Learning Representations for Limited and Heterogeneous Medical Data

by

Wei-Hung Weng

M.D., Chang Gung University (2011)
M.M.Sc., Harvard University (2017)

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

Doctor of Philosophy in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2022

Author .....
Department of Electrical Engineering and Computer Science
May 13, 2022

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

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

Learning Representations for Limited and Heterogeneous Medical Data

by

Wei-Hung Weng

Submitted to the Department of Electrical Engineering and Computer Science
on May 13, 2022, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Electrical Engineering and Computer Science

Abstract

Data insufficiency and heterogeneity are challenges of representation learning for machine learning in medicine due to the diversity of medical data and the expense of data collection and annotation. To learn generalizable representations from such limited and heterogeneous medical data, we aim to utilize various learning paradigms to overcome the issue. In this dissertation, we systematically explore the machine learning frameworks for limited data, data imbalance, and heterogeneous data, using cross-domain learning, self-supervised learning, contrastive learning, meta-learning, multitask learning, and robust learning. We present studies with different medical applications, such as clinical language translation, ultrasound image classification and segmentation, medical image retrieval, skin diagnosis classification, pathology metadata prediction, and lung pathology prediction.

We first focus on the limited data problem, which is common in medical domains. We learn cross-domain representations for clinical language translation with limited and unpaired medical language corpora using unsupervised embedding space alignment with identical anchors for word translation, and conduct sentence translation using statistical language modeling. Using metrics of clinical correctness and readability, the developed method outperforms a dictionary-based algorithm in both word- and sentence-level translation. For learning better data representations of limited numbers of ultrasound images, we then adopt the self-supervised learning technique and integrate the corresponding metadata as a multimodal resource to introduce inductive biases. We find that the representations learned by the developed approach yield better downstream task performance, such as ultrasound image quality classification and organ segmentation, compared with the standard transfer learning methods.

Next, we zoom into the data imbalance problem. We explore the utility of contrastive learning, specifically the Siamese network, to learn representations from an imbalanced fundoscopic imaging dataset for diabetic retinopathy image retrieval. Compared with the standard supervised learning setup, we obtain comparable but interpretable results using the representations learned from the Siamese network. We also utilize meta-learning for skin disease classification with an extremely imbalanced long-tailed skin image dataset. We find that model ensemble with meta-learning models and models trained with conventional class

imbalance techniques yields better prediction performance, especially for rare skin diseases.

Finally, for heterogeneous medical data, we develop a multimodal multitask learning framework to learn a shared representation for pathology metadata prediction. We use the multimodal fusion technique to integrate the slide image, free text, and structured metadata, and adopt a multitask objective loss to introduce the inductive bias while learning. This yields better prediction power than the standard single-modal single-task training setup. We also apply robust training techniques to learn representations that can tackle a distributional shift across two chest X-ray datasets. Compared with standard training, we find that robust training provides better tolerance when the shift exists, and learns a robust representation for lung pathology prediction.

The investigation in this dissertation is not exhaustive but it introduces an extensive understanding of utilizing machine learning in helping clinical decision making under the limited and heterogeneous medical data setting. We also provide insights and caveats to motivate future research directions of machine learning with low-resource and high-dimensional medical data, and hope to make a positive real-world clinical impact.

Thesis Supervisor: Peter Szolovits
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

I want to express my acknowledgment to many people who significantly influenced me at the finish line of my PhD journey.

First and foremost, I would like to thank my research advisor and life mentor, Peter Szolovits, who provides me great freedom in pursuing my career, but always supports my decisions and helps me overcome difficulties with his insights and wisdom. I turned out to be one of those "few masochistic doctors who want to obtain a computer science PhD as well as their MD" that Pete mentioned in our earliest correspondence. Eventually, I finished my PhD in Computer Science at MIT with Pete! I am immensely grateful to have Pete as my advisor. Thank you, Pete, for having me as your student!

I am also grateful to my PhD thesis committee members John Guttag, Leo Celi, Po-Hsuan Chen, and my RQE committee, David Sontag and Jerry Sussman. With their knowledge, support and guidance, I can make this thesis happen.

My labmates in the MIT Clinical Decision-Making Group, a.k.a. MEDG, have been incredible during my PhD life. They gave me great opportunities for research collaboration with talented people. They are also so friendly that they provide me with a sense of belonging in this country. Thank you so much, Emily Alsentzer, Willie Boag, Geeticka Chauhan, Marzyeh Ghassemi, Harry Tzu-Ming Hsu, Eric Lehman, Di Jin, Matthew McDermott, Tiffany So Yeon Min, Tristan Naumann, and Elena Sergeeva, to share your thoughts and a great time with me.

My research collaborators across MIT, Harvard, and Google also turn me from an amateur into a real researcher in machine learning. Thank you, Yu-An Chung, Schrasing Tong, Szu-Yeu Hu, Richard Chen, Tsui-Wei Weng, Hongzi Mao, Ryo Uchimido, Yuannan Cai, Yuan Liu, Vivek Natarajan, Gamaleldin Elsayed, Jon Deaton, Tom Pollard, Alistair Johnson, Jesse Raffa, Jen-Tang Lu, Preethi Raghavan, and Brandon Westover, for your efficient collaborations and inspiring discussions. I hope to have more opportunities to collaborate with you in the future. I also thank Fern Keniston, Janet Fischer, and Leslie Kolodziejski for helping me with all the academic and administrative support.

I also thank my friends around Cambridge, Boston, and Taiwan, for their help and

company. You all make me feel at home and make my PhD life colorful.

My PhD journey was supported and sponsored by Rolf G. Locher Graduate Fellowship, MIT-IBM Watson AI Lab, Bayer AG, and MIT-Takeda Program. I appreciate them for the generous funding support.

Finally, I would like to thank my parents, Shu-Chen Chen and Huan-Chen Weng, for their love and mental support. Without them, I would not be here to pursue my vocation and dream.

Contents

List of Figures

List of Tables

Glossary

Abbreviation	Full Form
ACGAN	Auxiliary Classifier Generative Adversarial Network
AI	Artificial Intelligence
AUC-ROC	Area Under the Receiver Operator Characteristic Curve
BDI	Bilingual Dictionary Induction
BERT	Bidirectional Encoder Representations from Transformers
BI	Bias Initialization
CBMIR	Content-based Medical Image Retrieval
CBP	Compact Bilinear Pooling
CE	Cross Entropy
CHV	Consumer Health Vocabulary
CNN	Convolutional Neural Network
CSL	Conventional Supervised Learning
CSLS	Cross-Domain Similarity Local Scaling
CSP	Count Sketch Projection Function
CT	Computed Tomography
C&W	Carlini and Wagner Attacks
CXR	Chest X-ray
DICOM	Digital Imaging and Communications in Medicine
DL	Deep Learning
DNN	Deep Neural Network
DR	Diabetic Retinopathy

Abbreviation	Full Form
EaaS	Ecosystem as a Service
EHR	Electronic Health Record
ELMo	Embeddings from Language Models
FFPE	Formalin-fixed Paraffin-embedded Fixation
FFT	Fast Fourier Transformation
FGSM	Fast Gradient Sign Method
FL	Focal Loss
FSL	Few-shot Learning
GAN	Generative Adversarial Network
GCN	Graph Convolution Network
H&E	Hematoxylin and Eosin Staining
HIPAA	Health Insurance Portability and Accountability Act
IBP	Interval Bound Propagation
ICU	Intensive Care Unit
IFW	Inverse Frequency Weighting
IRB	Institutional Review Boards
ISIC	International Skin Imaging Collaboration
k-NN	k-Nearest Neighbor
LSTM	Long Short-Term Memory
MAML	Model-Agnostic Meta-Learning
MAP	Mean Average Precision
ML	Machine Learning
MOS	Mean Opinion Score
MRI	Magnetic Resonance Imaging
MRR	Mean Reciprocal Rank
MT	Machine Translation
MTL	Multitask Learning
NLP	Natural Language Processing
PGD	Projected Gradient Descent
QA	Question Answering

Abbreviation	Full Form
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Network
SCNN	Siamese Convolutional Neural Network
Seq2Seq	Sequence-to-sequence
SGD	Stochastic Gradient Descent
SHAP	Shapley Additive Explanations
SMT	Statistical Machine Translation
SVD	Singular Value Decomposition
TCGA	The Cancer Genome Atlas
t-SNE	t-Distributed Stochastic Neighbor Embedding
UMLS	Unified Medical Language System
VQA	Visual Question Answering

Chapter 1

Introduction

1.1 Learning Better Data Representations for Medical Machine Learning

Medicine has become one of the key applied machine learning research domains due to the increase of digitized healthcare data and the increasing power of computation [Charles et al., 2013, Topol, 2019]. Researchers have framed various medical and healthcare-related challenges as machine learning tasks and adopted various algorithms to tackle them with massive amounts of medical data [Topol, 2019, Weng, 2020]. Some examples of commonly-seen topics in machine learning for medicine and healthcare research include diagnosis support [Lipton et al., 2016, Choi et al., 2016b, Gulshan et al., 2016, Esteva et al., 2017], outcome and risk prediction [Ghassemi et al., 2014, Futoma et al., 2015, Choi et al., 2016a, Xiao et al., 2018, Girkar et al., 2018], patient phenotyping [Miotto et al., 2016, Baytas et al., 2017], optimal decision-making [Raghu et al., 2017, Weng et al., 2017a, Komorowski et al., 2018, Dalal et al., 2020], clinical question answering and text summarization [Jin et al., 2021, Rawat et al., 2020, Weng et al., 2020a, Alsentzer and Kim, 2018], and workflow improvement [Horng et al., 2017, Chen et al., 2019b]. Researchers have utilized various types of healthcare data to address these tasks, such as lab measurements [Pivovarov et al., 2015], claims data [Doshi-Velez et al., 2014, Pivovarov et al., 2015, Choi et al., 2016c], clinical narratives [Pivovarov et al., 2015, Weng et al., 2017b], medical images [Gulshan et al., 2016, Esteva et al., 2017, Bejnordi et al.,

2017, Liu et al., 2017, Poplin et al., 2018, Nagpal et al., 2019], and waveform signals [Lehman et al., 2018]. Many efforts use multiple such modalities of available data.

For medical and healthcare applications, it is critical to develop robust techniques that can not only yield excellent performance on given tasks but also provide efficiency, reliability, and explainability [Szolovits and Pauker, 1978, Szolovits, 1982, Jin et al., 2022b], to improve the likelihood of their practical clinical deployment [Chen et al., 2019c]. For example, applying an attention mechanism or interpretable models better explains the model behavior or the prediction [Bahdanau et al., 2014, Ribeiro et al., 2016, Lundberg and Lee, 2017]. Designing models with a robust optimization to tolerate adversarial examples improves the model reliability [Madry et al., 2018]. Preprocessing data appropriately and making better data representations for algorithms allow us to develop models with better performance and interpretability.

A good representation organizes the data so that machine learning algorithms can learn models with good performance [Weng and Szolovits, 2019]. It may also transform the data into a form that provides human interpretability given a suitable model design. For example, the radial domain folding algorithm, an unsupervised multivariate clustering method developed by Joshi and Szolovits [2012], abstracts the patient states and summarizes the patient's physiology from vital signs, laboratory tests, and clinical categorical data to a dense but rich representation using domain knowledge. The resulting model outperforms classical clinical scoring systems on the critical patient mortality prediction task while retaining the human understandability of the representation. A good representation may also be derived from high-dimensional, multimodal data sources [Suresh et al., 2017, Weng et al., 2019a, Ghassemi et al., 2020]. Suresh et al. [2017] preprocessed, transformed, and represented the raw data from different modalities (static variables such as demographics, time-varying variables like vital signs and labs, and clinical narrative notes) into a representation for clinical intervention prediction tasks. They transformed the clinical notes into a low-dimensional vector of topic distributions to preserve the human interpretability of the representation. Therefore, having appropriate representations is essential for modeling since it provides the structural organization of the data in both a machine and human-understandable language [Bengio et al., 2013].

1.2 Challenges of using Medical Data

Even though modern machine learning methods, such as deep learning techniques, are promising as potential approaches to tackle various medical problems with little or no feature engineering, they are data-hungry [Topol, 2019]. These techniques usually rely on a significant amount of high-quality and high-fidelity hand-labeled training data for clinical applications. However, data insufficiency and heterogeneity are usually obstacles while learning better data representations from medical data via machine learning algorithms.

1.2.1 Limited Data Problem

It is sometimes infeasible to have enough high-quality medical data for applying machine learning techniques due to the expense of data collection and expert annotation [Ma et al., 2019]. Limited data availability can further result in a class imbalance problem. It is a very common challenge in machine learning for medicine since we usually have a relatively small number of cases with the disease compared with healthy or normal cases [Johnson and Khoshgoftaar, 2019]. We may also have a very skewed class distribution for some medical domains, such as dermatology [Liu et al., 2020b].

The challenge of digitizing medical data has been mitigated after the wide adoption of electronic health records (EHR) in our healthcare systems. However, collecting the appropriately curated high-quality, task-specific medical data is still challenging due to the protection of patient data law, i.e., Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule (Privacy Rule). It is also time-consuming and complicated to acquire raw data from hospitals, insurance companies, and the government, since approval from Institutional Review Boards (IRB) is required. Furthermore, some organizations often believe that they can extract values from their own data rather than allowing the data to be used by external organizations and competitors.

After having the raw data, the high-quality labeled data (annotations) are even harder to obtain since task-specific annotations require domain experts to annotate instances manually, and the expert time can be costly [McDermott et al., 2020]. The process is usually time and cost-intensive, and sometimes further adjudication is needed for reliable labeling [Chen

et al., 2019c].

Researchers in the machine learning community have developed various approaches to address the limited resource problem. For instance, data augmentation [Chen et al., 2020b], weakly-supervised and self-supervised learning [Chen et al., 2020b], transfer learning with pre-trained representations [Raghu et al., 2019], multimodal learning [Baltrušaitis et al., 2018], multitask learning [Ruder, 2017], or making constraints by leveraging prior knowledge [Che et al., 2015], are all popular and critical techniques in the context of learning better representations under limited data settings.

Here, we take transfer learning as an example to demonstrate how researchers utilize these techniques to approach domain-specific and data-limited scenarios, such as machine learning for medical problems. In the field of natural language processing (NLP), well-learned latent representations of larger corpora can serve as general pre-trained language models, such as Embeddings from Language Models (ELMo) and Bidirectional Encoder Representations from Transformers (BERT) [Peters et al., 2018, Devlin et al., 2019], for transfer learning across different machine learning tasks. We can fine-tune the model by starting with pre-trained language models trained on vast general-purpose corpora, and then incrementally fine-tuning these models using the typically much smaller data sets available from corpora such as clinical notes.

