease 2019
CRP	C-Reactive Protein
CT	Computed Tomography
DCG	Discounted Cumulative Gain

DCR	Dental Coherence Reward
DNorm	Disease Name Normalization
DRL	Deep Reinforcement Learning
ED	Emergency Department
EDW	Enterprise Data Warehouse
EMA	Exponential Moving Average
EMD	Earthmover's Distance
EMRs	Electronic Medical Records
ESR	Erythrocyte Sedimentation Rate
FDI	Fédération Dentaire Internationale
FedAvg	Federated Averaging
FL	Federated Learning
GANs	Generative Adversarial Networks
GLD-v2	Google Landmarks Dataset V2
GQAP	Generalized Quadratic Assignment Problem
HRGR-Agent	Hybrid Retrieval-Generation Reinforced Agent
ICC	Intraclass Correlation
ICD-9	International Classification of Diseases, Ninth Edition
ICU	Intensive Care Unit
ICU	Intensive Care Unit
IID	independent and identically distributed
IoU	Intersection over Union
LDA	Latent Dirichlet Allocation
LG	Language Generation
LiTS	Liver Tumor Segmentation
LL	Lateral
LSTM	Long Short Term Memory
MGH	Massachusetts General Hospital
MI	Myocardial Infarction
MIMIC	Medical Information Mart for Intensive Care

MIMIC-CXR	Medical Information Mart for Intensive Care - Chest X-Ray
MLG	Natural Language Generation
MNIST	Modified National Institute of Standards and Technology
MRI	Magnetic Resonance Imaging
MRR	Mean Reciprocal Rank
nDCG	Normalized Discounted Cumulative Gain
NLP	Natural Language Processing
NMS	Non-Maximum Suppression
OCT	Optical Coherence Tomography
OPG	Orthopantomogram
PA	Posteroanterior
PACS	Picture Archive And Communication System
PASCAL VOC	Pascal Visual Object Classes
PCA	Principle Component Analysis
PICC	Peripherally Inserted Central Catheter
PPV	Positive Predictive Value
R@P	Recall at Fixed Precision
RL	Reinforcement Learning
RNN	Recurrent Neural Network
ROI	Region of Interest
ROUGE	Recall-Oriented Understudy For Gisting Evaluation
RPN	Region Proposal Network
S&T	Show and Tell
SA&T	Show, Attend, and Tell
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
SAT	Subcutaneous Adipose Tissue
SCST	Self-Critical Sequence Training
SELU <sub>s</sub>	Scaled Exponential Linear Units
SGD	Stochastic Gradient Descent
SMI	Skeletal Muscle Index

TF-IDF	Term Frequency-Inverse Document Frequency
TieNet	Text-Image Embedding Network
VAT	Visceral Adipose Tissue
XML	Extended Markup Language

---

# Chapter 1

## Introduction

### 1.1 Modeling in Medical Imaging

Medical imaging constitutes a significant amount of the overall collected medical data in healthcare, and the majority of them are underutilized for research purposes as of 2019 (Landi, 2016; Healthcare, 2019). These medical images are manifestations of physical properties such as electromagnetic waves, magnetic resonance, nuclear radioactivity, visible light, and sound waves in the form of images, typically two-dimensional, three-dimensional, or spacial-temporal. Some commonly used imaging modalities in the clinics are X-ray radiography, computational tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), diffusion tensor imaging (DTI), and ultrasound imaging. These various imaging techniques have been providing unique evidence in clinical intervention in the form of disease finding discovery, treatment response confirmation, and surgical planning prior to or during invasive procedures.

The process of medical image interpretation often starts with clinicians ordering an imaging acquisition for the aforementioned clinical reasons. Radiologists, who specialize in medical image reading, then describe findings and then interpret them in the form of a radiology report. This report addresses the questions posted in the original imaging requisition. A final diagnosis is then given by the clinician based on the report and other observations and data from the electronic health record

(EHR) including clinical exam, history, and diagnostic test results. Treatment plans are designed in response to the diagnosis results and may change any time as the diagnosis changes.

This operating procedure, however, is often done under pressing time constraints. As a result, the interpretations are limited by inter-observer discrepancy and fatigue-induced error (Brady, 2017; Goddard et al., 2001; Siewert et al., 2008; Briggs et al., 2008; Quekel et al., 1999), with the majority of them being missed diagnosis (42%) (Kim and Mansfield, 2014). Other major categorization of radiological reading errors include failure to identify abnormality beyond the first one (22%) and faulty reasoning (9%). These factors, together with slow turnaround and inability to numerically quantify disease severity, hinder the scalable development of *traditional radiology*.

We welcomed a series of surprising breakthroughs in artificial intelligence (AI) technology starting in the year 2012 with the ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC 2012) (Deng et al., 2009; Russakovsky et al., 2015; Krizhevsky et al., 2012). Since then, deep learning-based modeling has taken over the field of machine learning, and in turn, inspired rapid development of *computational radiology* (Börger and Natterer, 1999). Now with the availability of AI modeling, we unlock the potential to perform large-scale automatic analysis of medical images with AI. The computational aspect of AI makes it inherently objective, consistent, fatigue-free, and available 24/7, which are the most desired improvements of traditional radiology.

Over the course of a decade, researchers have found multiple medical imaging applications suitable for AI modeling (Hosny et al., 2018; Suganyadevi et al., 2022; Zhou et al., 2021; Sonka et al., 2000; Prince and Links, 2006; Bankman, 2008; Zhou et al., 2019), including the ones described in the following sections.

### 1.1.1 Computer Aided Detection (CADe) and Diagnosis (CADx)

CADe (Chan et al., 2020) focuses on the localization characteristics of objects of interest, while CADx classifies the localized objects into discrete types. Commonly seen CADe and CADx applications include *lesion localization* (Yan et al., 2018; Drukker

et al., 2002; Ardila et al., 2019), *cancer diagnosis* (Han et al., 2017; Becker et al., 2018; Cheng et al., 2016; Wang et al., 2017a), *tumor classification* (Ge et al., 2020), and *medical implant localization* (Hsu and Wang, 2021; Lee et al., 2021a).

### 1.1.2 Quantification and Staging

In imaging-guided quantification and staging, researchers are looking for a more granular and continuous judgment of findings and diseases compared to simple classifications. Examples include *biometric measurement derivation* (Guo et al., 2019; Javaid et al., 2018; Moradi et al., 2019; Winkel et al., 2020; Li and Xia, 2020), *risk assessment* (Goehler et al., 2021; Hsu et al., 2021), and *longitudinal monitoring* (Hsu, 2020; Goehler et al., 2020)

### 1.1.3 Decision Support and Treatment Planning

Medical imaging is a critical component of *radiation therapy planning* (Javaid et al., 2019; Nguyen et al., 2019; Fan et al., 2019; Zhang et al., 2021a; Shen et al., 2019, 2020; Barkousaraie et al., 2019), *pre-surgical visualization of anatomy* (Tan et al., 2012; Fan et al., 2013), and *chronic disease treatment planning* (Watts et al., 2020; Nougaret et al., 2013).

In other cases, modeling of medical images can also support decision-making by enhancing the quality of images (Han, 2017; Kazemifar et al., 2019; Shen et al., 2018), or by providing alternative materials such as machine-interpreted radiology reports for clinicians (Liu et al., 2019; Boag et al., 2020; McDermott et al., 2020).

## 1.2 Challenges in Constrained Medical Imaging

Despite the powerful and promising technological advances brought by AI, specifically deep learning methods, the field of modern medical imaging is faced with some inherent challenges regarding the nature of medical imaging data, namely that (a) medical imaging data are dense in pixel (or voxel) counts, coming in all types of modalities,

and (b) the processing of the data is typically a series of expert-crafted pipelines that is complex and non-standardized.

For example, CT, which was hailed as one of the greatest inventions in radiology (Kalender, 2005), welcomed the first scanner to acquire more than 64 slices per rotation in the year of 2005 (Goldman, 2008), while scanners today typically acquire around 1,000 slices per imaging study. Not only did the *quantity* of raw medical imaging data increase dramatically with time, but the *velocity* had to at least keep up to retain a similar image acquisition time, if not shorter, for patient comfort and clinical workflow consistency. It is not hard to imagine that with the rising volume of medical imaging data, the downstream imaging analyses and statistical modeling tasks have become increasingly demanding in terms of resources and would require extensive domain knowledge to establish processing pipelines.

AI has been widely recognized for being data-hungry, and having an abundant data repository is without a doubt a prerequisite for AI modeling. Yet, different from natural images where annotators of data can be general mechanical turks (Paolacci et al., 2010), annotation of medical images is extremely costly (*e.g.*, expensive financially, time-consuming, and hard to find experts). To make matters worse, annotation of a single study can require the collaboration of clinicians of different expertise (*e.g.*, a radiologist summarizing the imaging study while the referring physician concludes the diagnosis), making the associated labels sparse across the data repository (Marlin et al., 2011; Mohan et al., 2013).

As medical imaging modeling scales with AI development, we now enter an era where inter-institution collaboration on modeling is possible. There are several major challenges to overcome, particularly that (a) acquisition protocols and instruments vary from site to site (Soin et al., 2022), and that (b) the underlying disease and demographic distributions can be substantially different (Abbasi-Sureshjani et al., 2020). Owing to these reasons, medical image datasets are isolated, creating a systematic roadblock for scalable medical imaging modeling.

Last but not least, even if we are in the era of computational radiography, due to the nature of medical data, we are still faced with similar issues to those used to



appear in traditional radiography. To be specific, (a) the underlying disease exhibits a long-tailed distribution (Roy et al., 2022), and (b) the raw imaging data is inherently information-sparse (*e.g.*, a lung nodule is typically not more than 0.01% of a chest CT in volume).

To summarize, data for modern medical imaging modeling is constrained by the aforementioned factors:

1. Medical images are dense in physical dimensions.
2. Processing of medical imaging data requires precise expertise and is complex.
3. Labels for medical images are sparse across studies.
4. Cross-site datasets are non-standardized and heterogeneous.
5. Disease ground truths are long-tailed.
6. Information is sparse in the imaging data.

Each of these factors is a complicated research problem on its own, and following the identification of problems, we highlight some of the promising methodological advances in the field of medical imaging, and how I propose to approach the solutions systemically in the dissertation.

### **1.2.1 Contributions & Organization**

This dissertation focuses on medical imaging problems. Specifically, I will take on several topics at the forefront of the field. Automatic disease quantification with derived biometric measurements from 3D imaging brings benefits into the domain due to its consistency, resilience to fatigue, and potential to be deployed as ambient intelligence (Ramos et al., 2008). Descriptive textual summary modeling from radiographic images is as well a heated and challenging topic as this aids clinicians in their daily routine, reducing the time needed to process their routine work in a more human-centric manner. Finally, dental imaging is a lesser-known area of research that

has emerged recently; I have assembled a dataset for segmentation and detection purposes, and ultimately derive useful information for dentists in their clinical workflow and patient education.

The proposed contribution is not solely constrained to the medical imaging community. From a methodological perspective, I will touch on several topics. Transfer learning, frequently presented as *domain adaptation*, has always been an obstacle for researchers to overcome as data distribution differences are everywhere. Prior to medical data, I have exposure to other aspects of domain adaptation including imbalanced class-distribution (Hsu et al., 2015) and heterogeneous data (Chen et al., 2016, 2019c), which are long-standing research topics in need of better solutions. Unsupervised learning, reinforcement learning, and cross-modal learning will also be explored in the thesis. Finally, federated learning, being an emergent technique extremely suitable for collaboration of medical data modeling under current regulation restrictions, will receive a close-up examination from a large-scale real-world perspective to ultimately benefit the overall computer vision domain including natural and medical imaging.

Below is a breakdown of my contributions in *developing modeling methodologies to tackle the problem of constrained data in medical imaging systematically*, and how they are organized in different chapters in the dissertation:

**Chapter 3** Body composition assessment has been shown to be capture cardio-metabolic risks better than anthropometric measurements such as body mass index, and yet manual body composition measurement is resource-intensive and AI-based measurements require sufficient learning data. We demonstrate how an efficient and performant automatic system can be built, and how the resulting body composition evaluation can be used to assess mortality in pancreatic cancer (Hsu et al., 2021).

**Chapter 4** In early phases of COVID-19 outbreak, researcher have started to look for quantifiers for COVID-19 severity. While body mass index was initially shown to be a decent indicator, we demonstrate that via automatic evaluation of body com-

position, the visceral fat quantity is able to correlate with COVID-19 severity better (Goehler et al., 2021).

**Chapter 5** Representation learning across medical images and texts has a great impact on downstream tasks. One specific technique of representation learning is embedding learning which projects data into a common embedding space, and such an embedding space benefits the medical machine learning community by enabling cross-domain retrieval, conditional generation of medical reports, and other applications that utilizes joint embeddings. We investigate how we can map the visual modality and textual modality into the same embedding space with a selection of algorithms, with and without pairing information provided as supervision. We find it surprising to retain decent performance on the task of cross-modal retrieval even with no supervision at all. (Hsu et al., 2018)

**Chapter 6** Automatic generation of radiology reports aims at describing a radiograph similarly to radiologists, and has the potential to accelerate clinical routine and improve patient care. A number of past works have tackled the problem, and yet only with language fluency in mind rather than also considering clinical efficacy. We explore how, by including clinical metrics into the generation of medical reports with reinforcement learning, we can produce a system capable of summarizing chest X-rays not only performant when evaluating with natural language perspectives but also clinical perspectives. (Liu et al., 2019)

**Chapter 7** Orthopantomogram is an essential first-line tool in dentistry due to its ease of acquisition. Dentists can acquire abundant information from these images, and thus they are perfect for automatic systems to provide assistance to reduce missed diagnoses and improve patient communication. Data collection for learning a system, however, is typically hard due to lengthy annotation time for dense pixel-wise maps. We explore an alternate annotation where only binary labels are required to improve learning, and show the efficacy of such an annotation type. (Hsu and Wang, 2021)

**Chapter 8** To enrich data quantity for machine learning while retaining privacy, researchers have developed federated learning where only learning signals are propagated rather than raw data. This is a perfect learning paradigm for medical data owing to legal boundaries around medical institutions, and yet how well federated learning algorithms perform under heterogeneous data distribution is not extensively studied. We resort to large-scale natural imaging data in this chapter, and benchmark existing federated learning algorithms under various heterogeneity settings. Finally we propose some improvements to existing algorithms which we show to boost learning performance. These improvements have the potential to also improve learning on medical images in the federated learning scenario. (Hsu et al., 2019, 2020)

## 1.2.2 Publications

Below is a list of publications relevant to this dissertation:

1. **Hsu, T.-M. H.**, Schawkat, K., Berkowitz, S. J., Wei, J. L., Makoyeva, A., Legare, K., DeCicco, C., Paez, S. N., Wu, J. S., Szolovits, P., et al. (2021). Artificial intelligence to assess body composition on routine abdominal ct scans and predict mortality in pancreatic cancer a recipe for your local application. *European Journal of Radiology*, 142:109834.
2. Goehler, A., **Hsu, T.-M. H.**, Seiglie, J. A., Siedner, M. J., Lo, J., Triant, V., Hsu, J., Foulkes, A., Bassett, I., Khorasani, R., et al. (2021). Visceral adiposity and severe COVID-19 disease: application of an artificial intelligence algorithm to improve clinical risk prediction. In *Open forum infectious diseases*, volume 8, page ofab275. Oxford University Press US.
3. **Hsu, T.-M. H.**, Weng, W.-H., Boag, W., McDermott, M., and Szolovits, P. (2018). Unsupervised multimodal representation learning across medical images and reports. *arXiv preprint arXiv:1811.08615*.
4. Liu, G.<sup>1</sup>, **Hsu, T.-M. H.**<sup>1</sup>, McDermott, M., Boag, W., Weng, W.-H., Szolovits,

---

<sup>1</sup> equal contribution.

- P., and Ghassemi, M. (2019). Clinically accurate chest X-ray report generation. In *Machine Learning for Healthcare Conference*, pages 249–269. PMLR.
5. **Hsu, T.-M. H.** and Wang, Y.-C. C. (2021). DeepOPG: Improving orthopantomogram finding summarization with weak supervision. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 366–376. Springer.
  6. **Hsu, T.-M. H.**, Qi, H., and Brown, M. (2019). Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*.
  7. **Hsu, T.-M. H.**, Qi, H., and Brown, M. (2020). Federated visual classification with real-world data distribution. In *European Conference on Computer Vision*, pages 76–92. Springer.
  8. **Hsu, T.-M. H.** (2020). Automatic longitudinal assessment of tumor responses. Masters dissertation, Massachusetts Institute of Technology.

Below, while not directly related, is a list of publications completed over the course of my Ph.D. program in reverse chronological order:

1. Feher, B., Kuchler, U., Schwendicke, F., Schneider, L., Eduardo Cejudo Grano de Oro, J., Xi, T., Vinayahalingam, S., **Hsu, T.-M. H.**, Brinz, J., Chaurasia, A., et al. (2022). Emulating clinical diagnostic reasoning for jaw cysts with machine learning. *Diagnostics*, 12(8):1968.
2. Ha, U., Leng, J., Khaddaj, A., Ma, Y., **Hsu, T.-M. H.**, Zhong, Z., and Adib, F. (2022). Methods and apparatus for radio frequency sensing in diverse environments. US Patent 11,308,291.
3. McDermott, M., Yap, B., **Hsu, T.-M. H.**, Jin, D., and Szolovits, P. (2021). Adversarial contrastive pre-training for protein sequences. *arXiv preprint arXiv:2102.00466*.

4. Ha, U., Leng, J., Khaddaj, A., Ma, Y., **Hsu, T.-M. H.**, Zhong, Z., and Adib, F. (2020). Methods and apparatus for radio frequency sensing in diverse environments. US Patent 10,872,209.
5. Goehler, A., **Hsu, T.-M. H.**, Lacson, R., Gujrathi, I., Hashemi, R., Chlebus, G., Szolovits, P., and Khorasani, R. (2020). Three-dimensional neural network to automatically assess liver tumor burden change on consecutive liver mris. *Journal of the American College of Radiology*, 17(11):14751484.
6. McDermott, M. B., **Hsu, T. M. H.**, Weng, W.-H., Ghassemi, M., and Szolovits, P. (2020). Chexpert++: Approximating the chexpert labeler for speed, differentiability, and probabilistic output. In *Machine Learning for Healthcare Conference*, pages 913–927. PMLR.
7. Boag, W., **Hsu, T.-M. H.**, McDermott, M., Berner, G., Alesentzer, E., and Szolovits, P. (2020). Baselines for chest X-ray report generation. In *Machine Learning for Health Workshop*, pages 126–140. PMLR.
8. Chen, W.-Y., **Hsu, T.-M. H.**, Tsai, Y.-H. H., Chen, M.-S., and Wang, Y.-C. F. (2019). Transfer neural trees: Semi-supervised heterogeneous domain adaptation and beyond. *IEEE Transactions on Image Processing*, 28(9):4620–4633.
9. Ha, U., Ma, Y., Zhong, Z., **Hsu, T.-M. H.**, and Adib, F. (2018). Learning food quality and safety from wireless stickers. In *Proceedings of the 17th ACM Workshop on Hot Topics in Networks*, pages 106–112.
10. Yao, S.<sup>2</sup>, **Hsu, T.-M. H.**<sup>2</sup>, Zhu, J.-Y., Wu, J., Torralba, A., Freeman, B., and Tenenbaum, J. (2018). 3d-aware scene manipulation via inverse graphics. *Advances in neural information processing systems*, 31.

---

<sup>2</sup> equal contribution.

# Chapter 2

## Related Works

Researchers in the field of medical imaging, over the years, have developed several lines of unique techniques designed to tackle the constrained data problem.

From a clinical perspective, to facilitate the AI learning scenario with constrained data, researchers have explored (a) surrogate modeling endpoints and (b) clinical knowledge-infused modeling.

From a computational perspective, due to parallel advances in the field of natural image modeling, medical imaging received tremendous benefits from (a) neural network architectural improvements, (b) annotation-efficient approaches, (c) federated learning, and (d) interpretability and uncertainty quantification methods.

Below are examples of the most relevant techniques introduced above to my proposed endeavor in the dissertation.

### 2.1 Clinical Motivations

#### 2.1.1 Surrogate Modeling Endpoints for Data Re-utilization

In recent years, the deep learning community has put enormous emphasis on *end-to-end learning* (Bojarski et al., 2016; Dieleman and Schrauwen, 2014), highlighting its ability to incorporate data pre-processing, feature extraction, prediction modeling, and output post-processing in a single black-box neural network for the maximum

possible optimization of prediction efficacy.

While there are similar pushes in the medical domain (Pati et al., 2021; Wu et al., 2018; Oh et al., 2021), we often lose out on various characteristics that clinicians would like to observe about AI models: interpretability and uncertainty quantifiability. On top of this undesired loss of traits, the collection of these specific annotations for end-to-end modeling is typically prohibitively difficult.

Researchers have then identified intermediate biometric measurements as a consequence of (a) not necessarily having direct patient outcome data, (b) having to wait for months, if not years, for the related outcome to be observed, or (c) the outcome being dependent on numerous factors that the imaging input cannot explain well from a medical knowledge perspective. The machine learning modeling process can associate the incoming imaging data and the biometric intermediates, while existing literature would then provide statistical linkage between the intermediates and the desired clinical outcomes. One of the most used surrogate intermediate endpoints is quantification values, with segmentation modeling being an easy path towards quantification. These segmentation models lead to many volumetric measurements that are inherently hard to directly predict from images (*i.e.*, model directly outputting a numeric value based on an input image), and predictive modeling for segmentation outputs have, in the recent years, grown enormously in terms of performance.

As an example, the biometric measurement *body composition* measures the amount of muscle and fat tissue areas at the axial image slice at the L3 vertebra. Bridge et al. (2018), Hashimoto et al. (2019), Weston et al. (2019), Burns et al. (2020), and Park et al. (2020) were the first works to develop deep learning systems upon individually locally annotated muscle/fat segmentation mask datasets to derive body composition measurements. Graffy et al. (2019) investigated muscle mass loss and density attenuation with the aforementioned mask annotations, albeit at a different clinical site. Pickhardt et al. (2021) further used the abdominal fat area from the L3 vertebra axial slice as the surrogate endpoint for metabolic syndrome and was above 80% in both sensitivity and specificity.

In the exploration for *liver diseases*, Ahn et al. (2020) annotated organ masks



in CT studies to automatically predict liver and spleen volumes, and liver-to-spleen ratio (Huang et al., 2014) for evaluation against liver disease occurrences. Haas et al. (2020) quantified liver fat (proton density fat fraction) before stratifying them into fatty/non-fatty liver patient cohorts for downstream genomics analyses.

### 2.1.2 Clinical Knowledge-Infused Modeling for Data Efficiency

Adapting modeling methods from natural imaging to medical imaging has been a mainstream pathway to improving modeling efficacy.

Wu et al. (2013) and Cheng et al. (2018) were the first to adopt neural networks for medical image *registration* on MRI/CT studies. Yang et al. (2015) detected landmarks in MRI studies with deep learning, and this primitive imaging task evolved into what is known today as *object detection* in medical images (De Vos et al., 2016). Brosch et al. (2013), Suk and Shen (2013), and Plis et al. (2014) investigated the first deep learning applications of *classification* on MRI for Alzheimer’s disease. Ciresan et al. (2012) brought segmentation to microscopy imagery and Stollenga et al. (2015) brought segmentation to 3D medical imaging.

These early works emphasized the similarity between medical and natural image modeling but had yet to apply clinical knowledge to the modeling process. Clinical knowledge originates from many sources, such as the physics of imaging acquisition, anatomy science, and the statistics of the population. Infusing this knowledge into the modeling process would allow the model to better utilize available data, making the data efficiency much higher. Taking chest X-ray disease classification as a demonstration, Li et al. (2019c), Li et al. (2020a), and Gozes and Greenspan (2018) modeled bone structure in chest X-rays and suppressed it before forwarding the images into lung disease classification models. Classification accuracy was shown to improve consistently with the same amount of data.

A more relevant task to this dissertation is *radiology report generation*, where models transform X-ray radiographs into summarizing textual reports describing findings in the images, similar to how a radiologist writes medical reports. The natural imaging counterpart is *image captioning* emerging from the Microsoft COCO challenge (Lin

et al., 2014a). Models compete for the most readable, accurate, and linguistically correct captions for a natural image, and several most successful works include *Show and Tell* (Vinyals et al., 2015), *Show, Attend, and Tell* (Xu et al., 2015), and *Self-Critical Sequence Training* (Rennie et al., 2017). These works have progressively added *recurrent neural network (RNN)*, *attention mechanism*, and *reinforcement learning (RL)* on top of the underlying *convolutional neural network (CNN)* feature extractor. Medical imaging researchers then transplanted the aforementioned techniques for medical report generation (Zhang et al., 2018a; Wang et al., 2018; Li et al., 2018), which showed initial seeming success: they produced fluent medical reports that appear to be written by medical professionals but failed to accurately describe correct disease findings (Boag et al., 2020).

To make these modeling methods useful in the medical domain, it is crucial for these modeling efforts to not only focus on the widely studied *language fluency* but also *clinical efficacy*. In this work, I aim to explore and advocate strategies to model radiology report generation that integrates medical prior knowledge into the model in order to (a) make the best use of the limited data, and to (b) enable clinical relevance for this line of research.

## 2.2 Computational Motivations

### 2.2.1 Annotation Reduction Approaches for Data Hunger Relief

To tackle the challenge with sparse labels and heterogeneous data in medical imaging, there is a wide array of methods developed to actively take advantage of representations learned across different data domains and clinical tasks. Several widely studied methods are (a) semi-supervised and unsupervised learning, (b) reinforcement learning, (c) cross-modal learning, and (d) transfer learning.

*Semi-supervised learning* operates by exploring both the relationships between the input image and the output annotations and the relationships among the input data

itself. The earliest instances in medical imaging are semi-supervised segmentation of cardiac MRI (Bai et al., 2017), retinal fundus images (Sedai et al., 2017), and biomedical imaging data (Gu et al., 2017). *Unsupervised learning* removes the reliance on any explicit annotation for modeling and hence is applicable to specific tasks such as medical image registration (Qin et al., 2019) and contrastive model pre-training (Hu et al., 2020).

*Reinforcement learning (RL)* focuses on indirectly learning models with typically non-differentiable objectives and sparser annotations. As an example, in a larger context of AI modeling, to teach a modeling agent to play Atari video games (Mnih et al., 2013), oftentimes the only supervision signal available is the total in-game score, which is typically a sum of several minor objectives weighted differently. As much as we desire to optimize the scores, in a non-differentiable environment we are unable to simply obtain a *gradient* of the reward with respect to the input as a way to indicate how we should finetune the model parameters. RL approaches this issue by allowing the agent to perform sampled or carefully chosen actions, each leading to distinct rewards, and by learning what actions lead to more optimal rewards, the modeling agent progressively becomes better even if supervision is sparse.

This reinforcement learning paradigm is effective in associating a large quantity of data with a smaller amount of annotations via systematically noisy modeling, which benefits segmentation greatly as conventionally segmentation annotations are labor-intensive. Yang et al. (2019) and Qin et al. (2020) used RL to optimize the augmentation pipeline and the segmentation models simultaneously. Liao et al. (2020) interestingly involved human experts in the loop to indicate whether segmentation is satisfactory in 3D MRI images.

*Cross-modal learning* introduces extra data modalities to the task at hand to improve model generalization by exploiting the distribution similarity between the domains. Medical textual reports have been commonly used to assist medical image analysis owing to their frequent co-occurrence with images. Weng et al. (2019) combined slide-level pathology images and case-level pathology reports by concatenating their features for multi-objective classification. Li et al. (2018) also merged both

chest X-ray image embedding and finding report textual embedding to jointly classify into thoracic disease types. [Chauhan et al. \(2020\)](#) leverages free-text reports in chest radiograph studies to predict pulmonary edema severity score more accurately than using the radiographs alone.

*Transfer learning* puts an emphasis on applying the knowledge learned from a source data domain to a target data domain, the distribution of which can be different from the source domain. One can learn a model from a publicly available large dataset and apply necessary fine-tuning to achieve satisfactory results on a local dataset ([Raghu et al., 2019b](#)). Med3D ([Chen et al., 2019b](#)) combined multiple datasets from organ to lesion segmentation to derive a generic pre-trained model suitable for all medical segmentation tasks. [Liu et al. \(2018\)](#) learned 2D feature encoding networks before transferring them to anisotropic 3D features.

The machine learning methods mentioned in this section have greatly reduced the annotations needed in terms of their quantity or types by intelligently utilizing the raw imaging data, conveniently coexisting data of other modalities, and expensive annotations.

### **2.2.2 Federated Learning for Data Collaboration**

Federated learning (FL) is not a new concept: the idea of establishing a multi-client system with locally available data but learning with a collaborative algorithm has been proposed as early as the 1990's ([Soueina et al., 1998](#)). The essential idea of FL is that multiple participating agents, or *clients*, can contribute to a central model without revealing their local data for data privacy arguments, as opposed to the conventional *centralized learning*, where all data is collected to a central data-center for learning. This learning paradigm is especially suitable for the collaboration of multi-center medical imaging modeling, as medical data access is subject to research ethics approval and various data use agreements that unavoidably impose frictions for medical machine learning.

Prior to deep learning, there are research works that built ad-hoc FL strategies for multivariate models ([Meeker et al., 2015](#)), Cox regression ([Lu et al., 2015](#)), and

logistic regression (Li et al., 2016b) in medical machine learning. In the year 2017, McMahan et al. (2017) reintroduced FL under a deep learning context by synthesizing benchmarking datasets and proposing an algorithm for efficient model learning. The natural imaging community quickly picked up this learning paradigm to tackle FL for *non-independent and identically distributed (non-IID)* data where the clients hold very dissimilar local data and bring challenging learning scenarios for existing FL algorithms (Zhao et al., 2018; Sattler et al., 2019).

There are early adopters of FL in medical imaging as well. Sheller et al. (2018) was the first to extend FL into medical image segmentation and reported comparable performance for the resulting model against centralized learning. Li et al. (2019b) explored privacy-preserving variants of FL algorithms with the same brain tumor segmentation dataset.

While natural imaging researchers made thrusts in synthesizing new challenging learning scenarios including imbalanced class distribution (Wang et al., 2021a; Yang et al., 2021) and uneven client sizes (Chou et al., 2021), medical imaging researchers are expanding the horizon of FL-enabled medical tasks including brain tumor segmentation (Sheller et al., 2020), COVID classification (Feki et al., 2021; Zhang et al., 2021b; Dou et al., 2021; Abdul Salam et al., 2021), and cancer diagnosis (Lee et al., 2021b).

Despite all these efforts and advances, there is a critical missing research area for FL in medical imaging – in all the research works presented above, the researchers are not *blind* to the global data distribution. They retain the capability to train, verify, and compare the models in a centralized setting to the ones obtained with FL. Obviously, this approach is sensible in a research sandbox environment to fully assess the underlying data and algorithms, and yet in a real-world medical FL application, we do not have the ability to *peek* at the centralized model, let alone determine whether we have a satisfactory FL model at the end of learning compared to the inaccessible centralized learning performance. It is essential to offer insights for real-world applications to suggest, prospectively, how much data should be collected for each of the participating parties, and how different they are allowed to be, to learn a

reasonably effective model in a federated manner.

# Chapter 3

## Transfer Learning for Cancer Mortality Assessment on Restricted Institutional Data

The body mass index (BMI) has been frequently used to define disease-related conditions such as obesity (WHO, 2000). While BMI is easy to obtain in the clinics for obesity characterization and tracking at a population level, it is often evidenced to capture lower association with cardiometabolic risk than muscle mass (Neeland et al., 2018). Visceral adipose tissue (VAT), subcutaneous adipose tissue (SAT), and lean muscle, on the other hand, are considered to have a direct linkage to cardiovascular disease, oncologic diseases (Neeland et al., 2019), and sarcopenia-induced mortality (Sayer et al., 2008).

These body composition measurements are prohibitively complex in nature to obtain from three-dimensional computed tomography (CT) images for clinicians, as all imaging slices require full pixel-wise annotations. While there are prior studies to estimate body composition measurements with convolutional neural networks (CNNs)

---

This chapter is adapted from the published article “*Artificial Intelligence to Assess Body Composition on Routine Abdominal CT Scans and Predict Mortality in Pancreatic Cancer – A Recipe for Your Local Application*” (Hsu et al., 2021) to which I have contributed as the first author.

producing intermediate segmentation outputs, and approximating composition measurements with single-slice results (Bridge et al., 2018; Burns et al., 2020; Hashimoto et al., 2019; Weston et al., 2019; Park et al., 2020), the studies were still using CT datasets with scan counts on the order of hundreds. In a more realistic research setup in smaller local medical imaging registries where there are only tens of CT scans and annotations available, having a publicly available dataset to bootstrap the training process first, then performing transfer learning to local data would be an appealing approach.

We use the liver tumor segmentation (LiTS) challenge dataset and a local pancreatic cancer registry dataset to verify the efficacy of models that predicted body composition from CT studies, trained and tested under various scenarios. We compare the Dice score of the segmentation and utilize the predicted fat/muscle measurements to characterize mortality rates using the Kaplan-Meier curve stratified by presence/absence of sarcopenia and high/low VAT measurements.

### 3.1 Overview

Body composition is associated with mortality; however its routine assessment is too time-consuming. To demonstrate the value of artificial intelligence (AI) to extract body composition measures from routine studies, we aim to develop a fully automated AI approach to measure fat and muscles masses, to validate its clinical discriminatory value, and to provide the code, training data and workflow solutions to facilitate its integration into local practice.

We develop a neural network that quantify the tissue components at the L3 vertebral body level using data from the LiTS Challenge and a pancreatic cancer cohort. We classify sarcopenia using accepted skeletal muscle index cut-offs and visceral fat based its median value. We use Kaplan Meier curves and Cox regression analysis to assess the association between these measures and mortality.

Applying the algorithm trained on LiTS data to the local cohort yields good agreement ( $>0.8$  intraclass correlation, ICC); when trained on both datasets, it has



excellent agreement ( $>0.9$  ICC). The pancreatic cancer cohort has 136 patients (mean age:  $67 \pm 11$  years; 54% women); 15% have sarcopenia; mean visceral fat is  $142 \text{ cm}^2$ . Concurrent with prior research, we find a significant association between sarcopenia and mortality (mean survival of  $15 \pm 12$  vs.  $22 \pm 12$  ( $p < 0.05$ ), adjusted HR of 1.58 (95% CI: 1.03 – 3.33)) but no association between visceral fat and mortality. The detector analysis takes  $1 \pm 0.5$  s.

AI body composition analysis can provide meaningful imaging biomarkers from routine exams demonstrating AI's ability to further enhance the clinical value of radiology reports.

## 3.2 Background

A rapidly growing literature on body composition has shown that anthropometric measurements, such as body mass index (BMI), are insufficient biomarkers to capture both cardiometabolic risk and muscle mass (Neeland et al., 2018). The measurement of body composition by imaging modalities such as computed tomography allows for a more precise quantitative assessment of these measures and provides insight into important clinical implications of visceral adipose tissue (VAT) and lean muscle mass. VAT, which is associated with proinflammatory activity, is considered an important risk factor for diabetes, cardiovascular disease, and several oncologic diseases (Neeland et al., 2019), whereas loss of lean muscle mass, as seen in sarcopenia, is associated with higher morbidity, disability, and mortality (Sayer et al., 2008) and is increasingly recognized as an important marker of poor prognosis in several neoplasms (Faron et al., 2020; Kamarajah et al., 2019; Lee and Giovannucci, 2018; Ojima et al., 2019).

Good approximation of these measures from a single axial slice CT or MRI slice at the level of the L3 vertebral body has been extensively validated (Schweitzer et al., 2015, 2016). While these imaging biomarkers can be readily obtained from any abdominal CT scans, it is too time-consuming to assess quantitatively in routine practice.

With the rapid development of convolutional neural networks (CNN) for image segmentation, several studies have presented mostly semiautomatic approaches to de-

living body composition measures (Bridge et al., 2018; Burns et al., 2020; Hashimoto et al., 2019; Weston et al., 2019; Park et al., 2020). However, for such a workflow to be fully automated, the algorithm must first identify a slice at the cranial aspect of the L3 and then segment the different tissue compartments at this level. Only one prior study (Bridge et al., 2018) included automation of both stages while others have manually selected the appropriate slice (Burns et al., 2020; Hashimoto et al., 2019; Weston et al., 2019; Park et al., 2020).

Given this, the first objective of this study is to develop a fully automated approach to body composition assessment for immediate use in the radiology community. Our second objective is to test the generalizability and clinical discriminatory value of our algorithm in a second independent dataset of patients with pancreatic cancer and specifically to assess how sarcopenia and visceral fat, both as quantified by the detector, predict mortality in this cohort. Finally, our third objective is to describe how we are able to integrate such a detector into our routine clinical workflow. By making development data and code available, other data scientists will have the opportunity to expand upon our work.

## 3.3 Materials and Methods

### 3.3.1 Data

We use two datasets to train and test the system.

#### **Liver Tumor Segmentation Challenge (LiTS) Dataset**

These are publicly available CT exams of the abdomen/pelvis or torso with contrast of 201 patients with colorectal cancer from multiple institutions (Bilic et al., 2019). No aggregated patient characteristics are available. Data resolution ranged from 0.6 to 1.0 mm in-plane and 0.5 to 6.0 mm in the  $z$ -direction.

## Local Pancreatic Cancer Registry Dataset

In this IRB-approved, HIPAA-compliant registry, we collect information on all patients with a newly diagnosed pancreatic cancer who presented to the multidisciplinary pancreatic clinic since 2014. Survival data are updated in quarterly intervals. The registry captures baseline demographics, including gender and age at diagnosis as well as the initial tumor stage. A sub cohort of this registry that was enrolled between January 2016 and December 2017, for a total of 136 patients, is used for this analysis. We use the initial staging CT abdomen/pelvis of the treatment naïve patients. Data resolution ranges from 0.6 to 1.0 mm in-plane and 2.0 to 5.0 mm in the  $z$ -direction.

### 3.3.2 Body Composition Measurement

For annotation, we randomly choose 40 CT volumes from the LiTS dataset and 40 from the local pancreatic registry. In both datasets, two board-certified radiologists (AM and KL) manually identify the target slice (uppermost level of the L3 vertebral body) and segment the tissue compartments (muscle, subcutaneous fat and visceral fat) at that level using ITK-snap (Yushkevich et al., 2006). Subsequently, they review each other’s work and resolve discrepancies by consensus.

### 3.3.3 Algorithm

Figure 3-1 depicts the overall workflow of the algorithm that consists of two stages. During stage one, the axial slice that contains (or is closest to) the cranial endplate of the L3 vertebral body is identified and extracted from the full CT volume. During stage two, the different body composition types (muscle, subcutaneous fat and visceral fat) are automatically annotated and quantified. Prior to training, the images are windowed to values between  $-150$  and  $250$  HU, prior to on-the-fly augmentation during which random brightness, random contrast adjustments (up to  $\pm 30\%$ ) and random rotations (up to  $\pm 20^\circ$ ) are added. During test time, only windowing is applied.

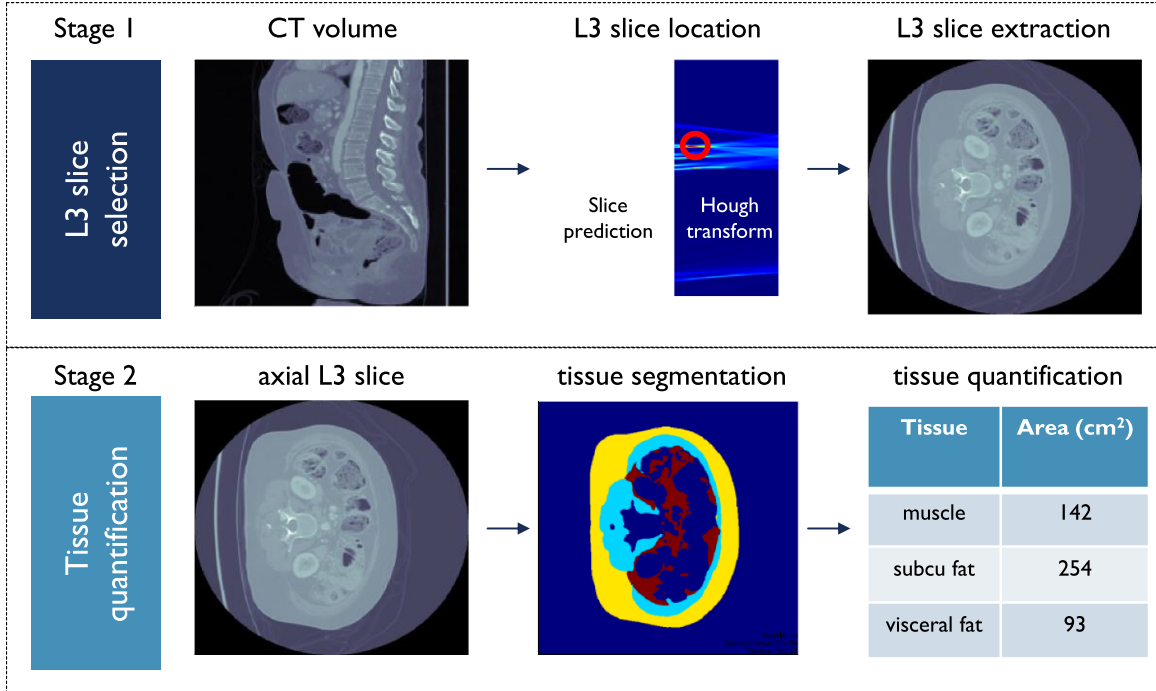


Figure 3-1: Overview of 2-stage, end-to-end algorithm.

### Stage One: Localization of the L3 Level

Beginning with the entire CT volume, the algorithm first identifies the cranial end-plate of the L3 vertebral body, formulated as a binary classification problem. Each axial slice in the volume is treated as an individual instance and labeled positive if cranial to the L3 slice and negative otherwise.

We develop a ResNet-18 model (He et al., 2016) to classify each slice. Passing the CT volume serially into the model allows us to obtain the probability of each slice’s location being cranial to the L3 slice. This curve closely approximates a sigmoid function  $p(z) = \sigma((z - z_0)/t)$ , where  $\sigma$  is the sigmoid function,  $z_0$  is the  $z$ -coordinate of the L3 slice,  $t$  is a fitted parameter, and  $z$  is the  $z$ -coordinate of the slice of interest. We adopt the Hough algorithm (Ballard, 1981) to transform the probabilities into likelihood values in the parameter space  $(z_0, t)$  in which a more likely set of parameters possessed a higher weight.

Due to potential noise in the probability array, we choose the most likely  $k$  parameter sets from the Hough algorithm, proposing  $k$  corresponding axial slices per

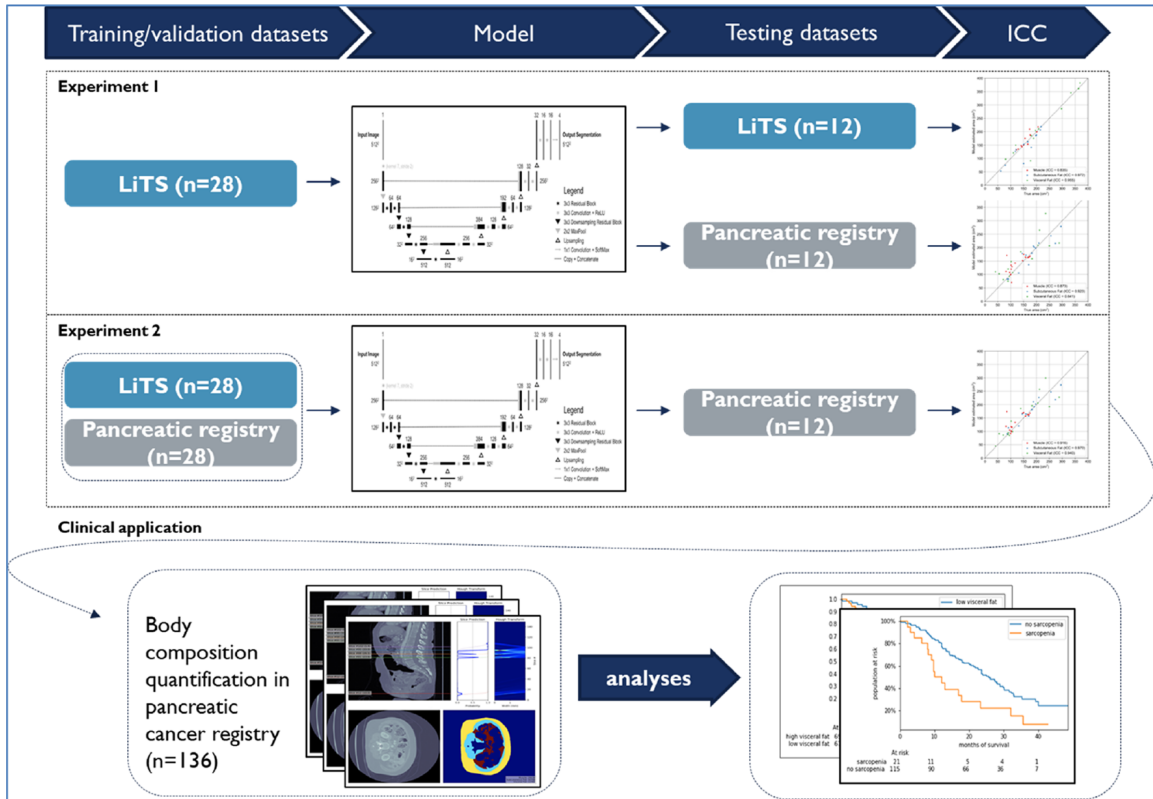


Figure 3-2: Overview of experimental set-up.

volume for the next stage of the pipeline.

## Stage Two: Segmentation of Individual Body Composition Components

We develop a 2D CNN based on the U-Net (Ronneberger et al., 2015b) architecture to segment the axial slice into the different compartments. The network consists of five down-sampling steps using ResNet-18 He et al. (2016) and five upsampling steps. It is optimized on cross-entropy, weighted inversely proportional to the class prevalence across the training set.

### 3.3.4 Analysis

#### Model Performance and Generalizability

**Experiment Setup** The experimental set-up consists of 3 steps that are illustrated in Figure 3-2. We first use the publicly available LiTS dataset to train our model

(“*Experiment 1*”). This will allow us make not only the model publicly available but also the training data. We randomly annotate 40 LiTS studies and randomly split them into 28 data to train and validate the model and 12 data to test the model. For *Experiment 2*, we annotate 40 datasets from the pancreatic registry and randomly split them in the same manner as the LiTS data. Subsequently, during the first part of experiment (“*Experiment 2a*”), we use the model developed on LiTS data (*Experiment 1*) and test it on the 12 test data from the pancreatic cohort. This will allow the reader to see how well the model generalizes to different datasets. During the second part of the experiment (“*Experiment 2b*”), we combine the training/validation datasets (for a total of 56) to train a new model (same architecture, just new weights) and then test it on the 12 pancreatic cancer test data. This will allow the reader to see the improved model accuracy if the available LiTS training data are *enriched* with local data. In the last step we use the model developed in *Experiment 2b* to predict the body composition measures for baseline CT scans for all 136 patients in the pancreatic cancer registry. We use these measures in the subsequent clinical analyses.

**Evaluation** The predictions are compared against the expert-labeled ground truth. To evaluate *Stage One*, we calculate the minimum localization error in the  $z$  direction across all the output slices. The mean error was then plotted against the number of proposed target slices.

To evaluate *Stage Two*, we compare the segmentation in terms of Dice score. For prediction region  $\mathcal{X}$  and truth region  $\mathcal{Y}$ , the Dice score is calculated as  $2 \times |\mathcal{X} \cap \mathcal{Y}| / (|\mathcal{X}| + |\mathcal{Y}|)$ .

Finally to evaluate the complete system, the algorithm identifies the most confident L3 slice location within the CT volume, and segments the respective tissue compartments in that slice. We compare the quantified areas of each tissue against the ground truth in terms of absolute error, absolute percentage error, and intraclass correlation coefficient (ICC) (Bartko, 1966). In this comprehensive evaluation step, these measures are preferred over Dice scores as the latter may consistently under/or

overestimate the true value, resulting in systematic error.

## Clinical Discriminatory Value

We use the values estimated by the detector trained in *Experiment 2*.

**Muscle Mass** We estimate the L3 skeletal muscle index (SMI) by dividing the cross-sectional skeletal muscle area ( $\text{cm}^2$ ) by the square of the height ( $\text{m}^2$ ). The cut-off values for sarcopenia are defined as  $43.75 \text{ cm}^2/\text{m}^2$  for men and  $38.5 \text{ cm}^2/\text{m}^2$  for women (Prado et al., 2009).

**Visceral Fat** We estimate the L3 visceral fat area ( $\text{cm}^2$ ) and then divide the cohort into two groups based on the median value for visceral fat, labeled as *low* (below the median) and *high* (above the median) visceral fat.

**Statistical Analyses** We compare proportions with a chi-squared test, age with a *t*-test and the other continuous variables with a Mann-Whitney-*U* test. Muscle mass and visceral fat are correlated to patient’s survival status with survival analysis stratified by presence of sarcopenia and compared using a log rank test. Cox regression analysis is used to control for confounding variables including age and tumor stage at diagnosis, gender and BMI.

## 3.4 Results

### 3.4.1 Model Performance and Generalizability

#### *Stage One Evaluation*

Stage One of the algorithm identifies the L3 vertebral body slice from the CT volume, hence the evaluation considers the location error; *i.e.*, deviation of the selected slice in the *z*-axis with regard to the upper edge of the L3 vertebral body. In *Experiment 1a* (trained on LiTS data and tested LiTS data), the absolute location error was 13.7 mm (95% CI: 0.0 – 28.7) when only a single proposed slice is considered. The minimum

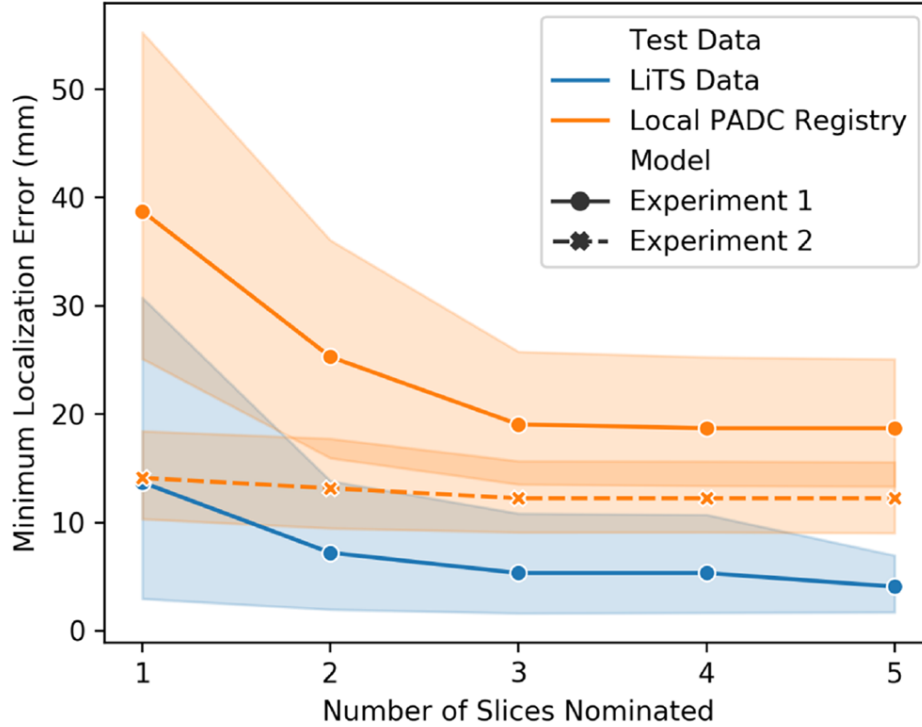


Figure 3-3: Performance of the slice localization. The minimum localization error among all identified slices is plotted against the number of nominated slices.

location error decreased with an increasing number of proposed slices as shown in Figure 3-3, and is 4.1 mm (95% CI: 1.2 – 7.0) when five slices are nominated. The same model yielded consistently higher errors when tested on the local pancreatic cancer registry data ranging from 38.7 mm (95% CI: 22.7 – 54.8) for a single slice to 18.7 mm (95% CI: 12.6 – 24.8) when 5 slices are proposed (*Experiment 1b*). In *Experiment 2*, when the model is trained on data combining LiTS and local pancreatic cancer registry data the errors decreased from 14.1 mm (95% CI: 9.9 – 18.3) to 12.2 mm (95% CI: 9.0 – 15.5) with increasing slice proposals.

### ***Stage Two Evaluation***

In *Experiment 1*, Dice scores for muscle, subcutaneous fat, and visceral fat are 0.92 (95% CI: 0.91 – 0.93), 0.93 (95% CI: 0.90 – 0.95), and 0.89 (95% CI: 0.86 – 0.92) for when a model developed on LiTS data is tested on LiTS data, and 0.83 (95% CI: 0.80 – 0.86), 0.90 (95% CI: 0.88 – 0.93), and 0.76 (95% CI: 0.70 – 0.81) when the same model is tested on data from the local pancreatic cancer registry. In *Experiment 2*,



these are 0.85 (95% CI: 0.83 – 0.88), 0.92 (95% CI: 0.91 – 0.93), and 0.80 (95% CI: 0.77 – 0.83).

## Full System Evaluation

Table 3.1 reports the evaluation of the complete (end-to-end) system which is most relevant to the clinical implications of the algorithm. For each of three experiments, we report the three body composition compartments; we report absolute and relative errors in area segmentation as well as the intraclass coefficient. Overall, the algorithm that is only trained in the LiTS data (*Experiment 1*) performs worse on the pancreatic test dataset than when tested in the LiTS test dataset. However, when the algorithm is trained on data from both datasets (*Experiment 2*), its performance on the pancreatic test dataset was similar to the performance on the LiTS test dataset (Table 3.1). Note that the data augmentation in *Experiment 2* increases not only the data diversity, but also the effective training set size. Both aspects benefit the performance for the resulting model.

### 3.4.2 Muscle Mass and Visceral Fat – Their Clinical Discriminatory Value

The pancreatic cancer cohort includes 136 patients with a mean age of  $67 \pm 11$  years, 54% (73/136) women, 13% (17/136) with stage I, 31% (42/136) stage II, 18% (25/136) and 38% (52/136) stage IV disease. 15% (21/136) of the patients have sarcopenia at their baseline scan and there is no statistically significant difference in demographics and baseline staging parameters between the two groups as shown in Table 3.2. The median value for visceral fat is  $142 \text{ cm}^2$ . There was a higher proportion of women in the low visceral fat group (38% versus 15%,  $p < 0.01$ ) but baseline staging parameters are not statistically different between the two groups (Table 3.2).

Mean follow-up is  $21 \pm 12$  months, and 65% (89/136) of patients expired during the follow up. The mean survival in the non-sarcopenia group is  $22.2 \pm 12.0$  versus  $14.6 \pm 11.7$  months in the sarcopenia group ( $p < 0.01$ , Figure 3-4a). At the same time,

Table 3.1: Performance of the overall system in quantifying the 3 different tissue components (compared to the manually annotated ground truth). 95% CI labeled in brackets.

Tissue	Area		
	Muscle	Subcutaneous Fat	Visceral Fat
<b>Experiment 1a (algorithm trained on LiTS data, tested on LiTS data)</b>			
Mean Absolute Error (cm <sup>2</sup> )	10.86 (5.20 – 16.51)	14.91 (3.83 – 25.99)	17.84 (5.07 – 30.62)
Mean Absolute Percentage Error (%)	6.66 (3.36 – 9.96)	10.22 (3.09 – 17.35)	11.06 (3.07 – 19.04)
Intraclass Correlation Coefficient (ICC)	0.835 (0.43 – 0.95)	0.972 (0.90 – 0.99)	0.955 (0.85 – 0.99)
<b>Experiment 1b (algorithm trained on LiTS data, tested on pancreatic data)</b>			
Mean Absolute Error (cm <sup>2</sup> )	21.10 (10.36 – 31.83)	22.29 (9.79 – 34.80)	36.95 (22.86 – 51.05)
Mean Absolute Percentage Error (%)	20.85 (8.05 – 33.66)	10.46 (6.31 – 14.60)	34.89 (13.02 – 56.77)
Intraclass Correlation Coefficient (ICC)	0.873 (0.62 – 0.96)	0.923 (0.77 – 0.97)	0.841 (0.53 – 0.95)
<b>Experiment 2 (algorithm trained on LiTS and pancreatic data, tested on pancreatic data)</b>			
Mean Absolute Error (cm <sup>2</sup> )	16.65 (5.42 – 27.88)	20.18 (7.98 – 32.39)	27.10 (15.26 – 38.94)
Mean Absolute Percentage Error (%)	16.64 (3.13 – 30.15)	9.39 (5.54 – 13.24)	19.67 (10.96 – 28.39)
Intraclass Correlation Coefficient (ICC)	0.916 (0.75 – 0.97)	0.970 (0.91 – 0.99)	0.940 (0.82 – 0.98)

Table 3.2: Baseline characteristics of pancreas cohort stratified by sarcopenia status.

	No Sarcopenia <i>N</i> = 115	Sarcopenia <i>N</i> = 21	<i>p</i>	Low Visceral Fat <i>N</i> = 68	High Visceral Fat <i>N</i> = 68	<i>p</i>
Age in years (mean $\pm$ SD)	67.4 $\pm$ 11.7	67.0 $\pm$ 8.0	0.91	65.6 $\pm$ 14.4	69.1 $\pm$ 9.4	0.06
Women, <i>N</i> (%)	61 (44)	12 (9)	0.72	48(35)	25 (18)	<0.001
T-stage, <i>N</i> (%)			0.12			0.77
T1	14 (10)	3 (2)		9 (7)	8 (6)	
T2	40 (29)	2 (1)		19 (14)	23 (17)	
T3	19 (14)	6 (4)		12 (9)	13 (10)	
T4	42 (31)	10 (7)		29 (21)	23 (17)	
SMI (cm <sup>2</sup> muscle mass/m <sup>2</sup> height) (mean $\pm$ SD)	53.0 $\pm$ 11.1	32.6 $\pm$ 11.8	<0.001	46.4 $\pm$ 11.1	53.3 $\pm$ 14.7	<0.01
Visceral fat (cm <sup>2</sup> ) (mean $\pm$ SD)	165 $\pm$ 111	127 $\pm$ 76	0.10	81 $\pm$ 3	241 $\pm$ 96	<0.001
BMI (kg/m <sup>2</sup> ) (mean $\pm$ SD)	27.8 $\pm$ 5.2	22.4 $\pm$ 3.8	<0.001	24.2 $\pm$ 4.3	29.7 $\pm$ 4.9	<0.001
Survival (months) (mean $\pm$ SD)	22.2 $\pm$ 12.0	14.6 $\pm$ 11.7	<0.01	20.8 $\pm$ 13.0	21.2 $\pm$ 11.5	0.40

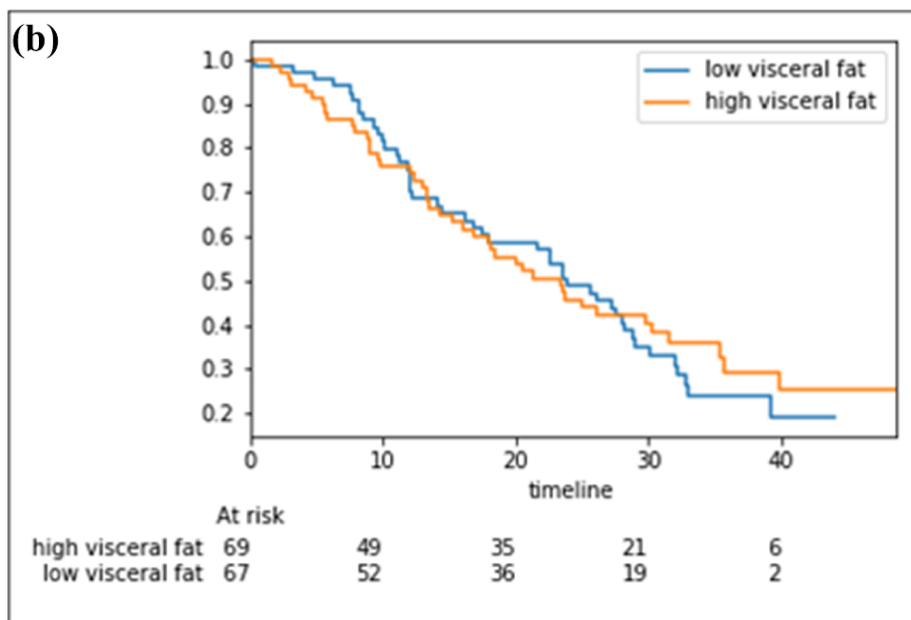
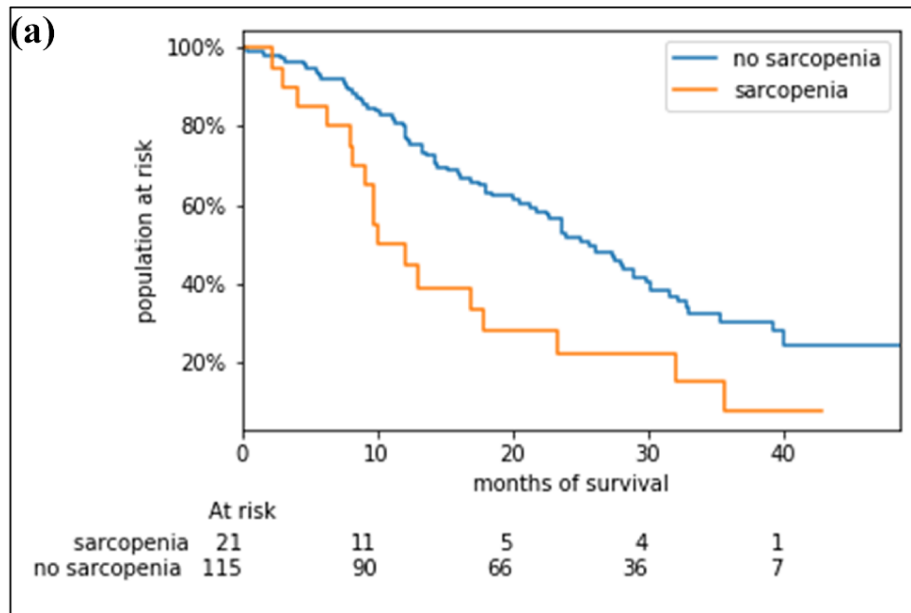


Figure 3-4: Kaplan Meier curve stratified by presence/absence of sarcopenia and low versus high visceral fat as derived by the AI algorithm.

there is no statistically significant difference in survival times between the *high* and *low* visceral fat groups (Figure 3-4b).

In a multivariate Cox regression analysis, sarcopenia is associated with a significantly increased risk of mortality of 1.85 (95% CI: 1.02 – 3.33) after controlling for

Table 3.3: Cox regression model, death

<b>Overall</b>	<b>Sarcopenia Model HR (95% CI)</b>	<b>Visceral fat Model HR (95% CI)</b>
Sarcopenia vs Non-sarcopenia	<b>1.85 (1.02 – 3.33)</b>	–
Visceral fat, high vs low	–	<b>0.93 (0.54 – 1.60)</b>
Age (years)	1.04 (1.01 – 1.06)	1.01 (1.02 – 1.07)
Female versus male	0.98 (0.64 – 1.51)	0.92 (0.59 – 1.44)
BMI	1.01 (0.96 – 1.06)	1.00 (0.95 – 1.05)
T1	Reference	Reference
T2 vs T1	2.31 (0.87 – 6.09)	2.30 (0.87 – 6.07)
T3 vs T1	2.02 (0.72 – 5.62)	2.04 (0.73 – 5.68)
T4 vs T1	4.09 (1.60 – 10.5)	4.43 (1.74 – 11.3)

the patient’s age, gender, BMI, and baseline disease stage (Table 3.3). High versus low visceral fat is not associated with an increased hazard for mortality.

### 3.4.3 Clinical Implementation

The output of the system consists of a slightly simplified version of Figure 3-1 in which the Hough transformation display is absent and only the most likely slice location (instead of all 5) is displayed on the sagittal image. The average time for the entire system takes 0.5 – 1.0 second on a single CPU, depending on the number of slices in the CT volume. We set up a server within the our institution that hosts the detector. Axial images of all abdominal CTs performed at our institution are routed to this server immediately upon completion of the exam. The detector performs inference on all studies and stores the results. Any interpreting radiologist can access the analysis via a URL link from the workstation if s/he feels that providing this information is of clinical value.

### 3.5 Discussion

In this study, we develop a fully automatic system to derive body composition from CT exams of the abdomen/pelvis. We then demonstrate the ability of the algorithm to risk-stratify patients with newly diagnosed pancreatic cancer with regard to their mortality risk based on their imaging findings even when adjusted for other clinical and demographic risk factors. From experimental studies (with manual segmentation) the association between different body composition measures and poor outcomes has been established in many oncologic diseases (Kamarajah et al., 2019; Lee and Giovannucci, 2018). Specifically, a recent clinical study in patients with pancreatic cancer demonstrated the association between sarcopenia (measured manually) and poor clinical outcomes but did not find such an association for visceral fat (Babic et al., 2019). This is consistent with our findings.

The novelty of this work is that the same task could be accomplished using an end-to-end AI detector and performed within less than 2 seconds. By integrating the algorithm into the routine workflow of imaging data from the scanner to PACS, we are able to provide this information on all routine abdominal/pelvic CT scan. This information can be directly ascertained from the PACS workstation by the interpreting radiologist and can be incorporated into the report if clinically appropriate. We are planning on monitoring the number of reports that include body composition measures stratified by providers and clinical referrers to better understand a variation in adaption and possible reasons for it.

We make the code and development data publicly available with the goal to enable broad integration of these *imaging biomarker* into routine clinical workflow, allowing for endeavors such as sarcopenia screening (Graffy et al., 2019) or opportunistic metabolic syndrome screening (Pickhardt et al., 2020). We also describe the workflow for how such a detector can be easily integrated into the local clinical workflow that is PACS vendor agnostic and only requires a standard server in the local datacenter to host the detector in a self-contained, virtualized container that is easy to install and update. This is something that is available in nearly every medical center and

thus suggests that the reach of such AI-based detectors could be broad and have an important impact on the clinical value of radiology reporting.

Applying the algorithm that is trained on the public LiTS data to pancreatic cancer registry data yields a good agreement ( $>0.8$  intraclass correlation (ICC)). This first experiment simulates an *off the shelf* use in practice; *i.e.*, any investigator who is interested in this work can just use the pre-trained algorithm and apply it to local data. When the training dataset is expanded to also include the training data from the local pancreatic cancer dataset, the performance returned an excellent agreement ( $>0.9$  ICC) (*Experiment 2*). The latter underscores the feasibility to easily adapt the algorithm to local practice.

When compare to the only other complete end-to-end body composition detector (Bridge et al., 2018), the excellent ICC across the different body composition compartments is similar when tested against the test portion of the development dataset. When tested against local data, the model by Bridge et al. (2018) performs better than our non-refined model, likely due to our significantly smaller training dataset (506 versus 28). After enriching the primary training data to include 28 studies from the local pancreatic registry and retraining the model (*Experiment 2*), our algorithm performs equally well. These comparisons are based on the publication by Bridge et al. (2018) that did not release their development data or code.

Our work has several limitations. While verification at different stages of the models show promising performance in the development dataset, there is a generalization gap across datasets when tested on local pancreatic cancer data. Enriching the training data with local data mostly overcomes this issue. Note that in order to perform a better designed experiment, one would slowly augment the LiTS training dataset with a gradually growing number of data points: from the initial 28 LiTS studies to the final 56 (LiTS + local) studies. Over the continuous data augmentation curve, we can judge the efficacy of the added data better than in our current experiments.

Furthermore, we restrict the analysis to a single slice using a 2D U-net architecture even though volumetric assessment encompassing the entire torso may be more accurate and technically feasible either with a 3D model architecture (Çiçek et al.,

2016) or by stitching 2D slices (Weston et al., 2019). Yet, this is a broadly accepted proxy that can be further assessed in future studies. Finally, while all image labels are performed by two radiologists and disagreement is solved by consensus, we do not document the disagreement systematically. However, given that the anatomy is quite obvious, we can comment qualitatively that there is not a systematic disagreement between radiologists but rather small tracing inconsistencies.

### 3.6 Summary

In this chapter we develop a well-performing end-to-end body composition detector and demonstrate its clinical discriminatory value for sarcopenia in an exemplary cohort of patients with pancreatic cancer. In the mean time we show that by augmenting the publicly available data with insitutional-local data, the performance of the system can be improved reasonably.

By making training data and code available and outlining one possible way to integrate this detector into clinical workflow that is agnostic to the vendor and does not require resources beyond those available in nearly every academic medical center, we believe this work can be easily integrated into local practice. One of the immediate benefits of AI to improve patient care is to routinely provide these types of imaging biomarkers that are readily available from our information rich data but currently not fully appreciated due the time-consuming nature of obtaining these data manually.

Following the development of the body composition detector in this chapter, we will continue to demonstrate use cases that expands the usefulness of body composition measurements to COVID-19 patient cohort in the next chapter.



# Chapter 4

## Surrogate Endpoint from Limited Imaging Data for COVID-19 Severity Prediction

In the early phases of the COVID-19 outbreak, researchers have associated obesity with the risk of severe clinical outcomes, primarily using body mass index (BMI) as the indicator (Gao et al., 2020; Palaiodimos et al., 2020; Cai et al., 2020). While BMI is a convenient metric to measure, its link to metabolic health in patients is heterogeneous, thus making BMI an underperforming single-dimensional indicator of COVID risks. Visceral fat, or visceral adipose tissue (VAT), on the other hand, has gained increased evidence (Battisti et al., 2020; Kuk et al., 2006) that it provides a better explanation of COVID-19 severity.

Aside from the clinical effectiveness of VAT, due to the availability of (a) routinely acquired computed tomography (CT) images, (b) advances in artificial intelligence (AI) algorithms, and (c) visceral fat segmentation annotations with a high inter-rater agreement, we are able to utilize the visceral fat segmentation as a surrogate endpoint,

---

This chapter is adapted from the published article “*Visceral Adiposity and Severe COVID-19 Disease: Application of an Artificial Intelligence Algorithm to Improve Clinical Risk Prediction*” (Goehler et al., 2021) to which I have contributed as the second author.

and approximate VAT measurements with pixel count from the segmentation outputs.

## 4.1 Overview

Obesity has been linked to severe clinical outcomes among people who are hospitalized with coronavirus disease 2019 (COVID-19). We test the hypothesis that VAT is associated with severe outcomes in patients hospitalized with COVID-19, independent of BMI. We analyze data from the Massachusetts General Hospital (MGH) COVID-19 Data Registry, which includes patients admitted with polymerase chain reaction – confirmed severe acute respiratory syndrome coronavirus 2 infection from March 11 to May 4, 2020. We use a validated, fully automated AI algorithm to quantify VAT from CT scans during or before the hospital admission. VAT quantification takes an average of  $2 \pm 0.5$  seconds per patient. We dichotomize VAT as high and low at a threshold of  $\geq 100$  cm<sup>2</sup> and used Kaplan-Meier curves and Cox proportional hazards regression to assess the relationship between VAT and death or intubation over 28 days, adjusting for age, sex, race, BMI, and diabetes status.

A total of 378 participants have CT imaging in the dataset. Kaplan-Meier curves show that participants with high VAT had a greater risk of the outcome compared with those with low VAT ( $p < 0.005$ ), especially in those with BMI  $< 30$  kg/m<sup>2</sup> ( $p < 0.005$ ). In multivariable models, the adjusted hazard ratio (aHR) for high vs low VAT is unchanged (aHR is 1.97 with 95% CI: 1.24 – 3.09), whereas BMI is no longer significant (aHR for obese versus normal BMI is 1.14 with 95% CI: 0.71 – 1.82). High VAT is associated with a greater risk of severe disease or death in COVID-19 and can offer more precise information to risk-stratify individuals beyond BMI. AI offers a promising approach to routinely ascertain VAT and improve clinical risk prediction in COVID-19.

## 4.2 Background

People with obesity who become infected with the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) have a greater risk of severe clinical outcomes (Gao et al., 2020; Palaiodimos et al., 2020; Cai et al., 2020). In the United States, over 160 million Americans are overweight or obese and over 500,000 individuals have died of COVID-19 since the start of the pandemic. However, clinical outcomes due to this infection are not uniformly worse among those who have obesity, and the mechanisms that link body habitus and clinical outcomes in people with COVID-19 remain poorly understood.

While body mass index is a convenient measure to obtain in clinical practice, it is widely recognized as a remarkably heterogeneous parameter for assessing metabolic health (Neeland et al., 2019). As such, people with similar BMI measurements have shown meaningfully different levels of health risk, in part due to the fact that BMI is not a reliable measure of total body or central abdominal fat mass and does not well capture wide variation in VAT distribution between individuals (Neeland et al., 2019). There is a growing body of evidence that VAT may be an important conduit for the health risk associated with obesity. Macrophages have been shown to infiltrate the hypertrophied adipocytes that are characteristic of excess VAT; this is believed to result in increased inflammatory cytokines including both tumor necrosis factor  $\alpha$  and interleukin-6 in this tissue (Neeland et al., 2019). Given this, VAT has been proposed as a factor that may help to elucidate the relationship between body weight and COVID19 disease severity (Battisti et al., 2020; Kuk et al., 2006). While VAT measurement is usually manually assessed by a radiologist and thus time-consuming to perform, AI algorithms offer a novel approach to measuring VAT quickly and accurately in patients who have recently had CT imaging performed, regardless of indication (Choy et al., 2018). Moreover, segmentation of the tissue compartments from a single cross-sectional slice at 1 lumbar vertebra can well approximate VAT and subcutaneous adipose tissue (SAT), both of which can be ascertained in seconds using AI algorithms (Schweitzer et al., 2015; Hsu et al., 2021).

In this chapter, we test the hypotheses that VAT is associated with severe outcomes in patients who are hospitalized with COVID-19 and is a stronger predictor of such outcomes than BMI in adjusted models including both indicators. We assess VAT using a fully automated end-to-end AI algorithm that provides this measure from the CT scans of patients who were hospitalized at MGH with COVID-19 disease during the first surge of the 2020 pandemic. We then use the VAT measure and survival analysis to test hypotheses linking high VAT and poor clinical outcomes over 28 days from hospitalization for COVID-19.

## 4.3 Methods

### 4.3.1 Data Source

This study use data from the MGH COVID-19 Data Registry ([Bassett et al., 2020](#); [Seigle et al., 2020](#)). The registry includes all patients who presented to care, defined as the first contact with the health care system for evaluation of COVID-19 symptoms, and were subsequently hospitalized at MGH between March 11 and May 4, 2020. All participants in the registry had PCR-confirmed SARSCoV-2 infection. The data in the registry are collected in 2 ways. First, a manual chart review is performed to assess key aspects of the patient’s medical history and details of the hospitalization including the main outcomes of interest at 28 days after presentation to care ([Seigle et al., 2020](#)). This manual chart review also identifies comorbidities of interest in this study including history of coronary artery disease or myocardial infarction (CAD or MI), history of congestive heart failure (CHF), history of diabetes, history of renal disease, and history of chronic obstructive pulmonary disease (COPD). This chart review is undertaken by physicians, research nurses, and a team of research assistants trained in a standard operating procedure for data extraction. In addition, height, weight, and BMI, as well as key laboratory values that are measured and recorded during the index hospitalization, are obtained electronically through the Enterprise Data Warehouse (EDW), a repository that is derived from the Epic electronic medical records system.

There are no missing height or weight values in this sample. Imaging data of CT exams performed during or before the hospitalization for any indication that were used to ascertain the VAT measures are also obtained from hospital picture archive and communication system (PACS). This research is approved by the Massachusetts General Brigham Institutional Review Board protocol 2020P000829.

BMI is calculated as the weight in kilograms divided by the square of height in meters. BMI categories were defined using standard thresholds of  $<18.5$  kg/m<sup>2</sup> for underweight,  $18.5 - 24.9$  kg/m<sup>2</sup> for normal weight,  $25.0 - 29.9$  kg/m<sup>2</sup> for overweight, and  $\geq 30.0$  kg/m<sup>2</sup> for obese. Diabetes is defined by meeting at least one of the following criteria:

1. medical history of diabetes is documented in the medical record and manually retrieved on chart review,
2. HbA<sub>1c</sub>  $\geq 6.5\%$  during the index hospitalization, or
3. random blood glucose  $\geq 200$  mg/dL at admission to the hospital and supportive history by chart review.

For those cases in which only the third diagnostic criterion is met, a detailed chart review was performed by two board-certified endocrinologists; this procedure has been described in detail previously. Of note, registry participants with active malignancy are excluded from this study. Demographic and clinical characteristics are defined as previously described (Bassett et al., 2020; Seiglie et al., 2020).

### **4.3.2 Ascertaining VAT and SAT Using an AI-Based Body Composition Detector**

Body composition measures such as VAT can be ascertained from a single axial CT or magnetic resonance imaging (MRI) slice (Schweitzer et al., 2015). We use a previously validated, fully automated AI algorithm to quantify VAT from CT scans (Hsu et al., 2021). In brief, this application is written in Tensorflow 1.13 (Abadi et al., 2016) and uses an end-to-end 2-stage artificial neural network that first localizes a single axial

slice at the L1 vertebral body level and then quantifies VAT in that specific slice in  $\text{cm}^2$ . We choose the L1 vertebral body as it is routinely included in both CT scans of the chest and CT exams of the abdomen/pelvis and has an excellent correlation with overall VAT in prior studies (0.986) (Schweitzer et al., 2015, 2016). Moreover, the validation of the AI algorithm itself shows excellent agreement with manual measurement by a radiologist. We use this AI algorithm to measure VAT in all patients with either a CT scan of the chest or a CT scan of the abdomen/pelvis that had been performed during the index hospitalization within a median (interquartile range) of 17 (4 – 25) months before the index hospitalization date. If both exam types are available, we choose the one that was temporally closest to the index hospitalization.

### 4.3.3 Exposures, Outcomes, and Statistical Analysis

Our primary exposure of interest is cross-sectional VAT area, which we dichotomize as high and low at a threshold of  $100 \text{ cm}^2$  based on prior literature demonstrating a meaningful increased risk of metabolic derangements above this threshold (Nicklas et al., 2003; Pickhardt et al., 2012; Yang et al., 2020). The primary outcome of interest in the study is need for intubation or death within 28 days after presentation to care. We first compare the demographic and health characteristics of those patients in the registry who had a relevant clinical imaging study to those for whom no applicable imaging study was conducted during the period of interest to assess for selection bias. Next, we compare differences in the demographic and clinical characteristics of the analytic sample among those with high versus low visceral fat. Then, we depict differences in time to death or intubation over 28 days among those with high vs low visceral fat using Kaplan-Meier curves and log-rank testing. We perform this analysis first in the full sample and then stratify for those who were in the normal or overweight BMI category and separately for those who were in the obese BMI category. We conduct a score test for proportional hazards assumptions. Then, we fit adjusted Cox proportional hazards models to estimate the hazard of 28-day death or intubation including VAT, adjusted for age, sex, race, and diabetes diagnosis in the models. We provide these models with and without adjustment for BMI. We also

provide a separate model with adjustment for BMI but without VAT included. A  $p$  value  $<0.05$  in the BMI-adjusted test of the association of visceral fat with COVID-19 outcomes indicates statistical significance. We also examine these same relationships in registry participants with and without imaging and separately conduct a stratified analysis among those who had imaging performed before vs during the hospitalization. Finally, we also provide an analysis in which VAT was categorized in quintiles and display the adjusted hazard ratio of VAT over a range of alternative thresholds with respect to the outcome of interest, as a further empirical assessment of the chosen threshold. Confidence intervals for the latter analysis were obtained via bootstrapping ( $1000\times$ ).

## 4.4 Results

The MGH COVID-19 Data Registry includes 866 individuals, among whom 410 (47.3%) have an abdominal or chest CT imaging study available during or before the hospitalization. Among these, 32 (7.8%) are excluded due to the presence of active malignancy, leaving a final sample of 378 registry participants for this analysis. A total of 268 of 378 (70.9%) people have a CT of the abdomen and pelvis, while 110 (29.1%) people have a CT scan of the chest available. 198 studies (52%) are performed during the hospitalization, and 180 studies (48%) are performed before the hospitalization. The total time to execute the analysis of an individual CT scan using the algorithm is  $2 \pm 0.5$  seconds on a standard CPU desktop computer within the hospital system.

There are several differences in demographic and health characteristics of those participants for whom imaging data is available compared with those participants who do not have imaging data. Individuals with available imaging are older, more likely to be male, and have a higher number of comorbidities, including diabetes and a history of CAD or MI, but the distribution of BMI and other demographic and health characteristics do not differ between these groups (Table 4.1). Participants who have a VAT  $\geq 100$  cm<sup>2</sup> have higher rates of diabetes and a higher C-reactive protein (CRP)

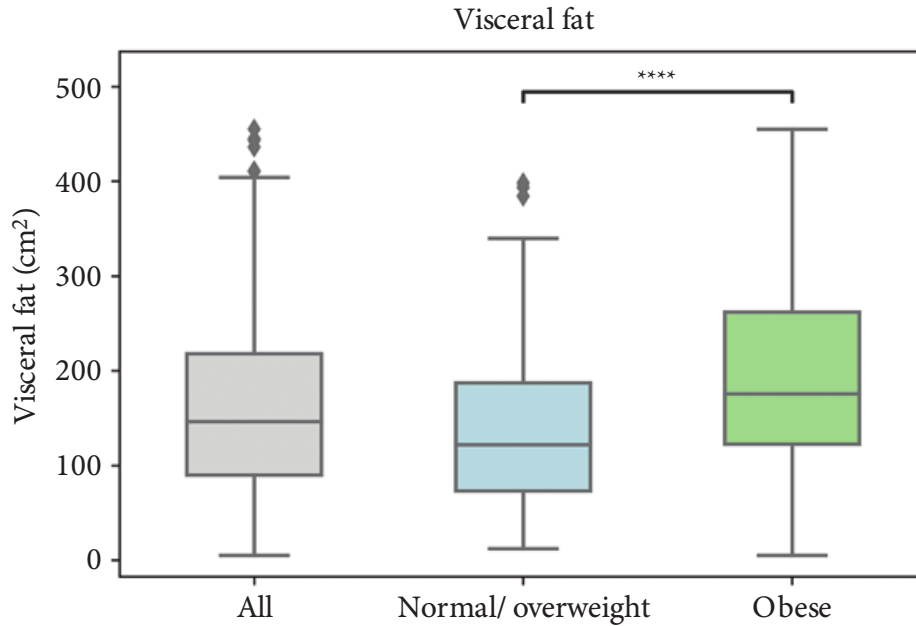


Figure 4-1: Visceral fat distribution, overall and by BMI category. \*\*\*\* $p < 0.0001$ .

on admission compared with those with a VAT  $< 100 \text{ cm}^2$ . (Table 4.2)

There are no significant differences in the rates of other key comorbidities, including CAD or MI and COPD, stratified by this VAT threshold. We find that the distribution of VAT differs significantly between those with a normal or overweight BMI compared with those with obesity ( $p < 0.0001$ ) (Figure 4-1). Specifically, the median VAT is greater among those in the BMI group with obesity compared with those in the normal or overweight BMI group (Figure 4-1). Exemplary VAT on body composition imaging by BMI status and gender is shown in Figure 4-3. In Figure 4-4, we also display the differences in the relationship between BMI and VAT by sex (Bredella, 2017).

There are 114 (38%) intubations and 54 (18%) deaths among 249 people by 28 days in the high-VAT group, compared with 15 (19%) intubations and 7 (9%) deaths among 129 people by 28 days in the low-VAT group. Kaplan-Meier curves from the total study sample show statistically significant differences in the risk of death or intubation over 28 days by VAT group (Figure 4-2). Those with high VAT have a greater risk of death or intubation over 28 days compared with those with low VAT ( $p < 0.001$ ). When stratifying the analysis into 2 groups defined by BMI (normal or



Table 4.1: **Differences in Demographic and Health Characteristics.** Comparison done between those with and without imaging data in the MGH COVID-19 Registry.

	Imaging Available ( <i>N</i> = 378)	No Imaging Available ( <i>N</i> = 488)	<i>p</i>
Age in years (mean ± SD)	63.3 ± 17.8	58.8 ± 18.2	<0.01
Male, %	61.7	52.7	<0.01
<b>Race or Ethnicity, %</b>			0.42
White	33.7	28.6	
Hispanic	24.1	27.0	
Black	9.0	8.0	
Other	33.2	36.4	
<b>Body Mass Index (BMI), %</b>			0.21
Normal	23.7	18.8	
Overweight	33.2	35.0	
Obese	43.1	46.2	
Diabetes	43.0	32.5	< 0.01
CAD or MI	21.8	10.6	< 0.01
COPD/Asthma	26.8	19.4	0.02
CHF	14.6	6.8	< 0.01
Renal disease	23.7	12.6	< 0.01
ESR (mean ± SD)	40.9 ± 27.1	45.6 ± 29.0	0.03
CRP (mean ± SD)	89.1 ± 80.8	97.2 ± 80.7	0.16

Abbreviations: BMI, body mass index; CAD, coronary artery disease; MI, myocardial infarction; COPD, chronic obstructive pulmonary disease; CHF, congestive heart failure; ESR, erythrocyte sedimentation rate; CRP, C-reactive protein; VAT, visceral adipose tissue.

Table 4.2: Differences in Demographic and Health Characteristics. Comparison done on 378 COVID-19 registry participants, overall and by VAT group.

	Overall ( <i>N</i> = 378)	VAT < 100cm <sup>2</sup> ( <i>N</i> = 115)	VAT ≥ 100cm <sup>2</sup> ( <i>N</i> = 263)	<i>P</i>
VAT (mean ± SD)	195 ± 107	63 ± 26	204 ± 80	< 0.01
Age in years (mean ± SD)	63.3 ± 17.8	62.2 ± 18.5	63.8 ± 16.0	0.41
Male, %	61.7	37.2	72.2	< 0.01
<b>Race or Ethnicity, %</b>				0.04
White	33.7	28.7	35.9	
Hispanic	24.1	21.7	25.2	
Black	9.0	15.7	9.0	
Other	33.2	33.9	29.9	
<b>Body Mass Index (BMI), %</b>				< 0.01
Normal	23.7	44.4	15.6	
Overweight	33.2	32.2	33.2	
Obese	43.1	23.4	51.2	
Diabetes	43.0	33.9	47.0	0.02
CAD or MI	21.8	20.9	22.1	0.78
COPD/Asthma	26.8	28.7	26.0	0.58
CHF	14.6	17.4	13.4	0.30
Renal disease	23.7	18.4	26.0	0.11
ESR (mean ± SD)	40.9 ± 27.1	37.5 ± 24.3	42.4 ± 28.1	0.13
CRP (mean ± SD)	89.4 ± 80.8	70.3 ± 70.8	97.5 ± 83.7	< 0.01

Abbreviations: BMI, body mass index; CAD, coronary artery disease; MI, myocardial infarction; COPD, chronic obstructive pulmonary disease; CHF, congestive heart failure; ESR, erythrocyte sedimentation rate; CRP, C-reactive protein; VAT, visceral adipose tissue.

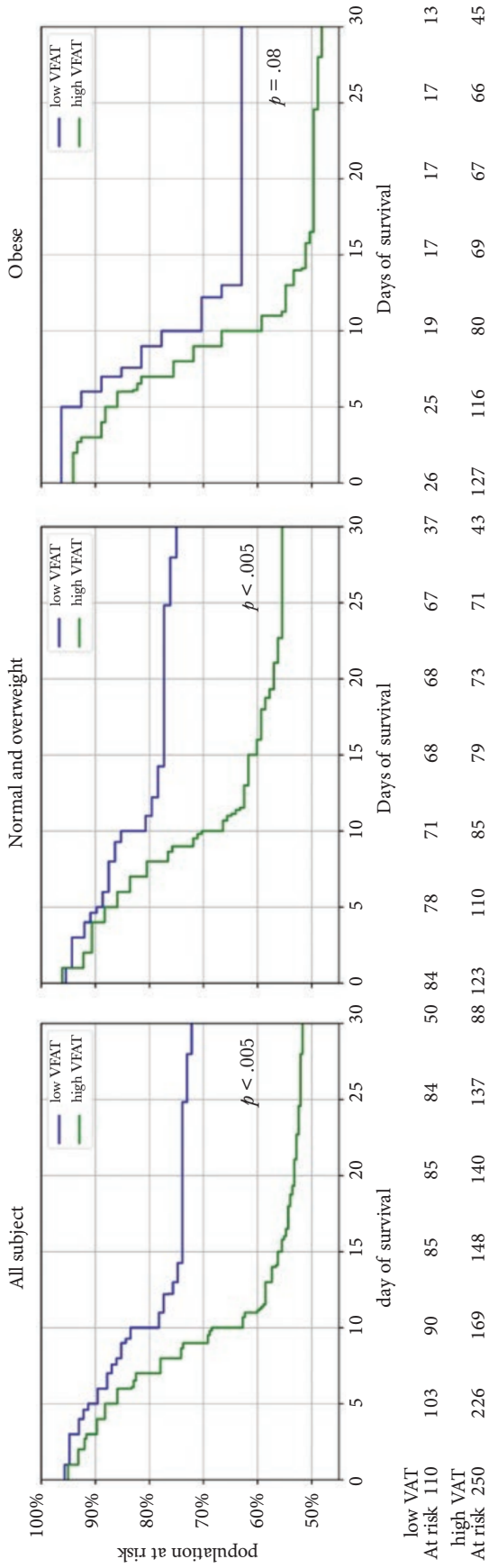


Figure 4-2: Kaplan-Meier curves for intubation or death within 28 days. Abbreviation: VAT, visceral adipose tissue.

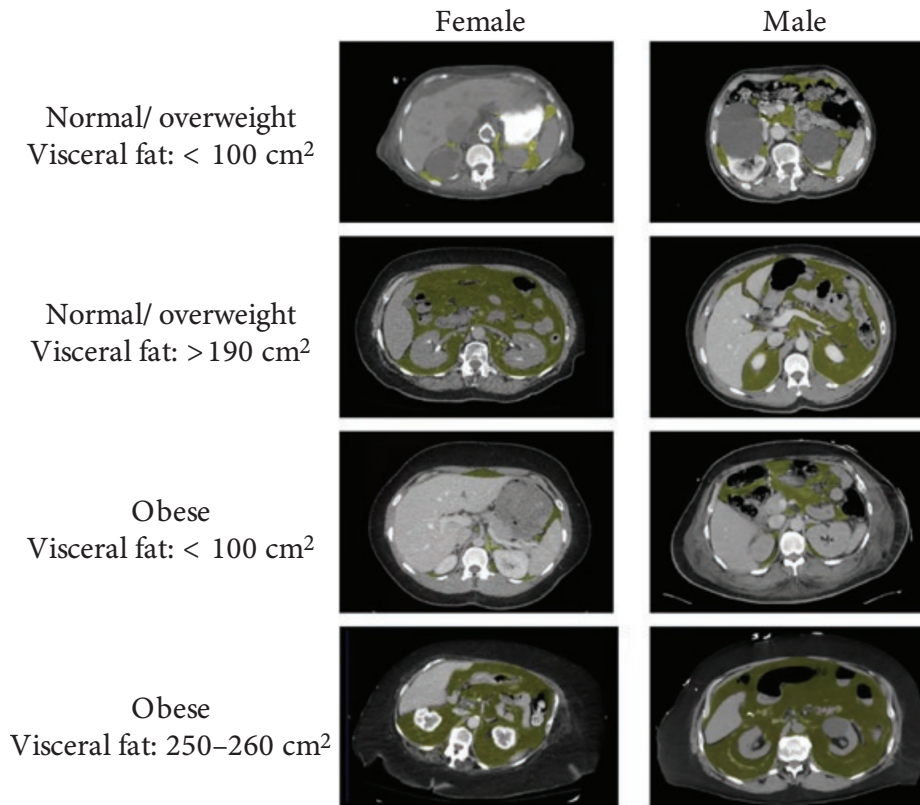


Figure 4-3: Exemplary visceral fat body compositions by BMI status and gender.