In clinical NLP, researchers have applied pre-trained ELMo to medical texts for de-identification and other clinical NLP benchmark tasks [Khin et al., 2018, Zhu et al., 2018, Uzuner et al., 2011]. Lee et al. [2019] released the BioBERT model, which is trained on a general domain corpus and fine-tuned on biomedical text such as PubMed. Alsentzer et al. [2019] took one more step toward EHR by pre-training clinically oriented BERT models with clinical notes in the MIMIC-III (Medical Information Mart for Intensive Care) database [Johnson et al., 2016], either all notes or focusing on discharge summaries, on top of BERT and BioBERT models, and demonstrating that the specialized clinical BERT models outperformed others in the clinical NLP tasks. Huang et al. [2020] also developed a clinical BERT model by fine-tuning the BERT model on EHR for the hospital readmission task. Such improvement in specific clinical tasks may result from the difference in linguistic features between general, biomedical, and clinical narratives. Si et al. [2019] investigated the

capability of a traditional word- or subword-level approach, e.g., word2vec, GloVe, fastText, and the contextualized methods like ELMo and BERT on a clinical concept extraction task and demonstrated that the contextualized practices achieve better performance on various benchmark tests in the i2b2 and SemEval datasets.

1.2.2 Data Heterogeneity Problem

Medical data are also known to be multi-source and multimodal [Ghassemi et al., 2020]. For example, individual patient data can include structured demographic, laboratory test data, vital sign data, clinical narratives, various kinds of medical imaging modalities such as chest X-ray, CT (computed tomography) and MRI (magnetic resonance imaging), genomic information, or even audio data [Topol, 2019]. To better utilize these information types to approach medical questions like clinical experts, it is critical to understand how to integrate the heterogeneity across data due to vastly different statistical properties and varying levels of noise inside data, in order to learn useful representations that can be generalizable to more unseen and unlabeled medical data.

The main idea of integrating heterogeneous data is to learn their representations in a common embedding space [Baltrušaitis et al., 2018]. We may achieve this goal via learning coordinated or joint representations, where we can learn it via cross-domain learning and multimodal learning, respectively. In cross-domain learning, we use data from different domains in the same data modality to learn a more general but useful coordinated representation. In multimodal learning, we instead use completely different data modalities to bias the learned language representation (e.g., visual question answering problem). The goal of both cross-domain and multimodal learning is to learn a data representation capturing both shared and independent information from data in different distributions [Baltrušaitis et al., 2018].

There have been many successes of cross-domain or multimodal learning in different tasks such as visual question answering (VQA) [Fukui et al., 2016, Kim et al., 2017, 2018, Gamper and Rajpoot, 2021], sentiment analysis [Zadeh et al., 2017], and survival analysis [Mobadersany et al., 2018, Chen et al., 2020a, 2021b]. Weng and Szolovits [2018] and Weng et al. [2019b] applied the bilingual dictionary induction algorithm to align two natural language

embeddings of different clinical language styles [Conneau et al., 2018], which are independently trained on non-parallel corpora and performed cross-domain professional to consumer clinical language translation. Such a framework has proven effective even in a cross-modal setting between speech and text corpora [Chung et al., 2018, 2019]. Mobadersany et al. [2018], Chen et al. [2020a] and Chen et al. [2021a] used different multimodal learning algorithms, from vector concatenation, Kronecker Product fusion, to the co-attention transformer in order to integrate pathology whole slide images and genomic-based features for cancer survival outcome prediction. Researchers also learned the multimodal representation between image and text for a similar report retrieval task [Hsu et al., 2018], as well as for text generation from image input [Liu et al., 2019a].

1.3 Contributions

To tackle the challenges of applying machine learning techniques to medical data, in this dissertation, we explore different strategies to learn representations under such limited and heterogeneous data settings effectively.

The dissertation is intended to contribute to the medical machine learning community by exploring the possible approaches to the common problems in this field, which are data insufficiency and heterogeneity, and providing insights and caveats for adopting them.

The limited and heterogeneous data problem is pervasive in such domain-specific machine learning. We demonstrate several examples of machine learning for medical applications by investigating various methods and learning paradigms, such as cross-domain multimodal learning, contrastive learning, self-supervised learning, meta-learning, multimodal multitask learning, and robust training, to thoroughly utilize the medical data for learning better representations, and achieve better model performance and interpretability. Overall, the rationales of method selection and contributions of each case study of the thesis are as follows:

- Clinical language translation is critical for reducing misunderstanding and miscommunication between clinicians and patients. Yet, it has the problem of unparalleled and very limited annotated data for developing the language translation model. We develop

the cross-domain learning approach, which is fully-unsupervised and statistics-based, to translate the professional medical language domain to layman-understandable language domain at a word- and sentence-level. The approach overcomes the unpaired data and limited annotation problems and yields better translation based on the newly designed metrics that consider both clinical correctness and readability [Weng et al., 2019b, Weng and Szolovits, 2018].

- Learning representations of ultrasound images can also be challenging due to limited data size. We use the self-supervised learning scheme, context encoder, to tackle the limited data problem, and further integrate its corresponding DICOM (Digital Imaging and Communications in Medicine) metadata for multimodal information integration to provide the inductive bias and improve the performance of various downstream tasks using the learned representations [Hu et al., 2020a,b].

- Class imbalance is one of the critical challenges in machine learning for medicine that results from limited data availability, especially for the tasks with enormous numbers of normal cases or with complicated label space. Learning representations for diabetic retinopathy image retrieval and skin disease prediction both share the class imbalance issue. We approach the problem using contrastive learning and meta-learning paradigms since these paradigms are known to tackle skewed, non-symmetric data distributions. With some caveats, we find the contrastive learning, more specifically, Siamese neural network, and the ensemble of meta-learning model and conventional class imbalance model can be helpful to learn a better representation for data with the class imbalance problem [Chung and Weng, 2017, Weng et al., 2020b].

- Pathology image metadata prediction helps establish and organize the archived data sources. However, heterogeneous data sources are a challenge while developing a prediction model. We develop a multimodal multitask learning framework with a multitask objective function to learn generalizable representations. The multimodal fusion integrates images, free texts, and structured data, and the multitask learning captures the interactions and utilizes the inductive bias from four metadata prediction tasks. The model improves the prediction performance compared with the standard single modal

single task framework [Weng et al., 2019a].

- Finally, we use the task of lung pathology classification across datasets to demonstrate how we tackle the dataset shift problem, which results from the data heterogeneity across datasets. We consider the adversarial and certified robust training techniques due to their capability and tolerance of noisy data and data perturbation. We find that the robustly-trained models obtain better adversarial accuracy and certified accuracy under the dataset shift setup compared with the standard, non-robust models [Weng and Weng, 2021].

1.4 Organization

In this dissertation, we develop and evaluate several strategies for learning generalizable data representations by achieving different medical machine learning tasks, in order to overcome the data insufficiency and heterogeneity issues in machine learning for medicine. The rest of this thesis is organized as follows. For the first four studies, we focus on limited data and class imbalance problems, and for the last two studies, we focus on data heterogeneity and dataset shift issues:

- Learning cross-domain representations using limited data via embeddings alignment with minimal supervision and unparalleled data for clinical language translation (chapter 2)

- Adopting self-supervised learning to learn representations with limited distant-labeled multimodal data for ultrasound image segmentation and classification (chapter 3)

- Maximizing the mutual information in the highly class imbalanced dataset using contrastive learning to improve diabetic retinopathy image retrieval performance (chapter 4)

- Utilizing the meta-learning framework for few-shot rare skin diagnosis classification with extremely unbalanced data (chapter 5)

- Learning multimodal representations using high-dimensional heterogeneous data by a multitask learning framework for pathology metadata prediction (chapter 6)
- Using robust learning techniques to tackle the dataset shift problem for chest X-ray diagnosis classification (chapter 7)
- Conclusion, where we summarize our findings and provide thoughts and outlook for future directions (chapter 8)

1.5 Publications

This thesis primarily relates to the following publications:

- **Weng, W. H.**, Chung, Y. A., & Szolovits, P. (2019). Unsupervised clinical language translation. KDD 2019.
- **Weng, W. H.**, & Szolovits, P. (2018). Mapping unparalleled clinical professional and consumer languages with embedding alignment. KDD 2018 Workshop on Machine Learning for Medicine and Healthcare.
- Hu, S. Y., Wang, S., **Weng, W. H.**, Wang, J., Wang, X., Ozturk, A., ... & Samir, A. E. (2020). Self-Supervised Pretraining with DICOM metadata in Ultrasound Imaging. MLHC 2020.
- Hu, S. Y., Wang, S., **Weng, W. H.**, Wang, J., Wang, X., Ozturk, A., ... & Samir, A. E. (2020). Weakly Supervised Context Encoder using DICOM metadata in Ultrasound Imaging. ICLR 2020 AI for Affordable Healthcare Workshop.
- **Weng, W. H.***, & Chung, Y. A.* (2017). Learning deep representations of medical images using siamese CNNs with application to content-based image retrieval. NeurIPS 2017 Machine Learning for Health Workshop & Medical Imaging meets NeurIPS 2017.
- **Weng, W. H.**, Deaton, J., Natarajan, V., Elsayed, G. F., & Liu, Y. (2020). Addressing the Real-world Class Imbalance Problem in Dermatology. NeurIPS 2020 Machine Learning for Health Workshop.
- **Weng, W. H.**, Cai, Y., Lin, A., Tan, F., & Chen, P. H. C. (2019). Multimodal multitask representation learning for pathology biobank metadata prediction. NeurIPS 2019 Machine Learning for Health Workshop.

*equally contributed

- **Weng, W. H.**, Szolovits, P., Weng, T. W. (2021). Preserving Model Robustness for Dataset Shift in Medical Imaging. Submitted to MLHC.

- **Weng, W. H.**, & Szolovits, P. (2019). Representation learning for electronic health records. arXiv preprint arXiv:1909.09248.

- **Weng, W. H.** (2020). Machine learning for clinical predictive analytics. Leveraging Data Science for Global Health.

While not directly related, the following articles have also been completed over the course of the PhD (ordered chronologically):

- **Weng, W. H.***, Jin, D.*, Sergeeva, E.*, Chauhan, G.*, & Szolovits, P. (2022). Explainable Deep Learning in Healthcare: A Methodological Survey from an Attribution View. WIREs Mechanisms of Disease (WSBM).

- Chen, R. J., Lu, M. Y., **Weng, W. H.**, Chen, T. Y., Williamson, D. F., Manz, T., ... & Mahmood, F. (2021). Multimodal Co-Attention Transformer for Survival Prediction in Gigapixel Whole Slide Images. ICCV 2021.

- Jin, D., Pan, E., Oufattole, N., **Weng, W. H.**, Fang, H., & Szolovits, P. (2021). What disease does this patient have? a large-scale open domain question answering dataset from medical exams. Applied Sciences, 11(14), 6421.

- Wei, M. L., Liao, Y. H., **Weng, W. H.**, Shih, Y. T., Sheen, Y. S., & Sun, C. K. (2021). A Study on Applying Slide-Free Label-Free Harmonic Generation Microscopy For Noninvasive Assessment of Melasma Treatments With Histopathological Parameters. IEEE Journal of Selected Topics in Quantum Electronics, 27(4), 1-10.

- Huang, S., Chen, L., **Weng, W. H.**, Wang, L., Cui, X., Feng, C., ... & Li, T. (2021). Artificial Intelligence Assisted Early Warning System for Acute Kidney Injury Driven by Multi-Center ICU Database. medRxiv, 2020-01.

- McDermott, M. B., Hsu, T. M. H., **Weng, W. H.**, Ghassemi, M., & Szolovits, P. (2020). Chexpert++: Approximating the chexpert labeler for speed, differentiability, and probabilistic output. MLHC 2020.

- Rawat, B. P. S., **Weng, W. H.**, Min, S. Y., Raghavan, P., & Szolovits, P. (2020). Entity-enriched neural models for clinical question answering. ACL 2020 BioNLP Workshop.

- Dalal, S., Hombal, V., **Weng, W. H.**, Mankovich, G., Mabotuwana, T., Hall, C. S., ... & Gunn, M. L. (2020). Determining follow-up imaging study using radiology reports. Journal of digital imaging, 33(1), 121-130.

- Amorim, E., Mo, S. S., Palacios, S., Ghassemi, M. M., **Weng, W. H.**, Cash, S. S., ... & Westover, M. B. (2020). Cost-effectiveness analysis of multimodal prognostication in cardiac arrest with EEG monitoring. Neurology, 95(5), e563-e575.

- Chen, Y. C., Li, X., Zhu, H., **Weng, W. H.**, Tan, X., Chen, Q., ... & Fan, X. (2020). Monitoring neuron activities and interactions with laser emissions. ACS Photonics, 7(8), 2182-2189.

- **Weng, W. H.**, Chung, Y. A., & Tong, S. (2020). Clinical Text Summarization with Syntax-Based Negation and Semantic Concept Identification. arXiv preprint arXiv:2003.00353.

- Mao, H., Negi, P., Narayan, A., Wang, H., Yang, J., Wang, H., ..., **Weng, W. H.** & Alizadeh, M. (2019). Park: An open platform for learning augmented computer systems. NeurIPS 2019.

- Ma, R., Chen, P. H. C., Li, G., **Weng, W. H.**, Lin, A., Gadepalli, K., & Cai, Y. (2019). Human-centric Metric for Accelerating Pathology Reports Annotation. NeurIPS 2019 Machine Learning for Health Workshop.

- Liu, G., Hsu, T. M. H., McDermott, M., Boag, W., **Weng, W. H.**, Szolovits, P., & Ghassemi, M. (2019). Clinically accurate chest x-ray report generation. MLHC 2019.

- Alsentzer, E., Murphy, J. R., Boag, W., **Weng, W. H.**, Jin, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical BERT embeddings. NAACL-HLT 2019 Clinical Natural Language Processing Workshop.

- Chung, Y. A., **Weng, W. H.**, Tong, S., & Glass, J. (2019, May). Towards unsupervised speech-to-text translation. ICASSP 2019.

- Hu, S. Y., **Weng, W. H.**, Lu, S. L., Cheng, Y. H., Xiao, F., Hsu, F. M., & Lu, J. T. (2019). Multimodal volume-aware detection and segmentation for brain metastases radiosurgery. MICCAI 2019 Workshop on Artificial Intelligence in Radiation Therapy.

- Girkar, U., Uchimido, R., Lehman, L. W. H., Szolovits, P., Celi, L., & **Weng, W. H.** (2019). Predicting Blood Pressure Response to Fluid Bolus Therapy Using Neural Networks with Clinical Interpretability. Circulation Research, 125(Suppl_1), A448-A448.

- Li, X., Qin, Y., Tan, X., Chen, Y. C., Chen, Q., **Weng, W. H.**, ... & Fan, X. (2019). Ultrasound modulated droplet lasers. ACS Photonics, 6(2), 531-537.

- Chen, Y. C., Li, X., Zhu, H., **Weng, W. H.**, Tan, X., Chen, Q., ... & Fan, X. (2019). Laser Recording of Subcellular Neuron Activities. bioRxiv, 584938.

- Chung, Y. A., **Weng, W. H.**, Tong, S., & Glass, J. (2018). Unsupervised cross-modal alignment of speech and text embedding spaces. NeurIPS 2018.

- Hsu, T. M. H., **Weng, W. H.**, Boag, W., McDermott, M., & Szolovits, P. (2018). Unsupervised multimodal representation learning across medical images and reports. NeurIPS 2018 Machine Learning for Health Workshop.

- Girkar, U. M., Uchimido, R., Lehman, L. W. H., Szolovits, P., Celi, L., & **Weng, W. H.** (2018). Predicting blood pressure response to fluid bolus therapy using attention-based neural networks for clinical interpretability. NeurIPS 2018 Machine Learning for Health Workshop.

- **Weng, W. H.**, Gao, M., He, Z., Yan, S., & Szolovits, P. (2017). Representation and reinforcement learning for personalized glycemic control in septic patients. NeurIPS 2017 Machine Learning for Health Workshop.

Chapter 2

Cross-domain Learning for Limited Data

2.1 Overview

As patients' access to their doctors' clinical notes becomes common, translating professional, clinical jargon to layperson-understandable language is essential to improve patient-clinician communication. Such translation yields better clinical outcomes by enhancing patients' understanding of their own health conditions, and thus improving patients' involvement in their own care. Existing research has used dictionary-based word replacement or definition insertion to approach the need. However, these methods are limited by expert curation, which is hard to scale and has trouble generalizing to unseen datasets that do not share an overlapping vocabulary. In contrast, we approach the clinical word and sentence translation problem completely unsupervised. We show that a framework using representation learning, bilingual dictionary induction (BDI) and statistical machine translation (MT) yields the best precision at 10 of 0.827 on professional-to-consumer word translation, and mean opinion score (MOS) of 4.10 and 4.28 out of 5 for clinical correctness and layperson readability, respectively, on sentence translation. Our fully-unsupervised strategy overcomes the curation problem, and the clinically meaningful evaluation reduces biases from inappropriate evaluators, which are critical in clinical machine learning.

2.2 Background

Effective patient-clinician communication yields better clinical outcomes by enhancing patients' understanding of their own health conditions and participation in their own care [Ross and Lin, 2003]. Patient-clinician communication happens not only during in-person clinical visits but also through health records sharing. However, the records often contain professional jargon and abbreviations that limit their efficacy as a form of communication. Statistics show that only 12% of adults are proficient in clinical language, and most consumers can't understand commonly used clinical terms in their health records [Lalor et al., 2018]. For example, the sentence "On floor pt found to be hypoxic on O2 4LNC O2 sats 85 %, CXR c/w pulm edema, she was given 40mg IV x 2, nebs, and put on a NRB with improvement in O2 Sats to 95 %" is easy for a trained clinician to understand, yet would not be obvious to typical healthcare consumers, normally patients and their families.

Clinicians usually provide discharge instructions in consumer-understandable language while discharging patients. Yet these instructions include very limited information, which does not well represent the patient's clinical status, history, or expectations of disease progression or resolution. Thus the consumers may not obtain needed information only from these materials. To understand more about their clinical conditions for further decision making—for example, seeking a second opinion about treatment plans—it is necessary to dive into the other sections of a discharge summary, which are written in professional language. However, consumers may have difficulty clearly understanding domain-specific details written in professional language without domain knowledge and training. Such poor understanding can cause anxiety, confusion, and fear about unknown domain knowledge [Giardina and Singh, 2011], and further result in poor clinical outcomes [Sudore et al., 2006]. Thus, translating clinical professional to consumer-understandable language is essential to improve clinician-consumer communication and assist consumers' decision making and awareness of their illness.

Traditionally, clinicians need to specifically write down the consumer-understandable information in the notes to explain the domain-specific knowledge. Such a manual approach is acceptable for a small number of cases, but presents a burden for clinicians since

the process isn't scalable as patient loads increase. An appealing alternative is to perform automated translation. Researchers have attempted to map clinical professional to appropriate consumer-understandable words in clinical narratives using an expert-curated dictionary [Zielstorff, 2003, Zeng and Tse, 2006, Zeng-Treitler et al., 2007, Kandula et al., 2010], as well as pattern-based mining [Vydiswaran et al., 2014]. However, such methods are either labor-intensive to build dictionaries or raise data reliability and quality issues, limiting their performance.

Through advances in representation learning, modern natural language processing (NLP) techniques can learn the semantic properties of a language without human supervision not only in the general domain [Mikolov et al., 2013, Bojanowski et al., 2017, Peters et al., 2018, Artetxe et al., 2018b, Conneau et al., 2018, Chung et al., 2018], but also in clinical language [Weng et al., 2017b, Wang et al., 2018, Weng and Szolovits, 2018]. We aim to advance the state of clinician-patient communication by translating clinical notes to layperson-accessible text. Specifically, we make the following contributions:

1. We first design and apply the fully-unsupervised BDI and statistical MT framework for the non-parallel clinical cross-domain (professional-to-consumer) language translation.

2. We utilize the identical strings in non-parallel corpora written in different clinical languages to serve as anchors to minimize supervision.

3. We design a clinically meaningful evaluation method that considers both correctness and readability for sentence translation without ground truth reference.

2.3 Related Works

2.3.1 Clinical Professional-Consumer Languages

Recent studies mapped clinical narratives to patient-comprehensible language using the Unified Medical Language System (UMLS) Metathesaurus combined with the consumer health vocabulary (CHV) to perform synonym replacement for word translation [Zeng and Tse, 2006, Zeng-Treitler et al., 2007]. Elhadad and Sutaria [2007] adopted the corpus-driven method and

UMLS to construct professional-consumer term pairs for clinical MT. Researchers also utilized external data sources, such as MedlinePlus, Wikipedia, and UMLS, to link professional terms to their definitions for explanation [Polepalli et al., 2013, Chen et al., 2018]. However, these dictionary-based approaches have limitations. Studies show that expert-curated dictionaries don't include all professional words that are commonly seen in the clinical narratives (e.g., "lumbar" is not seen in CHV) [Keselman et al., 2008, Chen et al., 2017]. In contrast, the layman terms are not covered well in the UMLS [Elhadad and Sutaria, 2007]. Many professional words also don't have corresponding words in consumer language (e.g., "captopril"), or the translated words are still in the professional language (e.g., "abd" $\rightarrow$ "abdomen"). Such issues limit the utility of dictionaries like CHV to be useful for evaluation but not for training the professional-to-consumer language translation model due to lack of appropriate translation pairs. Additionally, the definitions of some complex medical concepts in the ontology or dictionary are not self-explanatory. Consumers may still be confused after translation with unfamiliar definitions. Finally, such dictionary curation and expansion are expert-demanding and challenging to scale up.

Vydiswaran et al. [2014] applied a pattern-based method on Wikipedia using word frequency with human-defined patterns to explore the relationship between professional and consumer languages. The approach is more generalized, yet Wikipedia is not an appropriate proxy for professional language that physicians commonly use in clinical narratives. For example, clinical abbreviations such as "qd" (once per day) and "3vd" (three-vessel coronary artery disease), may not be correctly represented in Wikipedia. Wikipedia also has great challenges of quality and credibility, even though patients trust it. The patterns used to find translation pairs also require human involvement, and the coverage is questionable. Furthermore, none of the above methods can perform sentence translation that considers the semantics of the context and sentence readability without human supervision, which is a common but critical issue for clinical machine learning.

Clinical Language Representations

Recent progress in machine learning has exploited continuous space representations of discrete variables (i.e., tokens in natural language) [Mikolov et al., 2013, Bojanowski et al., 2017,

Peters et al., 2018]. In the clinical domain, such learned distributed representations (from word to document embeddings) can capture semantic and linguistic properties of tokens in the clinical narratives. One can directly adopt pre-trained embeddings trained on the general corpus, the biomedical corpus (PubMed, Merck Manuals, Medscape) [Pyysalo et al., 2013], or clinical narratives [Choi et al., 2016c], for downstream clinical machine learning tasks. We can also train the embedding space by fine-tuning the pre-trained model [Hsu et al., 2018], or even from scratch—learning the embedding space from one's own corpus [Weng et al., 2017b, Weng and Szolovits, 2018]. Learned language embedding spaces can also be aligned for cross-domain and cross-modal representation learning by BDI algorithms [Conneau et al., 2018, Artetxe et al., 2018b, Chung et al., 2018]. Researchers have applied such techniques to clinical cross-domain language mapping and medical image-text cross-modal embedding spaces alignment [Weng and Szolovits, 2018, Hsu et al., 2018]. We apply the concepts of the cross-domain embedding spaces alignment to our translation task.

Unsupervised Machine Translation

MT has been shown to have near human-level performance with large annotated parallel corpora such as English to French translation. However, one big challenge of current MT frameworks is that most language pairs, such as clinical language translation, are low-resource in this sense. To make the frameworks more generalizable to low-resource language pairs, it is necessary to develop techniques for fully utilizing monolingual corpora with less bilingual supervision [Lample et al., 2018a, Artetxe et al., 2018c, Chung et al., 2019].

Researchers have developed state-of-the-art neural-based MT frameworks [Lample et al., 2018a, Artetxe et al., 2018c], which first construct a synthetic dictionary using unsupervised BDI [Conneau et al., 2018, Artetxe et al., 2018b]. Then the dictionary is used to initialize the sentence translation. Next, the language model is trained and serves as a denoising autoencoder when applied to the encoder-decoder translator to refine the semantics and syntax of the noisy, rudimentary translated sentence [Sutskever et al., 2014, Bahdanau et al., 2014, Vincent et al., 2008]. Finally, iterative back-translation is adopted to generate parallel sentence pairs [Sennrich et al., 2016].

Apart from the neural-based approaches, statistical frameworks, such as phrase-based

statistical MT (SMT) [Koehn et al., 2003], do not require co-occurrence information to learn the language representations and therefore usually outperform neural-based methods when the dataset and supervision are limited, especially for low-resource language translation. In Lample et al. [2018b], they applied the same principles that researchers used in the neural-based MT framework to the SMT system and outperformed the neural-based frameworks in some conditions. We adopt the unsupervised BDI with the SMT framework to achieve word and sentence translations.

2.4 Methods

The two-step framework is built on several unsupervised techniques for NLP. First, we develop a word translation system that translates professional words into consumer-understandable words without supervision. Next, we adopt a state-of-the-art SMT system, which uses language models and back-translation to consider the contextual lexical and syntactic information for a better quality of translation. The framework follows Figure 2-1.

2.4.1 Learning Word Embedding Spaces

We apply the unsupervised skip-gram algorithm to learn the embedding space of the words that preserve the semantic and linguistic properties [Mikolov et al., 2013]. The skip-gram model is trained to maximize, for each token $w(n)$ in a corpus, the probability of tokens $\{w_{n-k}, ..., w_{n-1}, w_{n+1}, ..., w_{n+k}\}$ within a window of size k given $w(n)$. Word-level representations can also be learned by adding subword information, namely character-level n-gram properties that capture more lexical and morphological features in the corpus [Bojanowski et al., 2017]. We investigate the qualities of learned embedding spaces trained by the skip-gram algorithm with or without subword information.

The assumption of good BDI for translation is that the embedding spaces of source and target languages should be as similar as possible. Since human languages use similar semantics for similar textual representations [Barone, 2016], the nearest neighbor graphs derived from word embedding spaces in different languages are likely to be approximately isomorphic. Thus, it is theoretically possible to align embedding spaces trained by the same

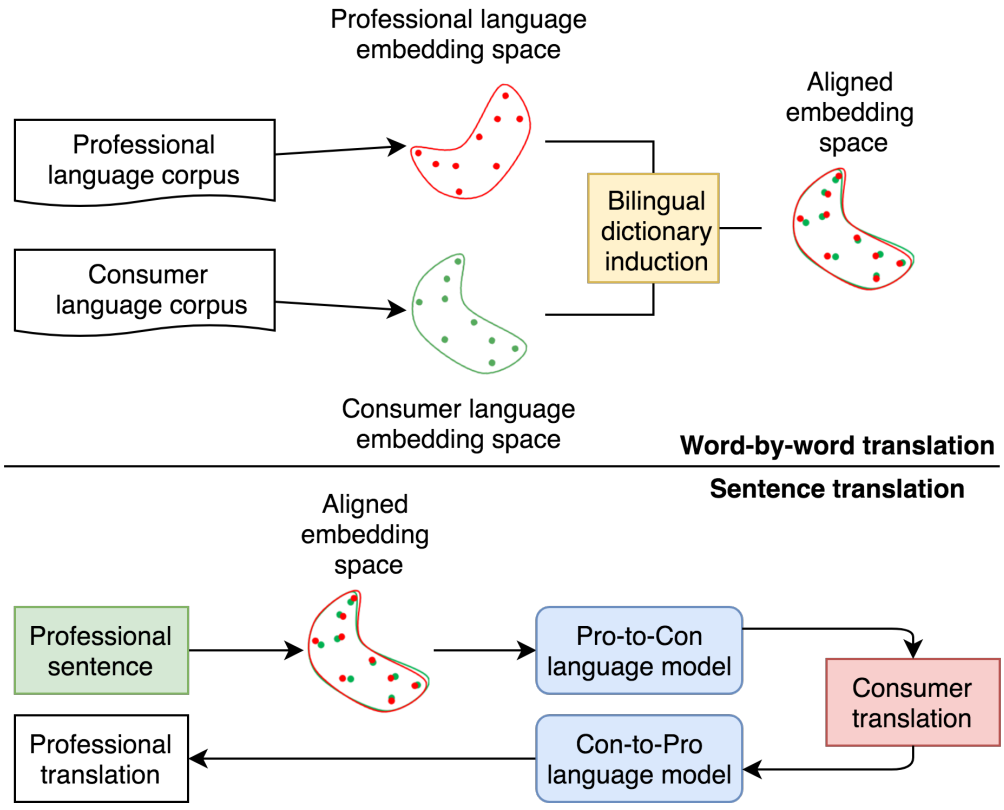


Figure 2-1: Overview of our framework. The framework is composed of two steps: (1) word translation through unsupervised word representation learning and bilingual dictionary induction (BDI), and (2) sentence translation, which is initialized by the BDI-aligned word embedding spaces and refined by a statistical language model and back-translation.

algorithm if they have similar shapes of distributions. To evaluate the similarity between embedding spaces, we compute the eigenvector score between them [Søgaard et al., 2018]. A higher eigenvector score indicates that the given two embedding spaces are less similar. Derived from the eigenvalues of Laplacian matrices, the eigenvector score can be computed as follows:

- Derive the nearest neighbor graphs, G_1, G_2, from the learned embedding spaces, then compute $L_1 = D_1 - A_1$ and $L_2 = D_2 - A_2$, where L_i, D_i, A_i are the Laplacian matrices, degree matrices, and adjacency matrices of G_i, respectively.
- Search for the smallest value of k for each graph such that the sum of the largest k Laplacian eigenvalues is smaller than 90% of the summation of all Laplacian eigenvalues.
- Select the smallest k across two graphs and compute the squared differences, which is the eigenvector score, between the largest k eigenvalues in two Laplacian matrices.

2.4.2 Bilingual Dictionary Induction for Word Translation

Unsupervised BDI algorithms can be applied to learn a mapping dictionary for the alignment of embedding spaces. We investigate two state-of-the-art unsupervised BDI methods: (1) iterative Procrustes process (`MUSE`) [Conneau et al., 2018] and (2) self-learning (`VecMap`) [Artetxe et al., 2018b]. The goal of alignment is to learn a linear mapping matrix W. To minimize supervision, we don't use any mapping dictionaries, such as CHV, but leveraged the characteristics of two English corpora to use identical strings in two corpora to build a synthetic seed dictionary.