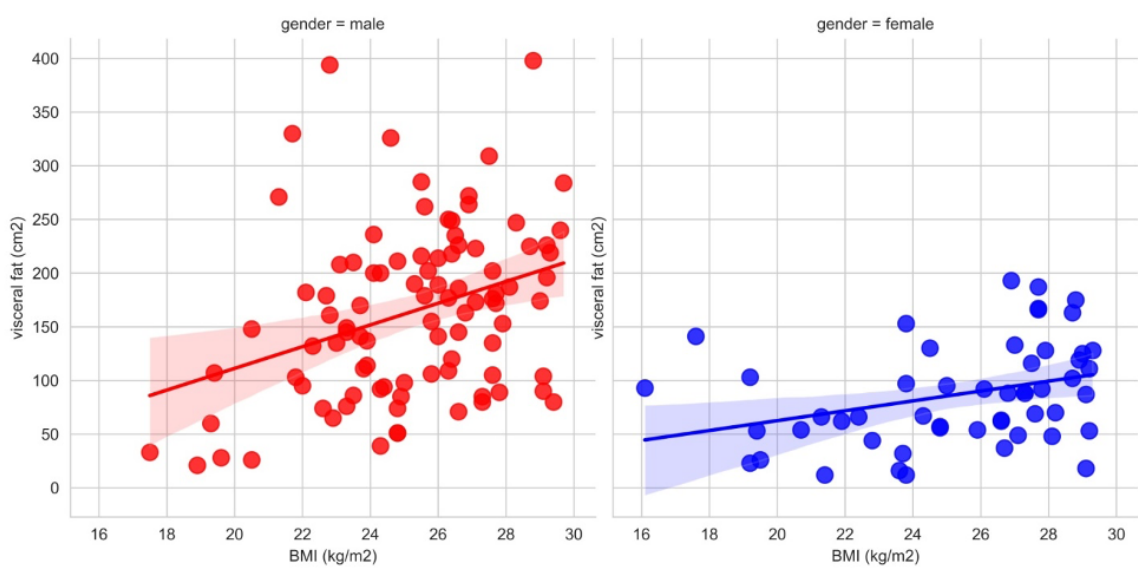


Figure 4-4: Scatterplot of VAT by BMI, stratified by sex.

overweight compared with obese), this same relationship is preserved among those who are normal or overweight ( $p < 0.005$ ). The differences are similar in magnitude

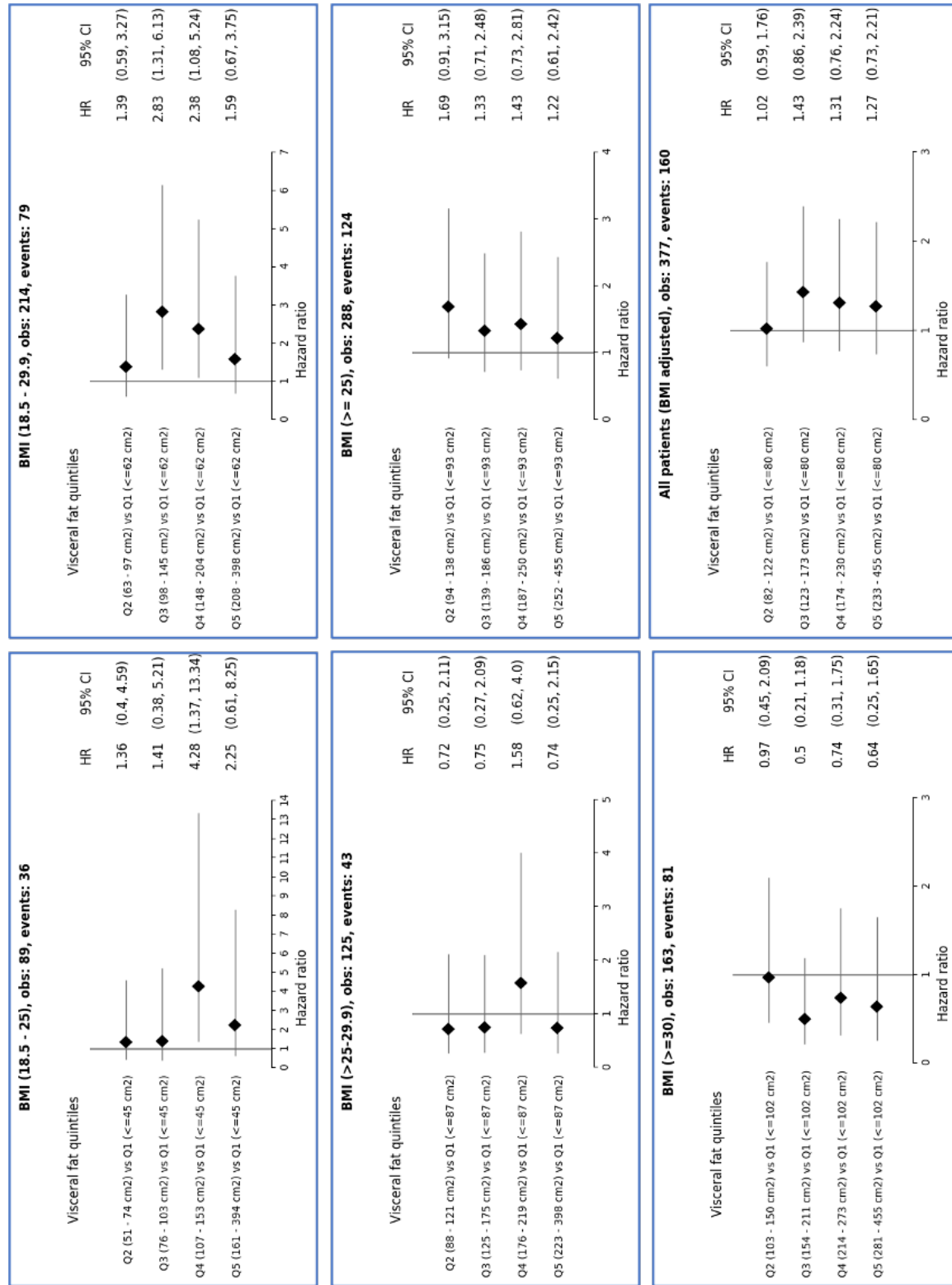


Figure 4-5: Cox proportional hazards model for the relationship between quintile of VAT and death or intubation within 30 days (adjusted for age, gender, and diabetes status).

Table 4.3: Multivariate Adjusted Hazard Ratio for Death or Intubation Within 28 Days From Hospitalization.

aHR (95% CI)	VAT Only	BMI + VAT	BMI Only
VAT $\geq$ 100 cm <sup>2</sup>	2.00 (1.32 – 3.02)	1.97 (1.24 – 3.09)	–
Age in years	1.00 (0.99 – 1.01)	1.00 (0.99 – 1.01)	1.00 (0.99 – 1.01)
Male	1.21 (0.85 – 1.72)	1.22 (0.85 – 1.76)	1.51 (1.07 – 2.13)
Diabetes	1.27 (0.93 – 1.74)	1.20 (0.87 – 1.66)	1.21 (0.88 – 1.67)
<b>Body Mass Index (BMI)</b>			
Normal	–	Reference	Reference
Overweight	–	0.76 (0.47 – 1.21)	0.95 (0.61 – 1.49)
Obese	–	1.14 (0.71 – 1.82)	1.57 (1.02 – 2.40)
<b>Race or Ethnicity</b>			
White	Reference	Reference	Reference
Hispanic	1.05 (0.67 – 1.63)	1.07 (0.69 – 1.68)	1.09 (0.70 – 1.70)
Black	1.88 (1.08 – 3.27)	1.95 (1.11 – 3.40)	1.67 (0.97 – 2.90)
Other	1.05 (0.71 – 1.54)	1.03 (0.70 – 1.52)	1.01 (0.68 – 1.49)

Abbreviations: aHR, adjusted hazard ratio; BMI, body mass index; VAT, visceral adipose tissue.

but do not reach statistical significance in the group with obesity ( $p = 0.08$ ). In Cox proportional hazards regression analyses, individuals with high VAT have an adjusted hazard ratio of 2.00 (95% CI, 1.32 – 3.02) of death or intubation at 28 days when adjusting for age, sex, race, and diabetes. Following additional adjustment for BMI, the adjusted hazard ratio for high VAT is unchanged at 1.97 (95% CI, 1.24 – 3.09) (Table 4.3). In a model with BMI but without VAT, the adjusted hazard ratio for obese vs normal BMI category is 1.57 (95% CI, 1.02 – 2.40); once VAT is included in the model, this declines to an adjusted hazard ratio of 1.14 (95% CI, 0.71 – 1.82). More analyses reveal no clear dose – response effect in the relationship between quintile of VAT and death or intubation within 30 days (Figure 4-5). Furthermore, a consideration of alternative dichotomous thresholds empirically reinforces the choice to use 100 cm<sup>2</sup>, as depicted in Figure 4-6.

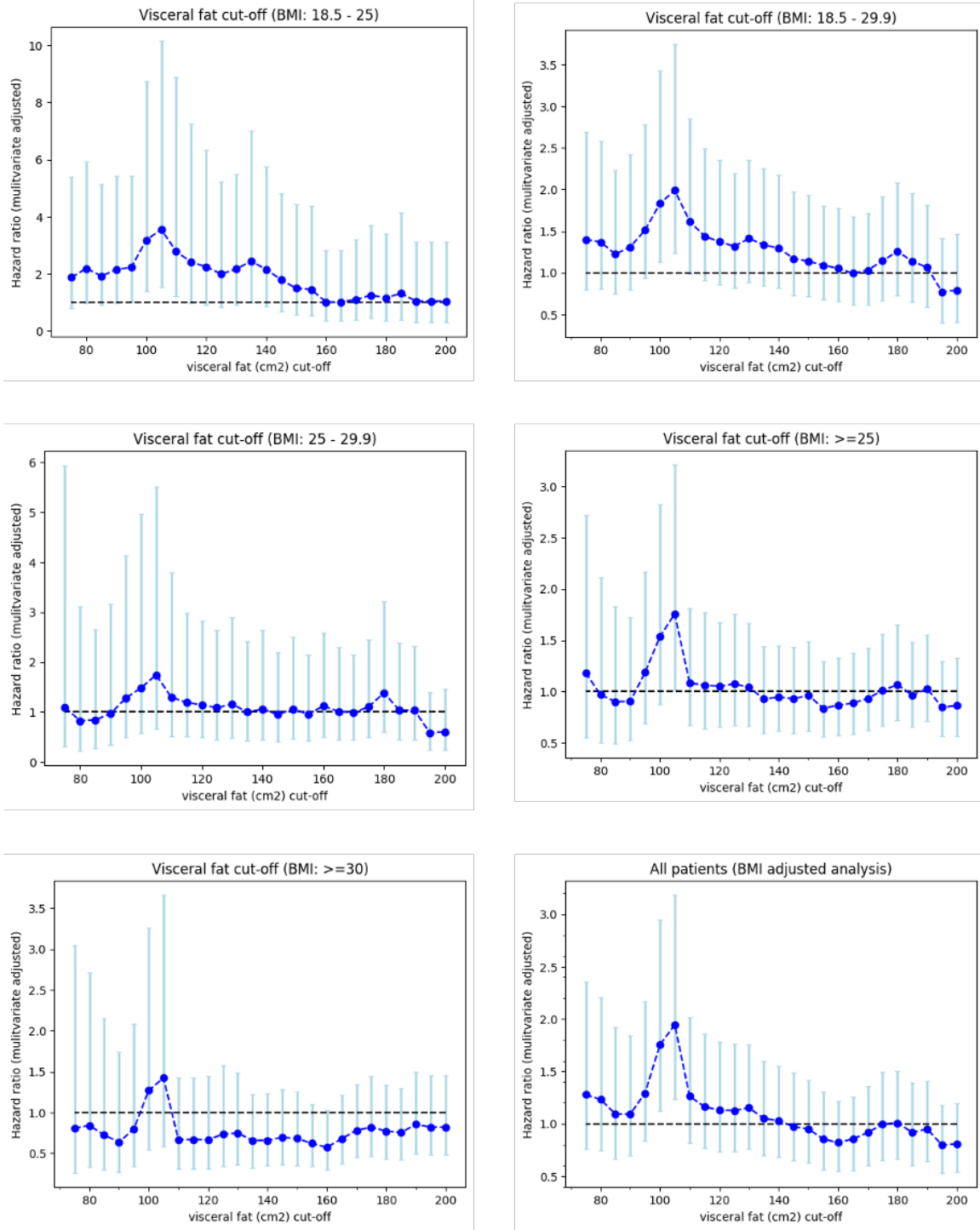


Figure 4-6: Adjusted hazard ratio for outcome of death or intubation within 30 days, overall and by BMI group (adjusted for age, gender, and diabetes status).

## 4.5 Discussion

We find that patients hospitalized with COVID-19 who have high VAT ( $\geq 100$  cm<sup>2</sup>) as ascertained by an AI algorithm from chest or abdominal CT scans have twice the risk of dying or being intubated within 28 days of admission than those with low VAT. This risk persists after adjusting models for BMI, suggesting that VAT may have a stronger and more precise relationship with severe COVID-19. This finding reflects the hypothesized biological significance of VAT as a more precise measure of differences in adipose tissue distribution and the health risk associated with obesity than BMI. These data support the possible use of VAT to risk-stratify hospitalized individuals with COVID-19 for severe clinical outcomes. Moreover, the AI algorithm used could be used by clinical teams to ascertain this measure quickly and automatically from imaging studies performed for other indications.

These findings are important for several reasons. First, there is an ample body of literature regarding risk prediction for severe COVID-19 outcomes that has not regularly included VAT as a consideration, though it may offer a more precise approximation of the metabolic risk associated with obesity when compared with BMI (Seigle et al., 2020; Longmore et al., 2021). This study provides evidence that measurement of VAT in hospitalized patients could be used to improve COVID-19 risk prediction. Second, as has been suggested previously, VAT may serve as a distinct driver of poor outcomes in COVID-19. The underlying mechanism to explain this relationship is not clear but may include the angiotensin-converting enzyme 2 (ACE-2) receptor as a possible link. This receptor facilitates cellular entry of SARS-CoV-2 and has been shown to have high expression in VAT (Zhang et al., 2018b; Al-Benna, 2020). Additionally, as detailed, VAT is metabolically active and secretes a variety of adipokines and pro-inflammatory cytokines that are hypothesized to play a role in severe COVID-19 (Neeland et al., 2019). As such, VAT may serve as a pro-inflammatory reservoir that could contribute to increased severity of COVID-19 among individuals with high VAT (Petersen et al., 2020).

One fundamental innovation of this study is the AI algorithm that is applied to



ascertain VAT in a fully automated fashion and with a precise and well-validated two-dimensional measure of this value. This is particularly unique, as many studies use a one-dimensional *VAT thickness* that is measured manually by a radiologist and lacks validation in the body composition literature. This AI algorithm has been applied and validated in several independent data sets and can facilitate opportunistic collection of both VAT and SAT from routine clinical imaging studies. Given that many hospitalized patients with severe COVID-19 have a CT scan of the chest or abdomen performed as part of their clinical workup, it would be possible to adapt this technology and automate collection of this measure using this publicly available algorithm. If performed in this way, the role of VAT in driving outcomes could be better understood and used to enhance prediction of risk for severe outcomes in real time.

This finding is largely consistent with three smaller studies from China and Europe that have suggested that adipose tissue distribution may be associated with outcomes in COVID-19 disease. The first study to explore this relationship consisted of a single-center cohort of 143 patients with confirmed COVID-19 who were hospitalized in Wuhan, China, between January and March 2020. These individuals all had abdominal CT scans from which radiologists manually measured VAT and several other measures of adipose tissue distribution (Yang et al., 2020). The rate of critical illness was almost double in people with higher VAT in this context, and in multivariate logistic regression models high VAT was associated with two times the odds of their severe disease end point. However, the sample represented a very small and select fraction of the total patients hospitalized with COVID-19 at this institution during this period, and their models did not adjust for BMI. These findings were reinforced by a second study of 144 patients who were consecutively admitted to the emergency department (ED) of a public hospital in Bufalini, Cesena, Italy, between February and April of 2020. All of these patients were found to have PCR-confirmed SARS-CoV-2 infection (Battisti et al., 2020). Upper abdominal VAT was assessed on sagittal images from chest CTs in all study participants. The primary outcome of interest in this study was admission to the intensive care unit (ICU). Those who

Table 4.4: Cox proportional hazards model for death or intubation, stratified for individuals with and without imaging.

aHR (95% CI)	Patients with Imaging	Patients without Imaging
Age in years	1.00 (0.99 – 1.01)	1.04 (1.03 – 1.06)
Male	1.51 (1.07 – 2.13)	0.94 (0.63 – 1.41)
Diabetes	1.21 (0.88 – 1.67)	1.56 (1.04 – 2.35)
<b>Body Mass Index (BMI)</b>		
Normal	Reference	Reference
Overweight	0.95 (0.61 – 1.49)	1.77 (0.95 – 3.31)
Obese	1.57 (1.02 – 2.40)	2.19 (1.19 – 4.04)
<b>Race or Ethnicity</b>		
White	Reference	Reference
Hispanic	1.09 (0.70 – 1.70)	1.05 (0.57 – 1.92)
Black	1.67 (0.97 – 2.90)	0.47 (0.14 – 1.54)
Other	1.01 (0.68 – 1.49)	1.65 (1.00 – 2.73)

were admitted to the ICU had a 30% higher VAT ( $p < 0.001$ ) and a 30% lower SAT ( $p = 0.011$ ), independent of age and sex. The latter findings were confirmed in similar studies in Rome, Italy, and a cohort of 30 patients in Berlin, Germany (Petersen et al., 2020; Watanabe et al., 2020).

Our study in this chapter has several important limitations. First, the study utilizes *opportunistic* imaging studies from people hospitalized with COVID-19 to estimate adipose distribution, and thus these parameters are only available in a subset of those hospitalized with COVID-19 during the study period. This design introduces important questions about how the inclusion of imaging may introduce additional selection bias in the sample of interest in this study. As described, those individuals who have a CT scan available during or within 2 years before their hospitalization for COVID-19 are older, more likely to be male, and have a higher prevalence of several important comorbidities, namely diabetes, though the distributions of BMI are similar in the two groups. As an additional analysis, we explore the relationship between BMI and diabetes among those with and without imaging. In these stratified Cox proportional hazards models (Table 4.4), we find that the relationships between

Table 4.5: Cox proportional hazards models for death or intubation within 30 days, overall and stratified by imaging during or prior to admission.

aHR (95% CI)	Overall ( <i>N</i> = 378)	Imaging During Admission ( <i>N</i> = 198)	Imaging Prior to Admission ( <i>N</i> = 180)
VAT $\geq$ 100 cm <sup>2</sup>	1.97 (1.24 – 3.09)	1.60 (1.00 – 2.56)	3.38 (1.44 – 7.91)
Age in years	1.00 (0.99 – 1.01)	1.00 (0.98 – 1.01)	1.01 (0.99 – 1.03)
Male	1.22 (0.85 – 1.76)	1.22 (0.81 – 1.89)	1.17 (0.63 – 2.19)
Diabetes	1.20 (0.87 – 1.66)	1.47 (1.00 – 2.16)	1.17 (0.68 – 2.01)
<b>Race or Ethnicity</b>			
White	Reference	Reference	Reference
Hispanic	1.07 (0.69 – 1.68)	1.09 (0.63 – 1.88)	0.90 (0.40 – 2.03)
Black	1.95 (1.11 – 3.40)	2.53 (1.28 – 4.97)	1.26 (0.47 – 3.38)
Other	1.03 (0.70 – 1.52)	1.11 (0.67 – 1.82)	0.87 (0.45 – 1.68)

diabetes and *obese* BMI and the outcome of interest are slightly attenuated in those with imaging compared with those for whom imaging is not available, but overall these relationships do not differ substantially. Given the lack of imaging in one group, differences in the relationship between VAT and the outcomes cannot be explored in this secondary analysis. This selection of higher-risk patients into the study likely limits power to detect differences in outcomes according to comorbidities known to associate with COVID risk, including those we previously identify. Second, the timing of imaging collection is a second source of heterogeneity that could also introduce selection effects.

In an additional analysis in Table 4.5 stratified by those with an imaging study and corresponding VAT measurement acquired during the index hospitalization and separately, those without a study and corresponding measurement that precede the hospitalization, we find that in both groups VAT is associated with severe disease, though the magnitude of the effect is greater among those who have the imaging study performed before admission. This difference in magnitude may indicate potential unmeasured confounding, for instance, related to the health condition that prompts the imaging study preceding the index hospitalization, but the relationship between

VAT and the outcomes is preserved in both groups and the small sample in each of the two groups after stratification makes it difficult to determine with certainty the importance of this potential limitation. Future research with larger cohorts should further interrogate these differences.

Beyond the potential limitations associated with selection bias, as detailed above, it is important to also state that these data are derived from a single center and thus may not be widely generalizable to other populations of individuals with COVID-19. Moreover, while we standardize data collection as much as possible through training of those performing chart review, the assignment of comorbid diagnoses other than diabetes and high BMI may have been subject to some variability across chart reviewers. Finally, the utility of this parameter is inherently dependent on the availability of a recent imaging study from which VAT may be measured and thus may be less widely used in people who do not routinely undergo imaging at presentation with COVID-19, for instance, younger people.

## 4.6 Summary

In this chapter, we successfully develop an automated AI system that ingests a CT imaging study from a patient, and predicts VAT quantity in numeric values as a surrogate endpoint for COVID-19 severity/mortality. The primary supporting reasonings are that (a) the segmentation modeling methods are comprehensively studied, (b) body composition data and annotations are readily available due to its extensive use cases, and that (c) there have been existing evidence linking metabolic descriptors to disease severity.

Following the observations, we present robust evidence that VAT can be used to stratify patients hospitalized with COVID-19 regarding their risk of severe disease or death and may be more precise and closely linked to poor outcomes than BMI. We have done this in the largest cohort and first US-based study of this relationship to date. We utilize an AI algorithm for ascertainment of adipose tissue distribution that automates collection of these data from routine clinical imaging studies. This

approach is promising because it is potentially scalable for use in real-world clinical settings and could improve prediction of poor outcomes among people who require hospitalization for COVID-19.



# Chapter 5

## Cross-Modal Representation

### Learning under Sparse Supervision

Cross-modal representation learning focuses on tackling data of two or more modalities, often also of very different dimensionalities, and attempts to unify them in an embedding space where associations among modalities can be reconstructed. There have been works (Chung et al., 2018, 2019) that aligned speech data with textual data via embedding spaces alignment, and in our case, we focus on the alignment of medical imaging data and their associated textual medical report. We hypothesize that there are distribution similarities between the modalities we are able to learn without explicitly providing the pairing information between the images and the reports.

#### 5.1 Overview

Joint embeddings between medical imaging modalities and associated radiology reports have the potential to offer significant benefits to the clinical community, ranging from cross-domain retrieval to conditional generation of reports to the broader goals of multimodal representation learning. In this chapter, we establish baseline joint em-

---

This chapter is adapted from the published article “*Unsupervised Multimodal Representation Learning across Medical Images and Reports*” (Hsu et al., 2018) to which I have contributed as the first author.

bedding results measured via both local and global retrieval methods on the Medical Information Mart for Intensive Care – Chest X-ray (MIMIC-CXR) dataset (Johnson et al., 2019) dataset consisting of both chest X-ray images and the associated radiology reports.

By removing the pairing information and learning the alignment of image and text embeddings unsupervisedly, we are able to verify whether, with limited information, representation learning could still be meaningful using adversarial domain adaptation (Tzeng et al., 2017) and Procrustes refinement (Conneau et al., 2017).

We evaluate the representation learning methods using retrieval-based metrics that count instances with the same International Classification of Diseases, Ninth Edition (ICD-9) codes as positive, denoting their similar indication of diseases. We also continuously add back pairing supervision to observe how the metrics are affected by the added information, and show that for document retrieval tasks with the learned representations, only a limited amount of supervision is needed to yield results comparable to those of fully-supervised methods. The objectives are to explore the effectiveness of cross-modal information on representation learning.

## 5.2 Introduction

Medical imaging is one of the most compelling domains for the immediate application of artificial intelligence tools. Recent years have seen not only tremendous academic advancements (Esteva et al., 2017; Gulshan et al., 2016; Rajpurkar et al., 2017) but additionally a breadth of applied tools (Marr, 2017; Walter, 2018; Lagasse, 2018; EnvoyAI, 2017).

There has been some emerging attention on joint processing of medical images and radiological free-text reports. Wang et al. (2018) used the public NIH Chest X-ray 14 dataset (Wang et al., 2017b) linked with the non-public associated reports to both improve disease classification performance and for automatic report generation. Gale et al. (2018) attempted to generate radiology reports while Shin et al. (2016) generated disease/location/severity annotations. Liu (2018) generated notes, includ-



ing radiology reports for the Medical Information Mart for Intensive Care (MIMIC) dataset using non-image modalities such as demographics, previous notes, labs, and medications. These works used annotations from either machines (Wang et al., 2017b) or humans. However, with a huge influx of imaging data beyond human capacity, parallel records from both imaging and text are not always readily available. We thus would like to bring up the question of whether we can take advantage of unannotated but massive imaging datasets and learn from the underlying distribution of these images.

One natural area that remains unexplored is representation learning across images and reports. The idea of representation learning in a joint embedding space can be realized in multiple ways. Some (Pan et al., 2011; Chen et al., 2016) explored statistical and metrical relevance across domains, and some (Ganin et al., 2016) realized it as an adversarially determined domain-agnostic latent space. Shen et al. (2017) and Mor et al. (2018) both used a the latent space for style transfer, in language sentiment and music style, respectively. Reed et al. (2016a) learned joint spaces of images and their captions, which Reed et al. (2016b) later used for caption-driven image generation. Conneau et al. (2017) and Grave et al. (2018) also used similar ideas to perform both supervised and unsupervised word-to-word translation tasks. (Chung et al., 2018) further aligned cross-modal embeddings through semantics in speech and text for spoken word classification and translation tasks.

A recent dataset, MIMIC-Chest X-ray (MIMIC-CXR) (Johnson et al., 2019), carries paired records of X-ray images and radiology reports, and the imaging modality has been explored by Rubin et al. (2018) and Quigley et al. (2022). In this study, we explore both the text and image modalities with joint embedding spaces under a spectrum of supervised and unsupervised methods. In particular, we make the following contributions:

1. We establish baseline results and evaluation methods for jointly embedding radiological images and reports via retrieval and distance metrics.
2. We profile the impact of supervision level on the quality of representation learn-

ing in joint embedding spaces.

3. We characterize the influence of using different sections from the report on representation learning.

## 5.3 Methodology

### 5.3.1 Data

All experiments in this study used the MIMIC-CXR dataset<sup>1</sup> (Johnson et al., 2019). MIMIC-CXR is the largest radiology dataset to date and consists of 473,057 chest X-ray images and 206,563 reports from 63,478 patients. Among these images, 240,780 are of anteroposterior (AP), 101,379 are of posteroanterior (PA), and 116,023 are of lateral (LL) views, and we focus on in AP images this work. Further, we eliminate all duplicated radiograph images with adjusted brightness or contrast (commonly produced for clinical needs), leaving a total of 95,242/87,353 images/reports, which we subdivide into a train set of 75,147/69,171 and a test set of 19,825/18,182 images/reports, with no overlap of patients between the two. Radiological reports are parsed into sections and we use either the *impression* or the *findings* sections.

For evaluation, we aggregate a list of unique International Classification of Diseases (ICD-9) codes from all patient admissions and ask a clinician to pick out a subset of codes that are related to thoracic diseases. Records with ICD-9 codes in the subset are then extracted, including 3,549 images from 380 patients. This population serves as a disease-related evaluation for retrieval algorithms. Note that this disease information is never provided during training in any setting.

### 5.3.2 Methods

Our overall experimental flow follows Figure 5-1. Notes are featurized via (1) term frequency-inverse document frequency (TF-IDF) over bi-grams, (2) pre-trained GloVe

---

<sup>1</sup>This work used an alpha version of MIMIC-CXR instead of the publicly released version. The main differences are the train/validation/test splitting, pre-processing, and artifact removals so that the publicly released version is more sanitized.

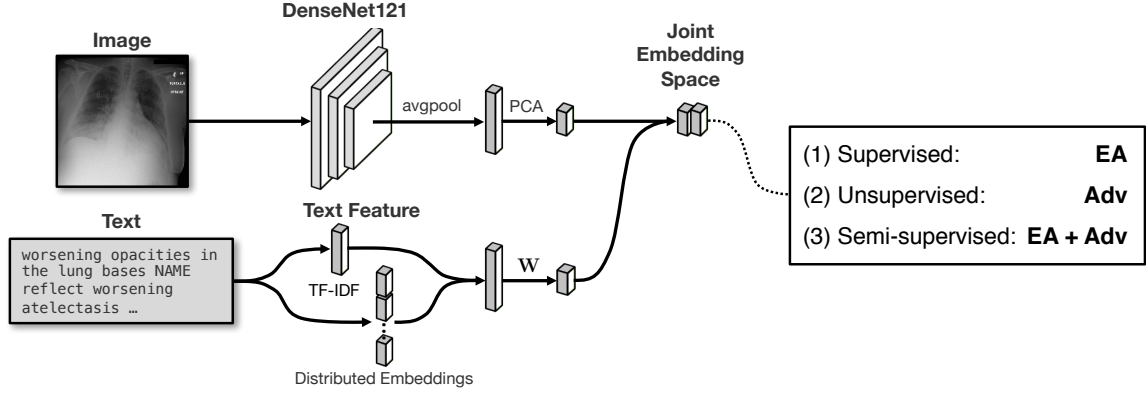


Figure 5-1: The overall experimental pipeline. EA: embedding alignment; Adv: adversarial training.

*word* embeddings (Pennington et al., 2014) averaged across the selected section of the report, (3) *sentence* embeddings, or (4) *paragraph* embeddings. In (3) and (4), we first perform sentence/paragraph splitting, and then fine-tune a deep averaging network (DAN) encoder (Bird and Loper, 2004; Cer et al., 2018; Iyyer et al., 2015) with the corpus. Embeddings are finally averaged across sentences/paragraphs. The DAN encoder is pretrained on a variety of data sources and tasks and fine-tuned on the context of report sections.

Images are resized to  $256 \times 256$ , then featurized to the last bottleneck layer of a pretrained DenseNet-121 model (Rajpurkar et al., 2017). Principle component analysis (PCA) is applied onto the 1024-dimension raw image features to obtain 64-dimension features (96.9% variance explained). Text features are projected into the 64-dimension image feature space. We use several methods regarding different objectives.

### Embedding Alignment (“EA”)

Here, we find a linear transformation between two sets of matched points  $\mathbf{X} \in \mathbb{R}^{d_X \times n}$  and  $\mathbf{Y} \in \mathbb{R}^{d_Y \times n}$  by minimizing  $\mathcal{L}_{EA}(\mathbf{X}, \mathbf{Y}) = \|\mathbf{W}^\top \mathbf{X} - \mathbf{Y}\|_F^2$ .

## Adversarial Domain Adaption (“*Adv*”)

Adversarial training pits a *discriminator*,  $D$ , implemented as a 2-layer (hidden size 256) neural network using scaled exponential linear units (SELUs) (Klambauer et al., 2017), against a projection matrix  $\mathbf{W}$ , as the *generator*.  $D$  is trained to classify points in the joint space according to source modality, and  $\mathbf{W}$  is trained adversarially to fool  $D$ . Alternatively,  $D$  minimizes

$$\mathcal{L}_{\text{Adv}}^D(\mathbf{X}, \mathbf{Y}) = \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim p(\mathbf{X}, \mathbf{Y})} [-\log D(\mathbf{W}^\top \mathbf{x}) - \log(1 - D(\mathbf{y}))] \quad (5.1)$$

when  $\mathbf{W}$  minimizes

$$\mathcal{L}_{\text{Adv}}^{\mathbf{W}}(\mathbf{X}, \mathbf{Y}) = \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim p(\mathbf{X}, \mathbf{Y})} [-\log(1 - D(\mathbf{W}^\top \mathbf{x}))]. \quad (5.2)$$

## Procrustes Refinement (“*Adv + Proc*”)

On top of adversarial training, we also use an unsupervised Procrustes induced refinement as in Conneau et al. (2017).

## Semi-Supervised

We also assess how much supervision is necessary to ensure strong performance on these modalities by randomly subsampling our data into supervised and unsupervised samples. We then combine the *embedding alignment* objective and *adversarial training* objective functions as  $\mathcal{L} = \mathcal{L}_{\text{EA}}(\mathbf{X}, \mathbf{Y}) + \lambda \mathcal{L}_{\text{Adv}}(\mathbf{X}, \mathbf{Y})$  and train simultaneously as we vary the fraction trained. Preliminary experiments suggests  $\lambda = 0.1$ .

## Orthogonal Regularization

Smith et al. (2017), Conneau et al. (2017), and Xing et al. (2015) all showed that imposing orthonormality on linear projections leads to better performance and stability in training. However, Brock et al. (2018) suggested orthogonality (*i.e.*, not constraining the norms) can perform better as a regularization. Thus on top of the objectives, we add  $\mathcal{R}_{\text{ortho}} = \beta \|\mathbf{W}^\top \mathbf{W} \odot (\mathbf{e}\mathbf{e}^\top - \mathbf{I})\|_F^2$ , where  $\odot$  denotes element-wise

product and  $\mathbf{e}$  denotes a column vector of all ones. The addition of the regularization term aims at suppressing correlation for off-diagonal terms. Scanning through a range shows  $\beta = 0.01$  yields good performance.

### 5.3.3 Evaluation

We evaluate via cross domain retrieval in the test set  $Q$ : querying in the joint embedding space for closest neighboring images using a report,  $T \rightarrow I$ , or vice-versa,  $I \rightarrow T$ . For direct pairings, we compute the *cosine similarity*, and *mean reciprocal rank*  $MRR = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{\text{rank}_q}$  where  $\text{rank}_q$  is the rank of the first true pair for  $q$  (e.g., the first paired image or text corresponding to the query  $q$ ) in the retrieval list. For thoracic disease induced pairings, we first define the relevance  $\text{rel}_{pq} \in [0, 1]$  between two entries  $p$  and  $q$  as the intersection-over-union of their respective set of ICD-9 codes. Then we calculate the normalized discounted cumulative gain (Järvelin and Kekäläinen, 2002)  $nDCG@k = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{IDCG_q} \sum_{p=1}^k \frac{2^{\text{rel}_{pq}} - 1}{\log_2(p+1)}$ , where  $IDCG_q$  denotes the ideal DCG, or discounted cumulative gain value, for  $q$  using a perfect retrieval algorithm. All experiments are repeated with random initial seeds for at least 5 times. Means and 95% confidence intervals are reported in the following section.

## 5.4 Results

### Retrieval with/without Supervision

Table 5.1 compares four types of text features and supervised/unsupervised methods. We find that unsupervised methods can achieve comparable results on disease-related retrieval tasks on a large scale ( $nDCG@100$ ) without the need for labeling the chest X-ray images. Experiments show uni-, bi-, and tri-grams yield very similar results and we only include bi-gram in the table. Additionally, we find that the high-level *sentence* and *paragraph* embeddings approach underperform the bi-gram text representation. Although having generalizability (Cer et al., 2018), sentence and paragraph embeddings learned from the supervised multi-task pre-trained model may not be able

Text Feature	Method	Similarity	MRR( $\times 10^{-3}$ )		nDCG@1		nDCG@10		nDCG@100	
			T $\rightarrow$ I	I $\rightarrow$ T	T $\rightarrow$ I	I $\rightarrow$ T	T $\rightarrow$ I	I $\rightarrow$ T	T $\rightarrow$ I	I $\rightarrow$ T
<i>chance</i>										
bi-gram	EA	<b>0.613</b> <sub>.000</sub>	<b>7.33</b> <sub>.04</sub>	<b>11.65</b> <sub>.07</sub>	<b>0.147</b> <sub>.001</sub>	<b>0.162</b> <sub>.001</sub>	<b>0.148</b> <sub>.000</sub>	<b>0.159</b> <sub>.000</sub>	<b>0.225</b> <sub>.000</sub>	<b>0.231</b> <sub>.000</sub>
word	EA	0.542 <sub>.000</sub>	2.00 <sub>.01</sub>	4.52 <sub>.02</sub>	0.096 <sub>.002</sub>	0.128 <sub>.001</sub>	0.116 <sub>.000</sub>	0.130 <sub>.000</sub>	0.202 <sub>.000</sub>	0.205 <sub>.000</sub>
sentence	EA	0.465 <sub>.000</sub>	1.08 <sub>.00</sub>	2.74 <sub>.02</sub>	0.073 <sub>.001</sub>	0.101 <sub>.000</sub>	0.100 <sub>.000</sub>	0.111 <sub>.000</sub>	0.189 <sub>.000</sub>	0.177 <sub>.000</sub>
paragraph	EA	0.505 <sub>.000</sub>	1.57 <sub>.01</sub>	2.53 <sub>.01</sub>	0.082 <sub>.001</sub>	0.134 <sub>.000</sub>	0.107 <sub>.000</sub>	0.124 <sub>.000</sub>	0.195 <sub>.000</sub>	0.196 <sub>.000</sub>
bi-gram	Adv	0.218 <sub>.073</sub>	0.77 <sub>.23</sub>	0.85 <sub>.33</sub>	0.095 <sub>.006</sub>	0.090 <sub>.003</sub>	0.101 <sub>.004</sub>	0.098 <sub>.003</sub>	0.171 <sub>.005</sub>	0.166 <sub>.004</sub>
bi-gram	Adv + Proc	0.221 <sub>.074</sub>	0.77 <sub>.24</sub>	0.87 <sub>.32</sub>	0.094 <sub>.006</sub>	0.091 <sub>.004</sub>	0.102 <sub>.004</sub>	0.099 <sub>.002</sub>	0.171 <sub>.005</sub>	0.166 <sub>.004</sub>
word	Adv	0.268 <sub>.016</sub>	0.65 <sub>.12</sub>	0.54 <sub>.12</sub>	0.096 <sub>.006</sub>	0.091 <sub>.003</sub>	0.105 <sub>.004</sub>	0.099 <sub>.003</sub>	0.176 <sub>.003</sub>	0.165 <sub>.004</sub>
word	Adv + Proc	<b>0.269</b> <sub>.013</sub>	0.64 <sub>.11</sub>	0.57 <sub>.07</sub>	<b>0.098</b> <sub>.006</sub>	0.092 <sub>.002</sub>	<b>0.107</b> <sub>.005</sub>	0.099 <sub>.003</sub>	<b>0.179</b> <sub>.003</sub>	0.165 <sub>.004</sub>
sentence	Adv	0.265 <sub>.010</sub>	0.64 <sub>.08</sub>	<b>1.07</b> <sub>.24</sub>	0.095 <sub>.007</sub>	0.094 <sub>.002</sub>	0.103 <sub>.006</sub>	0.100 <sub>.001</sub>	0.176 <sub>.006</sub>	0.167 <sub>.001</sub>
sentence	Adv + Proc	0.266 <sub>.012</sub>	0.68 <sub>.10</sub>	1.07 <sub>.21</sub>	0.096 <sub>.005</sub>	0.094 <sub>.004</sub>	0.105 <sub>.006</sub>	0.100 <sub>.002</sub>	0.178 <sub>.005</sub>	0.166 <sub>.002</sub>
paragraph	Adv	0.045 <sub>.136</sub>	0.69 <sub>.03</sub>	0.70 <sub>.04</sub>	0.062 <sub>.025</sub>	<b>0.123</b> <sub>.029</sub>	0.082 <sub>.015</sub>	<b>0.118</b> <sub>.017</sub>	0.163 <sub>.013</sub>	<b>0.169</b> <sub>.003</sub>
paragraph	Adv + Proc	0.225 <sub>.061</sub>	<b>1.15</b> <sub>.60</sub>	0.77 <sub>.21</sub>	0.093 <sub>.057</sub>	0.092 <sub>.011</sub>	0.090 <sub>.034</sub>	0.103 <sub>.008</sub>	0.163 <sub>.023</sub>	0.166 <sub>.005</sub>

Table 5.1: Comparison among supervised (upper) and unsupervised (lower) methods. Subscripts show the half width of 95% confidence intervals. **Bold** denotes the best performance in each group. *Chance* is the expected value if we randomly yield retrievals. Higher is better for all metrics.

to represent the domain-specific radiological reports well due to the lack of medical domain tasks in the pre-training process. Unsupervised procrustes refinement is occasionally, but not universally helpful. Note that MRR is comparatively small since reports are in general highly similar for radiographs with the same disease types.

### **The Impact of Supervision Fraction**

We define the *supervision fraction* as the fraction of pairing information provided in the training set. Note the ICD-9 codes are not provided for training even in the fully supervised setting. Figure 5-2 shows our evaluation metrics for models trained using bi-gram text features and the semi-supervised learning objective for various supervision fractions. A minimal supervision as low as 0.1% provided can drastically improve the alignment quality, especially in terms of cosine similarity and nDCG. More annotations further improve the performance measures, but one would almost require exponentially many data points in exchange for a linear increase. That implies the possibility of concatenating a well-annotated dataset and a large but unannotated dataset for a substantial performance boost.

### **Using Different Sections of the Report**

We investigate the effectiveness of using different sections for the embedding alignment task. All models in Figure 5-3 run with a supervision fraction of 1%. The models trained on the *findings* section outperform the models trained on the *impression* section using cosine similarity and MRR. This makes sense from a clinical perspective since the radiologists usually only describe image patterns in the *findings* section and thus they would be aligned well. On the other hand, they make radiological-clinical integrated interpretations in the *impression* section, which means that the both the image-uncorrelated clinical history and findings were mentioned in the *impression* section. Since nDCG is calculated using ICD-9 codes, which carry disease-related information, it naturally aligns with the purpose of writing an *impression* section. This may explain why the models trained on *impression* section worked better for nDCG.

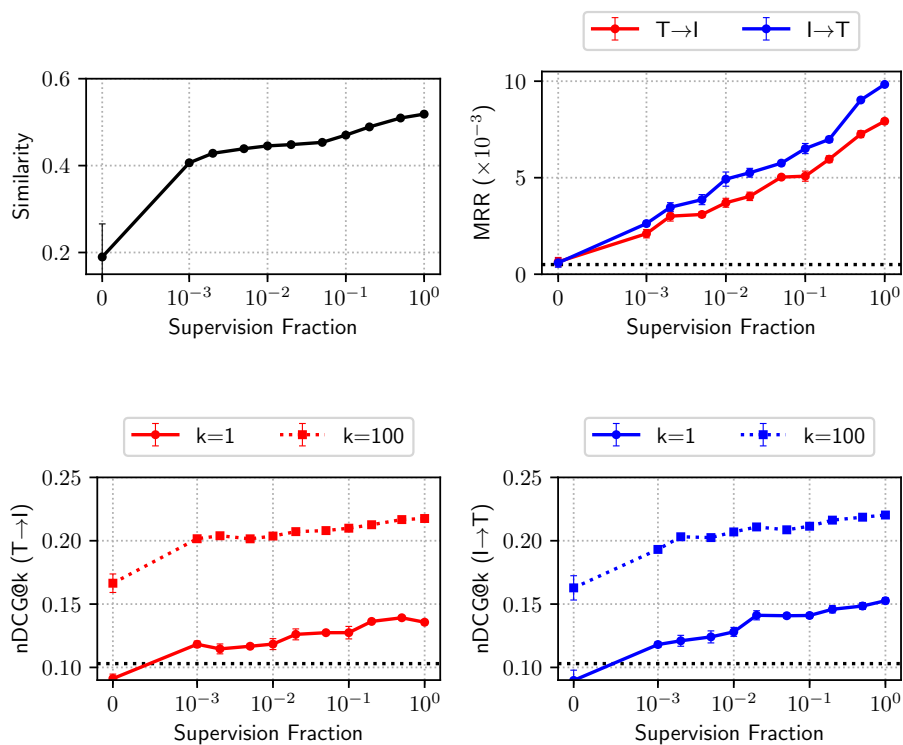


Figure 5-2: Performance measures of retrieval tasks at  $k$  retrieved items as a function of the supervision fraction. Higher is better. Note the  $x$ -axis is in log scale. Unsupervised is on the left, increasingly supervised to the right. Dashed lines indicate the performance by chance. Vertical bars indicate the 95% confidence interval, and some are too narrow to be visible.

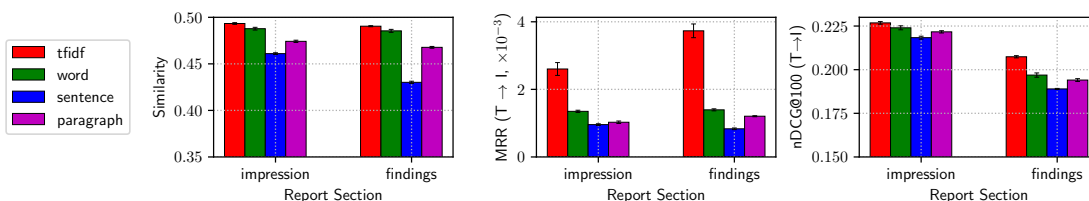


Figure 5-3: Different metrics for retrieval on either the *impression* or *findings* section using four types of features. 95% confidence intervals are indicated on the bars.

## 5.5 Summary

MIMIC-CXR is the largest publicly available imaging dataset consisting of both medical images and paired radiological reports to date, promising myriad applications that can make use of both modalities together. We establish baseline results using super-



vised and unsupervised joint embedding methods along with local (direct pairs) and global (ICD-9 code groupings) retrieval evaluation metrics. Results show a possibility of incorporating more unsupervised data into training for minimal-effort performance increase. A further study of joint embeddings between these modalities may enable significant applications, such as text/image generation or the incorporation of other electronic medical records (EMR) modalities.

In this chapter, we have demonstrated how imaging data are used in a retrieval context against textual data. Under the extreme conditions where supervision is totally removed, it is surprising that the retrieval performance is significantly better than the chance baseline, thus providing very promising hint that extra data modality can benefit medical imaging in a constrained setup.

We will continue to explore, in the next chapter, the relationship between medical reports and chest X-ray with the task of *medical report generation* using the MIMIC-CXR dataset presented in this chapter.



## Chapter 6

# Knowledge-Infused Learning for Medical Report Generation from Radiograph

The infusion of clinical knowledge into machine learning models has been under-emphasized in the early era of deep learning-based medical imaging. Early works (Wu et al., 2013; Cheng et al., 2018; Yang et al., 2015; De Vos et al., 2016; Brosch et al., 2013; Suk and Shen, 2013; Plis et al., 2014; Ciresan et al., 2012; Stollenga et al., 2015) emphasized bringing the then-popular deep learning into the field of medical imaging and yet did not leverage existing knowledge about the inherent medical structures. As we gradually push the frontier of medical imaging, researchers have since recognized the need for larger datasets than before, and yet gathering more data would either be time or financially consuming. Hence we hereby explored infusing clinical knowledge into the construction of medical machine learning models, in order to increase the effective amount of data for learning.

---

This chapter is adapted from the published article “*Clinically Accurate Chest X-Ray Report Generation*” (Liu et al., 2019) to which I have contributed as the first author.

## 6.1 Overview

The automatic generation of radiology reports given medical radiographs may have significant potential to operationally improve clinical patient care. A number of prior works have focused on this problem, employing advanced methods from computer vision and natural language generation to produce readable reports. However, these works often fail to account for the particular nuances of the radiology domain, and, in particular, the critical importance of clinical accuracy in the resulting generated reports. In this chapter, we present a domain-aware automatic chest X-ray radiology report generation system which first predicts what topics will be discussed in the report, then conditionally generates sentences corresponding to these topics. The resulting system is fine-tuned using reinforcement learning, considering both readability and clinical accuracy, as assessed by the proposed *Clinically Coherent Reward* (CCR). We verify this system on two datasets, Open-I (Demner-Fushman et al., 2015) and MIMIC-CXR (Johnson et al., 2019), and demonstrate that our model offers marked improvements on both language generation metrics and CheXpert (Irvin et al., 2019) assessed accuracy over a variety of competitive baselines.

## 6.2 Background

A critical task in radiology practice is the generation of a free-text description, or *report*, based on a clinical radiograph (*e.g.*, a chest X-ray). Providing automated support for this task has the potential to ease clinical workflows and improve both the quality and standardization of care. However, this process poses significant technical challenges. Many traditional image captioning approaches are designed to produce far shorter and less complex pieces of text than radiology reports. Further, these approaches do not capitalize on the highly templated nature of radiology reports. Additionally, generic natural language generation (NLG) methods prioritize descriptive accuracy only as a byproduct of readability, whereas providing an accurate clinical description of the radiograph is the *first* priority of the report. Prior works in this

domain have partially addressed these issues, but significant gaps remain towards producing high-quality reports with maximal clinical efficacy.

In this chapter, we take steps to address these gaps through our novel automatic chest X-ray radiology report generation system. Our model hierarchically generates a sequence of unconstrained topics, using each topic to generate a sentence for the final generated report. In this way, we capitalize on the often-templated nature of radiology reports while simultaneously offering the system sufficient freedom to generate diverse, free-form reports. The system is finally tuned via reinforcement learning to optimize readability (via the *Consensus-based Image Description Evaluation*, or CIDEr (Vedantam et al., 2015), score) as well as clinical accuracy (via the concordance of CheXpert (Irvin et al., 2019) disease state labels between the ground truth and generated reports). We test this system on the MIMIC-CXR (Johnson et al., 2019) dataset, which is the largest paired image-report dataset presently available, and demonstrate that our model offers improvements on both NLG evaluation metrics (BLEU (Papineni et al., 2002), CIDEr (Vedantam et al., 2015), and ROGUE (Lin, 2004)) and clinical efficacy metrics (CheXpert concordance) over several compelling baseline models, including a re-implementation of TieNet (Wang et al., 2018), simpler neural baselines, and a retrieval-based baseline.