Using Anchors

The identical strings serve as anchors to learn W with `MUSE` or `VecMap`. `MUSE` adopts the technique of the Procrustes process, which is a linear transformation. Assuming that we have the x-word, d-dimensional professional language embedding $\mathcal{P} = \{p_1, p_2, \ldots, p_x\} \subseteq \mathbb{R}^d$ and the y-word, d-dimension consumer language embedding $\mathcal{C} = \{c_1, c_2, \ldots, c_y\} \subseteq \mathbb{R}^d$. We use k

anchors to build the synthetic mapping dictionary and learn W between the two embedding spaces, such that $p_i \in \mathcal{P}$ maps to the appropriate $c_j \in \mathcal{C}$ without supervision. Then we have:

$$W^\star = \underset{W \in \mathbb{R}^{d \times d}}{\operatorname{argmin}} \|WX - Y\|^2 \tag{2.1}$$

where $X = \{x_1, x_2, \ldots, x_k\} \subseteq \mathbb{R}^d$ and $Y = \{y_1, y_2, \ldots, y_k\} \subseteq \mathbb{R}^d$ are two aligned matrices of size $d \times k$ formed by k-word embeddings selected from $\mathcal{P}$ and $\mathcal{C}$.

An orthogonality constraint is added on W, where the above equation will turn into the Procrustes problem that can be solved by singular value decomposition (SVD) with a closed-form solution [Xing et al., 2015]:

$$W^\star = \underset{W \in \mathbb{R}^{d \times d}}{\operatorname{argmin}} \|WX - Y\|^2 = UV^T \tag{2.2}$$

where $U\Sigma V^T = \mathsf{SVD}(YX^T)$. The aligned output of the professional language input p_i, i.e., the best translation $c_j = \operatorname{argmax}_{c_j \in \mathcal{C}} \cos(Wp_i, c_j)$.

For the `VecMap` self-learning method, the idea includes two steps [Artetxe et al., 2018b]. First, using a dictionary D_{ij} to learn the mappings W_X, W_Y that will transform both X and Y to maximize the similarity for the given dictionary as follows:

$$\underset{W_X, W_Y}{\operatorname{argmax}} \sum_i \sum_j D_{ij}(W_X x_i \cdot W_Y y_j) \tag{2.3}$$

where the optimal result is given by $W_X \Sigma W_Y^T = \mathsf{SVD}(X^T D Y)$. We again utilize the identical strings to build the initial dictionary. Symmetric re-weighting of X, Y is applied before and after SVD [Artetxe et al., 2018a].

Next, we use W_X, W_Y to bidirectionally compute the updated dictionary over the similarity matrix of the mapped embedding, $XW_X W_Y^T Y^T$. The values in the updated dictionary are filled using Cross-Domain Similarity Local Scaling (CSLS), where the value equals 1 if translation $y_j = \operatorname{argmax}_{y_j \in Y}(W_X x_i \cdot W_Y y_j)$, else equals zero. The above two steps are trained iteratively until convergence.

Without Anchors

We adopt adversarial learning for the configurations if identical strings were not used. We first learn an approximate proxy for W using a generative adversarial network (GAN) to make $\mathcal{P}$ and $\mathcal{C}$ indistinguishable, then refined by the iterative Procrustes process to build the synthetic dictionary [Conneau et al., 2018, Goodfellow et al., 2014].

In adversarial learning, the discriminator aims to discriminate between elements randomly sampled from $\mathcal{P} = \{Wp_1, Wp_2, \ldots, Wp_x\}$ and $\mathcal{C}$. The generator, W, is trained to prevent the discriminator from making an accurate prediction. Given W, the discriminator parameterized by θ_D tries to minimize the following objective function (Pro = 1 indicates that it is in the professional language):

$$\mathcal{L}_D(\theta_D|W) = -\frac{1}{x}\sum_{i=1}^{x}\log \mathbb{P}_{\theta_D}(\text{Pro}=1|Wp_i) - \frac{1}{y}\sum_{j=1}^{y}\log \mathbb{P}_{\theta_D}(\text{Pro}=0|c_j) \tag{2.4}$$

Instead, W minimizes the following objective function to fool the discriminator:

$$\mathcal{L}_W(W|\theta_D) = -\frac{1}{x}\sum_{i=1}^{x}\log \mathbb{P}_{\theta_D}(\text{Pro}=0|Wp_i) - \frac{1}{y}\sum_{j=1}^{y}\log \mathbb{P}_{\theta_D}(\text{Pro}=1|c_j) \tag{2.5}$$

The optimizations are executed iteratively to minimize $\mathcal{L}_D$ and $\mathcal{L}_W$ until convergence.

We perform nearest neighbors word retrieval using CSLS instead of simple nearest neighbor (NN). The purpose of using CSLS is to reduce the problem of "hubness," that a data point tends to be the nearest neighbors of many points in a high-dimensional space due to the asymmetric property of nearest neighbors [Conneau et al., 2018, Artetxe et al., 2018b, Dinu et al., 2015].

$$\begin{aligned}\mathsf{CSLS}(Wp_i, c_j) = {} & 2\cos(WWp_i, c_j) \\ & - \frac{1}{k}\sum_{c_j \in \mathsf{NN}_Y(Wp_i)} \cos(Wp_i, c_j) - \frac{1}{k}\sum_{Wp_i \in \mathsf{NN}_X(c_j)} \cos(Wp_i, c_j)\end{aligned} \tag{2.6}$$

Word translation is done using BDI algorithms by a series of linear transformations. However, language translation requires not only the word semantics, but also the semantic and syntactic correctness at the sentence level. For instance, the ideal translation is not the

nearest target word but synonyms or other close words with morphological variants. Further refinement is, therefore, necessary for sentence translation.

2.4.3 Sentence Translation

The unsupervised phrase-based SMT includes three critical steps 2-2:

1. Careful initialization with word translation,
2. Language models for denoising,
3. Back-translation to generate parallel data iteratively.

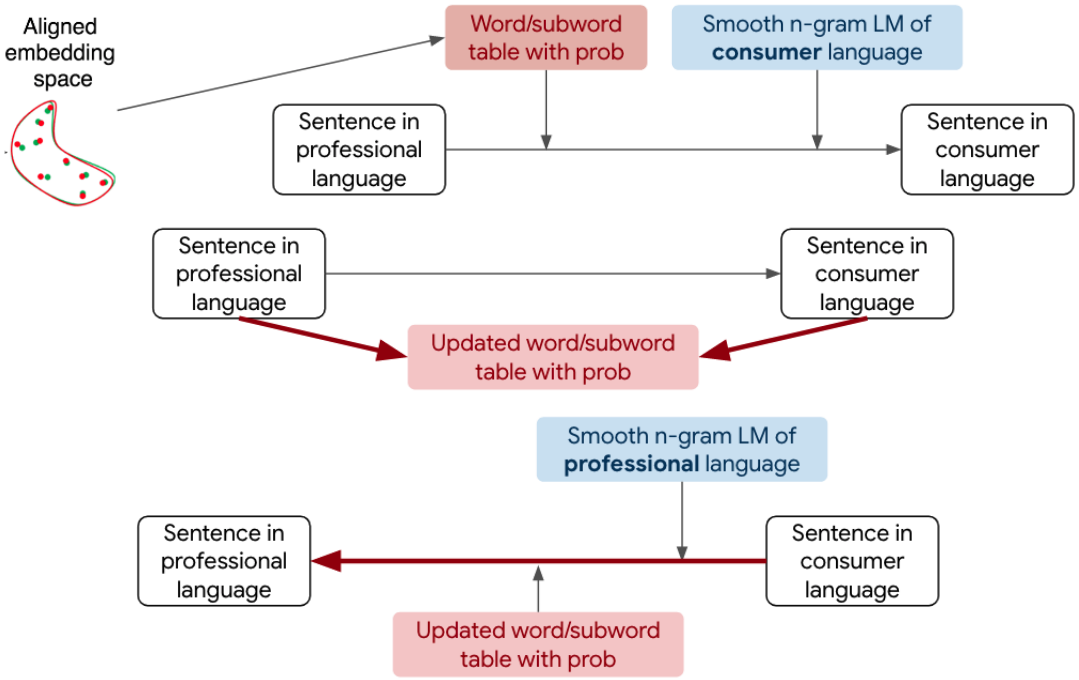


Figure 2-2: Overview of the statistical machine translation (SMT) framework. The framework is composed of three steps: (1) Initializing the word/subword table using the aligned embedding space obtained from word-level translation. (2) Learning language models from both professional and consumer corpora. (3) Using the word/subword table and the language model of consumer language to translate sentences in professional sentences, updating the word/subword table using generated translation pairs, then back-translating sentences in consumer language to the professional language, and doing the back-and-forth translation several times.

We initialize the sentence translation with the aligned word embedding spaces trained by unsupervised word representation learning and BDI algorithms.

To translate the word in professional language p_i to the word in consumer language c_j, the SMT scores c_j where $\text{argmax}_{c_j} \mathbb{P}(c_j|p_i) = \text{argmax}_{c_j} \mathbb{P}(p_i|c_j)\mathbb{P}(c_j)$. The $\mathbb{P}(p_i|c_j)$ is derived from the phrase tables and $\mathbb{P}(c_j)$ is from a language model [Lample et al., 2018b]. We use the mapping dictionary generated by the BDI algorithm as the initial phrase (word) table to compute the softmax scores, $\mathbb{P}(c_j|p_i)$, of the translation of a source word, where

$$\mathbb{P}(c_j|p_i) = \frac{\exp(T^{-1}\cos[W emb(p_i), emb(c_j)])}{\sum_k \exp(T^{-1}\cos[W emb(p_i), emb(c_k)])} \tag{2.7}$$

where $emb(x)$ is the embedding of word x, cos is the cosine similarity, and T is a hyperparameter for tuning the peakness of the distribution. We then learn smoothed n-gram language models using KenLM for both professional and consumer corpora [Heafield, 2011].

Next, we use the initial phrase table and language models mentioned above to construct the first rudimentary SMT system to translate the professional sentence into consumer language. Once we get the translated sentences, we are able to train a backward SMT from target to source language (back-translation) by learning new phrase tables and language models. Therefore, we can generate new sentences and phrase tables to update translation models in two directions, back and forth, for many iterations.

2.5 Data

Data for this study are collected from the MIMIC-III (Medical Information Mart for Intensive Care) database [Johnson et al., 2016], containing de-identified data on 58,976 intensive care unit (ICU) patients admissions to the Beth Israel Deaconess Medical Center (BIDMC), a large, tertiary care medical center in Boston, Massachusetts, USA. The database contains detailed information on patients admitted between 2001 and 2012, including hospital administrative data, vital signs, medications, laboratory test results, and survival data after hospital discharge.

We extract 59,654 free-text discharge summaries from MIMIC-III. These usually include the following sections: "Allergy", "Chief Complaint", "History of present illness", "Major Surgical or Invasive Procedure", "Past medical history", "Social history", "Family history",

"Brief hospital course", "Medications on Admissions", "Discharge medications", "Discharge diagnosis", "Discharge condition", "Discharge instruction", and "Followup instruction". For all discharge summaries, we extract and preprocess the sections of "History of present illness", "Brief hospital course", "Discharge instruction" and "Followup instruction". Clinical notes usually have many sections. Among all sections, we select the "History of present illness" and "Brief hospital course" sections to represent the content with professional jargon. These sections are usually the most narrative components with thoughts and reasoning for the communication between clinicians. In contrast, "Discharge instruction" and "Followup instruction" sections are written in consumer language for patients and their families. We omit other sections since they are usually not written in natural language but only lists of jargon terms, such as a list of medications or diagnoses.

For training word embeddings, there are 443,585 sentences in the clinical professional language set and 73,349 sentences in the consumer language set. There are 19,618 and 5,264 unique words in the professional and consumer term embeddings, respectively. Although the professional and consumer corpora are both from MIMIC-III, their contents are not parallel. However, we expect that there are identical strings across two corpora since both of them are written in English. We utilize 4,605 overlapping English terms as anchors to create a seed dictionary in BDI to minimize supervision.

We also collect additional consumer language data from the English version of the MedlinePlus corpus*. MedlinePlus is the patient and family-oriented information produced by the National Library of Medicine. The corpus is about diseases, conditions, and wellness issues and is written in consumer understandable language. We investigate whether the addition of the MedlinePlus corpus enhances the quality of BDI.

The statistics of the corpora used are shown in Table 2.1.

Corpus	#Sentence	#Vocabulary
MIMIC-professional language	443585	19618
MIMIC-consumer language	73349	5264
MIMIC-consumer + MedlinePlus	87295	6871

Table 2.1: The detailed statistics of the corpora.

*https://medlineplus.gov/xml.html

For data preprocessing, we remove all personal health information placeholders in the MIMIC corpora, then apply the Stanford CoreNLP toolkit and Natural Language Toolkit (NLTK) to perform document sectioning and sentence fragmentation [Manning et al., 2014].

To build the language models for sentence translation, we experiment with using either the MIMIC-consumer corpus or a general corpus, for which we use all sentences from the WMT English News Crawl corpora from the years 2007 through 2010, which include 38,214,274 sentences extracted online news publications.

2.6 Experiments

In this study, we consider MT in two parts: (1) word translation, and (2) sentence translation. We define the tasks and overview of evaluations in this section. The scripts are available at the project repository (`https://github.com/ckbjimmy/p2c`).

2.6.1 Word Translation

Learning Word Embeddings

We adopt the skip-gram algorithm to learn word embeddings. We train the word embeddings by setting the word window size $k = 5$. We consider all words that appear more than once, with a negative sampling rate of 10^{-5}. The models are trained by stochastic gradient descent (SGD) without momentum with a fixed learning rate of 0.1 for 5 epochs.

To include the subword information, we consider the length of character n-grams between 2 and 5, and used the fastText implementation for word representation learning for all experiments. We experiment on an embedding dimension of 50, 100, 200, 300 and 500 for clinician-designed word pairs evaluation, and 100, 200, 300, 500, and 1000 for CHV pairs evaluation.

Bilingual Dictionary Induction

When we use the identical character strings as anchors for unsupervised BDI, we do 5 iterations of the Procrustes process for `MUSE` [Conneau et al., 2018]. We adopt the default

setting for `VecMap` self-learning, which uses symmetric re-weighting before and after applying SVD, and bidirectional dictionary induction [Artetxe et al., 2018b].

Without anchors, we utilize adversarial learning [Conneau et al., 2018]. For the discriminator in adversarial training, we use a two-layer neural network of size 2048 without dropout, and a leaky rectified linear unit (ReLU) as the activation function. We train both the discriminator and W by SGD with a decayed learning rate from 0.1 to 10^{-6} with a decay rate of 0.98. We select the 1000 most frequent words for discrimination. For refinement, we also do five iterations of the Procrustes process.

Evaluation of Word Translation

The evaluation pairs are available through the project repository.

We investigate (1) whether adopting subword information to train word embedding spaces is beneficial, (2) if different BDI methods (`MUSE` or `VecMap`) matter, (3) whether integrating MedlinePlus to augment the consumer corpus is helpful, (4) what dimensionality of word embedding spaces is optimal.

We evaluate the quality of word translation through nearest neighbor words retrieval. Two evaluations were performed. First, we use a list of 101 professional-consumer word pairs developed by trained clinicians based on their commonly-used professional words. The word pairs list is further reviewed and approved by non-professionals with expert explanations. Several examples of the ground truth pairs include: (bicarbonate, soda), (glucose, sugar), (a-fib, fibrillation), (cr, creatinine), (qd, once/day). Since 14 out of 101 evaluation ground truth pairs do not appear in the training corpora, we use the matched 87 pairs for all quantitative evaluations. We also evaluate our method on CHV pairs, which include 17,773 unique word pairs. We choose the configurations and parameters for sentence translation based on the results of these two evaluations.

We show the performance by computing precision at k, where we used CSLS to query the nearest k words ($k = 1, 5, 10$) in the consumer language embedding space using the words in the aligned professional language embedding space.

2.6.2 Sentence Translation

The goal of sentence translation is to translate the sentence in the professional language domain into a sentence in the consumer language domain. We apply the SMT framework and examine the quality of translation by considering whether (1) subword information, (2) anchors for BDI, and (3) language model trained on a specific or general corpus, are helpful.

Language Modeling

We adopt Moses, a widely-used SMT engine that is used to train statistical translation models [Koehn et al., 2007]. In language modeling, we apply the default Moses smoothed n-gram language model with phrase reordering disabled during the very first generation [Lample et al., 2018b, Koehn et al., 2007]. We train the model iteratively while randomly picking source sentences for translation. The length of phrase tables is 4. The hyperparameter T for computing softmax scores is set to be 30.

We use the supervised, dictionary-based CHV professional-to-consumer word mapping and replacement as the strong baseline since the replacement mainly preserves clinical correctness. The Wikipedia pattern-based approach is not considered due to the issues of credibility and quality. Detailed configurations of SMT are shown in Table 2.2.

Configuration	Word embedding	Anchors	Language model
A	100d with subword	Y	WMT
B	100d with subword	Y	MIMIC-consumer
C	1000d with subword + augmentation	Y	WMT
D	1000d with subword + augmentation	Y	MIMIC-consumer
E	300d w/o subword	Y	WMT
F	300d w/o subword	Y	MIMIC-consumer
N	300d w/o subword	N	WMT

Table 2.2: Configurations of statistical MT (SMT) for sentence translation.

Evaluation of Sentence Translation

Since there is no ground truth reference for clinical professional-to-consumer sentence translation, using standard quantitative metrics such as BiLingual Evaluation Understudy (BLEU)

or Consensus-based Image Description Evaluation (CIDEr) score is not possible. Previously, researchers asked either clinical experts [Zeng-Treitler et al., 2007, Chen et al., 2018], or crowd-sourced Amazon Mechanical Turks (AMT) to score outputs or provide feedback on the readability of mapped terms [Lalor et al., 2018]. Instead, we not only invite non-clinicians to score and provide their comments on readability, but we also ask clinicians to evaluate the correctness of the translations before reaching out to the non-clinicians to evaluate readability. We adopt the two-step evaluation because clinical correctness is critical but hard to evaluate by non-clinicians; and by contrast, a judgment of readability may be biased for clinicians.

We recruit 20 evaluators—10 clinical professionals and 10 non-clinicians. For each evaluator, we randomly assign 20 sentence sets. Each set includes the professional sentence (PRO), the translated sentence using configuration A, B (or C, D), E, F, N, and CHV baseline. We ask evaluators to score the translated sentences.

We adopt the MOS to evaluate the quality of translation. In Figure 2-3, we demonstrate the workflow of the translated sentence evaluation.

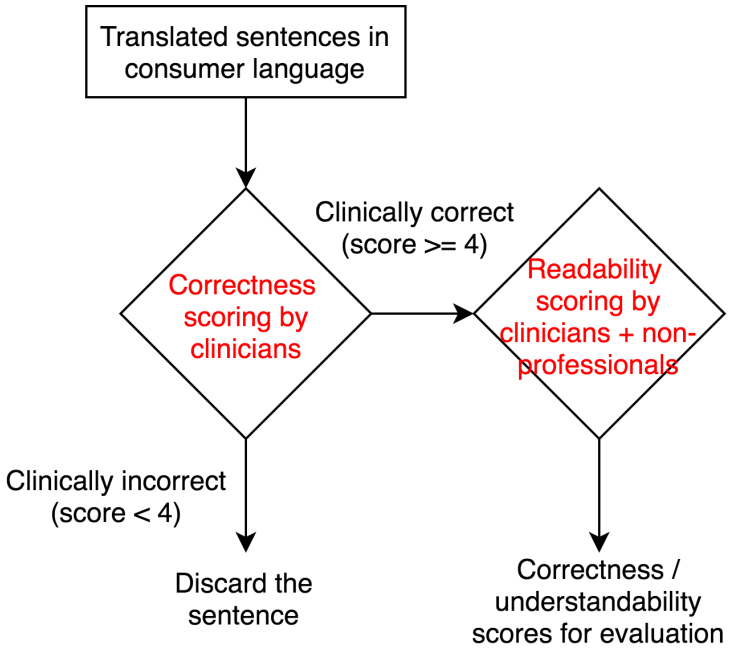


Figure 2-3: Evaluation process of sentence translation.

Our MOS evaluation includes two steps. The two-step sentence evaluation considers clinical correctness and consumer (both clinicians and non-clinicians) readability. We first

ask the clinicians to provide the correctness score of each translated sentence, score ranging from 1 (the worst) to 5 (the best). If the correctness score of the translated sentence is less than 4, the sentences will be discarded and not further scored by both professionals and non-professionals since the sentence is not clinically correct, as judged by professionals, and thus meaningless to score for readability. Otherwise, the sentences will be assigned to both clinicians and non-clinicians for readability scoring.

Criteria and examples for scoring correctness and readability are shown in Table 2.3.

For correctness, we consider negation error (yes/no, true/false, positive/negative), numeric/dosage error (e.g., 32 $\rightarrow$ 10), different named entities or semantics (e.g., ca $\rightarrow$ potassium, vancomycin $\rightarrow$ cefazolin, increase $\rightarrow$ decrease, hypertension $\rightarrow$ hypotension), or missingness of the critical information, as incorrectness. The errors of ambiguities that don't change the understanding of sentence (e.g., hospital $\leftrightarrow$ icu, pathology $\leftrightarrow$ biopsy, some lung signs such as opacity (which can be detected in x-ray) $\leftrightarrow$ x-ray), pronouns (he, she, they, you), typos, duplicated words, punctuation, and slightly incorrect grammar are acceptable. We only keep the sentences with a mean correctness score ≥ 4 for readability evaluation.

For readability, we consider whether the sentence is understandable for consumers. Thus, we accept grammatical and syntactic errors that don't confuse evaluators (e.g., a bone scan showed no on cancer disease .), and no need to consider whether it is a reasonable clinical condition since this should already be judged by professionals while doing correctness scoring. We also ask clinicians to score the readability, but the scores from non-clinicians and clinicians are calculated separately.

Before assigning the original and translated sentences, we roughly filter out the sentence sets with severe incompleteness, format errors (e.g., too many numbers, duplicated words, punctuation errors), or severe fragmented sentences. Minimal format errors are acceptable during evaluation. The sentences related to neonatal patients are excluded since they are very specific and we expect that the data distribution is very different from the sentences describing adult patients. Finally, we assign 1000 translated sentences to 20 evaluators (10 clinical and 10 non-clinical). The final MOS is computed by averaging all given valid scores.

Score	Criteria / Example
	Correctness (consider both original and translated sentences)
5	0 or 1 error. Original: chest tubes were discontinued without incident . Translation: chest tubes were stopped without complications .
4	2 errors, or the same as professional language. Original: it was stable on monitoring , and her stools were guaiac negative . Translation: they were stable , and were monitoring the black blood was negative .
3	3 errors, or with ambiguity that not easy to be inferred from clinical knowledge. Original: per family , she had + culture of a very resistant bacteria that is not mrsa . Translation: per family , you had no culture of a very different bacteria that is not cellulitis .
2	With the key clinical concepts but in the directions, numbers that don't make sense in clinical setting. Original: the patient had an o2 saturation in the 80s when they arrived , with a heart rate in the 130s , atrial fibrillation at that time and blood pressure with a systolic of 200 . Translation: the patient had an oxygen saturation in the 30 when they arrived to heart with a rate in the 30 to atrial fibrillation at that time and blood pressure with a systolic of 2000 .
1	Missing any critical information. Original: her wound was debrided at in the beginning of with several days of icu stay . Translation: the site was at the foot of the next few days for a hospital stay .
	Readability (consider only translated sentence)
5	0 or only 1 word can't understand. Translation: a scan was performed which showed a right upper small clots .
4	2 words. Translation: coumadin was held that night , and inr was therapeutic on .
3	3 words, or confusing about wordings. Translation: you were given coumadin 5 mg po qam and your inr was monitored daily .
2	More than 4 words, or need explanations to understand. Translation: improved sat with improving ms . .
1	Completely can't understand, semantically meaningless. Translation: person to person you service them and the pressures and neurosurgery .

Table 2.3: Criteria for correctness and readability scoring.

2.7 Results and Discussions

2.7.1 Word Translation

Bilingual Dictionary Induction Algorithm and Data Augmentation

In Table 2.4, we demonstrate that `MUSE` generally outperforms `VecMap`. We also identify a trend that the performance is better when consumer corpus augmentation is not used. The only exception is when we apply the corpus augmentation to subword embeddings while evaluating on CHV pairs.

The nature of MedlinePlus texts is very different from clinical narratives since the former are articles for general patient education whereas the latter is more specific to individual cases and colloquial. It is highly likely that MedlinePlus and MIMIC-consumer corpora have very different data distributions and therefore affect the quality of BDI. This also yields inferior performance when we did an evaluation on the clinician-designed word pairs since they are also in clinical narrative rather than literature style. In contrast, CHV covers many morphologically similar words that are shown in the literature but rare in clinical narratives, which results in better performance while using subword embeddings with MedlinePlus augmentation while evaluating on CHV pairs.

By computing the eigenvector score, we find that no augmentation yielded a smaller eigenvector score (smaller difference between embedding spaces) than with augmentation. Eigenvector scores increase from 0.035 to 0.177 (without subword information), and 0.144 to 0.501 (with subword information), after consumer corpus augmentation, which also indicates that adding MedlinePlus yields harder BDI. Since the embedding space similarity is higher without augmentation, `MUSE` can perform well in such conditions, as mentioned in previous literature [Artetxe et al., 2018b].

Subword Information and Dimensionality

Next, we search for the ideal dimensionality beyond the common parameterization. We use `MUSE` without MedlinePlus for the following experiments except for the embeddings with subword information evaluated on CHV pairs, for which we augment the consumer corpus

			Without subword		With subword	
			MUSE	VecMap	MUSE	VecMap
Clinician	P@1	aug(+)	15.61 (3.57)	15.73 (2.40)	15.49 (2.70)	15.24 (1.75)
		aug(-)	**17.78 (3.25)**	13.46 (1.69)	**20.62 (2.85)**	20.37 (4.09)
	P@5	aug(+)	39.02 (2.37)	37.19 (4.61)	42.56 (4.82)	40.97 (3.65)
		aug(-)	**42.84 (2.25)**	36.71 (3.76)	**48.15 (3.99)**	42.10 (2.88)
	P@10	aug(+)	46.95 (3.69)	45.85 (4.57)	54.76 (3.61)	52.56 (2.11)
		aug(-)	**53.86 (5.75)**	47.78 (5.73)	**58.27 (3.53)**	49.51 (3.37)
CHV	P@1	aug(+)	17.94 (2.54)	13.51 (1.97)	**22.11 (2.83)**	18.51 (1.54)
		aug(-)	**18.26 (2.72)**	13.09 (1.95)	21.29 (2.29)	16.70 (1.46)
	P@5	aug(+)	36.04 (4.01)	29.40 (2.38)	44.06 (3.61)	38.82 (2.15)
		aug(-)	**37.30 (3.87)**	29.39 (2.54)	**44.92 (3.07)**	37.01 (3.01)
	P@10	aug(+)	43.82 (4.15)	36.36 (2.61)	**53.73 (3.98)**	48.65 (3.65)
		aug(-)	**45.21 (3.59)**	37.99 (3.16)	53.42 (4.12)	47.27 (3.49)

Table 2.4: Performance of nearest neighbors retrieval using Cross-Domain Similarity Local Scaling (CSLS). Comparison between unsupervised Procrustes process (MUSE) and self-learning (VecMap), with or without augmented corpus (MedlinePlus) on clinician-designed pairs evaluation and CHV pairs evaluation. The word embeddings are trained by the fastText skip-gram. For subword information, we consider bigram to 5-gram. We choose a 100-dimensional parameterization, which is common for investigating BDI algorithms and data augmentation. The values reported are precision at k ($P@k$) $\times 100$ with standard deviation. Precision at 1 is equivalent to accuracy. Boldface values are the best combination of BDI and augmentation with or without subword information.

with MedlinePlus. For clinician-designed pairs, we find that the embeddings enriched with subword information have slightly superior performance to those without using subword information when the embedding space dimension is smaller (Table 2.5). However, embeddings without subword information yield better performance with higher embedding space dimensionality. In CHV pairs, embeddings with subword information yield superior performance to those without subword information. Such a finding is reasonable since there are many morphologically similar translation pairs in CHV pairs. For example, "asphyxiation" → "asphyxia". They also yield better performance with higher embedding space dimensionality.

The optimal embedding space dimensionality for the subword enriched embedding is 100-dimension in clinician-designed evaluation pairs, and 1000-dimension (with MedlinePlus) in CHV pairs evaluation. For the embeddings without subword information, 300-dimension usually yields better performance.

	Clinician-designed			CHV		
	Dim	No subword	Subword	Dim	No subword	Subword
P@1	50	12.22 (2.50)	12.84 (2.19)	100	18.26 (3.02)	22.11 (1.86)
	100	17.78 (3.25)	**20.62 (2.85)**	200	20.31 (2.57)	35.54 (2.53)
	200	21.11 (3.37)	19.26 (2.92)	300	**22.75 (2.98)**	46.68 (3.77)
	300	**20.62 (2.80)**	14.69 (2.05)	500	20.12 (2.12)	53.89 (4.13)
	500	20.37 (1.77)	12.96 (1.57)	1000	20.80 (1.95)	**54.55 (4.52)**
P@5	50	34.44 (3.21)	38.89 (3.78)	100	37.30 (3.43)	44.06 (3.81)
	100	42.84 (2.25)	48.15 (3.99)	200	39.16 (3.67)	61.34 (4.59)
	200	48.62 (7.11)	**49.51 (2.57)**	300	41.31 (3.61)	70.35 (5.02)
	300	48.08 (3.49)	46.91 (2.54)	500	**42.87 (4.01)**	77.81 (5.96)
	500	**51.21 (2.14)**	43.83 (1.95)	1000	40.82 (3.54)	**76.99 (5.72)**
P@10	50	44.07 (1.75)	47.41 (3.83)	100	45.20 (3.49)	53.73 (4.17)
	100	53.86 (5.75)	**58.27 (3.53)**	200	49.80 (3.28)	70.35 (5.10)
	200	59.75 (2.19)	57.78 (4.86)	300	50.49 (4.02)	76.49 (5.58)
	300	**60.25 (3.63)**	55.54 (3.62)	500	**51.27 (3.85)**	82.64 (6.10)
	500	59.01 (2.08)	53.95 (3.08)	1000	49.71 (3.16)	**82.71 (5.76)**

Table 2.5: Performance of word translation using iterative Procrustes process (`MUSE`) on clinician-designed pairs and CHV pairs evaluation. The word embeddings are trained by the fastText skip-gram algorithm. For subword information, we consider bigram to 5-gram. The values reported are precision at k ($P@k$) $\times 100$ with standard deviation. Precision at 1 is equivalent to accuracy.