## Clinical Relevance

This chapter focuses on generating a clinically useful radiology report from a chest X-ray image. This task has been explored multiple times, but directly transplanting natural language generation techniques onto this task only guarantees the reports to *look real* rather than to *predict right*. A more immediate focus for the report generation task is thus to produce accurate disease profiles to power downstream tasks such as diagnosis and care providing. Our goal is then minding the language fluency while also increasing the clinical efficacy of the generated reports.

## Technical Significance

We employ a hierarchical convolutional-recurrent neural network as the backbone for our proposed method. Reinforcement learning (RL) on a combined objective of both language fluency metrics and the proposed Clinically Coherent Reward (CCR) ensures we obtain a quality model on more correctly describing disease states. Our method aims to numerically align the disease labels of our generated report, as produced by a natural language labeler, with the labels from the ground truth reports. The reward function, though non-differentiable, can be optimized through policy gradient learning as promised by RL.

## 6.3 Related Works

### 6.3.1 Radiology



**Findings:**

There is no focal consolidation, effusion or pneumothorax. The cardiomediastinal silhouette is normal. There has been interval resolution of pulmonary vascular congestion since DATE.

**Impression:**

No pneumonia or pulmonary vascular congestion. Telephone notification to dr. NAME at TIME on DATE per request

Figure 6-1: A chest X-ray and its associated report written by a radiologist.

### Radiology Practice

Diagnostic radiology is the medical field of creating and evaluating radiological images (radiographs) of patients for diagnostics. Radiologists are trained to simultaneously identify various radiological findings (e.g., diseases), according to the details of the radiograph and the patient's clinical history, then summarize these findings and their overall impression in reports for clinical communication (Kahn Jr et al., 2009; Schwartz et al., 2011). A report typically consists of sections such as *history*,

*examination reason, findings, and impressions.* As shown in Figure 6-1, the *findings* section contains a sequence of positive, negative, or uncertain mentions of either disease observations or instruments including their detailed location and severity. The *impression* section, by contrast, summarizes diagnoses considering all report sections above and previous studies on the patient. In a good report (as characterized by an experienced radiologist), the impression section puts the findings into context and attempts to address questions raised in the image requisition, and in contrast, bad reports tend to regurgitate findings. Correctly identifying all abnormalities is a challenging task due to high variation and atypical cases (Rubin, 2015). Moreover, there is information overload inherent to some imaging modalities, such as X-ray scans.

This presents a strong intervention surface for machine learning techniques to help radiologists correctly identify the critical findings from a radiograph. The canonical way to communicate such findings in current practice would be through the free-text report, which could either be used as a *draft* report for the radiologists to extend or be presented to the physician requesting a radiological study directly (Schwartz et al., 2011).

## AI on Radiology Data

In recent years, several chest radiograph datasets, totalling almost a million X-ray images, have been made publicly available. A summary of these datasets is available in Table 6.1. Learning effective computational models through leveraging the information in medical images and free-text reports is an emerging field. Such a combination of image and textual data help further improve the model performance in both image annotation and automatic report generation (Litjens et al., 2017).

Schlegl et al. (2015) first proposed a weakly supervised learning approach to utilize semantic descriptions in reports as labels for better classifying the tissue patterns in optical coherence tomography (OCT) imaging. In the field of radiology, Shin et al. (2016) proposed a convolutional and recurrent network framework that jointly trained from image and text to annotate disease, anatomy, and severity in the chest X-ray images. Similarly, Moradi et al. (2018) jointly processed image and text signals to

Dataset	Source Institution	Disease Labeling	# Images	# Reports	# Patients
Open-I	Indiana Network for Patient Care	Expert	8,121	3,996	3,996
Chest-Xray8	National Institutes of Health	Automatic (DNorm + MetaMap)	108,948	0	32,717
CheXpert	Stanford Hospital	Automatic (CheXpert labeler)	224,316	0	65,240
PadChest	Hospital Universitario de San Juan	Expert + Automatic (Neural network)	160,868	206,222	67,625
MIMIC-CXR	Beth Israel Deacones Medical Center	Automatic (CheXpert labeler)	473,057	206,563	63,478

Table 6.1: A description of each available chest X-ray datasets. Open-I (Demner-Fushman et al., 2015), Chest-XRay8 (Wang et al., 2017b) which utilized DNorm (Leaman et al., 2015) and MetaMap (Aronson and Lang, 2010), CheXpert (Irvin et al., 2019), PadChest (Bustos et al., 2019), and MIMIC-CXR (Johnson et al., 2019).



produce regions of interest over chest X-ray images. Rubin et al. (2018) trained a convolutional network to predict common thoracic diseases given chest X-ray images. Shin et al. (2015), Wang et al. (2016), and Wang et al. (2017b) mined radiological reports to create disease and symptom concepts as labels. They first used Latent Dirichlet Allocation (LDA) to identify the topics for clustering, then applied the disease detection tools such as DNorm (Disease Name Normalization) (Leaman et al., 2013), MetaMap (Aronson, 2001), and several other Natural Language Processing (NLP) tools for downstream chest X-ray classification using a convolutional neural network. They also released the label set along with the image data.

Later on, Wang et al. (2018) used the same chest X-ray dataset to further improve the performance of disease classification and report generation from an image. For report generation, Jing et al. (2017) built a multi-task learning framework, which includes a co-attention mechanism module, and a hierarchical long short term memory (LSTM) module, for radiological image annotation and report paragraph generation. Li et al. (2018) proposed a reinforcement learning-based *Hybrid Retrieval-Generation Reinforced Agent (HRGR-Agent)* to learn a report generator that can decide whether to retrieve a template or generate a new sentence. Alternatively, Gale et al. (2018) generated interpretable hip fracture X-ray reports by identifying image features and filling text templates.

Finally, Hsu et al. (2018) trained the radiological image and report joint representation through unsupervised alignment of cross-modal embedding spaces for information retrieval.

### 6.3.2 Language Generation

Language generation (LG) is a staple of NLP research. LG comes up in the context of neural machine translation, summarization, question answering, image captioning, and more. In all these tasks, the challenges of generating discrete sequences that are realistic, meaningful, and linguistically correct must be met, and the field has devised a number of methods to surmount them. For many years, this was done through  $n$ -gram-based (Huang et al., 1993) or retrieval-based (Gupta and Lehal, 2010)

approaches.

Within the last few years, many have explored the very impressive results of deep learning for text generation. Graves (2013) outlined best practices for *Recurrent Neural Network*, or , RNN-based sequence generation. The following year, Sutskever et al. (2014) introduced the *sequence-to-sequence* paradigm for machine translation and beyond. However, Wiseman et al. (2017) demonstrated that while RNN-generated texts are often fluent, they have typically failed to reach human-level quality.

Reinforcement learning recently also come into play due to its capability to optimize for indirect target rewards, even if the targets themselves are often non-differentiable. Li et al. (2016a) used a crafted combination of human heuristics as the reward while Bahdanau et al. (2016) incorporated language fluency metrics. They were among the first to apply such techniques to neural language generation, but to date, training with log-likelihood maximization (Xie, 2017) has been the main working horse. Alternatively, Rajeswar et al. (2017) and Fedus et al. (2018) have tried using *Generative Adversarial Networks* (GANs) for text generation. However, Caccia et al. (2018) observed problems with training GANs and show that to date, they are unable to beat canonical sequence decoder methods.

## Image Captioning

We will also highlight some specific areas of exploration in image captioning, a specific kind of language generation which is conditioned on an image input. The canonical example of this task is realized in the *Microsoft Common Objects in Context (COCO)* (Lin et al., 2014b) dataset, which presents a series of images, each annotated with five human-written captions describing the image. The task, then, is to use the image as input to generate a readable, accurate, and linguistically correct caption.

This task has received significant attention with the success of *Show and Tell* (Vinyals et al., 2015) and its followup *Show, Attend, and Tell* (Xu et al., 2015). Due to the nature of the COCO competition, other works quickly emerged showing strong results: Yao et al. (2017) used boosting methods, Lu et al. (2017) employed adaptive

attention, and Rennie et al. (2017) introduced reinforcement learning as a method for fine-tuning generated text. Devlin et al. (2015) performed surprisingly well using a  $K$ -nearest neighbor method. They observed that since most of the true captions were simple, one-sentence scene descriptions, there was significant redundancy in the dataset.

### 6.3.3 Radiology Report Generation

Multiple recent works have explored the task of radiology report generation. Zhang et al. (2018a) used a combination of extractive and abstractive techniques to summarize a radiology report’s findings to generate an impression section. Due to limited text training data, Han et al. (2018) relied on weak supervision for a Recurrent-GAN and template-based framework for MRI report generation. Gale et al. (2018) used an RNN to generate template-generated text descriptions of pelvic X-rays.

More comparable to this work, Wang et al. (2018) used a CNN-RNN architecture with attention to generate reports that describe chest X-rays based on sequence decoder losses on the generated report. Li et al. (2018) generated chest X-ray reports using reinforcement learning to tune a hierarchical decoder that chooses (for each sentence) whether to use an existing template or to generate a new sentence, optimizing the language fluency metrics.

## 6.4 Methods

In this work we opt to focus on generating the *findings* section as it is the most direct annotation from the radiological images. First, we introduce the hierarchical generation strategy with a *CNN-RNN-RNN* architecture, and later we propose novel improvements that render the generated reports more clinically aligned with the true reports. Full implementation details, including layer sizes, training details, etc., are presented in Section 6.4.4.

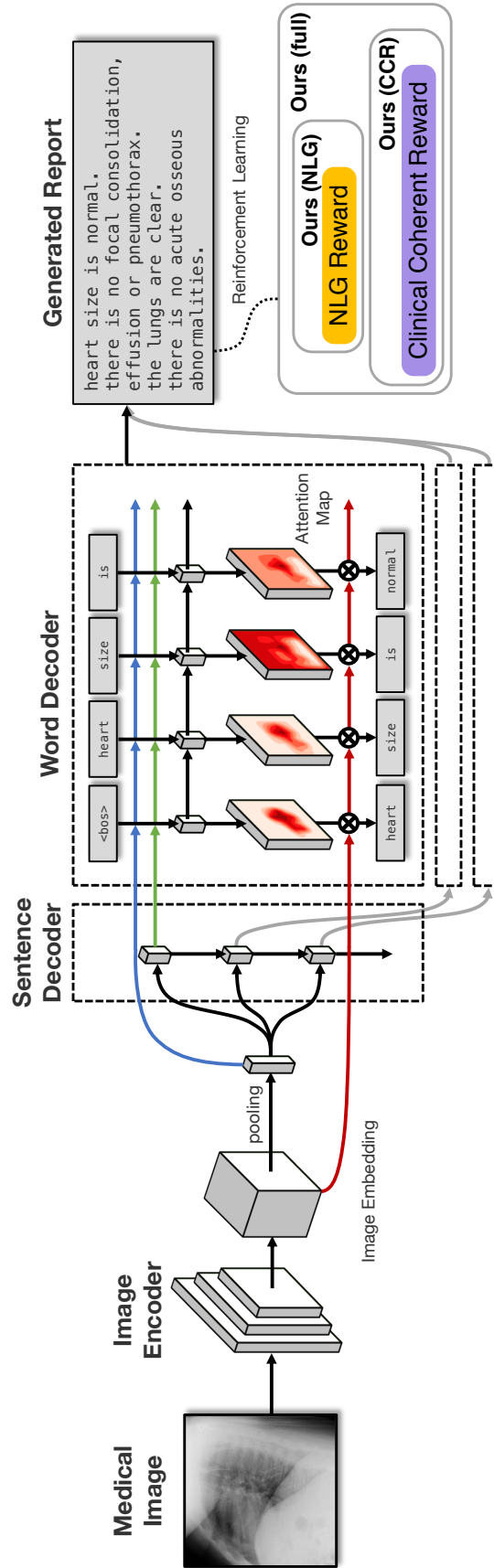


Figure 6-2: **The model for our proposed Clinically Coherent Reward.** Images are first encoded into image embedding maps, and a sentence decoder takes the pooled embedding to recurrently generate topics for sentences. The word decoder then generates the sequence from the topic with attention on the original images. NLG reward, clinically coherent reward, or combined, can then be applied as the reward for reinforcement policy learning.

### 6.4.1 Hierarchical Generation via CNN-RNN-RNN

As illustrated in Figure 6-2, we aim to generate a report as a sequence of sentences  $\mathbf{Z} = (\mathbf{z}_1, \dots, \mathbf{z}_M)$ , where  $M$  is the number of sentences in a report. Each sentence consists of a sequence of words  $\mathbf{z}_i = (z_{i1}, \dots, z_{iN_i})$  with words from a vocabulary  $z_{ij} \in \mathbb{V}$ , where  $N_i$  is the number of words in sentence  $i$ .

The image is fed through the *image encoder CNN* to obtain a visual feature map. The feature map is then taken by the *sentence decoder RNN* to recurrently generate vectors that represent the topic for each sentence. With the visual feature map and the topic vector, a *word decoder RNN* tries to generate a sequence of words and attention maps of the visual features. This hierarchical approach is in line with Krause et al. (2017) where they generate descriptive paragraphs for an image.

#### Image Encoder CNN

The input image  $I$  is passed through a CNN head to obtain the last layer before global pooling, and the feature is then projected to an embedding of dimensionality  $d$ , which is identical to the word embedding dimension. The resulting map  $\mathbf{V} = \{\mathbf{v}_k\}_{k=1}^K$  of spatial image features will be descriptive features for different spatial locations of an image. A mean visual feature is obtained by averaging all local visual features  $\bar{\mathbf{v}} = \frac{1}{K} \sum_k \mathbf{v}_k$ .

#### Sentence Decoder RNN

Given the mean visual feature  $\bar{\mathbf{v}}$ , we adopt Long-Short Term Memory (LSTM) and model the hidden state as

$$\mathbf{h}_i, \mathbf{m}_i = \text{LSTM}(\bar{\mathbf{v}}; \mathbf{h}_{i-1}, \mathbf{m}_{i-1}), \quad (6.1)$$

where  $\mathbf{h}_{i-1}$  and  $\mathbf{m}_{i-1}$  are the hidden state vector and the memory vector for the previous sentence ( $i - 1$ ) respectively. From the hidden state  $\mathbf{h}_i$ , we further generate two components, namely the topic vector  $\boldsymbol{\tau}_i$  and the stop signal  $u_i$  for the sentence,

as

$$\begin{aligned}\boldsymbol{\tau}_i &= \text{ReLU}(\mathbf{W}_\tau^\top \mathbf{h}_i + \mathbf{b}_\tau) \\ u_i &= \sigma(\mathbf{w}_u^\top \mathbf{h}_i + b_u),\end{aligned}\tag{6.2}$$

where  $\mathbf{W}$ 's and  $\mathbf{b}$ 's are trainable parameters, and  $\sigma$  is the sigmoid function. The stop signal acts as as the end-of-sentence token. When  $u > 0.5$ , it indicates the sentence decoder RNN should stop generating the next sentence.

### Word Decoder RNN

After we decode the sentence topics, we can start to decode the words given the topic vector  $\boldsymbol{\tau}_i$ . For simplicity, we drop the subscript  $i$  as this process applies to all sentences. We adopted the visual sentinel (Lu et al., 2017) that modulates the feature map  $\mathbf{V}$  with a sentinel vector. The hidden states and outputs are again modeled with LSTM, generating the posterior probability  $\mathbf{p}_j$  over the vocabulary with (1) the mean visual feature  $\bar{\mathbf{v}}$ , (2) the topic vector  $\boldsymbol{\tau}$ , and (3) the embedding of the previously generated word  $\mathbf{e}_{j-1} = \mathbf{E}_{z_{j-1}}$ , where  $\mathbf{E} \in \mathbb{R}^{d \times |\mathbb{V}|}$  is the trainable word embedding matrix. At training time, the next word is sampled from the probability  $z_j \sim p(z | \cdot) = (\mathbf{p}_j)_z$ , or the  $z$ -th element of  $\mathbf{p}_j$ .

We calculate the sentinel vector  $\mathbf{s}_j$ , the attention over the  $K$  regions of the images and the sentinel gate  $\hat{\boldsymbol{\alpha}}_j$ , the mixture context vector  $\hat{\mathbf{c}}_j$ , and the probability  $\mathbf{p}_j$  over the vocabulary as

$$\begin{aligned}\mathbf{h}_j, \mathbf{s}_j, \mathbf{m}_j &= \text{LSTM}([\bar{\mathbf{v}}, \boldsymbol{\tau}, \mathbf{e}_{j-1}]; \mathbf{h}_{j-1}, \mathbf{m}_{j-1}) \\ \boldsymbol{\alpha}_j &= \mathbf{w}_\alpha^\top \tanh([\mathbf{W}_{v\alpha}^\top \mathbf{V}, \mathbf{W}_{s\alpha}^\top \mathbf{s}_j] + (\mathbf{W}_{h\alpha}^\top \mathbf{h}_j) \mathbb{1}^\top) \\ \hat{\boldsymbol{\alpha}}_j &= \text{softmax}(\boldsymbol{\alpha}_j) \\ \hat{\mathbf{c}}_j &= [\mathbf{V}, \mathbf{s}_j] \hat{\boldsymbol{\alpha}}_j \\ \mathbf{p}_j &= \text{softmax}(\mathbf{W}_p^\top (\hat{\mathbf{c}}_j + \mathbf{h}_j)),\end{aligned}\tag{6.3}$$

where  $\mathbf{h}_{j-1}$  and  $\mathbf{m}_{j-1}$  again are the hidden state vector and the memory vector for

the previous step,  $\mathbf{W}$ 's are weights to be learned.  $[\cdot, \cdot]$  denotes matrix concatenation, and  $\mathbb{1}$  denotes a vector of all one's.

This formulation enables the model to look at different parts on the image while having the option of *looking away* at a sentinel vector. Note that this hierarchical encoder-decoder CNN-RNN-RNN architecture is fully differentiable.

### 6.4.2 Reinforcement Learning for Readability

As Rennie et al. (2017) showed, the automatic NLG metric CIDEr (Vedantam et al., 2015) is superior to other metrics such as BLEU (Papineni et al., 2002), and ROUGE (Lin, 2004). We consider the case of self-critical sequence training (SCST) (Rennie et al., 2017) which utilizes REINFORCE (Williams, 1992) algorithm, and minimizes the negative expected reward as a function of the network parameters  $\theta$ , as

$$\mathcal{L}_{\text{NLG}}(\theta) = -\mathbb{E}_{(u, \mathbf{Z}) \sim p_{\theta}(\cdot, \cdot)} [r_{\text{NLG}}(\mathbf{Z}, \mathbf{Z}^*) - r_{\text{NLG}}(\mathbf{Z}^g, \mathbf{Z}^*)], \quad (6.4)$$

where  $p_{\theta}$  is the distribution over output spaces,  $r_{\text{NLG}}$  is a metric evaluation function acting as a reward function that takes a sampled report  $\mathbf{Z}$  and a ground truth report  $\mathbf{Z}^*$ . The baseline in SCST has been replaced with the reward obtained with testing time greedily decoded report  $\mathbf{Z}^g$ .

Note that REINFORCE is effectively a *policy gradient* method in the reinforce learning (RL) space, where the *states* are the probabilities predicted by the networks, and *actions* are the multinomial sampling.

### 6.4.3 Novel Reward for Clinically Accurate Reinforcement Learning

One major downside with the approach outlined so far, unfortunately, is that in the clinical context, aiming for a good automatic metric such as CIDEr is not enough to correctly characterize the disease states. Negative judgments on diseases are critical components of the reports, by which radiologist indicates that the patient might not

have those diseases that were of concern and among the reasons for the examination. Li et al. (2018) indicated that a good portion of chest X-ray reports are heavily templated in patterns such as *no pneumothorax or pleural effusion; the lungs are clear;* or *no focal consolidation, pneumothorax or large pleural effusion*. These patterns also suggest that most patients are disease-free, hence the signal of positive mentions of the disease will be sparse.

Simply optimizing the automatic LG metrics may misguide the model to mention only the disease names as opposed to correctly positively/negatively describe the disease states. For example, if the ground truth report reads *no pleural effusion*, the models would prefer the text *mild pleural effusion* over unrelated text or even an empty string, which means intelligent optimization systems could game these metrics at the expense of clinical accuracy.

We hence propose using a *Clinically Coherent Reward (CCR)*, which utilizes a rule-based disease mention annotator, CheXpert (Irvin et al., 2019), to optimize our generated report for clinical efficacy directly. CheXpert performs classification on 12 types of thoracic diseases or X-ray related diagnoses. The mentions for support devices are also labeled. For each label type  $t$ , there are four possible outcomes for the labeling: (1) positive, (2) negative, (3) uncertain, or (4) absent mention; or,  $l_t(\mathbf{Z}) \in \{\text{p, n, u, a}\}$ . This outcome can be used to model the positive/negative disease state  $s_t \in \{+, -\}$  as  $s_t \sim p_{s|l}(\cdot|l_t(\mathbf{Z}))$ , the value of which will be discussed further later. CCR is then defined, dropping the subscripts for distribution for convenience, as

$$r_{\text{CCR}}(\mathbf{Z}, \mathbf{Z}^*) = \sum_t r_{\text{CCR},t}(\mathbf{Z}, \mathbf{Z}^*) \equiv \sum_t \sum_{s \in \{+, -\}} p(s|l_t(\mathbf{Z})) \cdot p(s|l_t(\mathbf{Z}^*)), \quad (6.5)$$

aiming to maximize the correlation of distribution over disease states between the generated text  $\mathbf{Z}$  and the ground truth text  $\mathbf{Z}^*$ . Unfortunately, as the true diagnostic state  $s$  of novel reports is unknown, we need to make several assumptions regarding the performance of the rule based labeler, allowing us to infer the necessary conditional probabilities  $p(s|l)$ .

To motivate these assumptions, first note that these diseases are universally rare,



or,  $p(+)\ll p(-)$ . Presuming the rule based labeler has any discriminative power, we can thus conclude that if the labeler assigns a negative or an absent label ( $l^-$  is one of  $\{\mathbf{n}, \mathbf{a}\}$ ),  $p(+|l^-) < p(+)\ll p(-) < p(-|l^-)$ . For sufficiently rare conditions, a reasonable assumption and simplification is to therefore take  $p(+|l^-)\approx 0$  and  $p(-|l^-)\approx 1$ . We further assume that the rule based labeler has a very high precision, and thus  $p(+|\mathbf{p})\approx 1$ . However, given an uncertain mention  $\mathbf{u}$ , the desired output probabilities are difficult to assess. As such, we define a reward-specific hyperparameter  $\beta_{\mathbf{u}}\equiv p(+|\mathbf{u})$ , which in this work we take to be 0.5. All of these assumptions could be easily adjusted, but they perform well for us here.

We also wish to use a baseline for the reward  $r_{\text{CCR}}$ . Instead of using a single exponential moving average (EMA) over the total reward, we apply EMA separately to each term as

$$\mathcal{L}_{\text{CCR}}(\theta) = -\mathbb{E}_{(u, \mathbf{Z})\sim p_{\theta}(u, \mathbf{Z})} \left[ \sum_t r_{\text{CCR},t}(\mathbf{Z}, \mathbf{Z}^*) - \bar{r}_{\text{CCR},t} \right], \quad (6.6)$$

where  $\bar{r}_{\text{CCR},t}$  is an EMA over  $r_{\text{CCR},t}$  updated as  $\bar{r}_{\text{CCR},t} \leftarrow \gamma\bar{r}_{\text{CCR},t} + (1 - \gamma)r_{\text{CCR},t}(\mathbf{Z}, \mathbf{Z}^*)$ .

We wish to pursue both semantic alignment and clinical coherence with the ground truth report, and thus we combine the above rewards for reinforcement learning on our neural network in a weighted fashion. Specifically,  $\mathcal{L}(\theta) = \mathcal{L}_{\text{NLG}}(\theta) + \lambda\mathcal{L}_{\text{CCR}}(\theta)$ , where  $\lambda$  controls the relative importance.

Hence the derivative of the combined loss with respect to  $\theta$  is thus

$$\nabla_{\theta}\mathcal{L}(\theta) = -\mathbb{E}_{(u, \mathbf{Z})\sim p_{\theta}(u, \mathbf{Z})} \left[ [r_{\text{NLG}}(\mathbf{Z}, \mathbf{Z}^*) + \lambda r_{\text{CCR}}(\mathbf{Z}, \mathbf{Z}^*)] \nabla_{\theta} \sum_i \left( \log u_i + \sum_j \log(\mathbf{p}_{ij})_{z_{ij}} \right) \right], \quad (6.7)$$

where  $\mathbf{p}_{ij}$  is the probability over vocabulary. We can approximate the above gradient with Monte-Carlo samples from  $p_{\theta}$  and average gradients across training examples in the batch.

#### 6.4.4 Implementation Details

We briefly describe the details of our implementation in this section.

##### Encoder

The image encoder CNN takes an input image of size  $256 \times 256 \times 3$ . The last layer before global pooling in a DenseNet-121 are extracted, which has a dimension of  $8 \times 8 \times 1024$ , and thus  $K = 64$  and  $d_\phi = 1024$ . Densenet-121 (Iandola et al., 2014) has been shown to be state-of-the-art in the context of classification for clinical images. The image features are then projected to  $d = 256$  dimensions with a dropout of  $p = 0.5$ .

Since typically in the X-ray image acquisition we are provided with the view position indicating the posture of the patient related to the machine, we conveniently pass this into the model as well. Indicated by a one-hot vector, the view position embedding is concatenated with the image embedding to form an input to the later decoders.

##### Decoder

As previously mentioned, the input image embedding to the LSTM has a dimension of 256, and it is the same for word embeddings and hidden layer sizes. The word embedding matrix is pretrained with Gensim (Rehurek and Sojka, 2010) in an unsupervised manner.

##### Training Details

We implement our model on PyTorch (Paszke et al., 2017) and train on 4 GeForce GTX TITAN X GPUs. All models are first trained with cross-entropy loss with the Adam (Kingma and Ba, 2014) optimizer using an initial learning rate of  $10^{-3}$  and a batch size of 64 for 64 epochs. Other than the weights stated above, the models are initialized randomly. Learning rates are annealed by 0.5 every 16 epochs and we increase the probability of feeding back a sample from the posterior  $\mathbf{p}$  by 0.05

every 16 epochs. After this bootstrapping stage, we start training with REINFORCE for another 64 epochs. The initial learning rate for the second stage is  $10^{-5}$  and is annealed on the same schedule.

Indicated by Rennie et al. (2017), we adopt CIDEr-D (Vedantam et al., 2015) metric as the reward module used in  $r_{\text{NLG}}$ . For the baseline for CCR, we choose a exponential moving average (EMA) momentum  $\gamma = 0.95$ . A weighting factor  $\lambda = 10$  has been chosen to balance the scales of the rewards for our full model.

### 6.4.5 TieNet Re-implementation

Since the implementation for TieNet (Wang et al., 2018) is not released, we re-implement it with the descriptions provided by the original authors. The re-implementation details are described in this section.

#### Overview

TieNet stands for *Text-Image Embedding Network*. It consists of three main components: image encoder, sentence decoder with *Attention Network*, and *Joint Learning Network*. It computes a global attention encoded text embedding using hidden states from a sentence decoder and saliency weighted global average pooling using attention maps from the attention network. The two global representations are combined as an input to the joint learning network. Finally, it outputs the multi-label classification of thoracic diseases. The end products are automatic report generation for medical images and classification of thoracic diseases.

#### Encoder

An image of size  $256 \times 256 \times 3$  is taken by the image encoder CNN as an input. The last two layers of ResNet-101 (He et al., 2016) are removed since we are not classifying the image. The final encoding produced has a size of  $14 \times 14 \times 2048$ . We also fine-tune convolutional blocks `conv2` through `conv4` of our image encoder during training time.

## Decoder

We also include the view position information by concatenating the view position embedding with the image embedding to form input. The view position embedding is indicated by a one-hot vector. At each decoding step, the encoded image and the previous hidden state with a dropout of  $p = 0.5$  is used to generate weights for each pixel in the attention network. The previously generated word and the output from the attention network are fed to the LSTM decoder to generate the next word.

## Joint Learning Network

TieNet proposed an additional component to automatically classify and report thoracic diseases. The joint learning network takes hidden states and attention maps from the decoder and computes global representations for report and images, then combines the result as the input to a fully connected layer to output disease labels.

In the original paper,  $r$  indicates the number of attention heads, which we set as  $r = 5$ ;  $s$  is the hidden size for attention generation, which we set as  $s = 2000$ . One key difference from the original work is that we are classifying the joint embeddings into CheXpert (Irvin et al., 2019) annotated labels, and hence we have the class count  $M = 14$ . The disease classification cross-entropy loss  $L_C$  and the teacher-forcing report generation loss  $L_R$  are combined as  $L_{\text{overall}} = \alpha L_C + (1 - \alpha)L_R$ , in which  $L_{\text{overall}}$  is the loss for which the network optimizes. However, the value  $\alpha$  was not disclosed in the original work and we use  $\alpha = 0.85$ .

## Training

We implement TieNet on PyTorch (Paszke et al., 2017) and train on 4 GeForce GTX TITAN X GPUs. The decoder is trained with cross-entropy loss with the Adam (Kingma and Ba, 2014) optimizer using an initial learning rate of  $10^{-3}$  and a mini-batch size of 32 for 64 epochs. Learning rate for the decoder is decayed by a factor of 0.2 if there is no improvement of BLEU (Papineni et al., 2002) score on the development set in 8 consecutive epochs. The joint learning network is trained with

sigmoid binary cross-entropy loss with the Adam (Kingma and Ba, 2014) optimizer using a constant learning rate of  $10^{-3}$ .

## Result

Since we are not able to access the original implementation of TieNet and we additionally inject view position information to the model, we might have small variations in result between the original paper and our re-implementation. We only compare the report generation part of TieNet to our model.

## 6.5 Experiments

### 6.5.1 Datasets

In this work, we use two chest X-ray/report datasets: MIMIC-CXR (Johnson et al., 2019) and Open-I (Demner-Fushman et al., 2015).

MIMIC-CXR is the largest radiology dataset to date and consists of 473,057 chest X-ray images and 206,563 reports from 63,478 patients<sup>1</sup>. Among these images, 240,780 are of anteroposterior (AP), 101,379 are of posteroanterior (PA), and 116,023 are of lateral (LL) views.

Furthermore, we eliminate duplicated radiograph images with adjusted brightness level or contrast as they are commonly produced for clinical needs, after which we are left with 327,281 images and 141,783 reports. The radiological reports are parsed into sections, among which we extract the *findings* section. We then apply tokenization and keep tokens with at least 5 occurrences in the corpus, resulting in 5,571 tokens in total.

Open-I is a public radiography dataset collected by Indiana University with 7,471 chest X-ray images and 3,955 reports. The reports are in extended markup language (XML) format and include pre-parsed sections. We then exclude the entries without the *findings* section and are left with 6,471 images and 3,336 reports. Tokenization

---

<sup>1</sup>This work used an alpha version of MIMIC-CXR instead of the publicly released version where the images are more standardized and the split into official train/test sets.

is done similarly, but due to the relatively small size of the corpus, we keep tokens with 3 or more occurrences, ending up with 948 tokens.

Both datasets are partitioned by patients into a train/validation/test ratio of 7/1/2 so that there is no patient overlap between sets. Words that are excluded were replaced by an *unknown* token, and the word embeddings are pretrained separately for each dataset.

### 6.5.2 Evaluation Metrics

To compare with other models including prior state-of-the-art and baselines, we adopt several different metrics that focus on different aspects ranging from a natural language perspective to clinical adequacy.

Automatic LG metrics such as *Consensus-Based Image Description Evaluation (CIDEr-D)* (Vedantam et al., 2015), *Recall-Oriented Understudy for Gisting Evaluation (ROUGE-L)* (Lin, 2004), and *Bilingual Evaluation Understudy Score (BLEU)* (Papineni et al., 2002) measure the statistical relation between two text sequences. One concern with such statistical measures is that with a limited scope from the  $n$ -grams ( $n$  up to 4) we are unable to capture disease states, as negations are common in the medical corpus and oftentimes the negation cue words and disease words can be far apart in a sentence. As such, we also include medical abnormality detection as a metric. Specifically, we compare the CheXpert (Irvin et al., 2019) labeled annotations between the generated report and the ground truth report on 14 different categories related to thoracic diseases and support devices<sup>2</sup>. We evaluate the accuracy, precision, and recall for all models.

### 6.5.3 Models

We compare our methods with state-of-the-art image captioning and medical report generation models as well as some simple baseline models: (a) *1-NN*, in which we

---

<sup>2</sup>We decide not to include NegBio (Peng et al., 2018), a previous state-of-the-art disease labeling system, due to its significant performance gap with CheXpert as reported Irvin et al. (2019) and Johnson et al. (2019)

query in the image embedding space for the closest neighbor in the train set using a test image. The corresponding report of the neighbor is used as the output for this test image; (b) *Show and Tell* (S&T) (Vinyals et al., 2015); (c) *Show, Attend, and Tell* (SA&T) (Xu et al., 2015); and (d) *TieNet* (Wang et al., 2018). To allow comparable results in all models, we slightly modify previous models to also accept the view position embedding which encodes AP/PA/LL as a one-hot vector to utilize the extra information available at image acquisition. This includes *Show and Tell*, *Show, Attend, and Tell*, and our re-implementation of *TieNet*, which is detailed in Section 6.4.5 because the authors did not release their code.

We observed our model to sometimes repeat the findings multiple times. We apply post-hoc processing where we remove exact duplicate sentences in the generated reports. This proves to improve the readability but interestingly slightly degrades NLG metrics.

Additionally, we perform several ablation studies to inspect the contribution of various components of our model. In particular, we assess

1. *Ours (NLG)*: Use  $r_{\text{NLG}}$  only for reinforced learning, as often is the case with the prior state-of-the-art.
2. *Ours (CCR)*: Use  $r_{\text{CCR}}$  only and do not care about aligning the natural language metrics.
3. *Ours (full)*: Consider both rewards as formulated in Section 6.4.3.

In order to provide some context to the metric scores, we also train an unsupervised RNN language model which generates free text without conditioning on input radiograph images, which we denote as *Noise-RNN*. All recurrent models, including prior works and our models, use beam search with a beam size of 4.

## 6.6 Results

### 6.6.1 Quantitative Results

#### Natural Language Metrics

In Table 6.2 we show the automatic evaluation scores for baseline models, prior works, and variants of our models on the aforementioned test sets. Ours (NLG), that solely optimizes CIDEr score, achieves superior performance in terms of natural language metrics, but its clinical meaningfulness is not significantly above the *major class* in which we predict all patients to be disease-free. This phenomenon is common among all other models that do not consider the clinical alignment between the ground truth and the generated reports. On the other hand, in our full model, if we consider both natural language and clinical coherence, we can achieve the highest clinical disease annotation accuracy while still retaining decently high NLG metrics.

We also conducted the ablation study with the model variant Ours (CCR), where we use reinforcement learning on only the clinical accuracy. It is clear that we are unable to achieve higher clinical coherence, though readability might be sacrificed. We thus conclude that a combination of both NLG metrics and a clinically sensible objective is crucial in training a useful model in clinical practice.

One thing to note is that although *Noise-RNN* is not dependent on the image, its NLG metrics, especially ROUGE, are not far off from models learned with supervision. We also note that MIMIC-CXR is better for training such an encoder-decoder model not just for its larger volume of data, but also due to its higher proportion of positive disease annotations at 16.7% while Open-I only has 5.4%. This discrepancy leads to a 156 times difference in the number of images from diseased patients.

#### Clinical Efficacy Metrics

In Table 6.3 we can compare the labels annotated by CheXpert calculated over all test set generated reports. Note that the labeling process generates a discrete binary label as opposed to predicting continuous probabilities, and as such we are unable to obtain



Model	Natural Language					Clinical Accuracy
	CIDEr	ROUGE	BLEU-1	BLEU-2	BLEU-3	
<i>Major Class</i>	-	-	-	-	-	0.828
Noise-RNN	0.716	0.272	0.269	0.172	0.113	0.803
1-NN	0.755	0.244	0.305	0.171	0.098	0.818
S&T	0.886	0.300	0.307	0.201	0.137	0.837
SA&T	0.967	0.288	0.318	0.205	0.137	0.849
TieNet	1.004	0.296	0.332	0.212	0.142	0.848
Ours (NLG)	<b>1.153</b>	<b>0.307</b>	<b>0.352</b>	<b>0.223</b>	<b>0.153</b>	0.834
Ours (CCR)	0.956	0.284	0.294	0.190	0.134	<b>0.868</b>
Ours (full)	1.046	<b>0.306</b>	0.313	0.206	0.146	<b>0.867</b>
<i>Major Class</i>	-	-	-	-	-	0.911
Noise-RNN	0.747	0.291	0.233	0.130	0.087	0.914
1-NN	0.728	0.201	0.232	0.116	0.051	0.911
S&T	0.926	0.306	0.265	0.157	0.105	0.915
SA&T	1.276	0.313	0.328	0.195	0.123	0.908
TieNet	1.334	0.311	0.330	0.194	0.124	0.902
Ours (NLG)	<b>1.490</b>	<b>0.359</b>	<b>0.369</b>	<b>0.246</b>	<b>0.171</b>	0.916
Ours (CCR)	0.707	0.244	0.162	0.084	0.055	<b>0.917</b>
Ours (full)	1.424	0.354	0.359	0.237	0.164	<b>0.918</b>

Table 6.2: **Automatic Evaluation Scores.** The table is divided into natural language metrics and clinical finding accuracy scores. BLEU- $n$  counts up  $n$ -gram for evaluation, and accuracy is the averaged macro accuracy across all clinical findings. *Major class* always predicts negative findings.

MIMIC-CXR

Label	Count	1-NN	S&T	SA&T	TieNet	Ours (NLG)	Ours (CCR)	Ours (full)
Total	69031	-	-	-	-	-	-	-
No Finding	15677	0.432	0.299	0.349	0.339	0.339	<b>0.491</b>	0.405
Enlarged Cardiome-diastinum	6064	0.123	0.134	0.163	0.179	0.180	<b>0.202</b>	0.167
Cardiome-galy	19065	0.440	0.535	0.438	0.464	0.000	0.678	<b>0.704</b>
Lung Lesion	2447	0.064	<b>0.333</b>	0.223	0.000	0.000	0.000	0.000
Air-space Opacity	21972	0.432	0.607	0.592	0.571	0.453	<b>0.640</b>	0.460
Edema	6594	0.265	0.331	0.244	<b>0.405</b>	0.266	0.280	0.000
Consolidation	2384	0.076	0.013	<b>0.180</b>	0.151	0.089	0.037	0.000
Pneumonia	3068	0.065	0.106	0.091	0.082	0.075	0.000	<b>0.400</b>
Atelectasis	16161	0.374	0.490	0.496	0.470	0.385	0.476	<b>0.521</b>
Pneumothorax	2636	0.079	0.034	0.095	0.081	0.081	0.039	<b>0.098</b>
Pleural Effusion	15283	0.534	0.550	0.545	<b>0.735</b>	0.487	0.683	0.689
Pleural Other	1285	<b>0.039</b>	0.000	0.103	0.000	0.000	0.000	0.000
Fracture	2617	<b>0.059</b>	0.000	0.000	0.000	0.000	0.000	0.000
Support Devices	22227	0.534	0.823	0.847	0.827	0.794	0.849	<b>0.880</b>
Precision (macro)		0.253	0.304	<b>0.312</b>	0.307	0.225	<b>0.313</b>	0.309
Precision (micro)		0.383	0.414	0.430	0.473	0.419	<b>0.634</b>	0.586
Recall (macro)		<b>0.265</b>	0.173	0.232	0.220	0.209	0.126	0.134
Recall (micro)		<b>0.400</b>	0.276	0.367	0.355	0.360	0.227	0.237

Table 6.3: **Clinical Finding Scores.** The precision scores for each of the labels are listed and aggregated into the overall precision scores. Recall scores are shown in the last two rows. Macro denotes averaging the numbers in the table directly and micro accounts for class prevalence.

discriminative metrics such as the Area Under the Receiver Operator Characteristic (AUROC) or the Area Under the Precision-Recall Curve (AUPRC). Precision-wise, Ours (CCR) achieves the highest overall scores including macro-average and micro-average. The runner-up is Ours (full) model, which additionally considers language fluency. Note that the macro- metrics can be quite noisy as the per-class metric can be dependent on just a few examples. Many entries in the table are zeros, as they never yield positive predictions and we regard them as zeros to penalize such behavior. Regarding the recall metric, we are able to see a substantial drop in Ours (CCR) and Ours (full) as a result of optimizing for accuracy. Accuracy is closely associated with precision but overpursuing it might lead to harm in terms of recall. It is worthwhile to notice that the nearest neighbor *1-NN* has the highest recall, and this is no surprise since as shown before (Strobelt et al., 2019), generated sequences tend to follow the statistics and favor common words too much. Rare combinations of tokens in the corpus can be easily neglected, resulting in predictions of mostly major classes.

## 6.6.2 Qualitative Results

### Evaluation of Generated Reports

Table 6.4 demonstrates the qualitative results of our full model. In general, our models are able to generate descriptions that align with the logical flow of reports written by radiologists, which start from general information (such as views, previous comparison), positive, then negative findings, with the order of lung, heart, pleura, and others. TieNet also generates report descriptions with such logical flow but in slightly different orders. For the negative findings cases, both our model and TieNet do well on generating reasonable descriptions without significant errors. Regarding the cases with positive findings, TieNet and our full model both cannot identify all radiological findings. Our full model is able to identify the major finding in each demonstrated case. For example, cardiomegaly in the first case, pleural effusion, and atelectasis in the second case.

A formerly practicing radiologist reviewed a larger subset of our generated reports

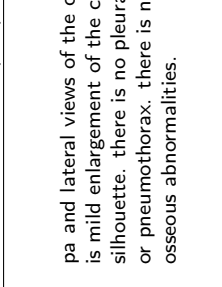
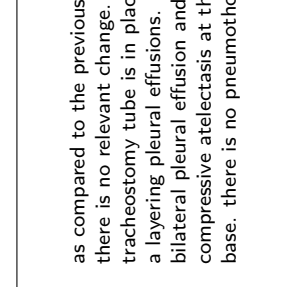
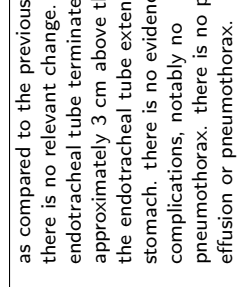
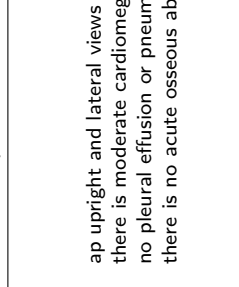
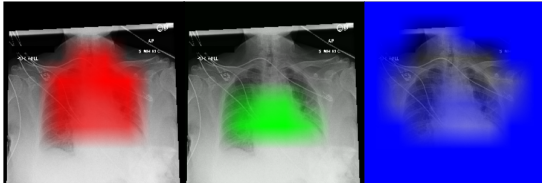
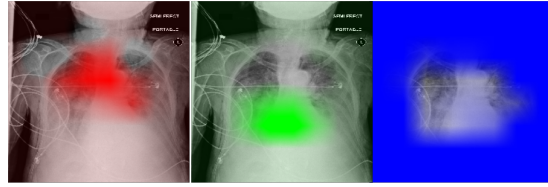
Ground Truth	TieNet	Ours (full)
 <p>cardiomegaly is moderate. bibasilar atelectasis is mild. there is no pneumothorax. a lower cervical spinal fusion is partially visualized. healed right rib fractures are incidentally noted.</p>	<p>ap portable upright view of the chest. there is no focal consolidation, effusion, or pneumothorax. the cardiomeastinal silhouette is normal. imaged osseous structures are intact.</p>	<p>pa and lateral views of the chest. there is mild enlargement of the cardiac silhouette. there is no pleural effusion or pneumothorax. there is no acute osseous abnormalities.</p>
 <p>as compared to the previous radiograph, the monitoring and support devices are unchanged. unchanged bilateral pleural effusions, with a tendency to increase, and resultant areas of atelectasis. the air collection in the bilateral soft tissues is slightly decreased. unchanged right picc line. no definite evidence of pneumothorax.</p>	<p>as compared to the previous radiograph, the patient has received a nasogastric tube. the course of the tube is unremarkable, the tip of the tube projects over the middle parts of the stomach. there is no evidence of complication, notably no pneumothorax. the other monitoring and support devices are constant. constant appearance of the cardiac silhouette and of the lung parenchyma.</p>	<p>as compared to the previous radiograph, there is no relevant change. the tracheostomy tube is in place. there is a layering pleural effusions. NAME bilateral pleural effusions and compressive atelectasis at the right base. there is no pneumothorax.</p>
 <p>the course of the dobbhoff feeding catheter is unremarkable, and its tip is seen projecting over the stomach. there is no evidence of complications, specifically no pneumothorax. as compared to the prior radiograph dated DATE, there has been no other significant interval change.</p>	<p>ap portable upright view of the chest. overlying ekg leads are present. there is no focal consolidation, effusion, or pneumothorax. the cardiomeastinal silhouette is normal. imaged osseous structures are intact.</p>	<p>as compared to the previous radiograph, there is no relevant change. the endotracheal tube terminates approximately 3 cm above the NAME. the endotracheal tube extends into the stomach. there is no evidence of complications, notably no pneumothorax. there is no pleural effusion or pneumothorax.</p>
 <p>interval placement of a left basilar pigtail chest tube with improving aeration in the left mid to lower lung and near complete resolution of the pleural effusion. there are residual patchy opacities within the left mid and lower lung as well as at the right base favoring resolving atelectasis. no pneumothorax is appreciated on this semi upright study. heart remains stably enlarged. mediastinal contours are stably widened, although this NAME be related to portable technique and positioning. this can be better evaluated on followup imaging. no pulmonary edema.</p>	<p>as compared to the previous radiograph, the patient has been extubated. the nasogastric tube is in unchanged position. the lung volumes remain low. moderate cardiomegaly with minimal fluid overload but no overt pulmonary edema. no larger pleural effusions. no pneumonia.</p>	<p>ap upright and lateral views of the chest. there is moderate cardiomegaly. there is no pleural effusion or pneumothorax. there is no acute osseous abnormalities.</p>

Table 6.4: Sample images along with ground truth and generated reports. Note that upper case tokens are results of anonymization.



ap upright and lateral views of the chest. there is moderate cardiomegaly. there is no pleural effusion or pneumothorax. there is no acute osseous abnormalities.

(a)



as compared to the previous radiograph, there is no relevant change. tracheostomy tube is in place. there is a layering pleural effusions. NAME bilateral pleural effusion and compressive atelectasis at the right base. there is no pneumothorax.

(b)

Figure 6-3: **Visualization of the generated report and image attention maps.** Different words are underlined with its corresponding attention map shown in the same color. Best viewed in color.

manually. They drew several conclusions. First, our full model tends to generate sentences related to pleural effusion, atelectasis, and cardiomegaly correctly—which is aligned with the clinical finding scores in Table 6.3. TieNet instead misses some positive findings in such cases. Second, there are significant issues in *all* generated reports, regardless of the source model, which include the description of supportive lines and tubes, as well as lung lesions. For example, TieNet is prone to generate nasogastric tube mentions while our model tends to mention tracheostomy or endotracheal tube, and yet both models have difficulty identifying some specific lines such as chest tube or PICC (peripherally inserted central catheter) line. Similarly, both systems do not generate the sentence with positive lung parenchymal findings correctly.