Qualitative Evaluation

In Table 2.6, we demonstrate that the BDI-learned mapping dictionaries are clinically meaningful through CSLS nearest neighbors retrieval. Four aligned embedding spaces (dictionaries), including the 100-dimension subword embedding, 300-dimension word embedding, 1000-dimension subword embedding with MedlinePlus augmentation, and 300-dimension word embedding without using anchors, are evaluated qualitatively. We retrieve the nearest top-5 neighbors in consumer language from the aligned embedding spaces using 12 professional words as queries. To ensure diversity, 3 anatomy-related, 3 disease-related, 2 procedure-related, 2 lab-related, and 2 medication-related professional words are used for querying. We find that the appropriate translations were shown in top-5 neighbors in most cases.

Subword embeddings utilize character n-gram information, which is helpful in capturing lexical and morphological patterns. The retrieved words from embeddings with subword information tend to retrieve a group of morphologically similar words, e.g., "mi" $\rightarrow$ "attack" and "attacks" (Table 2.6). However, the performance drops when the morphologically similar words are semantically incorrect, e.g., "ophthalmology" $\rightarrow$ "ob-gyn" and "ob/gyn", as in our case. Instead, word embeddings without subword information focus more on whole-word semantics, such as synonyms and antonyms, with different lexical morphologies. This is one of the reasons why subword embeddings can't always yield better results even though they utilize more information. Because they have different strengths in translation, we keep both for sentence translation.

Although the 1000-dimension subword embedding with MedlinePlus augmentation yields superior performance in CHV pairs evaluation, we don't see additional benefits in the qualitative evaluation compared to the 100-dimension version without augmentation, due to the tendency to retrieve professional-level words in the 1000-dimension version. This is because the translated words in CHV pairs are not always in consumer-understandable language, such as "abd" $\rightarrow$ "abdomen", and therefore using the parameters based on the CHV pairs evaluation is not reliable.

We also conclude that the embeddings without using anchors during BDI yield the worst performance in word translation. Using the eigenvector score, we find that adopting anchors

Rank	cr	cxr	ekg	hepatic	malignancy	mi	na	ophthalmology	qd	renal	sob	vancomycin
Embeddings with Subword Information, with Anchors (Clinician-designed Word Pairs)												
1	kidney	*x-ray*	echocardiogram	portal	*cancer*	stenosis	bun	ob/gyn	qday	*kidney*	shortnes	*antibiotic*
2	*creatinine*	*xray*	echocardiograms	biliary	carcinoma	*attack*	liter	ob-gyn	dailyname	*kidneys*	shortness	ceftriaxone
3	kidneys	csf	atrium	encephalopathy	hemoptysis	*attacks*	*sodium*	podiatry	qhs	dysfunction	pain/shortness	meropenem
4	baseline	ekg	emg	angiopathy	chemo	sclerosis	sat	opthamology	*once/day*	function	*breath*	daptomycin
5	renal	load	*ecg*	metastatic	diverticulosis	endocarditis	2l	*eye*	po/ng	adrenal	fevers	*antibiotics*
Embeddings without Subword Information, with Anchors (Clinician-designed Word Pairs)												
1	*creatinine*	*x-ray*	echocardiogram	*liver*	*cancer*	*attack*	*sodium*	*eye*	*daily*	*kidney*	*breath*	zosyn
2	potassium	*xray*	cardiac	cirrhosis	metastatic	infarction	moniter	opthamology	qday	liver	shortness	ceftriaxone
3	renal	resolution	lab	organ	lymphoma	myocardial	bun	ophthalmologist	po	*kidneys*	worsening	*antibiotic*
4	rose	lungs	infarction	attempt	represent	arrest	potassium	podiatry	mg/day	disease	*dyspnea*	cefepime
5	kidney	pleural	echo	portal	enlarged	nstemi	restriction	exam	twice	failure	palpitations	daptomycin
Embeddings with Subword Information and MedlinePlus Augmentation, with Anchors (CHV Word Pairs)												
1	renal	load	echocardiogram	pancreatic	malignant	*attack*	*sodium*	*eye*	*daily*	*kidney*	shortness	*antibiotic*
2	*creatinine*	ekg	echocardiograms	pancreatitis	*cancer*	infarct	gm	ophthalmologist	dailyyou	adrenal	shortnes	daptomycin
3	bun/creatinine	embolus	echocardiography	necrotic	polyps	atyou	bun	electrophysiology	coated	retinal	pain/shortness	ceftriaxone
4	potassium	edema	*ecg*	lymphatic	carcinoma	myocardial	potassium	neurologists	isosorbide	*kidneys*	lightheadedness	cefazolin
5	rapamune	xr	echo	pancreas	mass	mvp	serum	electrophysiologist	dailyplease	original	*breath*	*antibiotics*
Embeddings without Using Anchors												
1	report	made	relieved	operating	tell	status	sternal	tight	constipation	report	08:30	discharged
2	100.5	levofloxacin	continually	machinery	present	mental	temp	shown	quantity	notify	09:40	vital
3	greater	750	increasing	while	secretary	lethargy	pounds	reign	output	0.1	10:00	stablized
4	101	ciprofloxacin	headache	illicit	TRUE	confusion	101.5make	frontal	redness	fo	09:00	haven
5	sharp	regimen	ha	drowsy	expiratory	hallucinations	101.5	lacerations	bloating	watch	11:00	supplemental

Table 2.6: Examples of the top-5 nearest neighbor words (identical word excluded) in the consumer language embedding space queried by professional words (first row). Many commonly used appropriate corresponding consumer words for each queried clinical professional word are seen in the list. Italicizing words represent functionally correct word translation.

Configuration	Correctness (Clinicians)	Readability (Clinicians)	Readability (Non-clinicians)
A	2.85 (1.41)	4.48 (0.84)	4.02 (1.17)
B	2.89 (1.47)	4.35 (0.70)	3.60 (1.14)
C	2.95 (1.55)	4.56 (0.75)	4.26 (0.95)
D	3.33 (1.46)	4.03 (1.00)	3.81 (0.79)
E	3.57 (1.34)	4.55 (0.65)	4.28 (0.93)
F	4.10 (1.22)	4.59 (0.65)	4.25 (0.83)
N	1.18 (0.56)	-	-
Dictionary-based	4.13 (0.62)	3.57 (1.08)	3.55 (0.86)

Table 2.7: Performance of sentence translation using our unsupervised statistical machine translation (SMT) framework. The values are the average (standard deviation) of the mean opinion score (MOS) regarding the correctness and readability of translated sentences. The readability is accessed only for the sentences with a correctness score ≥ 4. For configurations, please refer to Table 2.2. Baseline is the supervised, dictionary-based CHV replacement method.

yielded higher embedding space similarity than without anchors. The eigenvector scores decrease 16.7% (0.54 $\rightarrow$ 0.45) in subword embeddings, and decreases 47.8% (0.46 $\rightarrow$ 0.24) in word embeddings after applying anchors.

For sentence translation, we decide to adopt (1) 100-dimension `MUSE`-aligned embedding spaces with subword information and without data augmentation, (2) 1000-dimension `MUSE`-aligned embedding spaces with subword information and with data augmentation, and (3) 300-dimension `MUSE`-aligned embedding spaces without subword information and without data augmentation.

2.7.2 Sentence Translation

In Table 2.7, we evaluate the translation by correctness as judged by clinicians, and readability by both clinicians and non-clinicians.

The supervised replacement baseline yields the best correctness score since it doesn't change the semantics of sentences too much. However, its readability scores are lower. This is because the CHV mapping doesn't align with the actual consumer language well, and still keeps many professional terms after replacement.

Selecting BDI-aligned Embedding Spaces

Configuration F, which uses the 300-dimension word embeddings with anchors in BDI and adopted the MIMIC-consumer corpus for language modeling, yields the highest correctness scores among all SMT configurations. Followed by configuration E, D, C, B, A, then N, we find that the most critical component for sentence translation is using identical strings as anchors for BDI. Mere sentences are correct and no sentences reach the threshold for readability evaluation (correctness score ≥ 4) if anchors are not used (configuration N). This emphasizes the critical role of anchors in language translation without supervision (Table 2.6).

Using word embeddings trained without subword information (configurations E, F) provides better correctness than those adopting subword information (configurations A, B, C, D). The reason is similar to word translation—training with subword information captures more morphologically similar words, yet the performance drops when those morphologically similar words are clinically incorrect. Instead, using word information captures more synonyms. The morphological errors can be corrected while applying language models during sentence translation.

Even though the 1000-dimension subword embeddings with MedlinePlus augmentation (configurations C, D) outperform other embeddings significantly on word-level CHV evaluation, their correctness in sentence translation is limited. This provides evidence that CHV pairs are less aligned with clinical narratives than the clinician-designed word pairs, and therefore the optimal setting for CHV is not the same as for real clinical narratives.

Language Modeling

Choosing the specific corpus (configurations B, D, F) for language modeling yields better correctness but inferior readability than using the general corpus (configurations A, C, E). Language models trained on general corpora tend to reshape professional words to more general terms and phrases. For example, mapping from "flagyl" (name of a kind of antibiotic) to "antibiotics" leads to better readability. Instead, using the more specific MIMIC corpus gives us more explanations of clinical professional terminologies. For example, "r femoral line" $\rightarrow$

"right central line" and "catheter" → "foley catheter". It also provides a better ability to expand the medical abbreviation to the completed word—which may also be helpful for the professional to consumer language translation. For instance, "afib" → "atrial fibrillation", "ppi" → "pantoprazole", "o2 sat" → "oxygen saturation", "na" → "sodium", "meds" → "medication", and "eval" → "evaluation". Such word and phrase identification and appropriate replacement are critical steps for professional-to-consumer translation, which is not seen in simply using a dictionary-based replacement method. Table 2.8 displays a few examples of sentence translation using configuration F.

Using different corpora for language modeling affects the quality of sentence translation, yet both general and specific corpora have their own limitations. We already know that a language model based on a general corpus helps to convert the specific terms into a general version. Such language generalization is helpful for readability, yet it sometimes results in oversimplification, ambiguity and vagueness that misses the important information (e.g., "hyponatremia" → "sodium", "troponin" → "cardiac"), and therefore reduces correctness. In contrast, language models using a specific corpus can better expand abbreviations and explanations. However, they may also make the translation be too specific, such as when "vancomycin" may be translated into "flagyl" or other antibiotics.

Some corpus-specific words are usually replaced during the sentence translation, such as pronouns, commonly-seen dosages, medications, and procedures. For example, language models using a general corpus tend to translate "pt" to "he", yet using the MIMIC-consumer corpus tends to translate it into "you", which makes sense since the MIMIC-consumer corpus contains colloquial instructions written for patient and family. Negation is also sometimes incorrect. Abbreviation ambiguity is also an issue: e.g., "pt" can be translated to "patient", "physical therapy", or "posterior tibial artery", but the results may be decided by the language model we use.

Choosing the ideal corpus, which considers the trade-off between general and specific, to build language models for SMT is a critical step for deployment. Other techniques to explore include introducing a copy mechanism or using biomedical ontologies to identify words and phrases that should be fixed or preserved (e.g., medication and procedure names). We may also need techniques to handle pronouns and negations correctly. For ambiguity, a

Original	the patient had an o2 saturation in the 80s when they arrived , with a heart rate in the 130s , atrial fibrillation at that time and blood pressure with a systolic of 200 .
SMT	the patient had an *oxygen saturation* in the 80s when they came , with a fast heart rate in the , atrial fibrillation at that time and with a *systolic blood pressure* of 200 .
Original	her hct was stable at 26-27 for many days prior to discharge .
SMT	your *hematocrit* was stable at stable for several days prior to discharge.
Original	food impaction within the esophagus .
SMT	*food stuck* in your esophagus .
Original	her diet was gradually advanced to puree consistency .
SMT	your diet was gradually advanced to *thickened liquids* .
Original	cxr from showed no infiltrate but did show volume overload .
SMT	from *x-ray* showed no pneumonia but did show volume overload .
Original	- continued beta blocker (toprol) at an increased dose) for htn control.
SMT	- continue beta blocker (metoprolol) at an increased dose) for *hypertension* control .
Original	the patient is not on a statin because of hepatic dysfunction .
SMT	the patient is not on a statin because of *liver function* .
Original	a bal was done on and was still showing yeast , by specimens were clear .
SMT	a *bronchoscopy* was done on and was still showed yeast , by *cultures* were clear .
Original	her fs levels were noted to be between 120-150 on tubefeeds and po intake .
SMT	your *glucose levels* were found to be between 140 *on nutrition* and po intake
Original	humalog iss discontinued on for good glycemic control .
SMT	*stopped on sliding scale for glucose control* good .
Original	pathology shows metastatic carcinoma morphologically consistent with breast origin , present at cauterized surgical margin .
SMT	*biopsy showed cancer morphologically consistent with breast cancer* , causing present at *clipped surgical area* .
Original	r femoral line placed .
SMT	*right central* line placed .
Original	a cystoscopy was performed intraop and a catheter was placed .
SMT	a *procedure* was performed and a foley catheter was placed .
Original	the patient developed an abdominal abscess .
SMT	the patient had an abdominal *infection* .
Original	lenis were negative for dvt .
SMT	*tests* were negative for anticoagulation .
Original	he has advanced well with his feeding .
SMT	you have *tolerated well with your nutrition* .
Original	hypoxia : patient desat to high 80s on ra while asleep .
SMT	*respiratory* : patient *oxygen* to high 80s on / while sleeping .
Original	the patient was subsequently reintubated secondary to pulmonary compromise .
SMT	the patient was *later intubated due to pulmonary injury* .
Original	she was treated with po levofloxacin with plan for a 14 day course .
SMT	you were treated with *oral antibiotics* with plan for a 14 day course .
Original	likely etiology was hypotension leading to underperfusion of coronary artery and possible conduction system abnormalities .
SMT	was likely due to transient hypotension *complication* of coronary artery and possible *rhythm system* abnormalities .

Table 2.8: Examples of unsupervised sentence translations. Italicizing words and phrases represent functionally correct translation, and underlining words and phrases represent inappropriate translation or the words/phrases require better translation.

contextualized word representation is an approach to be considered [Peters et al., 2018].