From this (small) sample, we are unable to draw a conclusion whether our model or TieNet truly outperforms the other since both present with significant issues and each has strengths the other lacks. Critically, neither of them can describe the majority of the findings in the chest radiograph well, especially for positive cases, even if the quantitative metrics demonstrate the reasonable performance of the models. This illustrates that *significant* progress is still needed in this domain, perhaps building on the directions we explore here before these techniques could be deployed in a clinical environment.

## Learning Meaningful Attention Maps

Attention maps have been a useful tool in visualizing what a neural network is attending to, as demonstrated by [Rajpurkar et al. \(2017\)](#). Figure 6-3 shows the intermediate attention maps for each word when it is being generated. As we can observe, the model is able to roughly capture the location of the indicated disease or parts, but we also find, interestingly, that the attention map tends to be the complement of the actual region of interest when the disease keywords follow a negation cue word. This might indicate that the model is actively looking at the rest of the image to ensure it does not miss any possible symptoms exhibited before asserting disease-free states. This behavior has not been widely discussed before, partially because attention maps for negations are not the primary focus of typical image captioning tasks, and most attention mechanisms employed in a clinical context were on classification tasks where they also do not specifically focus on negations.

## 6.7 Limitations & Future Work

Our work has several notable limitations and opportunities for future work. First and foremost, the post-processing step required to remove repeated sentences is an ugly necessity, and we endeavor to remove it in future iterations of this work. Promising techniques exist in NLG for the inclusion of greater diversity, which warrant further investigation here.

Secondly, our model operates using images in isolation, without consideration of whether these images are part of a series of ordered radiographs for a single patient, which might be summarized together. Using all available information, potentially including comorbidities, clinical impressions documented in notes, lab values, medication history, vital signs, and etc., has the potential to improve the quality of the generated reports, and should definitely be investigated further.

Lastly, we note that though our model yields very strong performance for CheXpert *precision*, its recall is much worse. Recall versus precision is favored to different degrees in differing clinical contexts. For example, for screening purpose, recall (sen-

sitivity) is an ideal metric since the healthy cases usually won't give positive findings. However, precision (positive predictive value, PPV) is much more critical for validating the clinical impression, which is common in an ICU (intensive care unit) setting where patients receive a radiological study on the basis of strong clinical suspicion. We believe that our system's poor recall is a direct result of the setup of our RL models and the CCR reward, which optimizes for accuracy and inherently boosts precision. It is the choice of optimization objectives that lead to the results. Depending on the actual clinical applications, we may, in turn, optimize *Recall at Fixed Precision* (R@P) or  $F_\beta$  score via methods described by Eban et al. (2016).

## 6.8 Reflections on Trends in the Field

In the course of this work, we also encounter several other larger points which are present not only in our study but also in many related studies in this domain and warrant further thought by the community.

### 6.8.1 System Generalizability

CheXpert used in our models is rule-based, which is harder to generalize to other datasets and to identify the implicit features inside the language patterns. CheXpert is also specialized to English and would require considerable work to re-code its rules for other natural languages. A more universal approach for subsequent research may use a learning-based approach for labeling to improve generalizability and extend to corpora in different languages; for example, PadChest in Spanish.

### 6.8.2 Be Careful What You Wish For

NLG metrics are known to be only limited substitutes for a true assessment of readability (Kilickaya et al., 2016; Liu et al., 2016). For radiology reports more specifically, this problem is even more profound, as prior works often use *readability* as a proxy for clinical efficacy. Additionally, we note that these NLG evaluation metrics are eas-

ily susceptible to gaming. In our results, our post-processing step of removing exact duplicates actually *worsens* our CIDEr score, which is the opposite of what should be desired for an NLG evaluation metric. Even if our proposed clinical coherence aims at resolving the unwanted misalignment between NLG and real practice, we are not able to obviously judge whether our system is better despite its performance on paper. This fact is especially troubling given the increasing trend of using reinforcement learning (RL) to directly optimize objectives, as has been done in prior work (Li et al., 2018) and as we do here. Though RL can offer marked improvements in these automatic metrics, which are currently the best the field can do, how well it translates to the real clinical efficacy is unclear. The careful design of improved evaluation metrics, specifically for radiology report generation, should be a prime focus for the field going forward.

## 6.9 Summary

In this chapter, we develop a chest X-Ray radiology report generation system which hierarchically generates topics from images, then words from topics. This structure gives the model the ability to use largely templated sentences (through the generation of similar topic vectors) while preserving its freedom to generate diverse text. The final system is also optimized with reinforcement learning for both readability (via CIDEr) and clinical correctness (via the novel Clinically Coherent Reward). Our system outperforms a variety of compelling baseline methods across readability and clinical efficacy metrics on both MIMIC-CXR and Open-I datasets.

It is not hard to observe that, by adding clinical heuristics to our system when constructing medical imaging solutions, we are able to achieve a much better model when evaluated from a clinical perspective. This insight is not surprising but is often overlooked in the process of modeling, as most AI solutions originate from the field of natural image/language processing where metrics and objectives are crafted for human perception of the natural world. Medical machine learning, on the other hand, holds the ultimate goal of teaching machines about the modern medicine under-



standing accumulated over the past decades, and is not aligned with simple everyday perception. Hence, it is very recommended that, during the modeling process, that AI scientists proactively collaborate with clinicians to allow a better integration of medical insights.



# Chapter 7

## Reinforcement Learning for Weakly Supervised Dental Imaging Data

The use of deep reinforcement learning (RL) concentrates on the cases where learning is done with deep learning models on top of non-differentiable targets. Conventionally, in deep learning, the targets to optimize for neural networks are designed to be differentiable to enable direct computation of gradients with back-propagation techniques. In some cases, however, the desired optimization targets are not easily computed, or the annotation provided by the dataset is not directly usable in the learning scenario in question.

In a medical imaging context, a limited number of prior works (Yang et al., 2019; Qin et al., 2020; Liao et al., 2020) has explored using RL in their learning pipelines. If there are alternative annotations, in the form of non-differentiable objectives, provided in a medical imaging dataset, the machine learning algorithm might be able to leverage the extra annotation to learn a better model, possibly using RL.

---

This chapter is adapted from the published article “*DeepOPG: Improving Orthopantomogram Finding Summarization with Weak Supervision*” (Hsu and Wang, 2021) to which I have contributed as the first author.

## 7.1 Overview

Clinical finding summaries from an orthopantomogram (OPG), or a dental panoramic radiograph, have significant potential to improve patient communication and speed up clinical judgments. While orthopantomogram is a first-line tool for dental examinations, no existing work has explored the summarization of findings from it. A finding summary has to find teeth in the imaging study and label the teeth with several types of past treatments. To tackle the problem, we develop DeepOPG that breaks the summarization process into functional segmentation and tooth localization, the latter of which is further refined by a novel dental coherence module. We also leverage weak supervision labels to improve detection results in a reinforcement learning scenario. Experiments show high efficacy of DeepOPG on finding summarization, achieving an overall AUC of 88.2% in detecting six types of findings. The proposed dental coherence and weak supervision are shown to improve DeepOPG by adding 5.9% and 0.4% to  $AP@IoU = 0.5$ .

## 7.2 Background

An orthopantomogram (OPG), or a dental panoramic radiograph, is a half-circle X-ray scanning of the oral region that compresses the complicated 3D structures to a 2D representation as shown in Figure 7-1. OPG has many advantages including short acquisition time and convenience of examination. Moreover, its capability to deliver rich information about the oral and maxillofacial regions makes it a first-line dental screening tool (Perschbacher, 2012). With that said, it is this structural complexity that unavoidably limits the interpretation of OPG to only dental experts (Henzler et al., 2018). Even for these dental experts, interpretation of findings can suffer from insufficient inter-rater agreement (Kweon et al., 2018) and low time efficiency in clinical practices (Plessas et al., 2019; Rozyło-Kalinowska, 2018). As such, an automatic system to provide finding summaries on the fly can be beneficial in terms of both patient communication and clinical assistance. The systematically collected

summaries can further provide an invaluable source for subsequent dental research and statistical analysis, which the current clinical workflow cannot offer.

There have been attempts to provide information about teeth in radiographs with convolutional neural networks (CNNs). In [Ronneberger et al. \(2015a\)](#), they offer pixel-wise segmentation maps that label seven different parts of teeth. [Miki et al. \(2017\)](#) classifies teeth images into eight categories but requires that the bounding boxes be manually annotated first. [Koch et al. \(2019\)](#) identifies silhouettes for natural teeth in OPG with semantic segmentation but treats all teeth as a single connected region. [Silva et al. \(2018\)](#) and [Jader et al. \(2018\)](#) use a novel OPG dataset and object detection to treat teeth as individual instances for object detection, yet they both do not number the teeth. [Tuzoff et al. \(2019\)](#) addresses both detection and numbering, but fails to include dental implants. [Kim et al. \(2020\)](#) provides detection of teeth, implant, and crowns but does not associate them with findings. Moreover, the vast majority of past research relies on annotations on dense attribute maps which is resource-intensive, and the use of weaker (and faster to collect) supervision has not been explored.

In this work, we aim to provide a summary of findings in an OPG image, including all teeth found in the image, their FDI (*Fédération Dentaire Internationale*) notations, and all the clinical findings on each. We propose DeepOPG, which breaks the finding summarization process into two sub-tasks: functional segmentation and tooth localization, the latter of which is further refined at inference-time by maximizing the novel *Dental Coherence Reward (DCR)*. DCR can also be used by reinforcement learning (RL) for training-time optimization, leveraging the *missing teeth annotation* that are quick for dentists to label as weak supervision. We curate a set of annotations on OPG including semantic segmentation, instance segmentation, and finding summaries for 298 studies on a dataset in the public domain. Our experiments show that DeepOPG achieves an overall AUC of 88.2% on finding detection, which is 1.6% higher than without weak supervision. The tooth/implant localization yields an average precision at zero intersection-over-union ( $AP@IoU = 0$ ) of 98.6%, which is 5.6% higher than without injecting dental domain knowledge and 0.9% higher than without

feeding in segmentation maps. The numbers demonstrate the effectiveness of each component of DeepOPG. To our knowledge, this is the first work to explore the summarization of findings in OPG images and to use weak supervision to improve finding summarization.

## 7.3 Methods

Our ultimate goal for the DeepOPG system is to generate a finding summary entailing six different types of findings on each tooth in an OPG. The resulting findings are formulated as binary attribute labels on the teeth found in the OPG. We decompose the problem into two main tasks: localizing the objects of interest (in this case, teeth and implants) and determining the visual features that result in the findings. Illustrated in Figure 7-1, there are three modules in DeepOPG, and they operate at original resolutions of the images. This is essential since some findings (e.g., fillings found in the root canal) are visually tiny, and any down-sampling would result in a loss of information. We combine the results from both tasks of localization and function determination to output predictive values for each of the finding types on individual teeth.

### 7.3.1 Model Architecture

First of all, the *functional segmentation* module as shown in Figure 7-1a consumes a radiographic image as the input and generates a map that shows the dental functionality of each pixel. The functional segmentation map and the original image is then concatenated to go into the *tooth localization* module in Figure 7-1b that picks out the individual dental object of interest including teeth and implants. The resulting detection outcomes are further refined by the *dental coherence* module where clinical heuristics are applied to ensure coherence with dental knowledge.

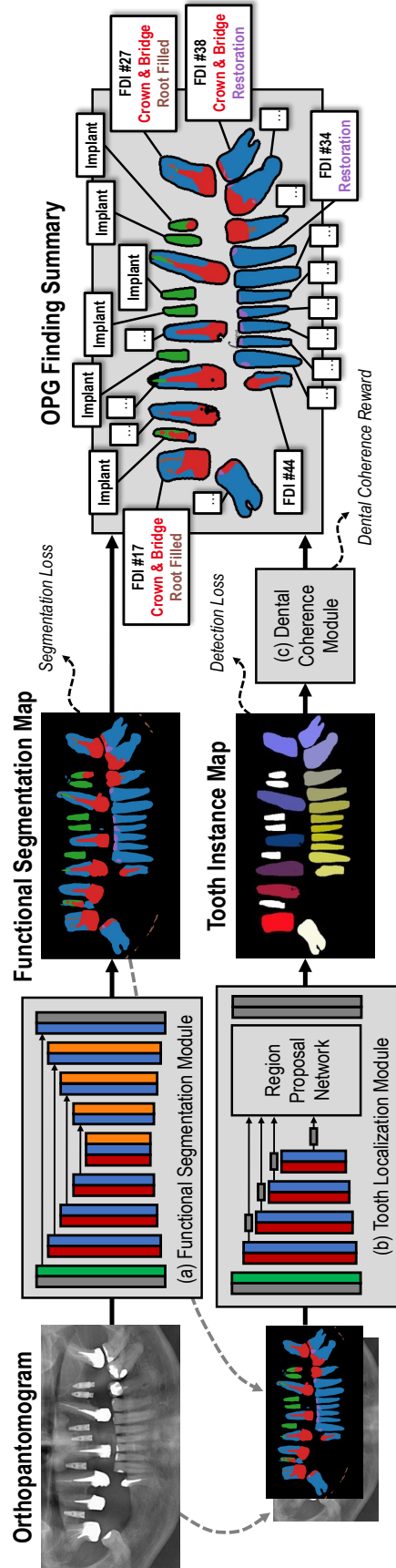


Figure 7-1: **System Overview of DeepOPG.** There are three modules in DeepOPG working together to obtain orthopantomogram finding summary on each tooth. (a) Functional Segmentation Module performs per-pixel classification, (b) Tooth localization Module localizes each tooth, and (c) Dental Coherence Module ensures that output is clinically reasonable. Ultimately, we combine the segmentation maps and teeth identity maps to produce findings with explainable predictive values.

## Functional Segmentation via Semantic Segmentation

Given the input gray-scale image  $I \in \mathbb{R}^{H \times W \times 1}$ , where  $H$  and  $W$  are the height and width of the image, we employ a network with an U-Net-like (Ronneberger et al., 2015a) structure to predict  $S \in \mathbb{R}^{H \times W \times C_{\text{seg}}}$ , a per-class probability for each pixel that determines its functional class  $c \in \{1, 2, \dots, C_{\text{seg}}\}$ , where  $C_{\text{seg}}$  is the number of classes. In our experiments,  $C_{\text{seg}} = 7$  and includes the following classes for finding summarization: (1) background, (2) normal (non-impacted) teeth, (3) impaction, (4) crown & bridge, (5) restoration, (6) root filling material, and (7) implant. Note that as these classes are mutually exclusive, a ground truth segmentation map  $S^{\text{gt}} \in \mathbb{R}^{H \times W \times C_{\text{seg}}}$  is one-hot encoded and the activation function of network output is thus softmax. Specifically, we choose ResNet-50 (He et al., 2016) to be the encoder and ResNet-18 with transposed convolutions to be the decoder.

## Tooth localization via Object Detection

The tooth localization module takes the concatenated image  $[I, S] \in \mathbb{R}^{H \times W \times (1 + C_{\text{seg}})}$  and produces  $N$  *detections*, each including a class probability vector  $\mathbf{p}_n \in \mathbb{R}^{C_{\text{det}}}$ , a region of interest (ROI)  $\mathbf{b}_n \in \mathbb{R}^4$ , and the class-wise masks  $M_n \in \mathbb{R}^{H \times W \times C_{\text{det}}}$ , where  $C_{\text{det}}$  is the number of classes in detection. Concretely, we adopt Mask-RCNN (He et al., 2017) that proposes a pool of candidate ROIs with a region proposal network (RPN) before using a small sub-network to derive the aforementioned detection properties.

As we are interested in not only natural teeth but also dental implants, in total there are  $C_{\text{det}} = 34$  classes representing the background, 32 different teeth in permanent dentition, and the implant. Hereafter the natural teeth are annotated using the FDI World Dental Federation notation as shown in Figure 7-4a.

## Inference-Time Dental Coherence Decoding

One major downside of directly using off-the-shelf detection algorithms is that they mostly consider the detection efficacy of individual objects rather than the conglomerate of several objects. As a result, in pilot experiments, we often observe the detection



module to output several objects of interest with the same FDI tooth number, which is highly unlikely in practice. Equally frequently, there are cases where an image patch can be detected as multiple different classes at the same time, with largely overlapping masks. Even with existing techniques such as non-maximum suppression (NMS) that filters out overlapping objects with lower confidence scores, we are only able to partially resolve the latter problem.

To this end, we propose to look at this problem from an optimization perspective and *decode* an assignment  $\mathbf{E}$  of teeth number to the detected objects by maximizing the Dental Coherence Reward (DCR) defined as

$$r_{\text{DCR}}(\mathbf{E}; \mathbf{P}, \mathbf{M}) \equiv \sum_{n,c} p_{nc} \cdot e_{nc} - \sum_{n,c,m,d} q_{ncmd} \cdot e_{nc} \cdot e_{md}, \quad (7.1)$$

subject to  $\sum_n e_{nc} \leq 1 \forall c \in \{1, 2, \dots, C\}$ , and  $e_{nc} \in \{0, 1\} \forall c \in \{1, 2, \dots, C\} \forall n \in \{1, 2, \dots, N\}$ , where  $p_{nc} = (\mathbf{p}_n)_c$  is the probability of object  $n$  belonging to class  $c$ ,  $q_{ncmd} = \frac{(M_n)_c \cap (M_m)_d}{(M_n)_c \cup (M_m)_d}$  is the intersection-over-union (IoU) between masks  $(M_n)_c$  (the class- $c$  mask of object  $n$ ) and  $(M_m)_d$ , and  $e_{nc}$  is an indicator whether we *assign* tooth  $c$  to object  $n$ . Note that an object  $n$  can be *suppressed* (*i.e.*, discarded) if  $\sum_c e_{nc} = 0$ . This formulation happens to be the *Generalized Quadratic Assignment Problem* (GQAP) (Lee and Ma, 2004) which is extensively studied in optimization theory and has solvers widely available. Implants are not modified in this module, and hence  $C = C_{\text{det}} - 1$  with implants excluded for optimization.

The idea to maximize DCR closely resembles how our dental experts parse an OPG, where they explain they would (1) identify all minimally overlapping objects and mentally assign each a number, followed by (2) ensuring that across a single image, no teeth share the same FDI number (obviously, multiple dental implants can still present simultaneously). While it is certainly possible in the clinics to observe the extremely rare cases where two natural teeth overlap on the OPG, oftentimes highly overlapping masks simply indicate that a tooth is independently recognized by two RPN proposals.

## Explainable OPG Finding Summary

We assemble the information from the semantic segmentation and the detection outputs to derive the finding summary. For each of the teeth or implants, we use its mask  $M$  to select the corresponding regions in the segmentation map  $S$  and calculate the percentage of pixel counts for each functional class  $c$  in that area as  $f_c = \sum_{i \in M} \frac{\mathbb{1}[S_i=c]}{|M|}$ , where  $\mathbb{1}[\cdot]$  is the indicator function. The percentage area  $f_c$  is then used as the predictive value for finding type  $c$  on that tooth. Doing so not only allows us to provide an explainable finding output that dentists can easily interpret, but we also can adjust the threshold on  $f_c$  based on our sensitivity/specificity requirements.

### 7.3.2 Improved Tooth Localization with Weakly Supervised Reinforcement Learning

The annotation for tooth localization usually requires that the dental experts carefully outline the silhouettes of each tooth and provide an FDI number for it. This type of annotation is labor-intensive and is usually not available at most data registries. What is more likely to be available is a description of whether a tooth is missing or not in a text report (*i.e.*,  $e_c^{\text{gt}} \equiv \sum_n e_{nc} = 1$  if the tooth  $c$  is present and 0 otherwise). We hereby are interested to find if weak supervision in the form of tooth missingness is helpful to train the tooth localization module in a reinforcement learning (RL) scenario.

We utilize the REINFORCE (Williams, 1992) algorithm where as long as a probability and a reward are defined for output, the network can learn to maximize the reward function. The algorithm is effectively a *policy gradient* method in the RL space, where the *states* are the probabilities predicted by the networks, and *actions* are the multinomial/binomial sampling. At training time, instead of decoding the GQAP problem, we sample a one-hot vector  $\hat{\mathbf{e}}_n = [\hat{e}_{n1}, \hat{e}_{n2}, \dots, \hat{e}_{nC}] \sim \mathbf{p}_n$  from the class distribution  $\mathbf{p}_n$  for each object  $n$  independently. As the random samples might violate the constraint that each FDI number cannot be taken by multiple teeth (*i.e.*,  $e_c^{\text{gt}} = \sum_n e_{nc} > 1$ ), we penalize this situation by setting the reward for extra teeth to

be negative

$$\hat{p}_{nc} = \begin{cases} +p_{nc} & \text{if tooth } c \text{ is present and } p_{nc} \text{ is the largest probability for it} \\ -\lambda p_{nc} & \text{otherwise (for extra teeth),} \end{cases} \quad (7.2)$$

and calculate  $r_{\text{DCR}}(\hat{\mathbf{E}}; \hat{\mathbf{P}}, \mathbf{M})$  on the samples.

Effectively, if we expand the augmented DCR with the sampled teeth,

$$\begin{aligned} r_{\text{DCR}}(\hat{\mathbf{E}}; \hat{\mathbf{P}}, \mathbf{M}) &= \sum_{n,c} \hat{p}_{nc} \cdot \hat{e}_{nc} - \sum_{n,c,m,d} q_{ncmd} \cdot \hat{e}_{nc} \cdot \hat{e}_{md} \\ &= \sum_{n,c} p_{nc} \cdot \hat{e}_{nc} - \sum_{n,c,m,d} q_{ncmd} \cdot \hat{e}_{nc} \cdot \hat{e}_{md} \\ &\quad + \lambda \sum_c \left( e_c^{\text{gt}} \left( \max_c p_{nc} \hat{e}_{nc} \right) - \sum_n p_{nc} \hat{e}_{nc} \right) \\ &= r_{\text{DCR}}(\hat{\mathbf{E}}; \mathbf{P}, \mathbf{M}) + \lambda \sum_c \left( e_c^{\text{gt}} \left( \max_c p_{nc} \hat{e}_{nc} \right) - \sum_n p_{nc} \hat{e}_{nc} \right), \end{aligned} \quad (7.3)$$

which can be derived from the vanilla DCR,  $r_{\text{DCR}}(\hat{\mathbf{E}}; \hat{\mathbf{P}}, \mathbf{M})$ .

The loss as given by REINFORCE is thus

$$\nabla_{\theta} \mathcal{L}_{\text{DCR}} = -\mathbb{E}_{\hat{\mathbf{E}} \sim p_{\theta}(\cdot)} \left[ r_{\text{DCR}}(\hat{\mathbf{E}}; \hat{\mathbf{P}}, \mathbf{M}) \nabla_{\theta} \sum_{n,c} \hat{e}_{nc} \log p_{nc} \right], \quad (7.4)$$

where  $p_{\theta}(\cdot)$  is the distribution characterized by the network. We can approximate the above gradient with Monte-Carlo samples and average gradients across training examples in the batch. Different from the aforementioned inference-time decoding, we can explicitly optimize the network for DCR with reinforcement learning here.

To learn DeepOPG, we employ a multi-stage learning procedure since the RPN in Mask-RCNN is non-differentiable. First, we train the functional segmentation module, optimizing the segmentation cross-entropy loss  $\mathcal{L}_{\text{seg}}$ . Following this, the tooth localization module learns using the inference-time predicted segmentation maps and minimizes a loss  $\mathcal{L}_{\text{det}}$  as detailed in He et al. (2017). Finally, we fine-tune the tooth localization module with DCR weak supervision, minimizing the joint loss  $\mathcal{L} = \mathcal{L}_{\text{det}} +$

$\mathcal{L}_{\text{DCR}}$ , freezing all network layers except for the last.

### 7.3.3 Training Details

We briefly describe the details of our implementation in this section.

All code implementations are in Tensorflow, run on four NVidia GTX 1080 Ti GPUs. All model training incorporates augmentations including random brightness, contrast, affine transformation, elastic transformation, and Gaussian blurring.

#### Functional Segmentation Module

The U-Net model is trained with cross-entropy loss on the Adam (Kingma and Ba, 2014) optimizer. The learning rate is  $10^{-5}$ , the weight decay is  $10^{-4}$ , and the batch size is 4. Models are train for 12,000 steps.

#### Tooth Localization Module

The Mask-RCNN (He et al., 2017) is trained similarly to the original work with the SGD (stochastic gradient descent) optimizer. The model is first trained with densely annotated masks only. The learning rate is set to  $10^{-3}$ , the weight decay is  $10^{-4}$ , and the batch size is 1. Models are trained until 100,000 steps.

In the fine-tuning reinforcement learning step with DCR, we reduce the learning rate to  $10^{-5}$  and only train the last network layer. No weight decay is applied, and the models are trained for another 150,000 steps. From each image, we draw 64 samples ( $\hat{\mathbf{E}}$ ) from each set of detection output and use the average reward across 64 samples as the baseline reward.

## 7.4 Experiments

In this section, we provide validation of individual modules as well as DeepOPG as a whole. First of all, we present a dataset with novel annotations on segmentation, detection, and finding summary. We then offer an overview of the finding summarization efficacy for each of the finding types. Following this, we provide an ablation

study on the tooth localization module including our proposed DCR decoding and reinforcement learning. Finally, we compare our DeepOPG with existing works under comparable settings.

### 7.4.1 Dataset

In this work, we use the UFBA-UESC (*Universidade Federal da Bahia – Universidade Estadual de Santa Cruz*) Dental Images Deep dataset (Silva et al., 2018) where there are 1,500 OPG images in total, 267 out of which are annotated for tooth localization (implant annotations are not provided in the original data). The OPG images can be split into four major categories: (1) studies with all permanent dentition present and no implants, (2) studies with missing teeth and no implants, (3) studies with implants, and (4) studies with mixed dentition (where primary and permanent teeth both present). We exclude all studies with mixed dentition and supernumerary teeth as they are outside the scope of this work.

To enrich the dataset for learning DeepOPG, we ask 3 board-certified dentists to provide additional annotations including (1) functional segmentation maps on 68 studies, (2) tooth/implant localization maps on 39 studies, (3) tooth/implant missingness summary (weak supervision, in the form of 32 binary labels per study) on 144 studies, and (4) finding summary (in the form of  $32 \times 6$  binary labels per study) on 47 studies.

To avoid overfitting the data, no study is annotated for two or more annotation types. It is important to note that segmentation/localization maps take, on average, 30 minutes to annotate per study, while the teeth missingness information only takes 30 seconds each. In each stage of DeepOPG learning, data is split into 70/30 training/test randomly, and the finding summary is exclusively used as test data.

Table 7.1: **AUC Comparisons.** We compare the AUCROC for six OPG finding types for two settings of DeepOPG. See Section 7.4.4 for method descriptions.

Method	AUCROC (%)						Macro Avg.
	Missing Teeth	Impacted Teeth	w/Crown & Bridge	w/Restoration	Root Filled	Implants	
DeepOPG (full)	<b>90.6</b>	<b>96.9</b>	86.5	<b>89.3</b>	<b>88.2</b>	77.6	<b>88.2</b>
w/o RL	87.6	96.5	<b>88.2</b>	86.4	82.9	<b>78.1</b>	86.6
Area Threshold @ max F1	24.2%	34.5%	25.9%	2.70%	0.33%	–	–

## 7.4.2 Overall Evaluation of DeepOPG for Findings Summarization

As mentioned before, DeepOPG combines the functional segmentation map and the tooth localization results by calculating the percentage area of each functional class for each tooth. Using the percentage area as the predictive value for the binary finding labels, we are able to evaluate the overall performance of DeepOPG by calculating the receiver operating characteristic (ROC) curve where we plot the true positive rate  $TPR = \frac{TP}{TP+FN}$  against the true negative rate  $TNR = \frac{TN}{TN+FP}$ . Note that for a finding prediction to be TP, it not only has to have enough pixels of that finding in the tooth, but the tooth number itself has to be correctly detected.

The ROC curves for the six types of findings are shown in detail in Figure 7-2. We can calculate the area under curve (AUC) for each of the findings as summarized in Table 7.1. In the table, we also compare a setting where the RL with DCR is disabled. It is clear that the weak supervisions with RL can improve the finding summarization. Of the six findings, impacted teeth with an AUC of 96.9% is the easiest task, possibly because it is a large object and that is often found at fixed locations such as the wisdom teeth. We also show, on the last row, the threshold on the percentage area at the operating point with the largest  $F1 = \frac{2 \times TP}{2 \times TP + FN + FP}$ . It is interesting to see that root-filled teeth only require 0.33% of the area to be finding-positive while impacted teeth require 34.5% of the tooth to be labeled impacted.

To highlight the usefulness of weak supervision, the “*w/o RL*” model (86.6% AUC) trains with 273 per-pixel annotations which take 136 expert hours to prepare. The

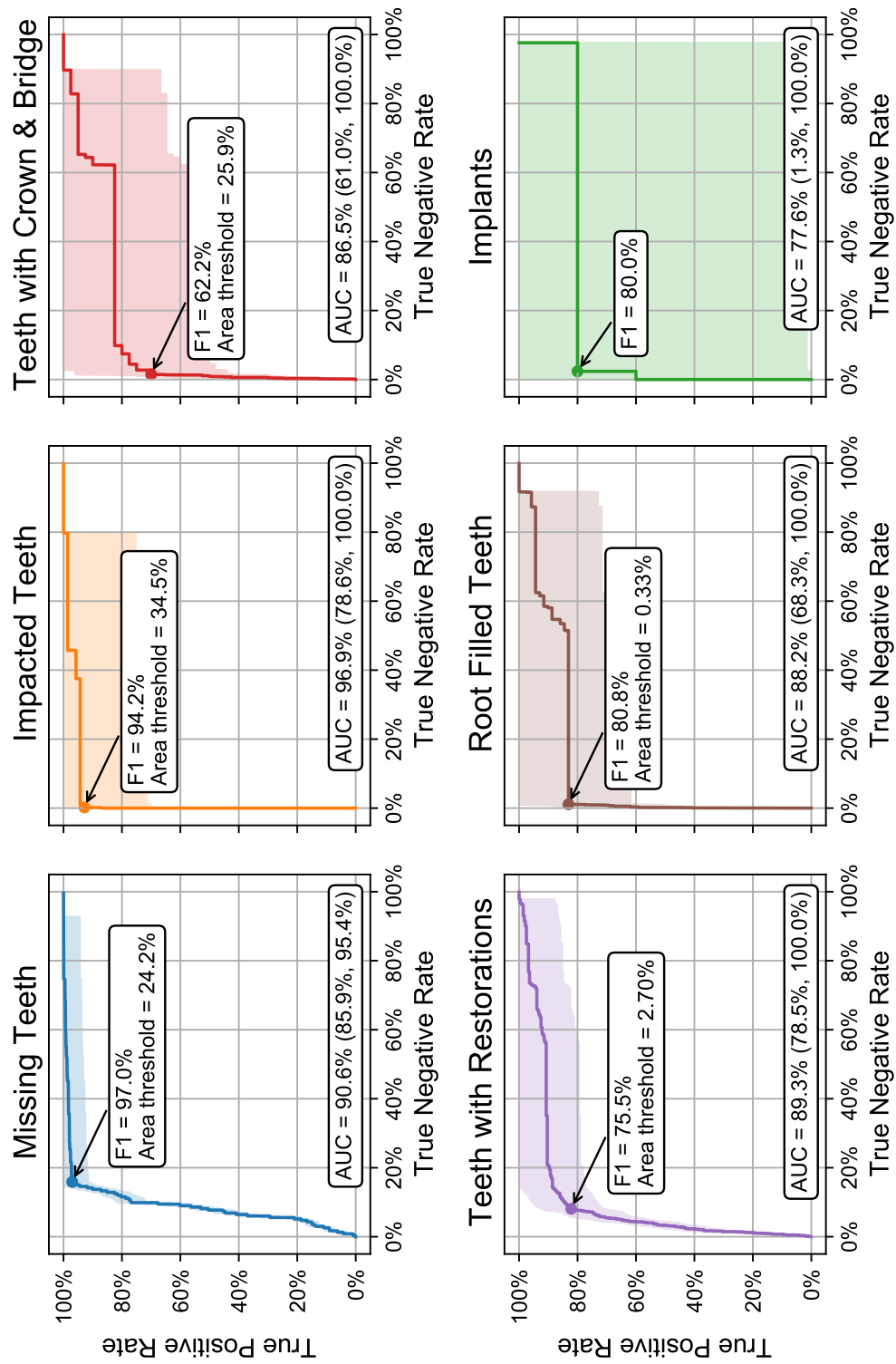


Figure 7-2: Receiver Operating Characteristic (ROC) Curves. Six finding types are considered. 95% confidence interval is annotated in shades and parentheses, and the operating points with the highest F1, along with the thresholds on percentage area, are labeled.

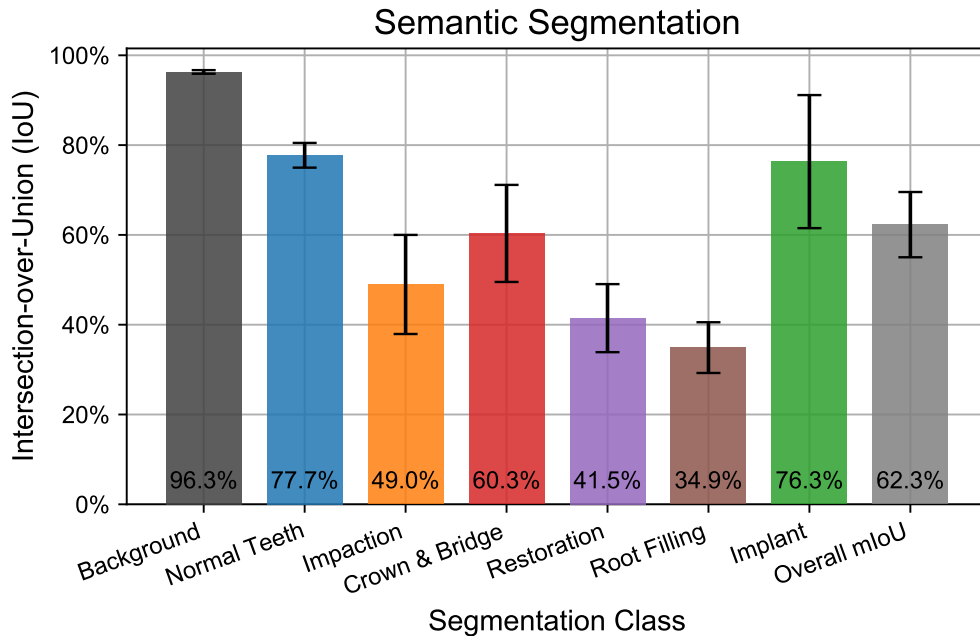


Figure 7-3: **Evaluation of the Functional Segmentation Module.** The IoU between the predicted segmentation and the ground truth segmentation are shown per class. Standard deviation is also labeled.

“*DeepOPG (full)*” model adds 100 weak supervision annotations which only take an additional 0.8 expert hours, but a gain of 1.6% overall AUC. This demonstrates weak supervision is effective in boosting AUC while requiring substantially less expert effort (<1% extra time) than per-pixel annotations.

### 7.4.3 Functional Segmentation

The macro-averaged intersection-over-union (IoU) is evaluated in Figure 7-3 for the seven classes in the functional segmentation module. It is interesting to see the labels with the least positive samples (root filling and restoration) to have the lowest metrics. In terms of performance stability, normal teeth segmentation retains the lowest standard deviation in IoU, as teeth are commonly presented in all studies. As a side note, IoU, by mathematical definition, is always smaller than the Dice coefficient.



Table 7.2: **Comparisons of Detection Metrics.** We show detection metrics for various settings of DeepOPG. We report the metric values and their standard errors.  $AP_x$  denotes  $AP@IoU = x$ . See Section 7.4.4 for method descriptions.

Method	Per-Object				Per-Image
	$AP_{0.0}$ (%)	$AP_{0.5}$ (%)	DA (%)	FA (%)	IoU (%)
DeepOPG (full)	<b>98.6</b> <sub>0.1</sub>	<b>97.6</b> <sub>0.3</sub>	<b>98.7</b> <sub>0.4</sub>	<b>97.5</b> <sub>0.6</sub>	<b>80.5</b> <sub>1.5</sub>
w/o RL	98.4 <sub>0.1</sub>	97.2 <sub>0.4</sub>	<b>98.7</b> <sub>0.4</sub>	<b>97.5</b> <sub>0.6</sub>	80.1 <sub>1.5</sub>
w/o RL and dental coherence	93.0 <sub>0.1</sub>	91.3 <sub>0.4</sub>	93.7 <sub>0.9</sub>	87.4 <sub>1.2</sub>	79.7 <sub>1.6</sub>
w/o segmentation	97.7 <sub>0.1</sub>	96.2 <sub>0.3</sub>	97.9 <sub>0.5</sub>	95.7 <sub>0.8</sub>	80.2 <sub>1.5</sub>

### 7.4.4 Tooth Localization with Dental Coherence

To verify the efficacy of the proposed modifications to the off-the-shelf object detection networks, we perform several ablation studies to inspect the contribution of these modifications. In particular, we assess

1. **DeepOPG (full):** We enable all model features, including feeding segmentation maps as the input for the tooth localization, using dental coherence module at inference, and training the model with reinforcement learning.
2. **w/o RL:** All model features, except training with RL.
3. **w/o RL and dental coherence:** We remove both the dental coherence module and the reinforcement learning components.
4. **w/o segmentation:** Segmentation maps are not fed into the tooth localization module in this case.

#### Metrics

The performance of different models are compared using various metrics, including the commonly used average precision (AP) defined in PASCAL Visual Object Classes (PASCAL VOC) (Everingham et al., 2010) for detection tasks, the detection accuracy  $DA \equiv \frac{TP+FN}{TP+FN+FP}$  and identification accuracy  $FA \equiv \frac{TP}{TP+FN+FP}$  (Cui et al., 2019). On a per-image level, we evaluate the intersection over union as  $IoU \equiv \frac{\sum_n M_n \cap M_n^{gt}}{\sum_n M_n \cup M_n^{gt}}$ .

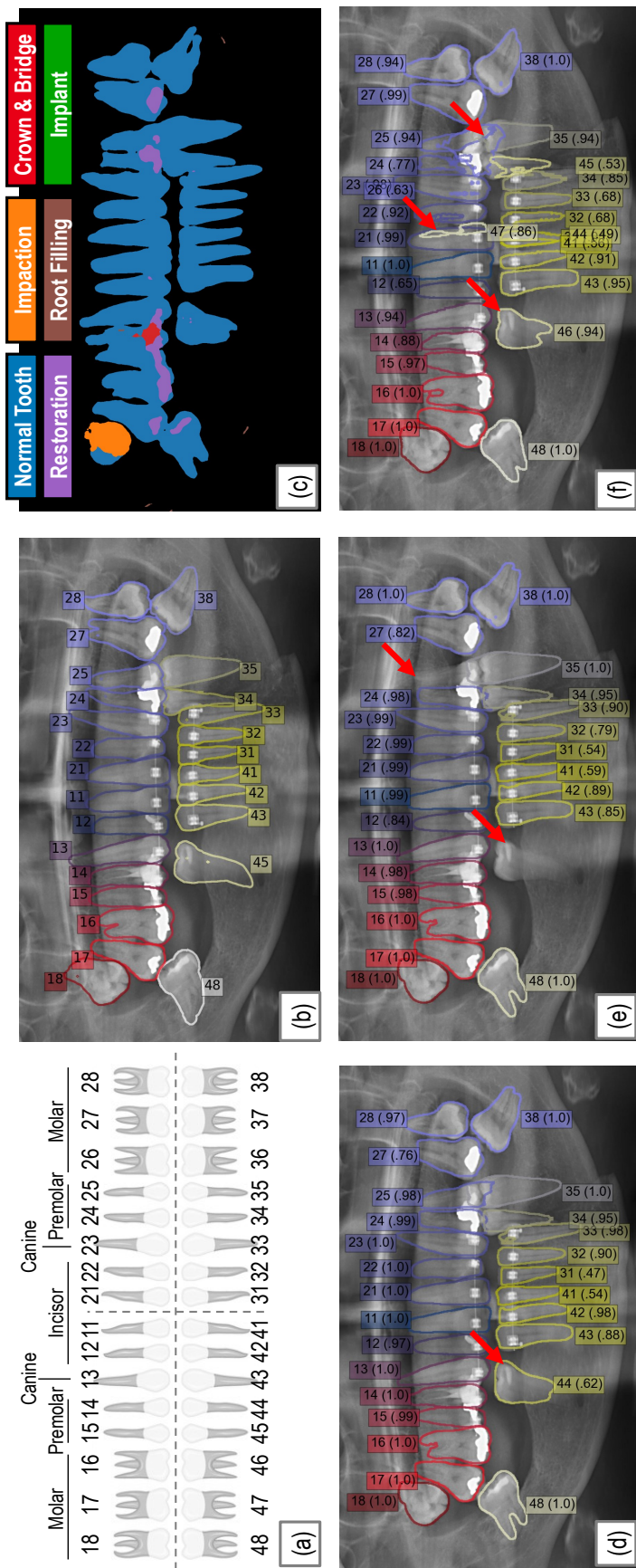


Figure 7-4: Illustrations of Tooth localization Results. (a) The FDI tooth numbering system, (b) input OPG with ground truth localization, (c) functional segmentation map, (d) localization for DeepOPG (full), (e) localization w/o reinforcement learning, and (f) localization w/o segmentation input.

In Table 7.2, we observe consistent gains in performance across all metrics when we incorporate different proposed features. Most notably, the dental coherence module constitutes most of the gains, providing +5.6% in AP@IoU = 0.0 and +5.0% in DA. Using segmentation maps also provides +0.9% gain in AP@IoU = 0.0 since segmentation maps carry more global information by nature. The weak supervision, while seemingly providing less compelling improvements, is, in fact, remarkable as annotating the teeth missingness summary is faster than annotating the localization maps by orders of magnitudes.

Figure 7-4 showcases localization results on a test image with three different configurations. This study contains a maloccluded tooth, on which all three configurations predict incorrectly. It is also worth noting that by removing the segmentation input, the localization depends totally on the input OPG and can be over-sensitive, as indicated by red arrows in Figure 7-4f.

### 7.4.5 Comparing Existing Works

Comparison of model performances across works suffers from not only dataset differences but also clinical task differences. While we are unable to obtain proprietary datasets from previous works for evaluation, we can set up DeepOPG to similar settings to allow fairer comparisons. For example, in Table 7.3, Wirtz et al. (2018) and Jader et al. (2018) tackled teeth-only segmentation, and hence we ignore error resulting from classes other than the teeth and the background in our segmentation module for a fair comparison. Tuzoff et al. (2019) and Kim et al. (2020) addressed detection of natural teeth and implants, and thus we compare only detection results. Across all tasks except for precision in tooth segmentation, we are able to show superior performance.

Finally, for the *missing teeth* finding summary, Kim et al. (2020) reached a sensitivity of 75.5% and a precision of 84.5% at a specificity of 80.4%. Under the same specificity, we have a sensitivity of 94.3% and a precision of 96.4%.

Table 7.3: **Comparison to Prior Works.** We compare our full model DeepOPG under similar conditions to prior works on various tasks. Note the evaluations are done on different dataset in each work. We report the metric values and their standard errors.  $AP_x$  denotes  $AP@IoU = x$ .

Method	Tooth Segmentation		
	Precision (%)	Recall (%)	F1 (%)
Wirtz et al. (2018)	79.0	82.7	80.3
Jader et al. (2018)	<b>94<sub>6</sub></b>	84 <sub>7</sub>	88 <sub>5</sub>
DeepOPG (Ours)	90.7 <sub>2.7</sub>	<b>90.6<sub>2.0</sub></b>	<b>90.6<sub>1.8</sub></b>

Method	Natural Tooth Detection			Implant Detection		
	Sensitivity <sup>1</sup> (%)	Precision <sup>1</sup> (%)	AP <sub>0.5</sub> (%)	AP <sub>0.7</sub> (%)	AP <sub>0.5</sub> (%)	AP <sub>0.7</sub> (%)
Tuzoff et al. (2019)	99.4	99.4	—	—	—	—
Kim et al. (2020)	—	—	96.7	75.4	45.1	26.6
DeepOPG (Ours)	<b>100.0</b>	<b>99.8</b>	<b>97.6<sub>0.3</sub></b>	<b>89.4<sub>1.0</sub></b>	<b>75.0<sub>0.1</sub></b>	<b>75.0<sub>0.1</sub></b>

<sup>1</sup>They considered detection of teeth as 32 one-vs-all sub-problems. Even when a tooth is mis-labelled, it still is correct on 30 problems, and hence the high metrics.

## 7.5 Summary

In this chapter, we provide an initial study, showing the possibilities to summarize findings for individual teeth from an orthopantomogram. By dividing the summarization process into two tasks: semantic segmentation and object detection, we can leverage weaker but faster-to-collect annotations to improve the detection model with reinforcement learning. The experiments demonstrate the efficacy of each module in the DeepOPG system.

In a scenario where conventional annotations are not sufficient, we have shown there may be alternate ways to leverage exotic annotation types. This trait is unique to medical imaging as the perception of disease progression or findings are inherently different from natural visual perception. Moreover, conventional annotations, while suitable and economical to obtain in natural imaging problems, might be infeasible in medical imaging settings. We hope to point the way for future works in this line and encourage both dental imaging research and the use of creative and effective annotations in constrained imaging data.



# Chapter 8

## Federated Learning for Heterogeneous Visual Classification

While all previous chapters discuss learning with limited data within a local data registry or leveraging external public data for better learning, federated learning (FL) (McMahan et al., 2017) is an emerging paradigm for multiple medical institutions to collaboratively learn models better than separately learned models. Essentially, the learning process consists of (a) participating clients receiving the latest model from the server, (b) the clients training the model with locally available data, (c) the clients sending the model gradients back to a server without revealing their data, and finally (d) the server aggregating the gradients to update the model. FL is naturally suitable for medical data due to their inherently private nature, and some works (Sheller et al., 2018; Li et al., 2019b; Sheller et al., 2020; Feki et al., 2021; Zhang et al., 2021b; Dou et al., 2021; Abdul Salam et al., 2021; Lee et al., 2021b) have already explored FL in a medical imaging setting.

---

This chapter is adapted from the published articles “*Measuring the Effects of Non-Identical Data Distribution for Federated Visual Classification*” (Hsu et al., 2019) and “*Federated Visual Classification with Real-World Data Distribution*” (Hsu et al., 2020) to both of which I have contributed as the first author.

While FL appears to be a great instrument when single institutions hold limited data, it remains a question of how various levels of data heterogeneity affect the training process. To fully understand the impact of data heterogeneity, we started with natural image classification as a proxy and curated one smaller (Krizhevsky et al., 2009) and two large-scale real-world datasets (Van Horn et al., 2018; Weyand et al., 2020b) specifically for FL. We benchmarked classic FL algorithms under a spectrum of data heterogeneity, and proposed improvements to resolve commonly observed FL challenges. The comparisons are made over a range of parameter grids to offer an in-depth insight, focusing on not only the final classification accuracy but also the convergence speed.

It is imperative that we derive sufficient insight from real-world datasets before proceeding to apply FL to many other medical applications, as more often than not, we are blind to the global distribution of the overall data from all federated parties.

## 8.1 Overview

Federated learning enables visual models to be trained on-device, bringing advantages for user privacy (data need never leave the device), but challenges in terms of data diversity and quality. Whilst typical models in the datacenter are trained using data that are independent and identically distributed (IID) with *distributed learning*, data at source are typically far from IID. Furthermore, differing quantities of data are typically available at each device (*imbalance*). In this study, we characterize the effect these real-world data distributions have on distributed learning, using as a benchmark the standard Federated Averaging (FedAvg) algorithm. To do so, we introduce two new large-scale datasets for species and landmark classification, with realistic per-user data splits that simulate real-world edge learning scenarios. We also develop two new algorithms (FedVC, FedIR) that intelligently resample and reweight over the client pool, bringing large improvements in accuracy and stability in training. The datasets are made available online.



## 8.2 Background

Federated learning (FL) is a privacy-preserving framework, introduced both with conventional learning algorithms (Li et al., 2016b; Lu et al., 2015; Meeker et al., 2015; Ohno-Machado et al., 2014) and deep learning McMahan et al. (2017), for training models from decentralized user data residing on devices at the edge. Models are trained iteratively across many federated rounds. For each round, every participating device (“*client*”), receives an initial model from a central server, performs stochastic gradient descent (SGD) on its local training data and sends back the gradients. The server then aggregates all gradients from the participating clients and updates the starting model. FL preserves user privacy in that the raw data used for training models never leave the devices throughout the process. In addition, differential privacy (McMahan et al., 2018) can be applied for a theoretically bounded guarantee that no information about individuals can be derived from the aggregated values on the central server.

Federated learning is an active area of research with a number of open questions (Li et al., 2019a; Kairouz et al., 2019) remaining to be answered. A particular challenge is the distribution of data at user devices. Whilst in centralized training, data can be assumed to be independent and identically distributed (IID), this assumption is unlikely to hold in federated settings. Decentralized training data on end-user devices will vary due to user-specific habits, preferences, geographic locations, etc. Furthermore, in contrast to the streamed batches from a central data store in the data center, devices participating in an FL round will have differing amounts of data available for training.

In this work, we study the effect these heterogeneous client data distributions have on learning visual models in a federated setting, and propose novel techniques for more effective and efficient federated learning. We focus in particular on two types of distribution shift: *Non-Identical Class Distribution*, meaning that the distribution of visual classes at each device is different, and *Imbalanced Client Sizes*, meaning that the number of data available for training at each device varies. Our key contributions

are:

- *We analyze the effect of learning with per-user data* in real-world datasets, in addition to carefully controlled setups with parametric (Dirichlet) and natural (geographic) distributions.
- *We propose two new algorithms* to mitigate per-client distribution shift and imbalance, substantially improving classification accuracy and stability.
- *We provide new large-scale datasets* with per-user data for two classification problems (natural world and landmark recognition) to the community.

This study is the first to our knowledge that attempts to train large-scale visual classification models for real-world problems in a federated setting. We expect that more is to be done to achieve robust performance in this and related settings, and are making our datasets available to the community to enable future research in this area.

## 8.3 Related Work

### 8.3.1 Synthetic Client Data

Several authors have explored the FedAvg algorithm on synthetic non-identical client data partitions generated from image classification datasets. McMahan et al. (2017) synthesize pathological non-identical user splits from the MNIST (*Modified National Institute of Standards and Technology*) dataset, sorting training examples by class labels and partitioning into shards such that each client is assigned 2 shards. They demonstrate that FedAvg on non-identical clients still converges to 99% accuracy, though taking more rounds than identically distributed clients. In a similar sort-and-partition manner, Zhao et al. (2018) and Sattler et al. (2019) use extreme partitions of the CIFAR-10 dataset to form a population consisting of 10 clients in total. In contrast to these pathological data splits, Yurochkin et al. (2019) and Hsu et al. (2019) synthesize more diverse non-identical datasets with Dirichlet priors.

### 8.3.2 Realistic Datasets

Other authors look at more realistic data distributions at the client. For example, [Caldas et al. \(2018\)](#) use the Extended MNIST dataset ([Cohen et al., 2017](#)) split over the writers of the digits and the CelebA dataset ([Liu et al., 2015](#)) split by the celebrity on the picture. The Shakespeare and Stack Overflow datasets ([Google, 2019b](#)) contain natural per-user splits of textual data using roles and online user ids, respectively. [Luo et al. \(2019\)](#) propose a dataset containing 900 images from 26 street-level cameras, which they use to train object detectors. These datasets are however limited in size, and are not representative of data captured on user devices in a federated learning context. Our work aims to address these limitations (see Section 8.5).

Variance reduction methods have been used in the federated learning literature to correct for the distribution shift caused by heterogeneous client data. [Sahu et al. \(2018\)](#) introduce a proximal term to client objectives for bounded variance. [Karimireddy et al. \(2019\)](#) propose to use control variates for correcting client gradient update drift in different communication rounds. Importance sampling is a classic technique for variance reduction in Monte Carlo methods ([Kahn and Marshall, 1953](#); [Hesterberg, 1995](#)) and has been used widely in domain adaption literature for countering covariate and target shift ([Saerens et al., 2002](#); [Zhang et al., 2013](#); [Ngiam et al., 2018](#)). In this work, we adopt a similar idea of importance reweighting in a novel federated setting resulting in augmented client objectives. Different from the classic setting where samples are drawn from one proposal distribution which has the same support as the target, heterogeneous federated clients form multiple proposal distributions, each of which has partially common support with the target.

## 8.4 Federated Visual Classification Problems

Many problems in visual classification involve data that vary around the globe ([Dorsch et al., 2015](#); [Hays and Efros, 2008](#)). This means that the distribution of data visible to a given user device will vary, sometimes substantially. For example, user observations in the citizen scientist app iNaturalist will depend on the underlying

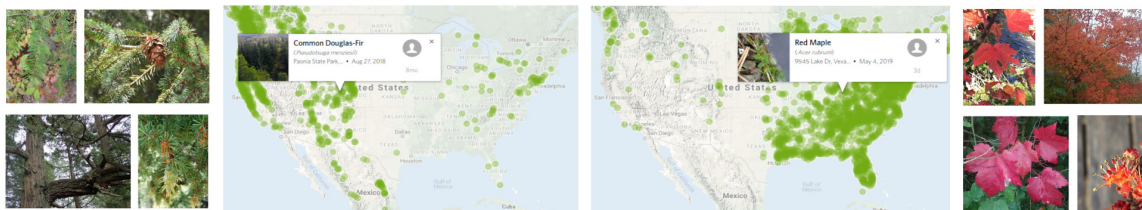


Figure 8-1: **iNaturalist Species Distribution.** Visualized here are the distributions of Douglas-Fir and Red Maple in the continental US within iNaturalist. In a federated learning context, visual categories vary with location, and users in different locations will have very different training data distributions.

species distribution in that region (see Figure 8-1). Many other factors could potentially influence the data present on a device, including the interests of the user, their photography habits, etc. For this study we choose two problems with an underlying geographic variation to illustrate the general problem of non-identical user data, *Natural Species Classification* and *Landmark Recognition*.

### 8.4.1 Natural Species Classification

We create a dataset and classification problem based on the iNaturalist 2017 Challenge (Van Horn et al., 2018), where images are contributed by a community of citizen scientists around the globe. Domain experts take pictures of natural species and provide annotations during field trips. Fine-grained visual classifiers could potentially be trained in a federated fashion with this community of citizen scientists without transferring images.

### 8.4.2 Landmark Recognition

We study the problem of visual landmark recognition based on the 2019 Landmark Recognition Challenge (Weyand et al., 2020a), where the images are taken and uploaded by Wikipedia contributors. It resembles a scenario where smartphone users take photos of natural and architectural landmarks (*e.g.*, famous buildings, monuments, mountains, and etc.) while traveling. Landmark recognition models could potentially be trained via federated learning without uploading or storing private

user photos at a centralized party.

Both datasets have data partitioning per user, enabling us to study a realistic federated learning scenario where labeled images were provided by the user and learning proceeds on-device. For experimentation in lab, we use a simulation engine for federated learning algorithms, similar to TensorFlow Federated (Google, 2019a).

## 8.5 Datasets

In the following section, we describe in detail the datasets we develop and analyze key distributional statistics as a function of user and geo-location. We have made these datasets available to the community<sup>1</sup>.

### 8.5.1 iNaturalist-User-120k and iNaturalist-Geo Splits

iNaturalist-2017 (Van Horn et al., 2018) is a large scale fine-grained visual classification dataset comprised of images of natural species taken by citizen scientists. It has 579,184 training examples and 95,986 test examples covering over 5,000 classes. Images in this dataset are each associated with a fine-grained species label, a longitude-latitude coordinate where the picture was originally taken, and authorship information.

The iNaturalist-2017 training set has a very long-tailed distribution over classes as shown in Figure 8-2a, while the test set is relatively uniform over classes. While studying learning robustly with differing training and test distributions is a topic for research (Van Horn and Perona, 2017) in itself, in our federated learning benchmark, we create class-balanced training and test sets with uniform distributions. This allows us to focus on distribution variations and imbalance at the *client level*, without correcting for overall domain shift between training and test sets.

To equalize the number of examples across classes, we first sort all class labels by their count and truncate tail classes with less than 100 training examples. This

---

<sup>1</sup>[https://github.com/google-research/google-research/tree/master/federated\\_vision\\_datasets](https://github.com/google-research/google-research/tree/master/federated_vision_datasets)

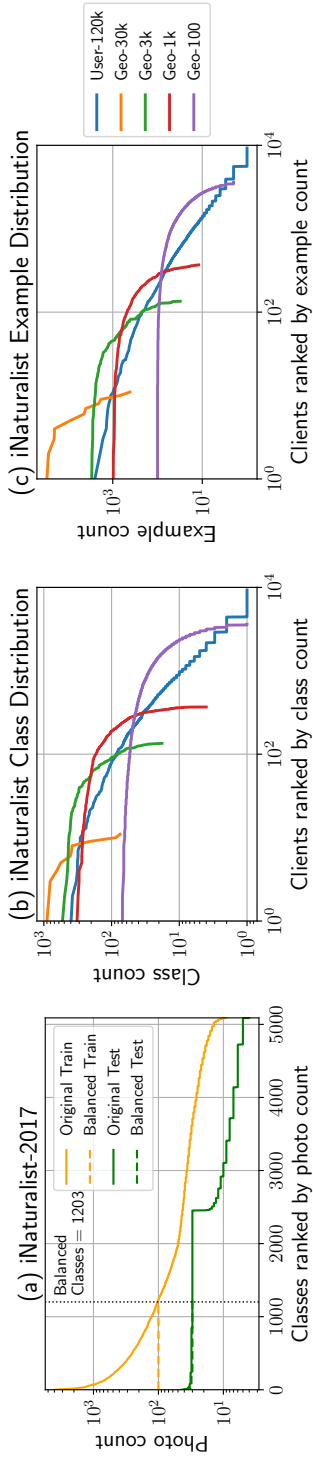


Figure 8-2: **iNaturalist Distribution**. In (a) we show the re-balancing of the original iNaturalist-2017 dataset. In (b) and (c) we show class and example counts vs clients for our 5 iNaturalist partitionings with varying levels of class distribution shift and size imbalance. The client count is different in each partitioning.

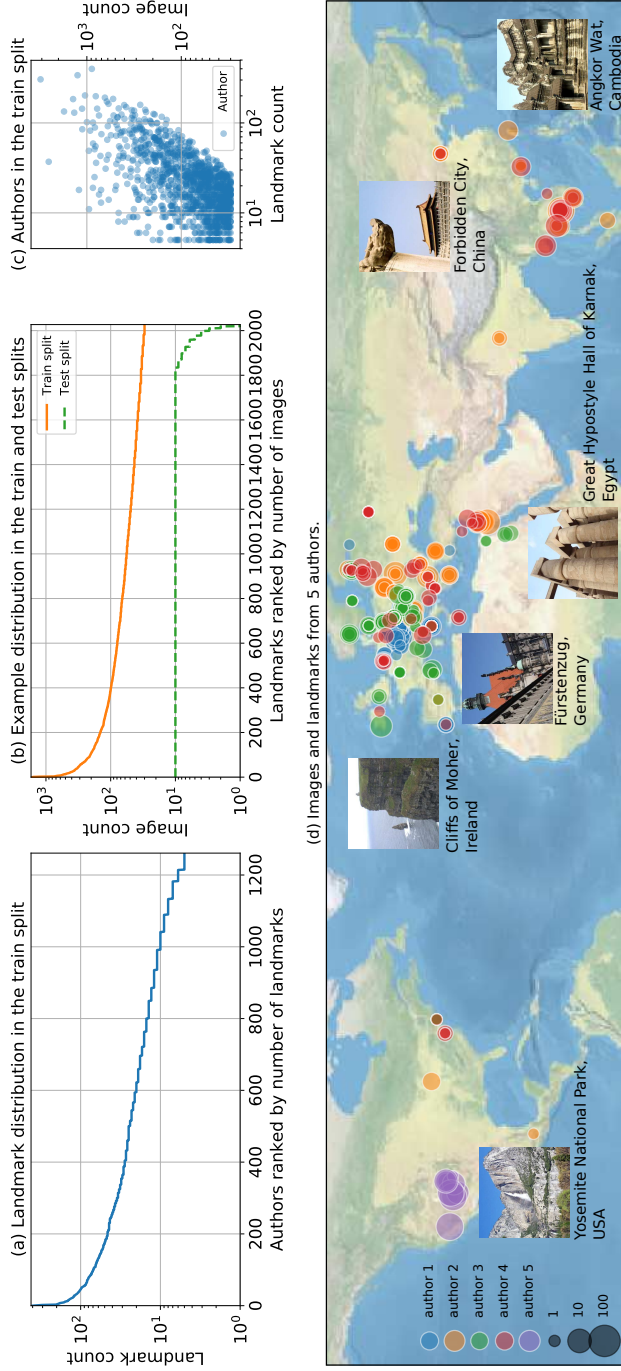


Figure 8-3: **Landmarks-User-160k Distribution**. Images are partitioned according to the authorship attribute from the GLD-v2 dataset. Filtering is applied to mitigate long tail in the train split.

is then followed by subsampling per-class until all remaining classes each have 100 examples. This results in a balanced training set consisting of 1,203 classes and 120,300 examples. We use this class-balanced iNaturalist subset for the remainder of the study.

The iNaturalist-2017 dataset includes user ids, which we use to partition the balanced training set into a population of 9,275 clients. We refer to this partitioning as *iNaturalist-User-120k*. This contributor partitioning resembles a realistic scenario where images are collected per-user.

In addition, to study the federated learning algorithms with client populations of varying levels of deviation from the global distribution, we also generate a *wide range* of populations by partitioning the dataset at varying levels of granularity according to the geographic locations.

To utilize the geo-location tags, we leverage the S2 grid system, which defines a hierarchical partitioning of the planet surface. We perform an adaptive partitioning similar to (Weyand et al., 2016). Specifically, every S2 cell is recursively subdivided into four finer-level cells until no single cell contains more than  $N_{\max}$  examples. Cells ending up with less than  $N_{\min}$  examples are discarded. With this scheme, we are able to control the granularity of the resulting S2 cells such that a smaller  $N_{\max}$  results in a larger client count. We use  $N_{\max} \in \{30k, 3k, 1k, 100\}$ ,  $N_{\min} = 0.01N_{\max}$  and refer to the resulting data partitionings as *iNaturalist-Geo- $\{30k, 3k, 1k, 100\}$* , respectively. Rank statistics of our geo- and per-user data splits are shown in Figures 8-2b and 8-2c.

### 8.5.2 Landmarks-User-160k

Google Landmarks Dataset V2 (GLD-v2) (Weyand et al., 2020a) is a large scale image dataset for landmark recognition and retrieval, consisting of 5 million images with each attributed to one of over 280,000 authors. The full dataset is noisy: images with the same label could depict landmark exteriors, historical artifacts, paintings or sculptures inside a building. For benchmarking federated learning algorithms on a well-defined image classification problem, we use the cleaned subset (GLD-v2-clean), which is a half the size of the full dataset. In this set, images are discarded if their

computed local geometric features cannot be matched to at least two other images with the same label (Ozaki and Yokoo, 2019).

For creating a dataset for federated learning with natural user identities, we partition the GLD-v2-clean subset according to the authorship attribute. In addition, we mitigate the long tail while maintaining realism by requiring every landmark to have at least 30 images and be visited by at least 10 users, meanwhile requiring every user to have contributed at least 30 images that depict 5 or more landmarks. The resulting dataset has 164,172 images of 2,028 landmarks from 1,262 users, which we refer to as the train split of *Landmarks-User-160k*.

Following the dataset curation in Weyand et al. (2020a), the test split is created from the leftover images in GLD-v2-clean whose authors do not overlap with those in the train split. The test split contains 19,526 images and is well-balanced among classes. 1,835 of the landmarks have exactly 10 test images, and there is a short tail for the rest of the landmarks due to insufficient samples (Figure 8-3).

### 8.5.3 CIFAR-10/100

#### Synthetic Clients with Dirichlet Prior

To generate non-identical client datasets from CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009) datasets, we partition each into 100 clients, with 500 training examples each. We assume every client  $k$  has their data independently drawn from the original dataset according to a multinomial distribution  $q_k(\cdot)$  of  $C$  classes ( $q_k(y) \geq 0$  and  $\sum_y q_k(y) = 1$ ).

To synthesize a population of non-identical clients, we draw a multinomial  $\mathbf{q}_k \sim \text{Dir}(\alpha\mathbf{p})$  from a Dirichlet distribution, where  $\mathbf{p}$  describes a prior class distribution over  $C$  classes, and  $\alpha > 0$  is a parameter controlling the *concentration*, or identicalness among all clients.  $\alpha$  can be used to control the overall homogeneity:  $\alpha \rightarrow \infty$  generates clients that are all identical to the prior  $\mathbf{p}$ , while  $\alpha \rightarrow 0$  generates clients that tend to hold very sparse labels. After drawing the class distributions  $\mathbf{q}_k$ , for every client  $k$ , we sample training examples from CIFAR-10/100 for each class according to  $\mathbf{q}_k$



*without replacement.* This is to ensure there are no overlapping examples between any two clients.

Note that by drawing examples without replacement, towards the end of the assignment process, some subset  $\mathcal{S}$  of classes can be exhausted earlier than other classes, ending up with a shorter list of available classes from which the client synthesis procedure can continue drawing samples. When this happens, we eliminate  $\mathcal{S}$  and enforce the remaining clients to only sample from classes  $\{1, 2, \dots, C\} \setminus \mathcal{S}$  with a multinomial distribution

$$\tilde{q}_k(y) = \begin{cases} 0, & y \in \mathcal{S} \\ q_k(y) / (1 - \sum_{s \in \mathcal{S}} q_k(s)), & y \notin \mathcal{S}. \end{cases} \quad (8.1)$$

For CIFAR-10, we use  $\alpha \in \{100, 10, 1, 0.5, 0.2, 0.1, 0.05, 0\}$ ; for CIFAR-100 we use  $\alpha \in \{1000, 100, 10, 5, 2, 1, 0.5, 0\}$ . Figure 8-4 illustrates populations drawn from the Dirichlet distribution with different concentration parameters. Summary statistics showing the class count over the client population in both datasets is given in Figure 8-5.

## 8.6 Methods

The datasets described above contain significant distribution variations among clients, which presents considerable challenges for efficient federated learning (Li et al., 2019a; Kairouz et al., 2019). In the following, we describe our baseline approach of Federated Averaging algorithm (FedAvg) (Section 8.6.1) and two new algorithms intended to specifically address the non-identical class distributions and imbalanced client sizes present in the data (Sections 8.6.2 and 8.6.3 respectively).

### 8.6.1 Federated Averaging and Server Momentum

A standard algorithm (McMahan et al., 2017) for FL, and the baseline approach used in this work, is Federated Averaging (FedAvg). See Algorithm 1. For every federated

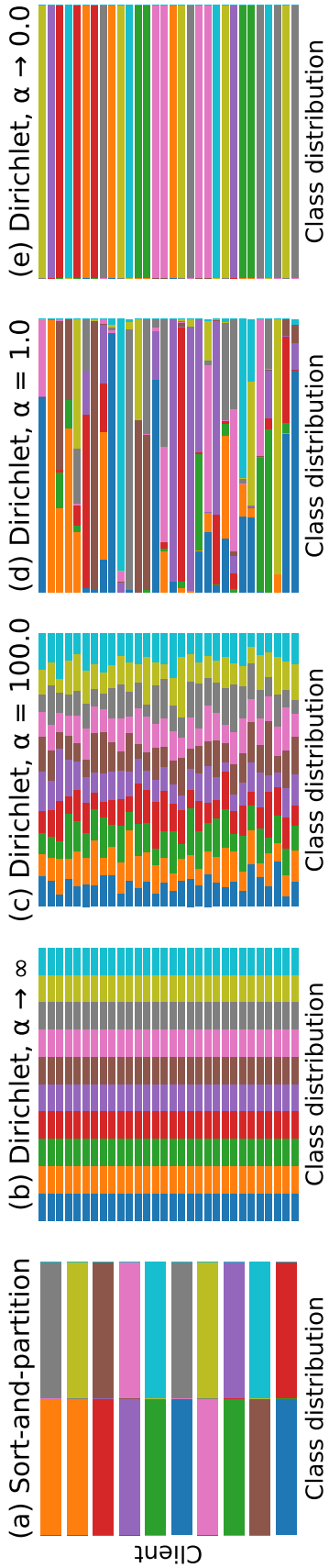


Figure 8-4: **Synthetic populations with non-identical clients.** Distribution among classes is represented with different colors, each standing for a class. (a) 10 clients generated from the sort-and-partition scheme, each assigned with 2 classes. (b-e) populations generated from Dirichlet distribution with different concentration parameters  $\alpha$  respectively, 30 random clients each.

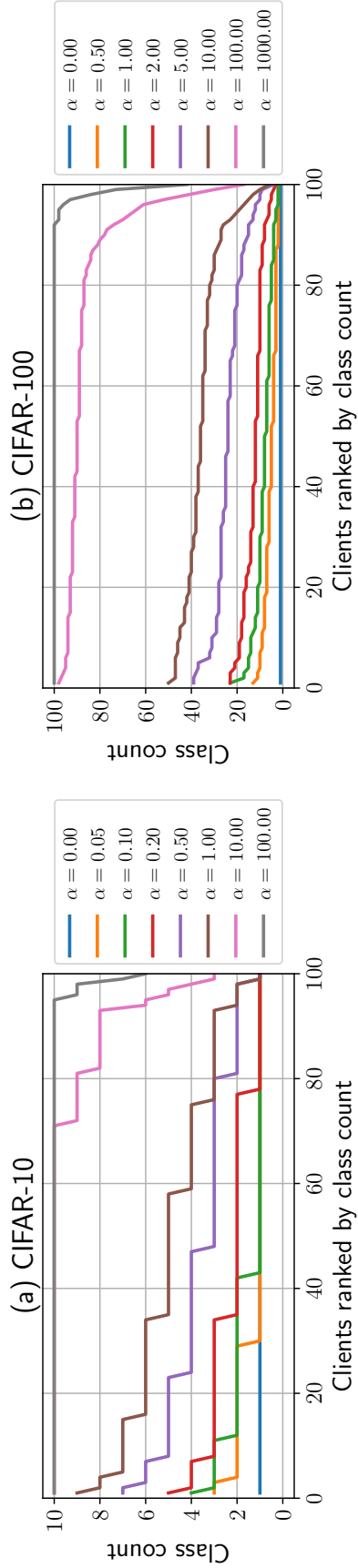


Figure 8-5: **CIFAR-10/100 Distribution.** Each curve represents the class counts of clients within a data partitioning synthesized using a Dirichlet concentration parameter  $\alpha$ .

---

**Algorithm 1:** A high-level overview of federated learning algorithms: FedAvg, FedAvgM, FedIR, and FedVC. Note that they share a similar structure except otherwise annotated.

---

**Server training loop:** ;  
 Initialize  $\theta_0$  ;  
**for** each round  $t = 0, 1, \dots$  **do**  
 | Subset of  $K$  clients  $\leftarrow$  SelectClients( $K$ ) ;  
 | **for** each client  $k = 1, 2, \dots, K$  **do in parallel**  
 | |  $\Delta\theta_t^k \leftarrow$  ClientUpdate( $k, \theta_t$ ) ;  
 | **end**  
 |  $\bar{g}_t \leftarrow$  AggregateClient( $\{\Delta\theta_t^k\}_{k=1}^K$ ) ;  
 |  $\theta_{t+1} \leftarrow \theta_t - \gamma\bar{g}_t$  ; ▷  $\theta_{t+1} \leftarrow \theta_t - \gamma v_t$ , where  $v_t \leftarrow \beta v_{t-1} + \bar{g}_t$   
**end**

SelectClients( $K$ ):  
 | **return**  $K$  clients sampled uniformly ; ▷ with probability  $\propto n_i$  for client  $i$

ClientUpdate( $k, \theta_t$ ):  
 |  $\theta \leftarrow \theta_t$  ;  
 | **for** each local mini-batch  $b$  over  $E$  epochs **do** ▷ over  $S$  steps  
 | |  $\theta \leftarrow \theta - \eta \nabla L(b; \theta)$  ; ▷  $\nabla \tilde{L}(b; \theta)$  in Eq.8.4  
 | **end**  
 | **return**  $\Delta\theta \leftarrow \theta_t - \theta$  to server

AggregateClient( $\{\Delta\theta_t^k\}_{k=1}^K$ ):  
 | **return**  $\sum_{k=1}^K \frac{n_k}{n} \Delta\theta_t^k$ , where  $n = \sum_{k=1}^K n_k$  ; ▷  $\frac{1}{K} \sum_{k=1}^K \Delta\theta_t^k$

---

round,  $K$  clients (the *report goal*) are randomly selected with uniform probability from a pool of active clients. Selected clients, indexed by  $k$ , download the same starting model  $\theta_t$  from a central server and perform local SGD optimization, minimizing an empirical loss  $L(b)$  over local mini-batches  $b$  with learning rate  $\eta$ , for  $E$  epochs before sending the accumulated model update  $\Delta\theta_t^k$  back to the server. The server then averages the updates from the reporting clients  $\bar{g}_t = \sum_{k=1}^K \frac{n_k}{n} \Delta\theta_t^k$  with weights proportional to the sizes of clients' local data and finishes the federated round by applying aggregated updates to the starting model  $\theta_{t+1} \leftarrow \theta_t - \gamma\bar{g}_t$ , where  $\gamma$  is the server learning rate. Given this framework, alternative optimizers can be applied. FedAvgM (Hsu et al., 2019) has been shown to improve robustness to non-identically distributed client data. It uses a momentum optimizer on the server with the update rule  $\theta_{t+1} \leftarrow \theta_t - \gamma v_t$ , where  $v_t \leftarrow \beta v_{t-1} + \bar{g}_t$  is the exponentially weighted moving

average of the model updates with powers of  $\beta$ .

## 8.6.2 Importance Reweighted Client Objectives

Now we address the non-identical class distribution shift in federated clients. Importance reweighting is commonly used for learning from datasets distributed differently from a target distribution. Given that the distribution variation among clients is an inherent characteristic of FL, we propose the following scheme.

Consider a target distribution  $p(x, y)$  of images  $x$  and class labels  $y$  on which a model is supposed to perform well (*e.g.*, a validation dataset known to the central server), and a predefined loss function  $\ell(x, y)$ . The objective of learning is to minimize the expected loss  $\mathbb{E}_p[\ell(x, y)]$  with respect to the target distribution  $p$ . SGD in the centralized setting achieves this by minimizing an empirical loss on mini-batches of IID training examples from the same distribution, which are absent in the federated setting. Instead, training examples on a federated client  $k$  are sampled from a client-specific distribution  $q_k(x, y)$ . This implies that the empirical loss being optimized on every client is a *biased* estimator of the loss with respect to the target distribution, since  $\mathbb{E}_{q_k}[\ell(x, y)] \neq \mathbb{E}_p[\ell(x, y)]$ .

We propose an importance reweighting scheme, denoted FedIR, that applies importance weights  $w_k(x, y)$  to every client’s local objective as  $\tilde{\ell}(x, y) = \ell(x, y)w_k(x, y)$ , where  $w_k(x, y) = \frac{p(x, y)}{q_k(x, y)}$ . With the importance weights in place, an unbiased estimator of loss with respect to the target distribution can be obtained using training examples from the client distribution

$$\mathbb{E}_p[\ell(x, y)] = \sum_{x, y} \frac{\ell(x, y)p(x, y)}{q_k(x, y)} q_k(x, y) = \mathbb{E}_{q_k} \left[ \ell(x, y) \frac{p(x, y)}{q_k(x, y)} \right]. \quad (8.2)$$

Assuming that all clients share the same conditional distribution of images given a class label as the target, *i.e.*,  $p(x|y) \approx q_k(x|y) \forall k$ , the importance weights can be computed on every client directly from the class probability ratio

$$w_k(x, y) = \frac{p(x, y)}{q_k(x, y)} = \frac{p(y)p(x|y)}{q_k(y)q_k(x|y)} \approx \frac{p(y)}{q_k(y)}. \quad (8.3)$$

Note that this computation does not sabotage the privacy-preserving property of federated learning. The denominator  $q_k(y)$  is private information available locally at and never leaves client  $k$ , whereas the numerator  $p(y)$  does not contain private information about clients and can be transmitted from the central server with minimal communication cost:  $C$  scalars in total for  $C$  classes.

Since scaling the loss also changes the effective learning rate in the SGD optimization, in practice, we use self-normalized weights when computing loss over a mini-batch  $b$  as

$$\tilde{L}(b) = \frac{\sum_{(x,y) \in b} \ell(x, y) w_k(x, y)}{\sum_{(x,y) \in b} w_k(x, y)}. \quad (8.4)$$

This corresponds to the self-normalized importance sampling in the statistics literature (Hesterberg, 1995). FedIR does not change server optimization loops and can be applied together with other methods, such as FedAvgM. See Algorithm 1.

### 8.6.3 Splitting Imbalanced Clients with Virtual Clients

The number of training examples in users' devices vary in the real world. Imbalanced clients can cause challenges for both optimization and engineering practice. Previous empirical studies (McMahan et al., 2017; Hsu et al., 2019) suggest that the number of local epochs  $E$  at every client has crucial effects on the convergence of FedAvg. A larger  $E$  implies more optimization steps towards local objectives being taken, which leads to slow convergence or divergence due to increased variance. Imbalanced clients suffer from this optimization challenge even when  $E$  is small. Specifically, a client with a large number of training examples takes significantly more local optimization steps than another with fewer training examples. This difference in steps is proportional to the difference in the number of training examples. In addition, a client with an overly large training dataset will take a long time to compute updates, creating a bottleneck in the federated learning round. Such clients would be abandoned by a FL production system in practice, if failing to report back to the central server within a certain time window (Bonawitz et al., 2019).

We hence propose a new *Virtual Client* (FedVC) scheme to overcome both issues.

The idea is to conceptually split large clients into multiple smaller ones, and repeat small clients multiple times such that all *virtual* clients are of similar sizes. To realize this, we fix the number of training examples used for a federated learning round to be  $N_{\text{VC}}$  for every client, resulting in exactly  $S = N_{\text{VC}}/B$  optimization steps taken at every client given a mini-batch size  $B$ . Concretely, consider a client  $k$  with a local dataset  $\mathcal{D}_k$  with size  $n_k = |\mathcal{D}_k|$ . A random subset consisting of  $N_{\text{VC}}$  examples is uniformly resampled from  $\mathcal{D}_k$  for every round the client is selected. This resampling is conducted without replacement when  $n_k \geq N_{\text{VC}}$ ; with replacement otherwise. In addition, to avoid underutilizing training examples from large clients, the probability that any client is selected for a round is set to be proportional to the client size  $n_k$ , in contrast to uniform as in FedAvg. Key changes are outlined in Algorithm 1. It is clear that FedVC is equivalent to FedAvg when all clients are of the same size.

### 8.6.4 Implementation Details

We use MobileNetV2 (Sandler et al., 2018) pre-trained on ImageNet (Deng et al., 2009) for both iNaturalist and Landmarks experiments; for the latter, a 64-dimensional bottleneck layer between the 1280-dimensional features and the softmax classifier. We replaced BatchNorm with GroupNorm (Wu and He, 2018) due to its superior stability for FL tasks (Hsieh et al., 2019). During training, the image is randomly cropped then resized to a target input size of  $299 \times 299$  (iNaturalist) or  $224 \times 224$  (Landmarks) with scale and aspect ratio augmentation similar to Szegedy et al. (2015). A weight decay of  $4 \times 10^{-5}$  is applied. For CIFAR-10/100 experiments, we use a CNN similar to LeNet-5 (LeCun et al., 1998) which has two  $5 \times 5$ , 64-channel convolution layers, each precedes a  $2 \times 2$  max-pooling layer, followed by two fully-connected layers with 384 and 192 channels respectively and finally a softmax linear classifier. This model is not the state-of-the-art on the CIFAR datasets, but is sufficient to show the relative performance for our investigation. Weight decay is set to  $4 \times 10^{-4}$ . Unless otherwise stated, we use a learning rate of 0.01 and momentum of 0.9 in FedAvgM, kept constant without decay for simplicity. The client batch size is 32 in Landmarks and 64 for others.

Table 8.1: **Training Dataset Statistics.** Note that while CIFAR-10/100 and iNaturalist datasets each have different partitionings with different levels of identicalness, the underlying data pool is unchanged and thus sharing the same centralized learning baselines.

Dataset	Clients		Classes	Examples	Centralized Accuracy
	Count	Size Imbalance	Count	Count	
Synthetic					
CIFAR-10	100	✗	10	50,000	86.16%
CIFAR-100	100	✗	100	50,000	55.21%
iNaturalist Geo Splits	11 to 3606	✓	1,203	120,300	57.90%
Real-World					
iNaturalist-User-120k	9,275	✓	1,203	120,300	57.90%
Landmarks-User-160k	1,262	✓	2,028	164,172	67.05%

## 8.7 Experiments

We now present an empirical study using the datasets and methods of Sections 8.5 and 8.6. We start by analyzing the classification performance as a function of non-identical data distribution (Section 8.7.1), using the CIFAR10/100 datasets. Next we show how *Importance Reweighting* can improve performance in the more non-identical cases (Section 8.7.2). With real user data, where clients are also imbalanced, we show how this can be mitigated with *Federated Virtual Clients* in Section 8.6.3. Finally we present a set of benchmark results with the per-user splits of iNaturalist and Landmark datasets (Section 8.7.4). A summary of the datasets used is provided in Table 8.1. Implementation details are deferred to Section 8.6.4.

### Metrics

When using the same dataset, the performance of a model trained with federated learning algorithms is inherently upper bounded by that of a model trained in the centralized fashion. We evaluate the *relative accuracy*, defined as  $\text{Acc}_{\text{federated}}/\text{Acc}_{\text{centralized}}$ , and compare this metric under different types of budgets. The centralized training baseline uses the same configurations and hyperparameters for a fair comparison.

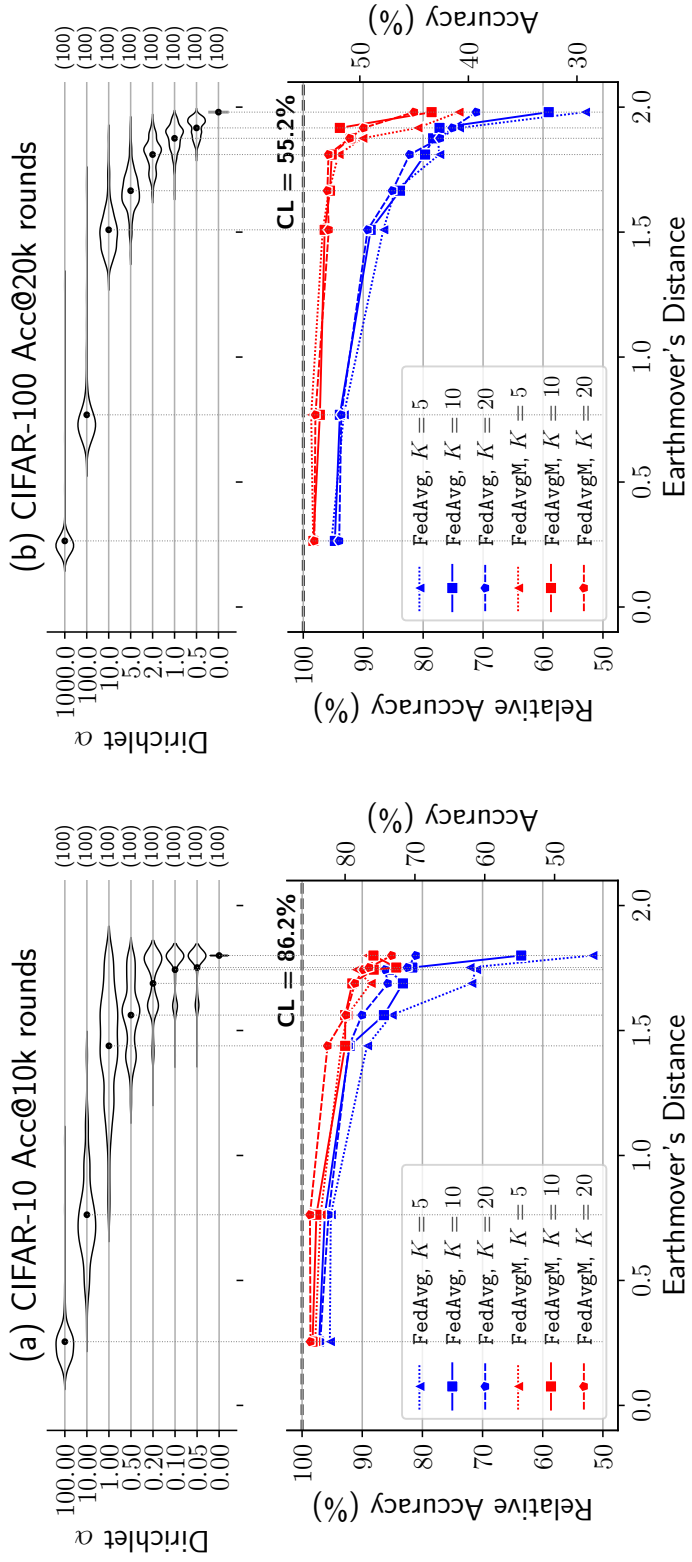


Figure 8-6: **Relative Accuracy vs. Non-identicalness.** Federated learning experiments are performed on (a) CIFAR-10 and (b) CIFAR-100 using local epoch  $E = 1$ . The top row demonstrates the distributions of EMD of clients with different data partitionings. Total client counts are annotated to the right, and the weighted average of all client EMD is marked. Data is increasingly non-identical to the right. The dashed line indicates the centralized learning performance. The best accuracies over a grid of hyperparameters are reported (see Section 8.7.5).



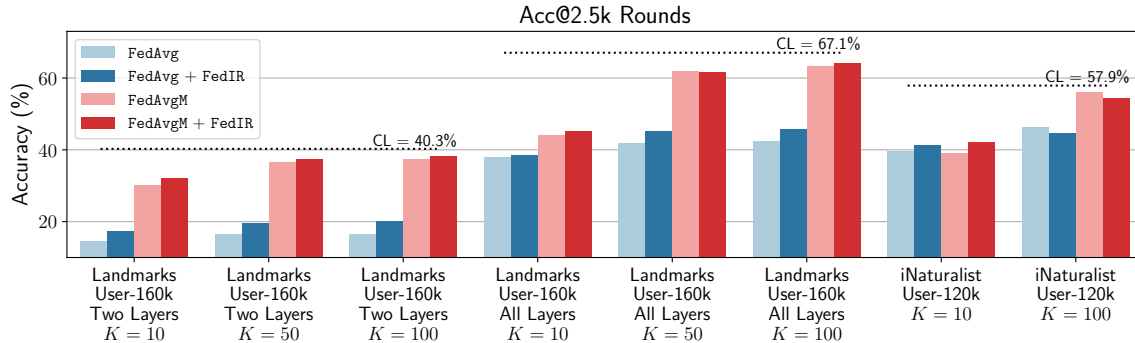


Figure 8-7: **Comparing Base Methods with and without FedIR.** Accuracy shown at 2.5k communication rounds. Centralized learning accuracy marked with dashed lines.

### 8.7.1 Classification Accuracy vs Distribution Non-Identicalness

Our experiments use CIFAR10/100 datasets to characterize classification accuracy with a continuous range of distribution non-identicalness. We follow the protocol described by Hsu et al. (2019) such that the class distribution of every client is sampled from a Dirichlet distribution with varying concentration parameter  $\alpha$ .

We measure distribution non-identicalness using an average *Earthmover’s Distance* (EMD) metric. Specifically, we take the discrete class distribution  $\mathbf{q}_i$  for every client, and define the population’s class distribution as  $\mathbf{p} = \sum_i \frac{n_i}{n} \mathbf{q}_i$ , where  $n = \sum_i n_i$  counts training samples from all clients. The non-identicalness of a dataset is then computed as the weighted average of distances between clients and the population:  $\sum_i \frac{n_i}{n} \text{Dist}(\mathbf{q}_i, \mathbf{p})$ .  $\text{Dist}(\cdot, \cdot)$  is a distance metric between two distributions, for which we, in particular, use  $\text{EMD}(\mathbf{q}, \mathbf{p}) \equiv \|\mathbf{q} - \mathbf{p}\|_1$ , bounded between  $[0, 2]$ .

Figures 8-6a and 8-6b show the trend in classification accuracy as a function of distribution non-identicalness (average EMD difference). We are able to approach centralized learning accuracy with data on the identical end. A substantial drop around an EMD of 1.7 to 2.0 is observed in both datasets. Applying momentum on the server, FedAvgM significantly improves the convergence under heterogeneity conditions for all datasets. Using more clients per round (larger report goal  $K$ ) is also beneficial for training but has diminishing returns.

### 8.7.2 Importance Reweighting

Importance Reweighting is proposed for addressing the per-client distribution shift. We evaluate FedIR with both FedAvg and FedAvgM on both two datasets with natural user splits: iNaturalist-User-120k and Landmarks-User-160k.

For Landmarks, we experiment with two different training schemes: (a) fine-tuning the entire network (*all layers*) end to end, (b) only training the last *two layers* while freezing the network backbone. We set the local epochs to  $E = 5$  and experiment with report goals  $K = \{10, 50, 100\}$ , respectively.

The result in Figure 8-7 shows a consistent improvement on the Landmarks-User-160k dataset over the FedAvg baseline. While FedAvgM gives the most significant improvements in all runs, FedIR further improves the convergence speed and accuracy, especially when the report goal is small (Figure 8-9).

Landmarks-User-160k (EMD = 1.94) has more skewed data distribution than iNaturalist-User-120k (EMD = 1.83) and benefits more from FedIR.

### 8.7.3 Federated Virtual Clients

We apply the Virtual Clients scheme (FedVC) to both FedAvg and FedAvgM and evaluate its efficacy using iNaturalist user and geo-location datasets, each of which contains significantly imbalanced clients. In the experiments, 10 clients are selected for every federated round. We use a mini-batch size  $B = 64$  and set the virtual client size  $N_{VC} = 256$ .

Figure 8-8 demonstrates the efficiency and accuracy improvements gained via FedVC when clients are imbalanced. The convergence of vanilla FedAvg suffers when clients perform excessive local optimization steps. In iNaturalist-Geo-3k, for example, clients can take up to 46 (*i.e.*,  $3000/64$ ) local steps before reporting to the server. To show that FedVC utilizes data efficiently, we report accuracy at fixed batch budgets in addition to fixed round budgets. Batch budget is calculated by summing the number of local batches taken for the largest client per round. As shown in Table 8.2, FedVC consistently yields superior accuracy on both FedAvg and FedAvgM. Learning curves

Table 8.2: **Accuracy of Federated Virtual Client on iNaturalist.** Acc@round denotes the accuracy at a FL communication round. Acc@batch denotes the batch count accumulated over the largest clients per round, and is a proxy for a fixed time budget.

Data	Method	FedVC	$K$	Acc@Round(%)					Acc@Batch(%)				
				1k	2.5k	5k	10k	25k	50k				
Geo-3k	FedAvg	✗	10	47.0	47.9	48.7	37.8	44.4	46.5				
	FedAvgM	✗	10	47.2	50.4	45.0	42.5	47.1	44.9				
	FedAvg	✓	10	37.4	46.2	52.8	46.2	53.1	55.5				
	FedAvgM	✓	10	<b>49.7</b>	<b>54.8</b>	<b>56.7</b>	<b>54.8</b>	<b>56.7</b>	<b>57.1</b>				
User-120k	FedAvg	✗	10	34.7	39.7	41.3	37.8	39.8	42.9				
	FedAvgM	✗	10	31.9	39.2	41.3	32.3	41.6	43.4				
	FedAvg	✓	10	31.3	39.7	43.9	39.7	48.9	52.8				
	FedAvgM	✓	10	<b>37.9</b>	<b>43.7</b>	<b>49.1</b>	<b>43.7</b>	<b>47.4</b>	<b>54.6</b>				
Centralized												57.9	

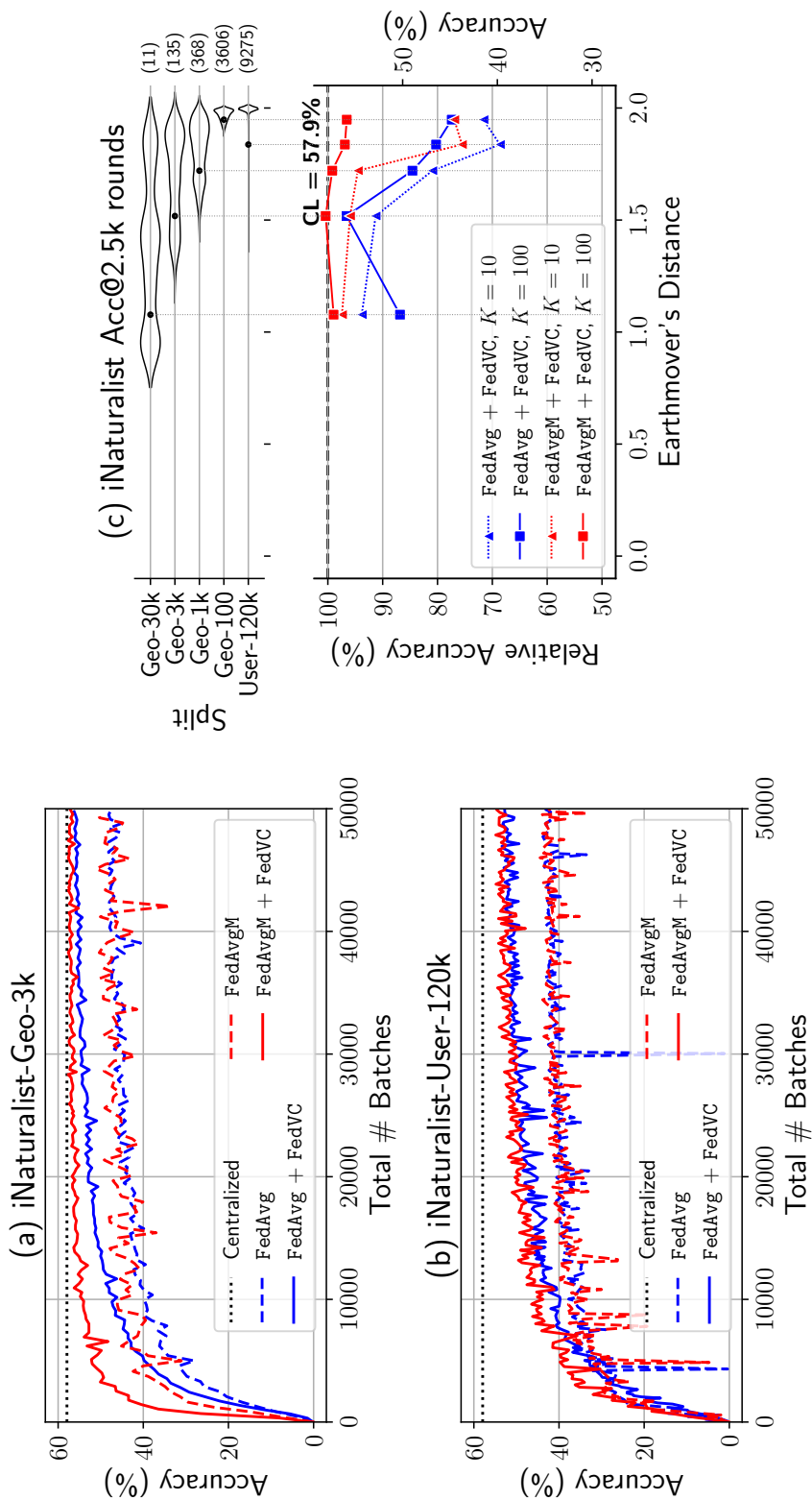


Figure 8-8: **Learning with Federated Virtual Clients.** Curves on the left are learned on the iNaturalist geo-partitioning Geo-3k and user split User-120k each with 135 clients and 9275 clients. Experiments on multiple iNaturalist partitionings are shown on the right, plotting relative accuracy at 2.5k communication rounds to mean EMD. Centralized learning achieves a 57.9% accuracy.