2.8 Summary

In this study, we recognize a strong need for a clinical professional to consumer language translation, which is a difficult task even using the current state-of-the-art MT method and system from the general domain of NLP. However, we demonstrate that our novel fully-unsupervised translation framework works in the setting of non-parallel corpora without the assistance of expert-curated knowledge, which is not possible using traditional approaches but essential for scalability while tackling real-world clinical data—since data availability and expert curation are always problems for clinical machine learning. We utilize unsupervised natural language representations, iterative Procrustes process with anchor information for BDI, and SMT using language modeling and obtain promising performance in both word and sentence-level translations evaluated by two quantitative tasks, and validated by human (both clinicians and non-clinicians) judgment. The newly-proposed two-step evaluation for sentence translation without ground truth reference, which considers both human experts (clinical correctness) and laymen (readability) judgments, is also helpful for practical use.

Some limitations in our study shed light on future directions. With the proposed framework, we find that the readability of the translation is superior but the correctness is slightly inferior to using supervised dictionary-based word replacement, which mainly results from the trade-off between over-simplification and over-specification. The loss of correctness or mistranslation in the clinical language translation may be harmful due to over-simplification.

Possible solutions for such obstacles to deployment can be identified in two directions. First, we can leverage the method by integrating domain knowledge with the idea of definition insertion from existing dictionaries. We may leverage the clinical concept-level information by just focusing on translating the words and phrases that match UMLS biomedical concepts. Linguistic features can also be considered in the process of translation. The proposed machine learning-based approach can be combined with linguistic characteristics of the corpus [Biran et al., 2011], in order to have better control of language simplification. Quantitative evaluation that considers linguistic features may help strengthen the interpretability [Feng et al.,

2010]. Adopting contextualized representations is also an alternative to improve the quality of embeddings and models. For generalizability, conducting more experiments and results on datasets in other domains with similar settings is also considered. Finally, we may expand the clinician-designed word pairs set and deploy the framework for public use as an online translator. The framework and results in this work could ultimately be helpful to improve patient engagement in their own health care, and toward the era of personalized medicine.

Chapter 3

Self-supervised Multimodal Learning for Limited Data

3.1 Overview

Modern deep learning algorithms gear towards clinical adoption usually rely on a large amount of high fidelity labeled data. Low-resource settings pose challenges like acquiring high fidelity data and become the bottleneck for developing artificial intelligence applications. Ultrasound images, stored in Digital Imaging and Communications in Medicine (DICOM) format, have additional metadata corresponding to ultrasound image parameters and medical exams. In this work, we leverage DICOM metadata from ultrasound images to help learn representations of the ultrasound image. We demonstrate that the developed method outperforms the approaches without using metadata across various downstream tasks.

3.2 Background

In recent years, deep learning algorithms have made forays into the clinical domain and have emerged as a successful technique in various medical imaging applications. They have shown the potential to automate disease detection, severity grading, and clinical diagnosis in different domains [Hu et al., 2019, Gulshan et al., 2016, Esteva et al., 2017]. However, clinically accepted deep learning algorithms require a considerable amount of annotated data.

For example, Gulshan et al. [2016] utilizes more than 100,000 images to train and validate the algorithm. Unfortunately, obtaining accurate annotations from clinicians is extremely expensive, constraining supervised learning approaches in limited data and low-resource settings.

Unsupervised or semi-supervised learning provides potential solutions to alleviate the problems by learning the data distribution without or with limited labels. Studies have shown that unsupervised pre-training can serve as a regularization method and lead to better generalization [Erhan et al., 2010]. Recently, weakly-supervised and self-supervised learning have also drawn significant attention to their ability to learn high-quality feature representations. In this chapter, we will explore one of the self-supervised techniques, the context encoder [Pathak et al., 2016], and use the metadata in medical imaging as the weak labels to reinforce its capability to learn representation features.

In most of the modern medical imaging acquisition devices, such as ultrasound imaging, the data is stored in DICOM format. Besides the image pixel data, the DICOM headers contain the metadata, such as the patient information, study descriptions, and the reported results. The abundant information encoded in DICOM format provides a unique opportunity for modern deep learning applications. Recent studies have shown that the metadata can be leveraged for series categorization using machine learning [Gauriau et al.]. Nevertheless, DICOM has not been a popular supervision target in machine learning. One major concern about DICOM is that its metadata is often noisy and may contain wrong tags [Gueld et al., 2002]. In practice, clinical personnel often adjust the examination protocol and imaging presets to improve the image quality, but these changes may not be reflected correctly in the DICOM tags. However, using DICOM metadata as weak labels may help incorporate valuable information into the deep learning algorithm while minimizing the noise.

In this work, we investigate weakly-supervised learning using metadata and develop a framework built on top of the self-supervised learning method. We show that incorporating DICOM metadata as weak labels can improve the quality of representation learning and improve the performance of the downstream segmentation and classification tasks.

3.3 Related Works

3.3.1 Pre-training Techniques

It is usually beneficial to train a model from pre-trained weights, rather than from random initialization, especially in the medical imaging field, where the labels are expensive to obtain [Erhan et al., 2010]. There are multiple ways for pre-training. The first is transfer learning, which first trains the model on a large amount of labeled data, and then tunes the pre-trained weights for new target tasks. ImageNet-pre-trained convolutional neural networks (CNN), which is arguably the most successful transfer learning model, have boosted the growth of the modern deep learning applications [Deng et al., 2009]. Even in medical imaging, the standard approach is to take an existing architecture trained on ImageNet and then fine-tune on the domain-specific data such as X-ray [Rajpurkar et al., 2017] or fundoscopic imaging [Abràmoff et al., 2016]. However, given the substantial difference between the natural images and medical imaging, recent studies raised questions about the precise effects of the pre-trained features and suggested that transfer learning does not constantly improve the final performance [Raghu et al., 2019, Kornblith et al., 2019, He et al., 2019].

While transfer learning relies on supervision from large-scale hand-labeled databases without employing the rich information presented in the image structure, unsupervised learning, another approach for pre-training, tries to build a useful feature representation using the data itself [Bengio, 2012]. For example, Hinton et al. [2006] presented a greedy layerwise unsupervised pre-training method to build representations of different levels. Variants of the autoencoder [Baldi, 2012], such as stacked denoising autoencoder [Vincent et al., 2010] or contractive autoencoder [Rifai et al., 2011], build the encoder to reconstruct the original image. In recent years, the generative adversarial network (GAN) also emerged as a robust framework for representation learning [Donahue et al., 2017, Donahue and Simonyan, 2019]. These methods train a network without labels, and the learned weights can be used either as high-level image feature inputs or as initialization for a target downstream task.

3.3.2 Self-supervised learning

Self-supervised learning is a unique form of supervised learning which eliminates the demand for manual labels. The key idea is to generate labels from the data itself and train the network in a supervised manner. Such methods, also known as pretext tasks, have proved to be an effective technique for representation learning, and have been widely used in natural language processing. For example, Bidirectional Encoder Representations from Transformers (BERT) [Devlin et al., 2019], one of the recent breakthroughs in language model pre-training, is trained to predict the masked words given the input sequences. In image-based tasks, many methods have also been proposed. Gidaris et al. [2018] randomly rotated the images while maintaining the semantic content unchanged, and the network was trained to predict the rotation angles. Noroozi and Favaro [2016] formulated the pretext task as a jigsaw puzzle and pre-trains the model by solving it. Contrastive predictive coding (CPC) learned an encoder to encode image patches and utilized an autoregressive decoder to predict the future vectors with a contrastive loss [Oord et al., 2018]. Chen et al. [2020b] further improved the training techniques of contrastive learning and had achieved performance close to that of supervised pre-training.

In this work, we employ the context encoder [Pathak et al., 2016] as the foundation of our proposed framework, in which the network is trained to predict the missing parts of the images. We leave the detailed descriptions of the context encoder to section 3.4.1.

3.3.3 Weakly-supervised learning

Weakly-supervised learning is another subclass of supervised learning, in which the labels can be either inexact or inaccurate. Inexact supervision usually involves annotations at a higher abstraction level. For example, Wang et al. [2017] and Yan et al. [2018] localized the location of pulmonary diseases in Chest X-rays with image-level classes (e.g., providing a bounding box of "atelectasis"); Hu et al. [2018] showed that the model trained on the position coordinates could improve the segmentation task.

Inaccurate supervision uses a large quantity of low-quality or noisy labels. One remarkable illustration is the work in Mahajan et al. [2018], which took advantage of billions of

Instagram hashtags for weakly-supervised pre-training to boost the ImageNet classification. Recently, Xie et al. [2020] leveraged noisy labels from a teacher-student framework and achieved the state-of-the-art classification accuracy on ImageNet.

Inspired by these works, we propose incorporating DICOM metadata, which has noisy labels embedded within the raw medical imaging data, for weakly-supervised pre-training.

3.3.4 Adversarial Training

An adversarial loss was added when training the context encoder to encourage realistic output. Adversarial training originates from the GAN [Goodfellow et al., 2014, Radford et al., 2016], which utilizes a discriminator network to distinguish generated images from real inputs. Beyond its great success in image generation, it also shows a substantial impact on other areas like domain adaptation [Ganin et al., 2016] or adversarial attack [Tramèr et al., 2018]. In typical tasks such as semantic segmentation, adversarial training can also boost the performance under semi-supervised [Hung et al., 2018] or unsupervised settings [Chen et al., 2019a].

A standard adversarial network does not require supervision, but recent studies have shown that the class labels can stabilize the training and improve image qualities. For example, Brock et al. [2019] fed the labels as the generator inputs to produce high-quality images. ACGAN [Odena et al., 2017] used the discriminator to classify the class labels as an auxiliary loss. Miyato and Koyama [2018] proposed a linear projection layer, which was also employed in Lučić et al. [2019] to generate high fidelity images with a limited number of labels.

3.4 Methods

We develop a new self-supervised representation learning framework, which incorporates the DICOM metadata as weak labels to improve the training. In particular, we employ the context encoder as the self-supervised pretext task. The overview of the framework is demonstrated in Figure 3-1.

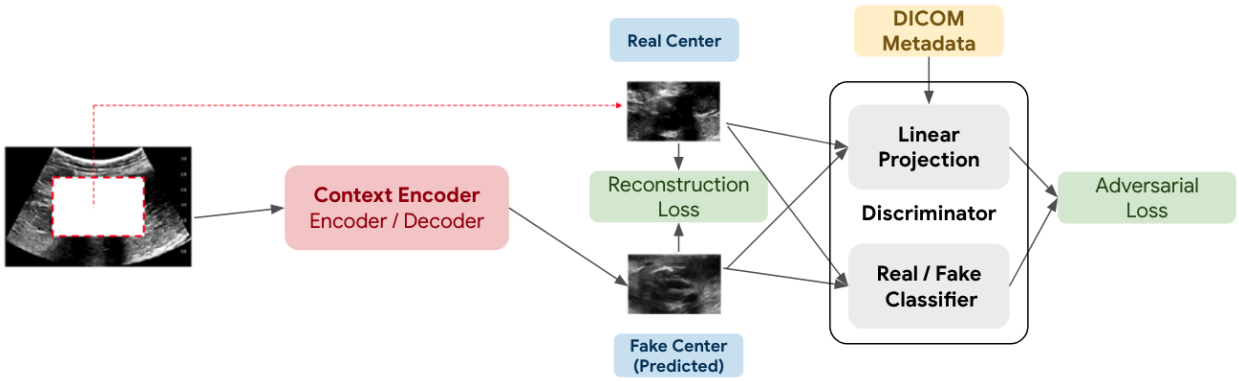


Figure 3-1: The developed framework for learning representation via self-supervised multi-modal learning with the context encoder and the DICOM metadata integration.

3.4.1 Context Encoder

The idea of the context encoder is that given an input image with intentionally masked out areas, we train a deep learning model to reconstruct the missing part (semantic in-painting) [Pathak et al., 2016]. The network utilizes an encoder-decoder structure. The encoder encodes the image context into a compact latent representation, and the decoder employs this to generate the missing image content. The network is trained to minimize the mean squared reconstruction loss.

In Pathak et al. [2016], it is proposed that the in-painting area can be either fixed or random blocks. Typically, models using random blocks tend to generalize better. However, due to the nature of ultrasound images, where the informative context is located in the central region, we crop a rectangular patch in the center of the image with a fixed size equal to half of the image width and height.

3.4.2 Discriminator with Linear Projection Layer

We also add the discriminator for adversarial training to encourage realistic output. The standard $\mathcal{L}_{adv}$ is formulated as:

$$\mathcal{L}_{adv} = \max_{D} \mathbb{E}_{x \in \mathcal{X}}[\log(D(x)) + \log(1 - D(F(\hat{x})))] \tag{3.1}$$

where F is the context encoder, D is the discriminator, x is an original image, and $\hat{x}$ is cropped input image. (Noted that F is often denoted as G in most of the GAN literature; here we use F to distinguish the context encoder and a regular GAN generator.)

To incorporate the DICOM metadata, we employ a linear projection layer as proposed in Miyato and Koyama [2018] and Lučić et al. [2019]. The discriminator is decomposed into a learned discriminator representation, $\tilde{D}(x)$, and the representation is then fed into two different parts: (1) A classifier C_{rf} to distinguish whether the image is real or fake; (2) A linear projection layer P, with a learned weight matrix W applied to a feature vector $\tilde{D}(x)$ and the encoded DICOM tags y as an input. The output of the discriminator becomes:

$$D(x, y) = C_{rf}(\tilde{D}(x)) + P(\tilde{D}(x), y)$$

where $P(\tilde{D}(x), y) = \tilde{D}(x)^{\top} W y$. Also, we adopt a hinge version of the adversarial loss. With the above modification, the loss function for context encoder F and the discriminator D can be rewritten as:

$$\mathcal{L}_D = -\mathbb{E}_{(x,y)\sim p(x,y)}[\min(0, -1 + D(x, y)))] - \mathbb{E}_{(\hat{x},y)\sim p(x,y)}[\min(0, -1 - D(F(\hat{x}), y))] \quad (3.2)$$

$$\mathcal{L}_F = -\lambda_{adv} \times \mathbb{E}_{(\hat{x},y)\sim p(x,y)}[D(F(\hat{x}), y)] + \lambda_{rec} \times \mathbb{E}_{(x,\hat{x})\sim p(x)}[(F(\hat{x}) - x)^2] \quad (3.3)$$

The second term of Equation(3.3) is the reconstruction loss (mean squared error). We include two hyperparameters λ_{adv} and λ_{rec} to balance the two different losses.

3.5 Data

For the pre-training task, a retrospective private database from Massachusetts General Hospital (MGH) was collected, from September 2018 to November 2019 after proper approval from the MGH Institutional Review Board (IRB). Informed consent was waived, and Health Insurance Portability and Accountability Act (HIPAA) compliance were ensured. A total of 12,267 images from 1,188 unique patients were collected. All images were previously acquired using Supersonic Aixplorer ultrasound machine (SuperSonic Imagine S.A., Aix-en-Provence,

France).