Table 8.3: **iNaturalist-User-120k accuracy**. Numbers reported at fixed communication rounds.  $K$  denotes the report goal per round.

Method	FedVC	FedIR	$K$	Accuracy@Rounds(%)		
				1k	2.5k	5k
FedAvg	✓	✗	10	31.3	39.7	43.9
FedAvg	✓	✗	100	36.9	46.5	51.4
FedAvg	✓	✓	10	30.1	41.3	47.5
FedAvg	✓	✓	100	35.5	44.8	49.8
FedAvgM	✓	✗	10	37.9	43.7	49.1
FedAvgM	✓	✗	100	<b>53.0</b>	<b>56.1</b>	<b>57.2</b>
FedAvgM	✓	✓	10	38.4	42.1	47.0
FedAvgM	✓	✓	100	51.3	54.3	56.2
Centralized					57.9	

in Figure 8-8 show that FedVC also decreases the learning volatility and stabilizes learning.

iNaturalist per-user and geo-location datasets reflect varying degrees of non-identicalness. Figure 8-8c, though noisier, exhibits a similar trend compared to Figure 8-6. The performance degrades as the degree of non-identicalness, characterized by EMD, increases.

#### 8.7.4 Federated Visual Classification Benchmarks

Having shown that our proposed modifications to FedAvg indeed lead to a speedup in learning on both iNaturalist and Landmarks, we wish to also provide some benchmark results on natural user partitioning with reasonable operating points. We hope that these datasets can be used for understanding real-world federated visual classification, and act as benchmarks for future improvements.

##### iNaturalist-User-120k

The iNaturalist-User-120k data has 9,275 clients and 120k examples, containing 1,203 species classes. We use report goals  $K = \{10, 100\}$  and FedVC samples  $N_{VC} = 256$  examples per client. A summary of the benchmark results is shown in Table 8.3.

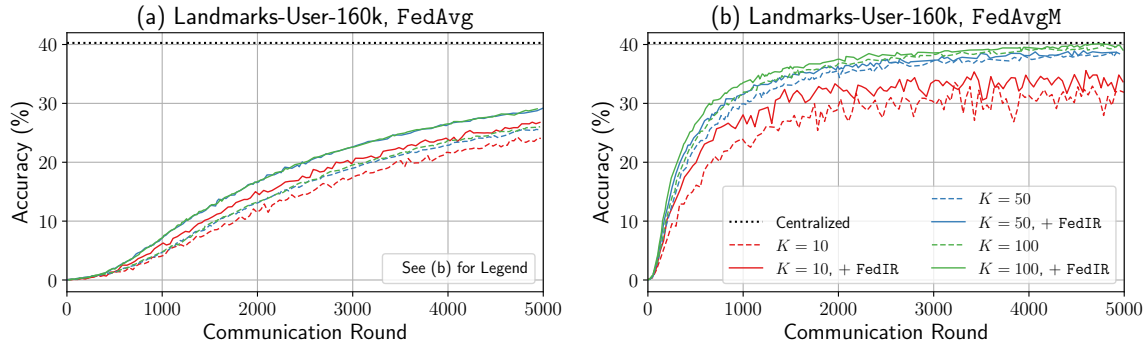


Figure 8-9: **Landmarks-User-160k Learning Curves.** Only the last two layers of the network are fine-tuned. FedIR is also shown due to its ability to address skewed training distribution as presented in this dataset.

Notice that FedAvgM with FedVC and a large report goal of  $K = 100$  has a 57.2% accuracy, almost reaching the same level as in centralized learning (57.9%). With that said, there is still plenty of room to improve performance with small reporting clients and round budgets. Being able to learn fast with a limited pool of clients is one of the critical research areas for practical visual FL.

### Landmarks-User-160k

The Landmarks-User-160k dataset comprises 164,172 images for 2,028 landmarks, divided among 1,262 clients. We follow the setup in Section 8.7.2 where we experiment with either training the whole model or fine-tuning the last two layers. Report goal  $K = \{10, 50, 100\}$  are used.

Similarly, FedAvgM with the  $K = 100$  is able to achieve 65.9% accuracy at 5k communication rounds, which is just 1.2% off from centralized learning. Interestingly, when we train only the last two layers with FL, the accuracy is as well not far off from centralized learning (39.8% compared to 40.3%)

### 8.7.5 Hyperparameter Sensitivity

To study how sensitive the hyperparameter tuning process is to different degrees of non-identicalness in FL settings, we perform experiments on CIFAR-10/100 datasets with a grid of hyperparameters. All CIFAR experiments in this study are tuned over

Table 8.4: Landmarks-User-160k Accuracy.

Method	FedIR	K	Accuracy@Rounds(%)								
			Two layers			All layers					
			1k	2.5k	5k	1k	2.5k	5k			
FedAvg	<b>X</b>	10	4.2	14.6	24.6	18.2	38.1	49.7			
FedAvg	<b>X</b>	50	4.5	16.5	26.0	20.9	42.0	53.3			
FedAvg	<b>X</b>	100	4.9	16.5	26.3	21.9	42.3	53.4			
FedAvg	<b>✓</b>	10	6.3	17.4	26.6	19.6	38.5	51.7			
FedAvg	<b>✓</b>	50	7.4	19.7	28.8	26.0	45.2	55.0			
FedAvg	<b>✓</b>	100	7.2	20.1	29.0	26.5	45.7	55.2			
FedAvgM	<b>X</b>	10	23.0	30.1	30.8	29.4	44.1	53.7			
FedAvgM	<b>X</b>	50	29.9	36.4	38.6	55.2	62.0	64.8			
FedAvgM	<b>X</b>	100	31.9	37.4	39.6	56.3	63.4	65.0			
FedAvgM	<b>✓</b>	10	26.5	32.1	31.3	27.9	45.1	53.5			
FedAvgM	<b>✓</b>	50	31.6	37.5	38.9	53.1	61.6	63.2			
FedAvgM	<b>✓</b>	100	<b>33.7</b>	<b>38.3</b>	<b>39.8</b>	<b>57.7</b>	<b>64.1</b>	<b>65.9</b>			
Centralized			40.27			67.05					

the the same grid. Following Shallue et al. (2018), we define the effective learning rate for FedAvgM as  $\eta_{\text{eff}} = \eta / (1 - \beta)$ . For all values of Dirichlet concentration  $\alpha$ , we sweep over learning rate  $\eta_{\text{eff}} \in \{10^{-3}, 10^{-2.5}, \dots, 10^0\}$  and momentum  $1 - \beta \in \{10^{-2.5}, 10^{-2}, \dots, 10^0\}$ .

In Figure 8-10 we show the effect of using different  $\eta_{\text{eff}}$  on the relative accuracy with each grid point showing the best result over all  $(\beta, \eta)$  combinations that give the same  $\eta_{\text{eff}}$ . We train for 10k/20k communication rounds with CIFAR-10/100 respectively.

Within each individual contour plot, it can be seen that the accuracy consistently drops with increased non-identicalness, and the set of hyperparameters yielding high performance becomes smaller. In general, we find an effective learning rate  $\eta_{\text{eff}} = 10^{-2}$  works well in many situations.

Across different report goals  $K$ , a larger  $K$  enables good performance over a wider range of  $\eta_{\text{eff}}$ . This result is unsurprising, since with more clients reporting in, the server observes more data and hence obtains gradients with less variance. The number of local epochs does not affect the choice of hyperparameters much in our experiments (see last two rows of Figure 8-10). Interestingly, while CIFAR-10 and CIFAR-100 have different numbers of classes and centralized learning accuracy, they exhibit very similar characteristics in terms of *relative accuracy* (the overall shape of plots in Figure 8-10 is similar).

### 8.7.6 The Effect of Pretraining

Pretraining large visual models (e.g., using ImageNet) is very common in centralized training. It is likely to be even more beneficial in federated settings, where extra computation rounds could be prohibitively time consuming. In some cases, however, it may be necessary or desirable to train from scratch. In this section, we investigate the feasibility of training large federated visual classification models without pretraining. Note that across this study, the smaller CIFAR10/100 experiments are trained from scratch, but the larger iNaturalist and Landmarks experiments use an ImageNet pretrained MobileNetV2..

We perform experiments using iNaturalist-Geo-3k with a combination of settings



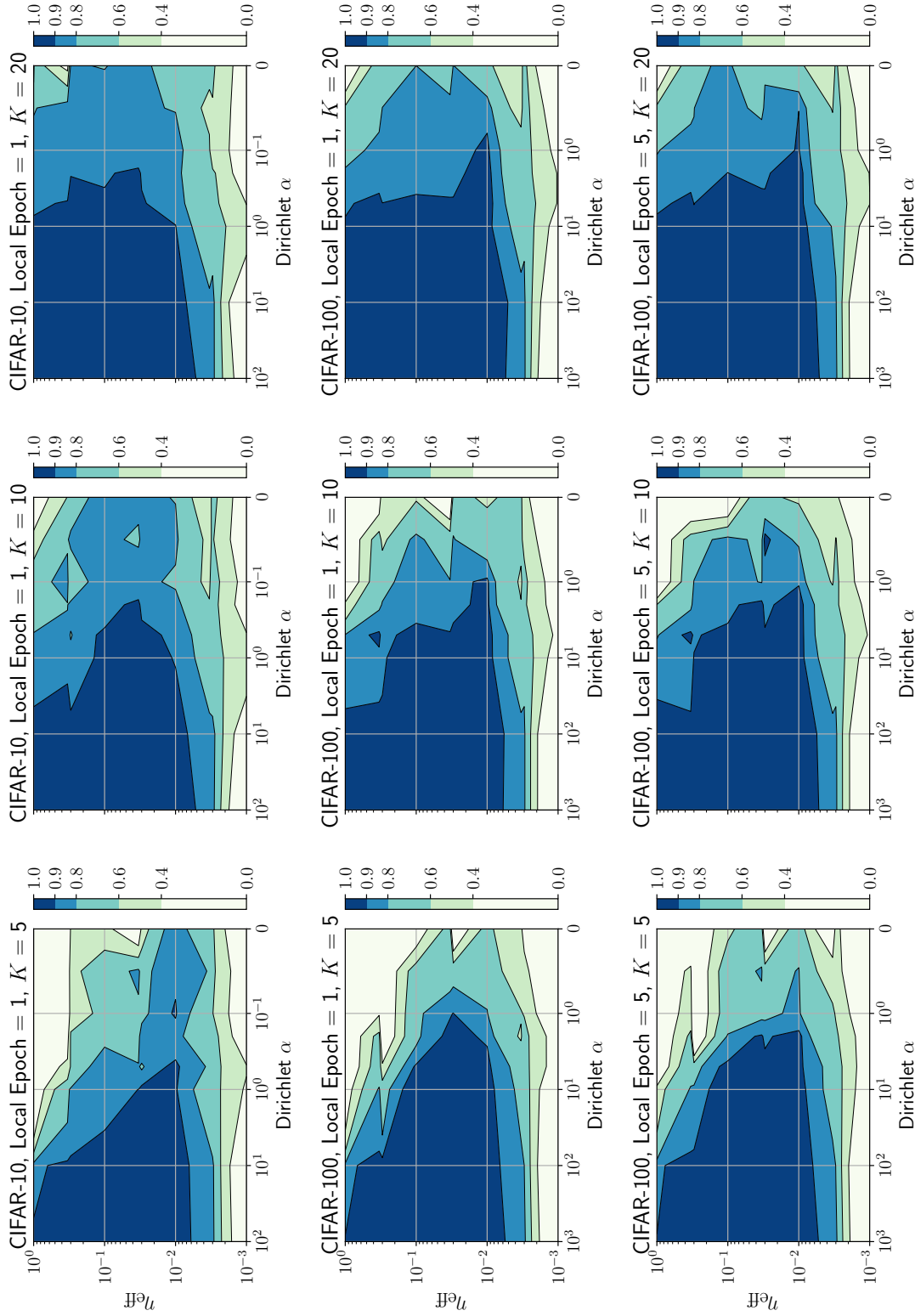


Figure 8-10: **Relative Accuracy of FedAvgM on CIFAR Datasets.** Darker shades denote regions of higher relative accuracy.  $\eta_{\text{eff}} = \eta / (1 - \beta)$  is the effective learning rate, and  $K$  is the reporting goal out of 100 clients. Note that data split is increasingly non-identical to the right.

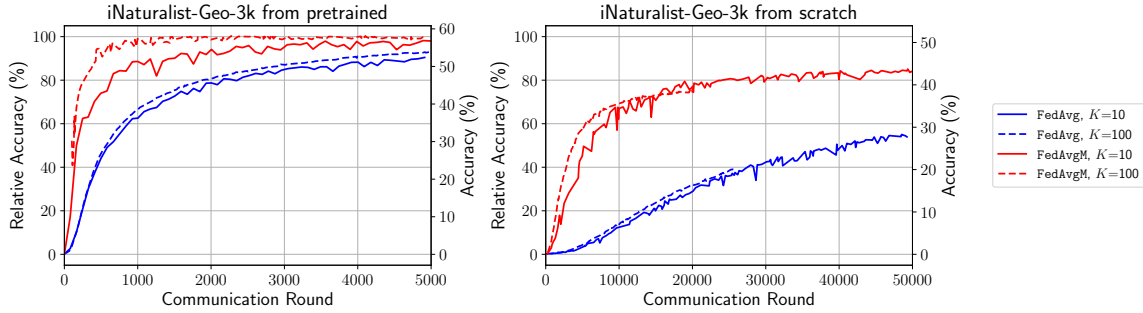


Figure 8-11: **Learning Curves from ImageNet Pretraining and from Scratch.** On the left vertical axis is the relative accuracy while on the right is the absolute accuracy. Two plots are rescaled to have the full span of 100% relative accuracy.

including the FL algorithm (FedAvg/FedAvgM) and report goal  $K$ . Since training from random initialization and from pretrained weights converge to different final test accuracy, we use *relative accuracy* for evaluating FL algorithms’ progress relative to the corresponding centralized learning upper bounds.

From Figure 8-11, we see that FL with pretraining requires orders of magnitude fewer communication rounds for convergence and yields higher final relative accuracy than training from scratch. Table 8.5 further shows the rounds needed to reach 10%, 50%, and 90% relative accuracy. We see that FedAvgM is able to accelerate convergence significantly, with a report goal  $K = 100$  it takes 94% ( $977 \rightarrow 60$ ) fewer rounds than FedAvg to reach 10% relative accuracy when starting from pretrained model weights. We also see that FedAvgM has a much steeper learning curve, reaching 90% relative accuracy in  $6.9\times$  the rounds needed to reach 10% (compared to  $20\times$  for FedAvg).

Whilst our results suggest that it is possible to train large federated visual classification models from scratch, doing so efficiently and effectively remains an open challenge with room for improvement.

### 8.7.7 Experiment Run Time

The federated learning experiments are carried out by simulation with a cluster of NVIDIA Tesla P100 GPUs in parallel. The experiment run time, while highly variable depending on the experimental setup (model complexity, dataset, local steps  $E$ , and

Table 8.5: **Communication Rounds to Reach Relative Accuracy.** Note that models have different centralized learning accuracy (51.4% from scratch and 57.9% from pretrained). The multipliers are calculated row-wise, using Rounds@10% as the baseline. Experiments that do not reach the target relative accuracy even after  $t$  rounds is marked  $> t$ .

Data	Method	Initialization	$K$	Rounds@Relative Accuracy		
				10 %	50 %	90 %
Geo-3k	FedAvg	pretrained	10	165 (1.0 $\times$ )	669 (4.1 $\times$ )	4912 (29.8 $\times$ )
	FedAvg	pretrained	100	165 (1.0 $\times$ )	567 (3.4 $\times$ )	3780 (22.9 $\times$ )
	FedAvgM	pretrained	10	79 (1.0 $\times$ )	249 (3.2 $\times$ )	1505 (19.1 $\times$ )
	FedAvgM	pretrained	100	<b>60</b> (1.0 $\times$ )	<b>116</b> (1.9 $\times$ )	<b>420</b> (6.9 $\times$ )
	FedAvg	scratch	10	9005 (1.0 $\times$ )	39236 (4.4 $\times$ )	> 50k
	FedAvg	scratch	100	7793 (1.0 $\times$ )	> 20k	> 20k
	FedAvgM	scratch	10	1463 (1.0 $\times$ )	5788 (4.0 $\times$ )	> 50k
	FedAvgM	scratch	100	977 (1.0 $\times$ )	3733 (3.8 $\times$ )	> 20k

reporting clients per round  $K$ ), is roughly 0.5 to 2.0 seconds per communication round per reporting client. This amounts to about 9 GPU-days for a run of Landmarks-User-160k experiment for 5000 rounds with  $K = 100$ .

## 8.8 Summary

We have shown that large-scale visual classifiers can be trained using a privacy-preserving, federated approach, and highlighted the challenges that per-user data distributions pose for learning. We provide two new datasets and benchmarks, providing a platform for other explorations in this space. We expect others to improve on our results, particularly when the number of participating clients and round budget is small. There remain many challenges for Federated Learning that are beyond the scope of this study: real world data may include domain shift, label noise, poor data quality and duplication. Model size, bandwidth and unreliable client connections also pose challenges in practice. We hope our work inspires further exploration in this area.

It is also worthwhile to note, in this chapter, we push the boundary of our understanding of large-scale visual learning under federated learning settings, which is never extensively explored before in either natural imaging or medical imaging. Studies like this are absolutely necessary since, in real-life applications, there is no way of peeking data from the respective institutions so much like we do in this study as to be able to measure centralized learning accuracy or relative accuracy. The most statistics we hold permissions to gather from participating parties are likely their label distribution or metadata distributions, and hence establishing a connection between the statistical descriptions of the distributions and the expected performance are beneficial to future federated learning application deployments.

# Chapter 9

## Conclusions

Throughout the dissertation, we systematically explore methodologies to deal with medical imaging problems when the data is constrained (*i.e.*, limited in type, quantity, quality, or else).

Concretely, medical images (a) are dense in physical dimensions, while (b) having sparse and concentrated information; (c) require specialized expertise to process them; (d) exhibit a long-tailed distribution in disease findings; (e) experience variances due to acquisition protocol and instruments; and (f) are heterogeneous from site to site for demographic and other factors. To combat these challenges, we inspect from both the clinical aspect as well as the computational aspect.

We introduce the concept of annotation reduction, where we are able to reduce the effective time and resources for data collection for medical imaging and retain similar modeling effectiveness. These methods include (a) semi-supervised and unsupervised learning, (b) reinforcement learning, (c) cross-modal learning, and (d) transfer learning; but note this is not an exhaustive list, and these are simply the ones we cover in the dissertation. Medical institutions can as well collaboratively learn AI models together while preserving privacy through federated learning paradigms.

It would be a shame if we only approach the constrained data problem from an algorithmic aspect, as there is existing knowledge about medicine and clinical practice that we can leverage. Hence we explore surrogate endpoint modeling which, although not directly predicting diagnostic endpoints, performs decently in predicting inter-

mediate variables that have been shown to correlate well with the target diagnostic endpoints in question. Aside from this two-step approach, we can also directly infuse the knowledge about diseases and findings directly into the modeling process. For example, our hearts are on the left side of the body; and we have 32 teeth in our permanent dentition.

## 9.1 Chapter Review

In Chapter 3, we look at the problem of automatic body composition evaluation, how transfer learning benefits the adaptation of models to institutional-local data, and how the resulting muscle quantity evaluation correlates with patient risks in pancreatic cancer better than the body mass index (BMI).

In Chapter 4, we further extend the scope of the body composition to using the resulting visceral fat evaluation as a surrogate endpoint to better predict severity for COVID-19 patients better than BMI.

In Chapter 5, we explore cross-modal representation learning between chest X-rays and their textual reports. Many algorithms are applied to associate image embeddings and text embeddings, and even when no supervision is provided between the two modalities, there is still effective distribution alignment, which can be evidenced through cross-modal retrieval.

In Chapter 6, the same X-ray dataset is reused for medical report generation from medical images. By minding the clinical efficacy of the output medical report, we are able to achieve much better clinical descriptive power while retaining similar language fluency.

In Chapter 7, we try to summarize, from dental panoramic images, relevant clinical findings about the teeth. We also experiment with a new label modality that allows teeth-wise binary labeling to augment the training process which typically only uses pixel-wise dense annotations. We are able to improve finding accuracy efficiently by supplying these binary labels that are efficient to label for clinicians.

In Chapter 8, we benchmark federated learning under heterogeneous data with

natural images as a precursor for medical imaging, and find *relative accuracy* and *earth-mover distance* to be excellent descriptors for the impact of heterogeneity on federated learning algorithms across multiple datasets.

## 9.2 Areas to Be Explored Before Landing AI

Though we have covered a wide range of modeling solutions in medical imaging, there are aspects that cannot be ignored before they are adopted for applications.

Specifically, in learning practical AI models with real-world data, we need to consider *missingness* in the training data. [Little and Rubin \(2019\)](#) classified missing data into three categories according to the driving mechanism: (a) missing completely at random (MCAR), where missingness occurs with a fixed probability; (b) missing at random (MAR), where missingness is conditional randomly on observed variables; and (c) missing not at random (MNAR), where missingness is conditional randomly on the missing variables themselves. For imaging data, the missingness can occur on the per-study level, or on a per-pixel/voxel level. To tackle MCAR and MAR, one can choose to drop or impute the data ([Mulugeta et al., 2017](#)), and for MNAR, there is not yet a well-studied solution due to its inherent statistical complexity, by definition.

Unsurprisingly, the source of missingness can as well be biases originating from the acquisition protocols, resource access differences, and even societal differences ([Rose, 2018](#); [Rajkomar et al., 2018](#)). The research findings suggest that comparing modeling outcomes across demographic groups is imperative, not only for fairness reasons ([Chen et al., 2019a](#)) but also as a sanity check for model generalizability ([Subbaswamy and Saria, 2018](#); [Subbaswamy et al., 2019](#)).

Finally, all of the AI systems we work on in this dissertation are targeted to be deployed as an augmenting tool for clinicians, and it would be ignorant if we neglect their perception and usage of such tools – *interpretability* and *uncertainty* (or confidence) being the most pressing needs. Interpretability allows researchers to dissect the internal workings of the algorithms, with the hope of improving modeling

performance; it also allows clinicians to reason why the model predictions are as such quantifiably. There is already a wide range of interpretability tools for medical imaging are gradient-weighted class activation mapping (Grad-CAM) (Selvaraju et al., 2017), guided backpropagation (Springenberg et al., 2014), and regression concept vectors (Graziani et al., 2019).

Uncertainty quantification focuses on describing how confident the model is about the prediction results, as a statistically uncorrelated measure to the predictions themselves. The AI models without uncertainty quantification can be confidently incorrect, and yet by estimating uncertainty, they can defer the decision to human experts when the certainty level is low enough regardless of the raw prediction levels. There are already plenty of works along this direction (Xia et al., 2020; Hiasa et al., 2019; Karimi et al., 2019; Jungo and Reyes, 2019; Joskowicz et al., 2018; Ghesu et al., 2019; Raghu et al., 2019a) in medical imaging and hence should be incorporated for consideration in any medical applications.

### 9.3 Novel Research Directions

Aside from the existing research fields to circumvent issues in medical imaging, we also want to bring up several interesting directions that have been barely systematically studied, at least to the best of our knowledge.

**Pre-Training Deep Neural Networks** Pre-training models with the ImageNet (Deng et al., 2009) competition dataset have been a conventional method to bootstrap any convolutional neural networks (CNN). While it makes perfect sense for natural imaging tasks, as the ImageNet data capture everyday objects and concepts quite extensively, it is inexplicably odd to also adopt this model to bootstrap medical imaging tasks. For example, a chest radiograph looks nothing like objects in real life. There is a limited number of studies that attempted to pre-train medical-generic (Alzubaidi et al., 2021) or 3-dimensional models (Chen et al., 2019b), and built a platform to collect medical imaging models (Li et al., 2017; Gibson et al., 2018).



We believe there are shared visual patterns in medical images, just like the case for natural images shown in past research (Garg et al., 2019). The visual patterns can be extracted by more recent techniques such as *contrastive learning* (Jiang et al., 2020; Wang et al., 2021b) in an unsupervised manner, from which medical imaging could benefit greatly.

**Metric Learning** Clinicians’ judgments are not always on an *absolute* scale, and more often than not they are on a *relative* scale. Concretely, when given two cases, they are better at describing the relative severity of the diseases, rather than giving the two cases severity levels as defined in handbooks (Hammon et al., 2014). As a result, collecting absolute disease staging annotation is time-consuming for clinicians, and even after doing so, the annotations have low inter-rater agreements due to the inherent nature of severity being on a spectrum.

Hence, researchers have introduced *metric learning* from outside the medical imaging community, and focus on using distance descriptions in learning models. In the context of medical imaging, that translates to using relative disease severity annotations to construct a continuous severity grading model (Li et al., 2020b; Akbar et al., 2022). Siamese neural networks (Koch et al., 2015) are often utilized in metric learning works due to their formulation that applies the same network on multiple inputs.

**Differential Learning** AI models are meant to augment clinicians and not replace them (Ghassemi et al., 2020). While medical AI has spent time on how to obtain more accurate models, we often neglect the fact that these models are meant to be used in a time-pressing environment for clinician decision support. Take dental imaging as an example: while dentists can take a single glimpse to understand how many dental implants are in the patients’ oral region, periapical lesions are harder targets to identify and might require the dentists to read for a few extra minutes to spot. If the AI models are configured in a way that they do not output apparent findings to clinicians, but only the *differential* findings that are less obvious, the signal-to-noise

ratio for such a system can be maximized for practicing clinicians. To the best of our knowledge, there are no existing studies that attempt to address the differential prediction setting specifically, or benchmark how the differential operating points should be tuned to optimize AI's ability to augment individual human clinicians.

## 9.4 Conclusions

AI is data-hungry. Data is thus an extremely important aspect of computer vision, and it is even more so for medical imaging owing to data privacy issues and resource-intensive interpretations. While many challenges may be present, it also implies that many unique opportunities exist for us to improve the current landscape of healthcare. We hope the several research topics showcased in this dissertation inspire further research and are able to fuel explorations down the line, ultimately benefiting humanity on a civilization scale.

# Bibliography

- Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., et al. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*. 69
- Abbasi-Sureshjani, S., Raumanns, R., Michels, B. E., Schouten, G., and Cheplygina, V. (2020). Risk of training diagnostic algorithms on data with demographic bias. In *Interpretable and Annotation-Efficient Learning for Medical Image Computing*, pages 183–192. Springer. 32
- Abdul Salam, M., Taha, S., and Ramadan, M. (2021). Covid-19 detection using federated machine learning. *PLoS One*, 16(6):e0252573. 45, 151
- Ahn, Y., Yoon, J. S., Lee, S. S., Suk, H.-I., Son, J. H., Sung, Y. S., Lee, Y., Kang, B.-K., and Kim, H. S. (2020). Deep learning algorithm for automated segmentation and volume measurement of the liver and spleen using portal venous phase computed tomography images. *Korean journal of radiology*, 21(8):987. 40
- Akbar, M. N., Wang, X., Erdogmus, D., and Dalal, S. (2022). PENet: Continuous-valued pulmonary edema severity prediction on chest X-ray using siamese convolutional networks. 185
- Al-Benna, S. (2020). Association of high level gene expression of ACE2 in adipose tissue with mortality of COVID-19 infection in obese patients. *Obesity medicine*, 19:100283. 80
- Alzubaidi, L., Santamaría, J., Manoufali, M., Mohammed, B., Fadhel, M. A., Zhang, J., Al-Timemy, A. H., Al-Shamma, O., and Duan, Y. (2021). MedNet: pre-trained convolutional neural network model for the medical imaging tasks. *arXiv preprint arXiv:2110.06512*. 184
- Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B., Reicher, J. J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G., et al. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature medicine*, 25(6):954–961. 31
- Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association. 105

- Aronson, A. R. and Lang, F.-M. (2010). An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236. **22, 104**
- Babic, A., Rosenthal, M. H., Bamlet, W. R., Takahashi, N., Sugimoto, M., Danai, L. V., Morales-Oyarvide, V., Khalaf, N., Dunne, R. F., Brais, L. K., et al. (2019). Postdiagnosis loss of skeletal muscle, but not adipose tissue, is associated with shorter survival of patients with advanced pancreatic cancerbody composition change and pancreatic cancer prognosis. *Cancer Epidemiology, Biomarkers & Prevention*, 28(12):2062–2069. **62**
- Bahdanau, D., Brakel, P., Xu, K., Goyal, A., Lowe, R., Pineau, J., Courville, A., and Bengio, Y. (2016). An actor-critic algorithm for sequence prediction. *arXiv preprint arXiv:1607.07086*. **106**
- Bai, W., Oktay, O., Sinclair, M., Suzuki, H., Rajchl, M., Tarroni, G., Glocker, B., King, A., Matthews, P. M., and Rueckert, D. (2017). Semi-supervised learning for network-based cardiac mr image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 253–260. Springer. **43**
- Ballard, D. H. (1981). Generalizing the Hough transform to detect arbitrary shapes. *Pattern recognition*, 13(2):111–122. **52**
- Bankman, I. (2008). *Handbook of medical image processing and analysis*. Elsevier. **30**
- Barkousaraie, A. S., Ogunmolu, O., Jiang, S., and Nguyen, D. (2019). A reinforcement learning application of guided monte carlo tree search algorithm for beam orientation selection in radiation therapy. In *MEDICAL PHYSICS*, volume 46, pages E236–E236. WILEY 111 RIVER ST, HOBOKEN 07030-5774, NJ USA. **31**
- Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological reports*, 19(1):3–11. **54**
- Bassett, I. V., Triant, V. A., Bunda, B. A., Selvaggi, C. A., Shinnick, D. J., He, W., Lu, F., Porneala, B. C., Cao, T., Lubitz, S. A., et al. (2020). Massachusetts general hospital covid-19 registry reveals two distinct populations of hospitalized patients by race and ethnicity. *PloS one*, 15(12):e0244270. **68, 69**
- Battisti, S., Pedone, C., Napoli, N., Russo, E., Agnoletti, V., Nigra, S. G., Dengo, C., Mughetti, M., Conte, C., Pozzilli, P., et al. (2020). Computed tomography highlights increased visceral adiposity associated with critical illness in COVID-19. *Diabetes Care*, 43(10):e129–e130. **65, 67, 81**
- Becker, A. S., Mueller, M., Stoffel, E., Marcon, M., Ghafoor, S., and Boss, A. (2018). Classification of breast cancer in ultrasound imaging using a generic deep learning analysis software: a pilot study. *The British journal of radiology*, 91(xxxx):20170576. **31**

- Bilic, P., Christ, P. F., Vorontsov, E., Chlebus, G., Chen, H., Dou, Q., Fu, C.-W., Han, X., Heng, P.-A., Hesser, J., et al. (2019). The liver tumor segmentation benchmark (lits). *arXiv preprint arXiv:1901.04056*. 50
- Bird, S. and Loper, E. (2004). NLTK: the natural language toolkit. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 31. Association for Computational Linguistics. 91
- Boag, W., Hsu, T.-M. H., McDermott, M., Berner, G., Alesentzer, E., and Szolovits, P. (2020). Baselines for chest X-ray report generation. In *Machine Learning for Health Workshop*, pages 126–140. PMLR. 31, 42
- Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L. D., Monfort, M., Muller, U., Zhang, J., et al. (2016). End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*. 39
- Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., Kiddon, C. M., Konečný, J., Mazzocchi, S., McMahan, B., Overvelde, T. V., Petrou, D., Ramage, D., and Roselander, J. (2019). Towards federated learning at scale: System design. In *SysML 2019*. 165
- Börger, C. and Natterer, F. (1999). *Computational radiology and imaging: therapy and diagnostics*, volume 110. Springer Science & Business Media. 30
- Brady, A. P. (2017). Error and discrepancy in radiology: inevitable or avoidable? *Insights into imaging*, 8(1):171–182. 30
- Bredella, M. A. (2017). Sex differences in body composition. *Sex and gender factors affecting metabolic homeostasis, diabetes and obesity*, pages 9–27. 72
- Bridge, C. P., Rosenthal, M., Wright, B., Kotecha, G., Fintelmann, F., Troschel, F., Miskin, N., Desai, K., Wrobel, W., Babic, A., et al. (2018). Fully-automated analysis of body composition from CT in cancer patients using convolutional neural networks. In *OR 2.0 Context-aware operating theaters, computer assisted robotic endoscopy, clinical image-based procedures, and skin image analysis*, pages 204–213. Springer. 40, 48, 50, 63
- Briggs, G. M., Flynn, P. A., Worthington, M., Rennie, I., and McKinstry, C. (2008). The role of specialist neuroradiology second opinion reporting: is there added value? *Clinical radiology*, 63(7):791–795. 30
- Brock, A., Donahue, J., and Simonyan, K. (2018). Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*. 92
- Brosch, T., Tam, R., Initiative, A. D. N., et al. (2013). Manifold learning of brain mris by deep learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 633–640. Springer. 41, 99

- Burns, J. E., Yao, J., Chalhoub, D., Chen, J. J., and Summers, R. M. (2020). A machine learning algorithm to estimate sarcopenia on abdominal ct. *Academic radiology*, 27(3):311–320. **40, 48, 50**
- Bustos, A., Pertusa, A., Salinas, J.-M., and de la Iglesia-Vayá, M. (2019). Padchest: A large chest x-ray image dataset with multi-label annotated reports. *arXiv preprint arXiv:1901.07441*. **22, 104**
- Caccia, M., Caccia, L., Fedus, W., Larochelle, H., Pineau, J., and Charlin, L. (2018). Language gans falling short. *arXiv preprint arXiv:1811.02549*. **106**
- Cai, Q., Chen, F., Wang, T., Luo, F., Liu, X., Wu, Q., He, Q., Wang, Z., Liu, Y., Liu, L., et al. (2020). Obesity and COVID-19 severity in a designated hospital in shenzhen, china. *Diabetes care*, 43(7):1392–1398. **65, 67**
- Caldas, S., Wu, P., Li, T., Konečný, J., McMahan, H. B., Smith, V., and Talwalkar, A. (2018). LEAF: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*. **155**
- Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., et al. (2018). Universal sentence encoder. *arXiv preprint arXiv:1803.11175*. **91, 93**
- Chan, H.-P., Hadjiiski, L. M., and Samala, R. K. (2020). Computer-aided diagnosis in the era of deep learning. *Medical physics*, 47(5):e218–e227. **30**
- Chauhan, G., Liao, R., Wells, W., Andreas, J., Wang, X., Berkowitz, S., Horng, S., Szolovits, P., and Golland, P. (2020). Joint modeling of chest radiographs and radiology reports for pulmonary edema assessment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 529–539. Springer. **44**
- Chen, I. Y., Szolovits, P., and Ghassemi, M. (2019a). Can AI help reduce disparities in general medical and mental health care? *AMA journal of ethics*, 21(2):167–179. **183**
- Chen, S., Ma, K., and Zheng, Y. (2019b). Med3d: Transfer learning for 3d medical image analysis. *arXiv preprint arXiv:1904.00625*. **44, 184**
- Chen, W.-Y., Hsu, T.-M. H., Tsai, Y.-H. H., Chen, M.-S., and Wang, Y.-C. F. (2019c). Transfer neural trees: Semi-supervised heterogeneous domain adaptation and beyond. *IEEE Transactions on Image Processing*, 28(9):4620–4633. **34**
- Chen, W.-Y., Hsu, T.-M. H., Tsai, Y.-H. H., Wang, Y.-C. F., and Chen, M.-S. (2016). Transfer neural trees for heterogeneous domain adaptation. In *European Conference on Computer Vision*, pages 399–414. Springer. **34, 89**

- Cheng, J.-Z., Ni, D., Chou, Y.-H., Qin, J., Tiu, C.-M., Chang, Y.-C., Huang, C.-S., Shen, D., and Chen, C.-M. (2016). Computer-aided diagnosis with deep learning architecture: applications to breast lesions in us images and pulmonary nodules in CT scans. *Scientific reports*, 6(1):1–13. 31
- Cheng, X., Zhang, L., and Zheng, Y. (2018). Deep similarity learning for multimodal medical images. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 6(3):248–252. 41, 99
- Chou, Y.-H., Hong, S., Sun, C., Cai, D., Song, M., and Li, H. (2021). Grp-fed: Addressing client imbalance in federated learning via global-regularized personalization. *arXiv preprint arXiv:2108.13858*. 45
- Choy, G., Khalilzadeh, O., Michalski, M., Do, S., Samir, A. E., Pianykh, O. S., Geis, J. R., Pandharipande, P. V., Brink, J. A., and Dreyer, K. J. (2018). Current applications and future impact of machine learning in radiology. *Radiology*, 288(2):318. 67
- Chung, Y.-A., Weng, W.-H., Tong, S., and Glass, J. (2018). Unsupervised cross-modal alignment of speech and text embedding spaces. *arXiv preprint arXiv:1805.07467*. 87, 89
- Chung, Y.-A., Weng, W.-H., Tong, S., and Glass, J. (2019). Towards unsupervised speech-to-text translation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7170–7174. IEEE. 87
- Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer. 63
- Ciresan, D., Giusti, A., Gambardella, L., and Schmidhuber, J. (2012). Deep neural networks segment neuronal membranes in electron microscopy images. *Advances in neural information processing systems*, 25. 41, 99
- Cohen, G., Afshar, S., Tapson, J., and Van Schaik, A. (2017). Emnist: Extending mnist to handwritten letters. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2921–2926. IEEE. 155
- Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2017). Word translation without parallel data. *arXiv preprint arXiv:1710.04087*. 88, 89, 92
- Cui, Z., Li, C., and Wang, W. (2019). Toothnet: Automatic tooth instance segmentation and identification from cone beam CT images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6368–6377. 145

- De Vos, B. D., Wolterink, J. M., De Jong, P. A., Viergever, M. A., and Išgum, I. (2016). 2d image classification for 3d anatomy localization: employing deep convolutional neural networks. In *Medical imaging 2016: Image processing*, volume 9784, pages 517–523. SPIE. **41**, **99**
- Demner-Fushman, D., Kohli, M. D., Rosenman, M. B., Shooshan, S. E., Rodriguez, L., Antani, S., Thoma, G. R., and McDonald, C. J. (2015). Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310. **22**, **100**, **104**, **117**
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. IEEE. **30**, **166**, **184**
- Devlin, J., Gupta, S., Girshick, R., Mitchell, M., and Zitnick, C. L. (2015). Exploring nearest neighbor approaches for image captioning. *arXiv preprint arXiv:1505.04467*. **107**
- Dieleman, S. and Schrauwen, B. (2014). End-to-end learning for music audio. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6964–6968. IEEE. **39**
- Doersch, C., Singh, S., Gupta, A., Sivic, J., and Efros, A. A. (2015). What makes Paris look like Paris? *Communications of the ACM*, 58(12):103–110. **155**
- Dou, Q., So, T. Y., Jiang, M., Liu, Q., Vardhanabhuti, V., Kaissis, G., Li, Z., Si, W., Lee, H. H., Yu, K., et al. (2021). Federated deep learning for detecting COVID-19 lung abnormalities in ct: a privacy-preserving multinational validation study. *NPJ digital medicine*, 4(1):1–11. **45**, **151**
- Drukker, K., Giger, M. L., Horsch, K., Kupinski, M. A., Vyborny, C. J., and Mendelson, E. B. (2002). Computerized lesion detection on breast ultrasound. *Medical physics*, 29(7):1438–1446. **30**
- Eban, E. E., Schain, M., Mackey, A., Gordon, A., Saurous, R. A., and Elidan, G. (2016). Scalable learning of non-decomposable objectives. *arXiv preprint arXiv:1608.04802*. **127**
- EnvoyAI (2017). EnvoyAI launches with 35 algorithms contributed by 14 newly-contracted artificial intelligence development partners. *EnvoyAI Blog*. **88**
- Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115. **88**
- Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338. **145**



- Fan, J., Wang, J., Chen, Z., Hu, C., Zhang, Z., and Hu, W. (2019). Automatic treatment planning based on three-dimensional dose distribution predicted from deep learning technique. *Medical physics*, 46(1):370–381. **31**
- Fan, X. C., Nemoto, T., Blatto, K., Mangiafesto, E., Sundberg, J., Chen, A., Foti, A., Holzhauser, M., Lahr, P., Snitzer, E., et al. (2013). Impact of presurgical breast magnetic resonance imaging (mri) on surgical planning—a retrospective analysis from a private radiology group. *The Breast Journal*, 19(2):134–141. **31**
- Faron, A., Sprinkart, A. M., Kuetting, D. L., Feisst, A., Isaak, A., Endler, C., Chang, J., Nowak, S., Block, W., Thomas, D., et al. (2020). Body composition analysis using CT and MRI: intra-individual intermodal comparison of muscle mass and myosteatosis. *Scientific reports*, 10(1):1–10. **49**
- Fedus, W., Goodfellow, I., and Dai, A. M. (2018). Maskgan: Better text generation via filling in the\_. *arXiv preprint arXiv:1801.07736*. **106**
- Feki, I., Ammar, S., Kessentini, Y., and Muhammad, K. (2021). Federated learning for COVID-19 screening from chest X-ray images. *Applied Soft Computing*, 106:107330. **45, 151**
- Gale, W., Oakden-Rayner, L., Carneiro, G., Bradley, A. P., and Palmer, L. J. (2018). Producing radiologist-quality reports for interpretable artificial intelligence. *arXiv preprint arXiv:1806.00340*. **88, 105, 107**
- Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. (2016). Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030. **89**
- Gao, F., Zheng, K. I., Wang, X.-B., Sun, Q.-F., Pan, K.-H., Wang, T.-Y., Chen, Y.-P., Targher, G., Byrne, C. D., George, J., et al. (2020). Obesity is a risk factor for greater COVID-19 severity. *Diabetes care*, 43(7):e72–e74. **65, 67**
- Garg, I., Panda, P., and Roy, K. (2019). A low effort approach to structured cnn design using pca. *IEEE Access*, 8:1347–1360. **185**
- Ge, C., Gu, I. Y.-H., Jakola, A. S., and Yang, J. (2020). Deep semi-supervised learning for brain tumor classification. *BMC Medical Imaging*, 20(1):1–11. **31**
- Ghassemi, M., Naumann, T., Schulam, P., Beam, A. L., Chen, I. Y., and Ranganath, R. (2020). A review of challenges and opportunities in machine learning for health. *AMIA Summits on Translational Science Proceedings*, 2020:191. **185**
- Ghesu, F. C., Georgescu, B., Gibson, E., Guendel, S., Kalra, M. K., Singh, R., Digmurthy, S. R., Grbic, S., and Comaniciu, D. (2019). Quantifying and leveraging classification uncertainty for chest radiograph assessment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 676–684. Springer. **184**

- Gibson, E., Li, W., Sudre, C., Fidon, L., Shakir, D. I., Wang, G., Eaton-Rosen, Z., Gray, R., Doel, T., Hu, Y., et al. (2018). NiftyNet: a deep-learning platform for medical imaging. *Computer methods and programs in biomedicine*, 158:113–122. **184**
- Goddard, P., Leslie, A., Jones, A., Wakeley, C., and Kabala, J. (2001). Error in radiology. *The British journal of radiology*, 74(886):949–951. **30**
- Goehler, A., Hsu, T.-M. H., Lacson, R., Gujrathi, I., Hashemi, R., Chlebus, G., Szolovits, P., and Khorasani, R. (2020). Three-dimensional neural network to automatically assess liver tumor burden change on consecutive liver mris. *Journal of the American College of Radiology*, 17(11):1475–1484. **31**
- Goehler, A., Hsu, T.-M. H., Seiglie, J. A., Siedner, M. J., Lo, J., Triant, V., Hsu, J., Foulkes, A., Bassett, I., Khorasani, R., et al. (2021). Visceral adiposity and severe COVID-19 disease: application of an artificial intelligence algorithm to improve clinical risk prediction. In *Open forum infectious diseases*, volume 8, page ofab275. Oxford University Press US. **31, 35, 65**
- Goldman, L. W. (2008). Principles of ct: multislice ct. *Journal of nuclear medicine technology*, 36(2):57–68. **32**
- Google (2019a). *TensorFlow Federated*. <https://www.tensorflow.org/federated>. **157**
- Google (2019b). *TensorFlow Federated Datasets*. [https://www.tensorflow.org/federated/api\\_docs/python/tff/simulation/datasets](https://www.tensorflow.org/federated/api_docs/python/tff/simulation/datasets). **155**
- Gozes, O. and Greenspan, H. (2018). Lung structures enhancement in chest radiographs via CT based fcnn training. In *Image Analysis for Moving Organ, Breast, and Thoracic Images*, pages 147–158. Springer. **41**
- Graffy, P. M., Liu, J., Pickhardt, P. J., Burns, J. E., Yao, J., and Summers, R. M. (2019). Deep learning-based muscle segmentation and quantification at abdominal ct: application to a longitudinal adult screening cohort for sarcopenia assessment. *The British journal of radiology*, 92(1100):20190327. **40, 62**
- Grave, E., Joulin, A., and Berthet, Q. (2018). Unsupervised alignment of embeddings with wasserstein procrustes. *arXiv preprint arXiv:1805.11222*. **89**
- Graves, A. (2013). Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*. **106**
- Graziani, M., Brown, J. M., Andrearczyk, V., Yildiz, V., Campbell, J. P., Erdogmus, D., Ioannidis, S., Chiang, M. F., Kalpathy-Cramer, J., and Müller, H. (2019). Improved interpretability for computer-aided severity assessment of retinopathy of prematurity. In *Medical Imaging 2019: Computer-Aided Diagnosis*, volume 10950, pages 450–460. SPIE. **184**

- Gu, L., Zheng, Y., Bise, R., Sato, I., Imanishi, N., and Aiso, S. (2017). Semi-supervised learning for biomedical image segmentation via forest oriented super pixels (voxels). In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 702–710. Springer. 43
- Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22):2402–2410. 88
- Guo, Z., Li, X., Huang, H., Guo, N., and Li, Q. (2019). Deep learning-based image segmentation on multimodal medical imaging. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 3(2):162–169. 31
- Gupta, V. and Lehal, G. S. (2010). A survey of text summarization extractive techniques. *Journal of emerging technologies in web intelligence*, 2(3):258–268. 105
- Haas, M. E., Pirruccello, J. P., Friedman, S. N., Emdin, C. A., Ajmera, V. H., Simon, T. G., Homburger, J. R., Guo, X., Budoff, M., Corey, K. E., et al. (2020). Machine learning enables new insights into clinical significance of and genetic contributions to liver fat accumulation. *medRxiv*. 41
- Hammon, M., Dankerl, P., Voit-Höhne, H. L., Sandmair, M., Kammerer, F. J., Uder, M., and Janka, R. (2014). Improving diagnostic accuracy in assessing pulmonary edema on bedside chest radiographs using a standardized scoring approach. *BMC anesthesiology*, 14(1):1–9. 185
- Han, X. (2017). Mr-based synthetic CT generation using a deep convolutional neural network method. *Medical physics*, 44(4):1408–1419. 31
- Han, Z., Wei, B., Leung, S., Chung, J., and Li, S. (2018). Towards automatic report generation in spine radiology using weakly supervised framework. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 185–193. Springer. 107
- Han, Z., Wei, B., Zheng, Y., Yin, Y., Li, K., and Li, S. (2017). Breast cancer multi-classification from histopathological images with structured deep learning model. *Scientific reports*, 7(1):1–10. 31
- Hashimoto, F., Kakimoto, A., Ota, N., Ito, S., and Nishizawa, S. (2019). Automated segmentation of 2d low-dose CT images of the psoas-major muscle using deep convolutional neural networks. *Radiological physics and technology*, 12(2):210–215. 40, 48, 50
- Hays, J. and Efros, A. A. (2008). IM2GPS: Estimating geographic information from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE. 155

- He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969. 136, 139, 140
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778. 52, 53, 115, 136
- Healthcare, G. (2019). One of the largest ai platforms in healthcare is one you’ve never heard of, until now. 29
- Henzler, P., Rasche, V., Ropinski, T., and Ritschel, T. (2018). Single-image tomography: 3d volumes from 2d cranial x-rays. In *Computer Graphics Forum*, volume 37, pages 377–388. Wiley Online Library. 132
- Hesterberg, T. (1995). Weighted average importance sampling and defensive mixture distributions. *Technometrics*, 37(2):185–194. 155, 165
- Hiasa, Y., Otake, Y., Takao, M., Ogawa, T., Sugano, N., and Sato, Y. (2019). Automated muscle segmentation from clinical CT using Bayesian U-net for personalized musculoskeletal modeling. *IEEE transactions on medical imaging*, 39(4):1030–1040. 184
- Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., and Aerts, H. J. (2018). Artificial intelligence in radiology. *Nature Reviews Cancer*, 18(8):500–510. 30
- Hsieh, K., Phanishayee, A., Mutlu, O., and Gibbons, P. B. (2019). The non-IID data quagmire of decentralized machine learning. *arXiv preprint arXiv:1910.00189*. 166
- Hsu, T.-M. H. (2020). Automatic longitudinal assessment of tumor responses. Master’s thesis, Massachusetts Institute of Technology. 5, 31
- Hsu, T. M. H., Chen, W. Y., Hou, C.-A., Tsai, Y.-H. H., Yeh, Y.-R., and Wang, Y.-C. F. (2015). Unsupervised domain adaptation with imbalanced cross-domain data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4121–4129. 34
- Hsu, T.-M. H., Qi, H., and Brown, M. (2019). Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*. 36, 151, 154, 163, 165, 169
- Hsu, T.-M. H., Qi, H., and Brown, M. (2020). Federated visual classification with real-world data distribution. In *European Conference on Computer Vision*, pages 76–92. Springer. 36, 151
- Hsu, T.-M. H., Schawkat, K., Berkowitz, S. J., Wei, J. L., Makoyeva, A., Legare, K., DeCicco, C., Paez, S. N., Wu, J. S., Szolovits, P., et al. (2021). Artificial intelligence to assess body composition on routine abdominal CT scans and predict mortality in pancreatic cancer—a recipe for your local application. *European Journal of Radiology*, 142:109834. 31, 34, 47, 67, 69

- Hsu, T.-M. H. and Wang, Y.-C. C. (2021). DeepOPG: Improving orthopantomogram finding summarization with weak supervision. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 366–376. Springer. [31](#), [35](#), [131](#)
- Hsu, T.-M. H., Weng, W.-H., Boag, W., McDermott, M., and Szolovits, P. (2018). Un-supervised multimodal representation learning across medical images and reports. *arXiv preprint arXiv:1811.08615*. [35](#), [87](#), [105](#)
- Hu, S.-Y., Wang, S., Weng, W.-H., Wang, J., Wang, X., Ozturk, A., Li, Q., Kumar, V., and Samir, A. E. (2020). Self-supervised pretraining with dicom metadata in ultrasound imaging. In *Machine Learning for Healthcare Conference*, pages 732–749. PMLR. [43](#)
- Huang, X., Alleva, F., Hon, H.-W., Hwang, M.-Y., Lee, K.-F., and Rosenfeld, R. (1993). The sphinx-ii speech recognition system: an overview. *Computer Speech & Language*, 7(2):137–148. [105](#)
- Huang, Y., Huang, B., Kan, T., Yang, B., Yuan, M., and Wang, J. (2014). Liver-to-spleen ratio as an index of chronic liver diseases and safety of hepatectomy: a pilot study. *World journal of surgery*, 38(12):3186–3192. [41](#)
- Iandola, F., Moskewicz, M., Karayev, S., Girshick, R., Darrell, T., and Keutzer, K. (2014). Densenet: Implementing efficient convnet descriptor pyramids. *arXiv preprint arXiv:1404.1869*. [114](#)
- Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghighi, B., Ball, R., Shpankaya, K., et al. (2019). Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *arXiv preprint arXiv:1901.07031*. [22](#), [100](#), [101](#), [104](#), [112](#), [116](#), [118](#)
- Iyyer, M., Manjunatha, V., Boyd-Graber, J., and Daumé III, H. (2015). Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1681–1691. [91](#)
- Jader, G., Fontineli, J., Ruiz, M., Abdalla, K., Pithon, M., and Oliveira, L. (2018). Deep instance segmentation of teeth in panoramic x-ray images. In *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 400–407. IEEE. [133](#), [147](#), [148](#)
- Järvelin, K. and Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446. [93](#)
- Javaid, U., Dasnoy, D., and Lee, J. A. (2018). Multi-organ segmentation of chest CT images in radiation oncology: comparison of standard and dilated unet. In *International conference on advanced concepts for intelligent vision systems*, pages 188–199. Springer. [31](#)

- Javaid, U., Souris, K., Dasnoy, D., Huang, S., and Lee, J. A. (2019). Mitigating inherent noise in monte carlo dose distributions using dilated u-net. *Medical Physics*, 46(12):5790–5798. 31
- Jiang, Z., Chen, T., Chen, T., and Wang, Z. (2020). Robust pre-training by adversarial contrastive learning. *Advances in Neural Information Processing Systems*, 33:16199–16210. 185
- Jing, B., Xie, P., and Xing, E. (2017). On the automatic generation of medical imaging reports. *arXiv preprint arXiv:1711.08195*. 105
- Johnson, A., Pollard, T., Mark, R., Berkowitz, S., and Horng, S. (2019). Mimic-cxr database. *PhysioNet10*, 13026:C2JT1Q. 22, 88, 89, 90, 100, 101, 104, 117, 118
- Joskowicz, L., Cohen, D., Caplan, N., and Sosna, J. (2018). Automatic segmentation variability estimation with segmentation priors. *Medical image analysis*, 50:54–64. 184
- Jungo, A. and Reyes, M. (2019). Assessing reliability and challenges of uncertainty estimations for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 48–56. Springer. 184
- Kahn, H. and Marshall, A. W. (1953). Methods of reducing sample size in Monte Carlo computations. *Journal of the Operations Research Society of America*, 1(5):263–278. 155
- Kahn Jr, C. E., Langlotz, C. P., Burnside, E. S., Carrino, J. A., Channin, D. S., Hovsepian, D. M., and Rubin, D. L. (2009). Toward best practices in radiology reporting. *Radiology*, 252(3):852–856. 102
- Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. (2019). Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*. 153, 161
- Kalender, W. A. (2005). Ct: the unexpected evolution of an imaging modality. *European Radiology Supplements*, 15(4):d21–d24. 32
- Kamarajah, S. K., Bundred, J., and Tan, B. H. (2019). Body composition assessment and sarcopenia in patients with gastric cancer: a systematic review and meta-analysis. *Gastric Cancer*, 22(1):10–22. 49, 62
- Karimi, D., Zeng, Q., Mathur, P., Avinash, A., Mahdavi, S., Spadinger, I., Abolmaesumi, P., and Salcudean, S. E. (2019). Accurate and robust deep learning-based segmentation of the prostate clinical target volume in ultrasound images. *Medical image analysis*, 57:186–196. 184

- Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S. J., Stich, S. U., and Suresh, A. T. (2019). Scaffold: Stochastic controlled averaging for on-device federated learning. *arXiv preprint arXiv:1910.06378*. 155
- Kazemifar, S., McGuire, S., Timmerman, R., Wardak, Z., Nguyen, D., Park, Y., Jiang, S., and Owringi, A. (2019). Mri-only brain radiotherapy: Assessing the dosimetric accuracy of synthetic CT images generated using a deep learning approach. *Radiotherapy and Oncology*, 136:56–63. 31
- Kilickaya, M., Erdem, A., Ikizler-Cinbis, N., and Erdem, E. (2016). Re-evaluating automatic metrics for image captioning. *arXiv preprint arXiv:1612.07600*. 127
- Kim, C., Kim, D., Jeong, H., Yoon, S.-J., and Youm, S. (2020). Automatic tooth detection and numbering using a combination of a cnn and heuristic algorithm. *Applied Sciences*, 10(16):5624. 133, 147, 148
- Kim, Y. W. and Mansfield, L. T. (2014). Fool me twice: delayed diagnoses in radiology with emphasis on perpetuated errors. *AJR Am J Roentgenol*, 202(3):465–470. 30
- Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. 114, 116, 117, 140
- Klambauer, G., Unterthiner, T., Mayr, A., and Hochreiter, S. (2017). Self-normalizing neural networks. In *Advances in Neural Information Processing Systems*, pages 971–980. 92
- Koch, G., Zemel, R., Salakhutdinov, R., et al. (2015). Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, page 0. Lille. 185
- Koch, T. L., Perslev, M., Igel, C., and Brandt, S. S. (2019). Accurate segmentation of dental panoramic radiographs with u-nets. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 15–19. IEEE. 133
- Krause, J., Johnson, J., Krishna, R., and Fei-Fei, L. (2017). A hierarchical approach for generating descriptive image paragraphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 317–325. 109
- Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images. Technical report, Citeseer. 152, 160
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105. 30
- Kuk, J. L., Katzmarzyk, P. T., Nichaman, M. Z., Church, T. S., Blair, S. N., and Ross, R. (2006). Visceral fat is an independent predictor of all-cause mortality in men. *Obesity*, 14(2):336–341. 65, 67

- Kweon, H. H.-I., Lee, J.-H., Youk, T.-m., Lee, B.-A., and Kim, Y.-T. (2018). Panoramic radiography can be an effective diagnostic tool adjunctive to oral examinations in the national health checkup program. *Journal of periodontal & implant science*, 48(5):317–325. 132
- Lagasse, J. (2018). FDA approves first AI tool for detecting retinopathy, NIH shows machine learning success in imaging | Healthcare Finance News. *Healthcare Finance*. 88
- Landi, H. (2016). Ibm unveils watson-powered imaging solutions at rsna. *RSNA*. 29
- Leaman, R., Islamaj Doğan, R., and Lu, Z. (2013). DNorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, 29(22):2909–2917. 105
- Leaman, R., Khare, R., and Lu, Z. (2015). Challenges in clinical natural language processing for automated disorder normalization. *Journal of biomedical informatics*, 57:28–37. 22, 104
- LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324. 166
- Lee, C.-G. and Ma, Z. (2004). The generalized quadratic assignment problem. *Research Rep., Dept., Mechanical Industrial Eng., Univ. Toronto, Canada*, page M5S. 137
- Lee, D. H. and Giovannucci, E. L. (2018). Body composition and mortality in the general population: A review of epidemiologic studies. *Experimental Biology and Medicine*, 243(17-18):1275–1285. 49, 62
- Lee, D.-W., Kim, S.-Y., Jeong, S.-N., and Lee, J.-H. (2021a). Artificial intelligence in fractured dental implant detection and classification: evaluation using dataset from two dental hospitals. *Diagnostics*, 11(2):233. 31
- Lee, H., Chai, Y. J., Joo, H., Lee, K., Hwang, J. Y., Kim, S.-M., Kim, K., Nam, I.-C., Choi, J. Y., Yu, H. W., et al. (2021b). Federated learning for thyroid ultrasound image analysis to protect personal information: Validation study in a real health care environment. *JMIR medical informatics*, 9(5):e25869. 45, 151
- Li, H., Han, H., Li, Z., Wang, L., Wu, Z., Lu, J., and Zhou, S. K. (2020a). High-resolution chest x-ray bone suppression using unpaired CT structural priors. *IEEE transactions on medical imaging*, 39(10):3053–3063. 41
- Li, J., Monroe, W., Ritter, A., Galley, M., Gao, J., and Jurafsky, D. (2016a). Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*. 106
- Li, M. D., Chang, K., Bearce, B., Chang, C. Y., Huang, A. J., Campbell, J. P., Brown, J. M., Singh, P., Hoebel, K. V., Erdoğmuş, D., et al. (2020b). Siamese neural networks for continuous disease severity evaluation and change detection in medical imaging. *NPJ digital medicine*, 3(1):1–9. 185



- Li, T., Sahu, A. K., Talwalkar, A., and Smith, V. (2019a). Federated learning: Challenges, methods, and future directions. *arXiv preprint arXiv:1908.07873*. 153, 161
- Li, W., Milletari, F., Xu, D., Rieke, N., Hancox, J., Zhu, W., Baust, M., Cheng, Y., Ourselin, S., Cardoso, M. J., et al. (2019b). Privacy-preserving federated brain tumour segmentation. In *International workshop on machine learning in medical imaging*, pages 133–141. Springer. 45, 151
- Li, W., Wang, G., Fidon, L., Ourselin, S., Cardoso, M. J., and Vercauteren, T. (2017). On the compactness, efficiency, and representation of 3d convolutional networks: brain parcellation as a pretext task. In *International conference on information processing in medical imaging*, pages 348–360. Springer. 184
- Li, Y., Jiang, X., Wang, S., Xiong, H., and Ohno-Machado, L. (2016b). Vertical grid logistic regression (vertigo). *Journal of the American Medical Informatics Association*, 23(3):570–579. 45, 153
- Li, Y., Liang, X., Hu, Z., and Xing, E. P. (2018). Hybrid retrieval-generation reinforced agent for medical image report generation. *Advances in neural information processing systems*, 31. 42, 43, 105, 107, 112, 128
- Li, Z., Li, H., Han, H., Shi, G., Wang, J., and Zhou, S. K. (2019c). Encoding CT anatomy knowledge for unpaired chest x-ray image decomposition. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 275–283. Springer. 41
- Li, Z. and Xia, Y. (2020). Deep reinforcement learning for weakly-supervised lymph node segmentation in CT images. *IEEE Journal of Biomedical and Health Informatics*, 25(3):774–783. 31
- Liao, X., Li, W., Xu, Q., Wang, X., Jin, B., Zhang, X., Wang, Y., and Zhang, Y. (2020). Iteratively-refined interactive 3d medical image segmentation with multi-agent reinforcement learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9394–9402. 43, 131
- Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*. 101, 111, 118
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014a). Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer. 41
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P., and Zitnick, L. (2014b). Microsoft coco: Common objects in context. In *ECCV. European Conference on Computer Vision*. 106
- Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., and Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88. 103

- Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons. 183
- Liu, C.-W., Lowe, R., Serban, I. V., Noseworthy, M., Charlin, L., and Pineau, J. (2016). How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*. 127
- Liu, G., Hsu, T.-M. H., McDermott, M., Boag, W., Weng, W.-H., Szolovits, P., and Ghassemi, M. (2019). Clinically accurate chest X-ray report generation. In *Machine Learning for Healthcare Conference*, pages 249–269. PMLR. 31, 35, 99
- Liu, P. J. (2018). Learning to write notes in electronic health records. *arXiv preprint arXiv:1808.02622*. 88
- Liu, S., Xu, D., Zhou, S. K., Pauly, O., Grbic, S., Mertelmeier, T., Wicklein, J., Jerebko, A., Cai, W., and Comaniciu, D. (2018). 3d anisotropic hybrid network: Transferring convolutional features from 2d images to 3d anisotropic volumes. In *International conference on medical image computing and computer-assisted intervention*, pages 851–858. Springer. 44
- Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*. 155
- Longmore, D. K., Miller, J. E., Bekkering, S., Saner, C., Mifsud, E., Zhu, Y., Safery, R., Nichol, A., Colditz, G., Short, K. R., et al. (2021). Diabetes and overweight/obesity are independent, nonadditive risk factors for in-hospital severity of COVID-19: an international, multicenter retrospective meta-analysis. *Diabetes Care*, 44(6):1281–1290. 80
- Lu, C.-L., Wang, S., Ji, Z., Wu, Y., Xiong, L., Jiang, X., and Ohno-Machado, L. (2015). Webdisco: a web service for distributed cox model learning without patient-level data sharing. *Journal of the American Medical Informatics Association*, 22(6):1212–1219. 44, 153
- Lu, J., Xiong, C., Parikh, D., and Socher, R. (2017). Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 375–383. 106, 110
- Luo, J., Wu, X., Luo, Y., Huang, A., Huang, Y., Liu, Y., and Yang, Q. (2019). Real-world image datasets for federated learning. *arXiv preprint arXiv:1910.11089*. 155
- Marlin, B. M., Zemel, R. S., Roweis, S. T., and Slaney, M. (2011). Recommender systems: missing data and statistical model estimation. In *Twenty-Second International Joint Conference on Artificial Intelligence*. 32
- Marr, B. (2017). First FDA approval for clinical cloud-based deep learning in health-care. *Forbes*. 88

- McDermott, M. B., Hsu, T. M. H., Weng, W.-H., Ghassemi, M., and Szolovits, P. (2020). Chexpert++: Approximating the chexpert labeler for speed, differentiability, and probabilistic output. In *Machine Learning for Healthcare Conference*, pages 913–927. PMLR. 31
- McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. 45, 151, 153, 154, 161, 165
- McMahan, H. B., Ramage, D., Talwar, K., and Zhang, L. (2018). Learning differentially private recurrent language models. In *International Conference on Learning Representations (ICLR)*. 153
- Meeker, D., Jiang, X., Matheny, M. E., Farcas, C., D’Arcy, M., Pearlman, L., Nookala, L., Day, M. E., Kim, K. K., Kim, H., et al. (2015). A system to build distributed multivariate models and manage disparate data sharing policies: implementation in the scalable national network for effectiveness research. *Journal of the American Medical Informatics Association*, 22(6):1187–1195. 44, 153
- Miki, Y., Muramatsu, C., Hayashi, T., Zhou, X., Hara, T., Katsumata, A., and Fujita, H. (2017). Classification of teeth in cone-beam CT using deep convolutional neural network. *Computers in biology and medicine*, 80:24–29. 133
- Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. (2013). Playing Atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*. 43
- Mohan, K., Pearl, J., and Tian, J. (2013). Graphical models for inference with missing data. *Advances in neural information processing systems*, 26. 32
- Mor, N., Wolf, L., Polyak, A., and Taigman, Y. (2018). A universal music translation network. *arXiv preprint arXiv:1805.07848*. 89
- Moradi, M., Madani, A., Gur, Y., Guo, Y., and Syeda-Mahmood, T. (2018). Bimodal network architectures for automatic generation of image annotation from text. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 449–456. Springer. 103
- Moradi, S., Oghli, M. G., Alizadehasl, A., Shiri, I., Oveisi, N., Oveisi, M., Maleki, M., and Dhooge, J. (2019). Mfp-unet: A novel deep learning based approach for left ventricle segmentation in echocardiography. *Physica Medica*, 67:58–69. 31
- Mulugeta, G., Eckert, M. A., Vaden, K. I., Johnson, T. D., and Lawson, A. B. (2017). Methods for the analysis of missing data in fMRI studies. *Journal of biometrics & biostatistics*, 8(1). 183
- Neeland, I. J., Poirier, P., and Després, J.-P. (2018). Cardiovascular and metabolic heterogeneity of obesity: clinical challenges and implications for management. *Circulation*, 137(13):1391–1406. 47, 49

- Neeland, I. J., Ross, R., Després, J.-P., Matsuzawa, Y., Yamashita, S., Shai, I., Seidell, J., Magni, P., Santos, R. D., Arsenault, B., et al. (2019). Visceral and ectopic fat, atherosclerosis, and cardiometabolic disease: a position statement. *The lancet Diabetes & endocrinology*, 7(9):715–725. 47, 49, 67, 80
- Ngiam, J., Peng, D., Vasudevan, V., Kornblith, S., Le, Q. V., and Pang, R. (2018). Domain adaptive transfer learning with specialist models. *arXiv preprint arXiv:1811.07056*. 155
- Nguyen, D., Jia, X., Sher, D., Lin, M.-H., Iqbal, Z., Liu, H., and Jiang, S. (2019). 3d radiotherapy dose prediction on head and neck cancer patients with a hierarchically densely connected u-net deep learning architecture. *Physics in medicine & Biology*, 64(6):065020. 31
- Nicklas, B. J., Penninx, B. W., Ryan, A. S., Berman, D. M., Lynch, N. A., and Dennis, K. E. (2003). Visceral adipose tissue cutoffs associated with metabolic risk factors for coronary heart disease in women. *Diabetes care*, 26(5):1413–1420. 70
- Nougaret, S., Reinhold, C., Mikhael, H. W., Rouanet, P., Bibeau, F., and Brown, G. (2013). The use of mr imaging in treatment planning for patients with rectal carcinoma: have you checked the “distance”? *Radiology*, 268(2):330–344. 31
- Oh, J.-H., Kim, H.-G., Lee, K. M., Ryu, C.-W., Park, S., Jang, J. H., Choi, H. S., and Kim, E. J. (2021). Effective end-to-end deep learning process in medical imaging using independent task learning: Application for diagnosis of maxillary sinusitis. *Yonsei medical journal*, 62(12):1125. 40
- Ohno-Machado, L., Agha, Z., Bell, D. S., Dahm, L., Day, M. E., Doctor, J. N., Gabriel, D., Kahlon, M. K., Kim, K. K., Hogarth, M., et al. (2014). pSCANNER: Patient-centered scalable national network for effectiveness research. *Journal of the American Medical Informatics Association*, 21(4):621–626. 153
- Ojima, Y., Harano, M., Sumitani, D., and Okajima, M. (2019). Impact of preoperative skeletal muscle mass and quality on the survival of elderly patients after curative resection of colorectal cancer. *Journal of the Anus, Rectum and Colon*, 3(4):143–151. 49
- Ozaki, K. and Yokoo, S. (2019). Large-scale landmark retrieval/recognition under a noisy and diverse dataset. *arXiv preprint arXiv:1906.04087*. 160
- Palaiodimos, L., Kokkinidis, D. G., Li, W., Karamanis, D., Ognibene, J., Arora, S., Southern, W. N., and Mantzoros, C. S. (2020). Severe obesity, increasing age and male sex are independently associated with worse in-hospital outcomes, and higher in-hospital mortality, in a cohort of patients with COVID-19 in the bronx, new york. *Metabolism*, 108:154262. 65, 67
- Pan, S. J., Tsang, I. W., Kwok, J. T., and Yang, Q. (2011). Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210. 89

- Paolacci, G., Chandler, J., and Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision making*, 5(5):411–419. [32](#)
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics. [101](#), [111](#), [116](#), [118](#)
- Park, H. J., Shin, Y., Park, J., Kim, H., Lee, I. S., Seo, D.-W., Huh, J., Lee, T. Y., Park, T., Lee, J., et al. (2020). Development and validation of a deep learning system for segmentation of abdominal muscle and fat on computed tomography. *Korean journal of radiology*, 21(1):88–100. [40](#), [48](#), [50](#)
- Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic differentiation in pytorch. [114](#), [116](#)
- Pati, S., Thakur, S. P., Bhalerao, M., Thermos, S., Baid, U., Gotkowski, K., Gonzalez, C., Guley, O., Hamamci, I. E., Er, S., et al. (2021). Gandlf: A generally nuanced deep learning framework for scalable end-to-end clinical workflows in medical imaging. *arXiv preprint arXiv:2103.01006*. [40](#)
- Peng, Y., Wang, X., Lu, L., Bagheri, M., Summers, R., and Lu, Z. (2018). NegBio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Summits on Translational Science Proceedings*, 2017:188. [118](#)
- Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543. [91](#)
- Perschbacher, S. (2012). Interpretation of panoramic radiographs. *Australian dental journal*, 57:40–45. [132](#)
- Petersen, A., Bressemer, K., Albrecht, J., Thieß, H.-M., Vahldiek, J., Hamm, B., Makowski, M. R., Niehues, A., Niehues, S. M., and Adams, L. C. (2020). The role of visceral adiposity in the severity of COVID-19: Highlights from a unicenter cross-sectional pilot study in Germany. *Metabolism*, 110:154317. [80](#), [82](#)
- Pickhardt, P., Graffy, P., Zea, R., Lee, S., Liu, J., Sandfort, V., et al. (2020). Opportunistic screening for metabolic syndrome in asymptomatic adults utilizing fully automated abdominal CT-based biomarkers. *AJR Am J Roentgenol*. [62](#)
- Pickhardt, P. J., Graffy, P. M., Zea, R., Lee, S. J., Liu, J., Sandfort, V., and Summers, R. M. (2021). Utilizing fully automated abdominal ct-based biomarkers for opportunistic screening for metabolic syndrome in adults without symptoms. *American Journal of Roentgenology*, 216(1):85–92. [40](#)

- Pickhardt, P. J., Jee, Y., O'Connor, S. D., and del Rio, A. M. (2012). Visceral adiposity and hepatic steatosis at abdominal CT: association with the metabolic syndrome. *American Journal of Roentgenology*, 198(5):1100–1107. 70
- Plessas, A., Nasser, M., Hanoach, Y., O'Brien, T., Delgado, M. B., and Moles, D. (2019). Impact of time pressure on dentists' diagnostic performance. *Journal of dentistry*, 82:38–44. 132
- Plis, S. M., Hjelm, D. R., Salakhutdinov, R., Allen, E. A., Bockholt, H. J., Long, J. D., Johnson, H. J., Paulsen, J. S., Turner, J. A., and Calhoun, V. D. (2014). Deep learning for neuroimaging: a validation study. *Frontiers in neuroscience*, 8:229. 41, 99
- Prado, C. M., Baracos, V. E., McCargar, L. J., Reiman, T., Mourtzakis, M., Tonkin, K., Mackey, J. R., Koski, S., Pituskin, E., and Sawyer, M. B. (2009). Sarcopenia as a determinant of chemotherapy toxicity and time to tumor progression in metastatic breast cancer patients receiving capecitabine treatment. *Clinical cancer research*, 15(8):2920–2926. 55
- Prince, J. L. and Links, J. M. (2006). *Medical imaging signals and systems*. Pearson Prentice Hall Upper Saddle River. 30
- Qin, C., Shi, B., Liao, R., Mansi, T., Rueckert, D., and Kamen, A. (2019). Unsupervised deformable registration for multi-modal images via disentangled representations. In *International Conference on Information Processing in Medical Imaging*, pages 249–261. Springer. 43
- Qin, T., Wang, Z., He, K., Shi, Y., Gao, Y., and Shen, D. (2020). Automatic data augmentation via deep reinforcement learning for effective kidney tumor segmentation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1419–1423. IEEE. 43, 131
- Quekel, L. G., Kessels, A. G., Goei, R., and van Engelshoven, J. M. (1999). Miss rate of lung cancer on the chest radiograph in clinical practice. *Chest*, 115(3):720–724. 30
- Quigley, K., Cha, M., Liao, R., Chauhan, G., Horng, S., Berkowitz, S., and Golland, P. (2022). RadTex: Learning efficient radiograph representations from text reports. *arXiv preprint arXiv:2208.03218*. 89
- Raghu, M., Blumer, K., Sayres, R., Obermeyer, Z., Kleinberg, B., Mullainathan, S., and Kleinberg, J. (2019a). Direct uncertainty prediction for medical second opinions. In *International Conference on Machine Learning*, pages 5281–5290. PMLR. 184
- Raghu, M., Zhang, C., Kleinberg, J., and Bengio, S. (2019b). Transfusion: Understanding transfer learning for medical imaging. *Advances in neural information processing systems*, 32. 44

- Rajeswar, S., Subramanian, S., Dutil, F., Pal, C., and Courville, A. (2017). Adversarial generation of natural language. *arXiv preprint arXiv:1705.10929*. 106
- Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., and Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 169(12):866–872. 183
- Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., et al. (2017). CheXnet: Radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv preprint arXiv:1711.05225*. 88, 91, 126
- Ramos, C., Augusto, J. C., and Shapiro, D. (2008). Ambient intelligence—the next step for artificial intelligence. *IEEE Intelligent Systems*, 23(2):15–18. 33
- Reed, S., Akata, Z., Lee, H., and Schiele, B. (2016a). Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 49–58. 89
- Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., and Lee, H. (2016b). Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*. 89
- Rehurek, R. and Sojka, P. (2010). Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Citeseer. 114
- Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., and Goel, V. (2017). Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024. 42, 107, 111, 115
- Ronneberger, O., Fischer, P., and Brox, T. (2015a). Dental x-ray image segmentation using a u-shaped deep convolutional network. In *International Symposium on Biomedical Imaging*. 133, 136
- Ronneberger, O., Fischer, P., and Brox, T. (2015b). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer. 53
- Rose, S. (2018). Machine learning for prediction in electronic health data. *JAMA network open*, 1(4):e181404–e181404. 183
- Roy, A. G., Ren, J., Azizi, S., Loh, A., Natarajan, V., Mustafa, B., Pawlowski, N., Freyberg, J., Liu, Y., Beaver, Z., et al. (2022). Does your dermatology classifier know what it doesn’t know? detecting the long-tail of unseen conditions. *Medical Image Analysis*, 75:102274. 33
- Rozylo-Kalinowska, I. (2018). Artificial intelligence in dentomaxillofacial radiology: Hype or future? *Journal of Oral and Maxillofacial Radiology*, 6(1):1–1. 132

- Rubin, G. D. (2015). Lung nodule and cancer detection in CT screening. *Journal of thoracic imaging*, 30(2):130. 103
- Rubin, J., Sanghavi, D., Zhao, C., Lee, K., Qadir, A., and Xu-Wilson, M. (2018). Large scale automated reading of frontal and lateral chest X-rays using dual convolutional neural networks. *arXiv preprint arXiv:1804.07839*. 89, 105
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252. 30
- Saerens, M., Latinne, P., and Decaestecker, C. (2002). Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural computation*, 14(1):21–41. 155
- Sahu, A. K., Li, T., Sanjabi, M., Zaheer, M., Talwalkar, A., and Smith, V. (2018). On the convergence of federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127*. 155
- Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4510–4520. 166
- Sattler, F., Wiedemann, S., Müller, K.-R., and Samek, W. (2019). Robust and communication-efficient federated learning from non-iid data. *IEEE transactions on neural networks and learning systems*, 31(9):3400–3413. 45, 154
- Sayer, A. A., Dennison, E. M., Syddall, H. E., Jameson, K., Martin, H. J., and Cooper, C. (2008). The developmental origins of sarcopenia: using peripheral quantitative computed tomography to assess muscle size in older people. *The Journals of gerontology series a: biological sciences and medical sciences*, 63(8):835–840. 47, 49
- Schlegl, T., Waldstein, S. M., Vogl, W.-D., Schmidt-Erfurth, U., and Langs, G. (2015). Predicting semantic descriptions from medical images with convolutional neural networks. In *International Conference on Information Processing in Medical Imaging*, pages 437–448. Springer. 103
- Schwartz, L. H., Panicek, D. M., Berk, A. R., Li, Y., and Hricak, H. (2011). Improving communication of diagnostic radiology findings through structured reporting. *Radiology*, 260(1):174–181. 102, 103
- Schweitzer, L., Geisler, C., Pourhassan, M., Braun, W., Glüer, C.-C., Bosy-Westphal, A., and Müller, M. J. (2015). What is the best reference site for a single MRI slice to assess whole-body skeletal muscle and adipose tissue volumes in healthy adults? *The American journal of clinical nutrition*, 102(1):58–65. 49, 67, 69, 70



- Schweitzer, L., Geisler, C., Pourhassan, M., Braun, W., Glüer, C.-C., Bosy-Westphal, A., and Müller, M. J. (2016). Estimation of skeletal muscle mass and visceral adipose tissue volume by a single magnetic resonance imaging slice in healthy elderly adults. *The Journal of nutrition*, 146(10):2143–2148. [49](#), [70](#)
- Sedai, S., Mahapatra, D., Hewavitharanage, S., Maetschke, S., and Garnavi, R. (2017). Semi-supervised segmentation of optic cup in retinal fundus images using variational autoencoder. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 75–82. Springer. [43](#)
- Seiglie, J., Platt, J., Cromer, S. J., Bunda, B., Foulkes, A. S., Bassett, I. V., Hsu, J., Meigs, J. B., Leong, A., Putman, M. S., et al. (2020). Diabetes as a risk factor for poor early outcomes in patients hospitalized with COVID-19. *Diabetes care*, 43(12):2938–2944. [68](#), [69](#), [80](#)
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626. [184](#)
- Shallue, C. J., Lee, J., Antognini, J., Sohl-Dickstein, J., Frostig, R., and Dahl, G. E. (2018). Measuring the effects of data parallelism on neural network training. *arXiv preprint arXiv:1811.03600*. [176](#)
- Sheller, M. J., Edwards, B., Reina, G. A., Martin, J., Pati, S., Kotrotsou, A., Milchenko, M., Xu, W., Marcus, D., Colen, R. R., et al. (2020). Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific reports*, 10(1):1–12. [45](#), [151](#)
- Sheller, M. J., Reina, G. A., Edwards, B., Martin, J., and Bakas, S. (2018). Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. In *International MICCAI Brainlesion Workshop*, pages 92–104. Springer. [45](#), [151](#)
- Shen, C., Gonzalez, Y., Chen, L., Jiang, S. B., and Jia, X. (2018). Intelligent parameter tuning in optimization-based iterative CT reconstruction via deep reinforcement learning. *IEEE transactions on medical imaging*, 37(6):1430–1439. [31](#)
- Shen, C., Gonzalez, Y., Klages, P., Qin, N., Jung, H., Chen, L., Nguyen, D., Jiang, S. B., and Jia, X. (2019). Intelligent inverse treatment planning via deep reinforcement learning, a proof-of-principle study in high dose-rate brachytherapy for cervical cancer. *Physics in Medicine & Biology*, 64(11):115013. [31](#)
- Shen, C., Nguyen, D., Chen, L., Gonzalez, Y., McBeth, R., Qin, N., Jiang, S. B., and Jia, X. (2020). Operating a treatment planning system using a deep-reinforcement learning-based virtual treatment planner for prostate cancer intensity-modulated radiation therapy treatment planning. *Medical physics*, 47(6):2329–2336. [31](#)

- Shen, T., Lei, T., Barzilay, R., and Jaakkola, T. (2017). Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems*, pages 6830–6841. 89
- Shin, H.-C., Lu, L., Kim, L., Seff, A., Yao, J., and Summers, R. M. (2015). Interleaved text/image deep mining on a very large-scale radiology database. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1090–1099. 105
- Shin, H.-C., Roberts, K., Lu, L., Demner-Fushman, D., Yao, J., and Summers, R. M. (2016). Learning to read chest X-rays: Recurrent neural cascade model for automated image annotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2497–2506. 88, 103
- Siewert, B., Sosna, J., McNamara, A., Raptopoulos, V., and Kruskal, J. B. (2008). Missed lesions at abdominal oncologic ct: lessons learned from quality assurance. *Radiographics*, 28(3):623–638. 30
- Silva, G., Oliveira, L., and Pithon, M. (2018). Automatic segmenting teeth in x-ray images: Trends, a novel data set, benchmarking and future perspectives. *Expert Systems with Applications*, 107:15–31. 133, 141
- Smith, S. L., Turban, D. H., Hamblin, S., and Hammerla, N. Y. (2017). Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *arXiv preprint arXiv:1702.03859*. 92
- Soin, A., Merkow, J., Long, J., Cohen, J. P., Saligrama, S., Kaiser, S., Borg, S., Tarapov, I., and Lungren, M. P. (2022). Chexstray: Real-time multi-modal data concordance for drift detection in medical imaging ai. *arXiv preprint arXiv:2202.02833*. 32
- Sonka, M., Fitzpatrick, J. M., et al. (2000). *Handbook of medical imaging. Volume 2, Medical image processing and analysis*. University of Iowa. 30
- Soueina, S. O., Far, B. H., Katsube, T., and Koono, Z. (1998). Mall: A multi-agent learning language for competitive and uncertain environments. *IEICE TRANSACTIONS on Information and Systems*, 81(12):1339–1349. 44
- Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. (2014). Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*. 184
- Stollenga, M. F., Byeon, W., Liwicki, M., and Schmidhuber, J. (2015). Parallel multi-dimensional lstm, with application to fast biomedical volumetric image segmentation. *Advances in neural information processing systems*, 28. 41, 99
- Strobelt, H., Gehrmann, S., and Rush, A. (2019). Giant language model test room. <http://gltr.io/dist/index.html>. 123

- Subbaswamy, A. and Saria, S. (2018). Counterfactual normalization: Proactively addressing dataset shift using causal mechanisms. In *UAI*, pages 947–957. 183
- Subbaswamy, A., Schulam, P., and Saria, S. (2019). Preventing failures due to dataset shift: Learning predictive models that transport. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3118–3127. PMLR. 183
- Suganyadevi, S., Seethalakshmi, V., and Balasamy, K. (2022). A review on deep learning in medical image analysis. *International Journal of Multimedia Information Retrieval*, 11(1):19–38. 30
- Suk, H.-I. and Shen, D. (2013). Deep learning-based feature representation for ad/mci classification. In *International conference on medical image computing and computer-assisted intervention*, pages 583–590. Springer. 41, 99
- Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112. 106
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9. 166
- Tan, N., Margolis, D. J., McClure, T. D., Thomas, A., Finley, D. S., Reiter, R. E., Huang, J., and Raman, S. S. (2012). Radical prostatectomy: value of prostate MRI in surgical planning. *Abdominal imaging*, 37(4):664–674. 31
- Tuzoff, D. V., Tuzova, L. N., Bornstein, M. M., Krasnov, A. S., Kharchenko, M. A., Nikolenko, S. I., Sveshnikov, M. M., and Bednenko, G. B. (2019). Tooth detection and numbering in panoramic radiographs using convolutional neural networks. *Dentomaxillofacial Radiology*, 48(4):20180051. 133, 147, 148
- Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. (2017). Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176. 88
- Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., and Belongie, S. (2018). The iNaturalist species classification and detection dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8769–8778. 152, 156, 157
- Van Horn, G. and Perona, P. (2017). The devil is in the tails: Fine-grained classification in the wild. *arXiv preprint arXiv:1709.01450*. 157
- Vedantam, R., Lawrence Zitnick, C., and Parikh, D. (2015). Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575. 101, 111, 115, 118

- Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164. 42, 106, 119
- Walter, M. (2018). Densitas gains FDA clearance for machine learning software that assesses breast density. *Radiology Business*. 88
- Wang, H., Zhou, Z., Li, Y., Chen, Z., Lu, P., Wang, W., Liu, W., and Yu, L. (2017a). Comparison of machine learning methods for classifying mediastinal lymph node metastasis of non-small cell lung cancer from 18f-fdg pet/ct images. *EJNMMI research*, 7(1):1–11. 31
- Wang, L., Xu, S., Wang, X., and Zhu, Q. (2021a). Addressing class imbalance in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10165–10173. 45
- Wang, X., Lu, L., Shin, H.-c., Kim, L., Nogues, I., Yao, J., and Summers, R. (2016). Unsupervised category discovery via looped deep pseudo-task optimization using a large scale radiology image database. *arXiv preprint arXiv:1603.07965*. 105
- Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, R. M. (2017b). ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 3462–3471. IEEE. 22, 88, 89, 104, 105
- Wang, X., Peng, Y., Lu, L., Lu, Z., and Summers, R. M. (2018). 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. 42, 88, 101, 105, 107, 115, 119
- Wang, X., Zhang, R., Shen, C., Kong, T., and Li, L. (2021b). Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3024–3033. 185
- Watanabe, M., Caruso, D., Tuccinardi, D., Risi, R., Zerunian, M., Polici, M., Pucciarelli, F., Tarallo, M., Strigari, L., Manfrini, S., et al. (2020). Visceral fat shows the strongest association with the need of intensive care in patients with COVID-19. *Metabolism*, 111:154319. 82
- Watts, J., Khojandi, A., Vasudevan, R., and Ramdhani, R. (2020). Optimizing individualized treatment planning for parkinson’s disease using deep reinforcement learning. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 5406–5409. IEEE. 31
- Weng, W.-H., Cai, Y., Lin, A., Tan, F., and Chen, P.-H. C. (2019). Multimodal multitask representation learning for pathology biobank metadata prediction. *arXiv preprint arXiv:1909.07846*. 43

- Weston, A. D., Korfiatis, P., Kline, T. L., Philbrick, K. A., Kostandy, P., Sakinis, T., Sugimoto, M., Takahashi, N., and Erickson, B. J. (2019). Automated abdominal segmentation of CT scans for body composition analysis using deep learning. *Radiology*, 290(3):669–679. 40, 48, 50, 64
- Weyand, T., Araujo, A., Cao, B., and Sim, J. (2020a). Google Landmarks Dataset v2 - a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. 156, 159, 160
- Weyand, T., Araujo, A., Cao, B., and Sim, J. (2020b). Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2575–2584. 152
- Weyand, T., Kostrikov, I., and Philbin, J. (2016). Planet-photo geolocation with convolutional neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 37–55. Springer. 159
- WHO (2000). Obesity: preventing and managing the global epidemic. 47
- Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256. 111, 138
- Winkel, D. J., Weikert, T. J., Breit, H.-C., Chabin, G., Gibson, E., Heye, T. J., Comaniciu, D., and Boll, D. T. (2020). Validation of a fully automated liver segmentation algorithm using multi-scale deep reinforcement learning and comparison versus manual segmentation. *European journal of radiology*, 126:108918. 31
- Wirtz, A., Mirashi, S. G., and Wesarg, S. (2018). Automatic teeth segmentation in panoramic x-ray images using a coupled shape model in combination with a neural network. In *International conference on medical image computing and computer-assisted intervention*, pages 712–719. Springer. 147, 148
- Wiseman, S., Shieber, S. M., and Rush, A. M. (2017). Challenges in data-to-document generation. *arXiv preprint arXiv:1707.08052*. 106
- Wu, D., Kim, K., Dong, B., and Li, Q. (2018). End-to-end abnormality detection in medical imaging. 40
- Wu, G., Kim, M., Wang, Q., Gao, Y., Liao, S., and Shen, D. (2013). Unsupervised deep feature learning for deformable registration of mr brain images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 649–656. Springer. 41, 99
- Wu, Y. and He, K. (2018). Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19. 166

- Xia, Y., Yang, D., Yu, Z., Liu, F., Cai, J., Yu, L., Zhu, Z., Xu, D., Yuille, A., and Roth, H. (2020). Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation. *Medical Image Analysis*, 65:101766. 184
- Xie, Z. (2017). Neural text generation: A practical guide. *arXiv preprint arXiv:1711.09534*. 106
- Xing, C., Wang, D., Liu, C., and Lin, Y. (2015). Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011. 92
- Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR. 42, 106, 119
- Yan, K., Wang, X., Lu, L., and Summers, R. M. (2018). Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of medical imaging*, 5(3):036501. 30
- Yang, D., Roth, H., Xu, Z., Milletari, F., Zhang, L., and Xu, D. (2019). Searching learning strategy with reinforcement learning for 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 3–11. Springer. 43, 131
- Yang, D., Zhang, S., Yan, Z., Tan, C., Li, K., and Metaxas, D. (2015). Automated anatomical landmark detection on distal femur surface using convolutional neural network. In *2015 IEEE 12th international symposium on biomedical imaging (ISBI)*, pages 17–21. IEEE. 41, 99
- Yang, M., Wang, X., Zhu, H., Wang, H., and Qian, H. (2021). Federated learning with class imbalance reduction. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pages 2174–2178. IEEE. 45
- Yang, Y., Ding, L., Zou, X., Shen, Y., Hu, D., Hu, X., Li, Z., and Kamel, I. R. (2020). Visceral adiposity and high intramuscular fat deposition independently predict critical illness in patients with SARS-CoV-2. *Obesity*, 28(11):2040–2048. 70, 81
- Yao, T., Pan, Y., Li, Y., Qiu, Z., and Mei, T. (2017). Boosting image captioning with attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4894–4902. 106
- Yurochkin, M., Agarwal, M., Ghosh, S., Greenewald, K., Hoang, N., and Khazaeni, Y. (2019). Bayesian nonparametric federated learning of neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 7252–7261. 154

- Yushkevich, P. A., Piven, J., Hazlett, H. C., Smith, R. G., Ho, S., Gee, J. C., and Gerig, G. (2006). User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *Neuroimage*, 31(3):1116–1128. 51
- Zhang, J., Wang, C., Sheng, Y., Palta, M., Czito, B., Willett, C., Zhang, J., Jensen, P. J., Yin, F.-F., Wu, Q., et al. (2021a). An interpretable planning bot for pancreas stereotactic body radiation therapy. *International Journal of Radiation Oncology\* Biology\* Physics*, 109(4):1076–1085. 31
- Zhang, K., Schölkopf, B., Muandet, K., and Wang, Z. (2013). Domain adaptation under target and conditional shift. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 819–827. 155
- Zhang, L., Shen, B., Barnawi, A., Xi, S., Kumar, N., and Wu, Y. (2021b). Feddpgan: federated differentially private generative adversarial networks framework for the detection of COVID-19 pneumonia. *Information Systems Frontiers*, 23(6):1403–1415. 45, 151
- Zhang, Y., Ding, D. Y., Qian, T., Manning, C. D., and Langlotz, C. P. (2018a). Learning to summarize radiology findings. *arXiv preprint arXiv:1809.04698*. 42, 107
- Zhang, Y., Somers, K. R., Becari, C., Polonis, K., Pfeifer, M. A., Allen, A. M., Kellogg, T. A., Covassin, N., and Singh, P. (2018b). Comparative expression of renin-angiotensin pathway proteins in visceral versus subcutaneous fat. *Frontiers in physiology*, 9:1370. 80
- Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., and Chandra, V. (2018). Federated learning with non-IID data. *arXiv preprint arXiv:1806.00582*. 45, 154
- Zhou, S. K., Greenspan, H., Davatzikos, C., Duncan, J. S., Van Ginneken, B., Madabhushi, A., Prince, J. L., Rueckert, D., and Summers, R. M. (2021). A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*. 30
- Zhou, S. K., Rueckert, D., and Fichtinger, G. (2019). *Handbook of medical image computing and computer assisted intervention*. Academic Press. 30