We evaluate the results on three different downstream tasks: (1) Quality score classification on a private dataset. (2) Liver and kidney segmentation on a private dataset. (3) Thyroid nodule segmentation on an open dataset. The two private datasets were retrospectively collected from the same institution. All the images were acquired using a GE Logiq E9 ultrasound machine (GE Healthcare, Chicago, IL, USA).

There is no overlap between our pre-training dataset and the downstream evaluation dataset. The description for the open dataset can be found in Pedraza et al. [2015]. The overview of the dataset is shown in Table 3.1.

Dataset	**Task**	**# of images**	**# Train**	**# Val**	**# Test**
Private	Semantic In-painting	12267	9814	2453	0
Private	Quality Classification	3226	2548	343	335
Private	Liver/Kindey Segmentation	591	391	100	100
Public	Thyroid Nodule Segmentation	466	298	74	94

Table 3.1: Description of the dataset.

3.5.1 DICOM Metadata

We select two DICOM tags as the target since they directly relate to the image semantic context:

- **Transducer data** (DICOM tag: (0018, 5010)), which indicates the probe type used for examination. There are three different transducer probes in the dataset— SC6-1, SL10-2, SL15-4, where S represents single crystal, C or L represents curvilinear or linear probe geometry, and the numbers represent the ultrasound frequency bandwidth in MHz. We classify the probes into two groups—linear (SL10-2, SL15-4) and curvilinear (SC6-1).

- **Study Description** (DICOM tag: (0008, 1030)). The study description illustrates the protocol when performing the ultrasound exam. For example, images of "US BIOPSY LIVER NONFOCAL" are acquired during an ultrasound-guided liver biopsy. Therefore, we can expect that these images are predominantly liver. We identified 45 differ-

ent study descriptions in our dataset (Table 3.2). Due to the spurious nature of the tags, we categorize the study series into eight different groups according to procedure type or site, including liver, kidney, thyroid, abdomen, chest, soft tissue, nodule, and drainage. The DICOM categorization is performed manually by a board-certified radiologist. Each study series can belong to more than one group. For example, the tag "US BIOPSY LIVER NONFOCAL" is mapped to two groups—liver and abdomen. We binarized the DICOM labels in a multi-label format.

Figure 3-2 demonstrates some image examples of the DICOM tags.

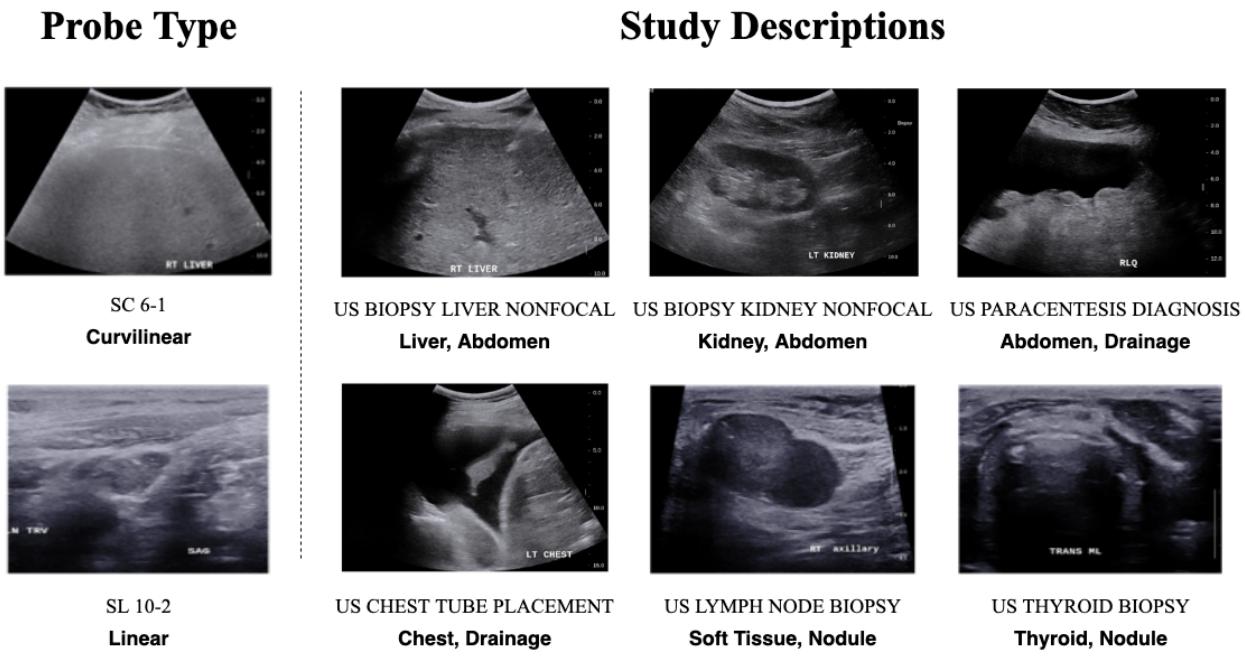


Figure 3-2: Examples of DICOM metadata.

Study Descriptions	Encoding
US BIOPSY LIVER NONFOCAL	liver,abdomen
US BIOPSY LIVER FOCAL	liver,abdomen
US LYMPH NODE BIOPSY	soft tissue,nodule
US BIOPSY KIDNEY NONFOCAL (EITHER SIDE)	kidney,abdomen
US PARACENTESIS THERAPEUTIC	abdomen,drainage
US BIOPSY TRANSPLANTED KIDNEY	kidney,abdomen
US PARACENTESIS DIAGNOSTIC AND THERAPEUTIC	abdomen,drainage
US THYROID BIOPSY	thyroid,nodule
US PARACENTESIS DIAGNOSTIC	abdomen,drainage

Study Descriptions	Encoding
US THORACENTESIS DIAGNOSTIC AND THERAPEUTIC	chest,drainage
US THYROID ASPIRATION/FNA	thyroid,nodule
US DRAINAGE INTERVENTION NOT OTHERWISE SPECIFIED	soft tissue,drainage
US DRAINAGE ABDOMEN	abdomen,drainage
US DRAINAGE GALLBLADDER (CHOLECYSTOSTOMY)	abdomen,drainage
US THORACENTESIS THERAPEUTIC (RIGHT)	chest,drainage
US THORACENTESIS THERAPEUTIC (LEFT)	chest,drainage
US BIOPSY MESENTERY	abdomen,drainage,soft tissue
US NECK SOFT TISSUE BIOPSY	soft tissue,nodule
US DRAINAGE CATHETER PLACEMENT	soft tissue,drainage
US DRAINAGE PELVIS	abdomen,drainage
US SOFT TISSUE BIOPSY	soft tissue,nodule
US BIOPSY KIDNEY NONFOCAL (LEFT)	kidney,abdomen
US CHEST TUBE PLACEMENT (RIGHT)	chest,drainage
US BIOPSY NOT OTHERWISE SPECIFIED	soft tissue,nodule,drainage
US ABDOMINAL PELVIC BIOPSY NOT OTHERWISE SPECIFIED	soft tissue,nodule,drainage
US CHEST TUBE PLACEMENT (LEFT)	chest,drainage
CT BIOPSY LIVER FOCAL	liver,abdomen
US BIOPSY KIDNEY FOCAL (LEFT)	liver,abdomen
US ASPIRATION ABDOMINAL COLLECTION	abdomen,drainage
CT LYMPH NODE BIOPSY	soft tissue,nodule
US DRAINAGE LIVER	liver,drainage,abdomen
US BIOPSY RETROPERITONEUM	abdomen
US LYMPH NODE ASPIRATION/FNA	soft tissue,nodule,drainage
US SOFT TISSUE ASPIRATION	soft tissue,drainage
US ASPIRATION PELVIS	abdomen,drainage
US THORACENTESIS DIAGNOSTIC (RIGHT)	chest,drainage
US THORACENTESIS DIAGNOSTIC (LEFT)	chest,drainage
US DRAINAGE KIDNEY/PARARENAL (RIGHT)	abdomen,kidney,drainage
US HEAD/NECK INTERVENTION NOT OTHERWISE SPECIFIED	soft tissue
US BIOPSY KIDNEY FOCAL (RIGHT)	kidney,abdomen
CT ABDOMINAL PELVIC BIOPSY NOT OTHERWISE SPECIFIED	abdomen
US DRAINAGE KIDNEY/PARARENAL (LEFT)	kidney,abdomen
IR PARACENTESIS (THERAPEUTIC)	abdomen,drainage
US PSEUDOANEURYSM THROMBIN INJECTION	soft tissue,nodule

Table 3.2: The full list of the DICOM Study Descriptions and the corresponding encoding.

3.6 Experiments

3.6.1 Context Encoder pre-training

Architecture

The developed framework consists of three parts—the encoder, the decoder, and the discriminator. For the encoder, we employ two different existing network architectures—VGG16 with batch normalization [Simonyan and Zisserman, 2015] and Resnet-50 [He et al., 2016] as the backbone. The decoder has four up-sampling blocks each with a 3×3 up-convolutional, a batch normalization, and a rectified linear unit (ReLU) layer. The discriminator also has four blocks, each with a 3×3 convolutional, a batch normalization, and a LeakyReLU layer as suggested by Radford et al. [2016].

Training

The dataset for semantic in-painting is split into training (80%, 9814 images) and validation set (20%, 2453 images) randomly while ensuring that all images from the same patient were within one set. All images are resized to an input size of 256 × 384 pixels, and Z-score is normalized before feeding into the network. Data augmentation is performed using random flipping, vertical and horizontal shifting.

We update the context encoder and discriminator parameters using the Adam optimizer [Kingma and Ba, 2014], minimizing $\mathcal{L}_F$ and $\mathcal{L}_D$ alternatively, with hyper-parameters set to $\beta_1 = 0.9$, $\beta_2 = 0.999$, batch size = 32, context encoder learning rate = 0.0001, and discriminator learning rate = 0.00001. The models are trained over 200 epochs without early stopping, and the ones with the lowest $\mathcal{L}_F$ on the validation set are selected for downstream evaluation. We don't split a held-out test set since we present our quantitative results on the downstream tasks.

3.6.2 Downstream Task Experiment Settings

For both VGG16 and ResNet50, we compare the following pre-training configurations:

- **Baseline**: The encoder is randomly initialized without pre-training.

- **ImageNet**: The encoder is trained on ImageNet classfication.
- **DCM**: The encoder is trained to predict the DICOM metadata directly.
- **CE**: The encoder is trained with the context encoder without the DICOM metadata.
- **CE + DCM**: The encoder is trained using our framework with DICOM metadata.
- **CE + DCM + F**: The encoder is trained using our framework with DICOM metadata, and the parameters of context encoder are frozen while training the downstream tasks.

3.6.3 Downstream Tasks Evaluation

Architecture

After the model is trained, only the encoder part is fine-tuned for downstream tasks. Downstream classification tasks using a classifier layer, consisting of a 1×1 convolutional, a dropout, a global average pooling, and a fully connected layer, is appended followed by a sigmoid activation function. For downstream segmentation tasks, we adopt an architecture similar to U-Net [Ronneberger et al., 2015], where the encoder arm is modified to be the pre-trained VGG16 or ResNet. We follow an implementation similar to Iglovikov and Shvets [2018], adding five up-convolutional blocks and skip connections to complete the network.

Classification

The downstream classification task, quality score classification, is to identify an optimal view for Morrison's pouch—an anatomic site between the right lobe of the liver and the right kidney. The images used in the classification task were reviewed by a board-certified radiologist and given five different rankings as the quality score (Figure 3-3).

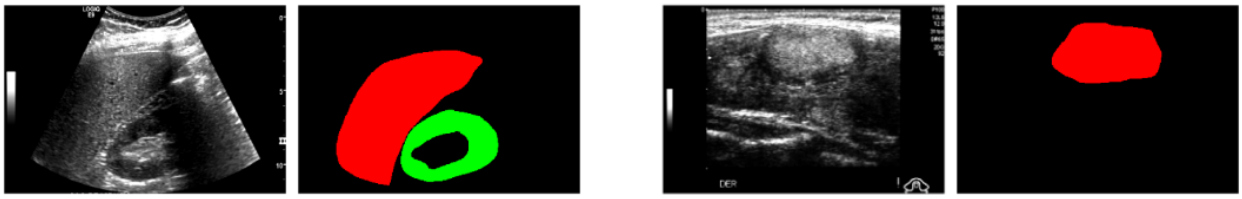
Figure 3-3: Quality score classification examples.

Class 0 indicates the view does not include the liver or the kidney, and should not be used; classes 1 and 2 are the correct Morrison's pouch view, but the anatomic structure is not clear enough for clinical applications; classes 3 and 4 are the clinically acceptable views, and class 4 represent the optimal Morrison's pouch view that will be used by an experienced operator. We use the ordinal encoding for the labels. (class 0: [0,0,0,0], class 1: [1,0,0,0], class 2: [1,1,0,0], class 3: [1,1,1,0], class 4: [1,1,1,1])

Clinically, the optimal view for Morrison's pouch is essential to identify ascites and hemoperitoneum when abnormal fluid accumulation is present. Furthermore, it is the reference view to estimate the severity of steatosis using the hepatorenal index. Therefore, quantifying the optimal view is crucial in an ultrasound examination.

Segmentation

We evaluate two different segmentation tasks. The first is to segment the kidney and liver in B-mode ultrasound imaging. This work is related to quality score classification. A board-certified radiologist select the images representing optimal Morrison's pouch view from the institutional database and manually annotated the kidney and liver anatomy (Figure 3-4 (a)). The second task is thyroid nodule segmentation, including cystic nodules, adenomas, and thyroid cancers, using an open access B-mode thyroid ultrasound image dataset. An example is shown in Figure 3-4 (b).



(a) Liver and Kidney Segmentation

(b) Thyroid Nodule Segmentation

Figure 3-4: Examples of the downstream segmentation tasks.

Training

For all three downstream tasks, the images follow the same pre-processing procedure described in section 3.6.1. The training hyper-parameters are summarized in Table 3.3. Mod-

els are trained without early stopping, and the epochs with the lowest validation loss are selected. All reported values are evaluated on the held-out test set.

	Quality Score	**Liver/Kidney Segmentation**	**Thyroid Segmentation**
Optimizer	Adam (($\beta_1 = 0.9$, $\beta_2 = 0.999$))	Adam (($\beta_1 = 0.9$, $\beta_2 = 0.999$))	Adam (($\beta_1 = 0.9$, $\beta_2 = 0.999$))
Batch Size	4	8	8
Training Epochs	300	500	500
Loss Function	Weighted Binary Cross Entropy	Soft Dice loss	Soft Dice Loss
Learning Rate	VGG16:0.00005 ResNet: 0.00005	VGG16: 0.0001 ResNet:0.00005	VGG16: 0.0001 ResNet:0.0001

Table 3.3: Training details for the downstream tasks.

3.7 Results

3.7.1 Context Encoder with DICOM

The qualitative results of the context encoder with and without DICOM tags are shown in Figure 3-5. We observe that training instances without DICOM tags are more prone to mode collapse in our experiments, making it challenging to obtain optimal results. The generated images look sharper with DICOM tags and can resemble the actual organ texture similar to liver and kidney. Figure 3-5 qualitatively shows that the joined weakly-supervised training with DICOM tags improved the prediction quality.

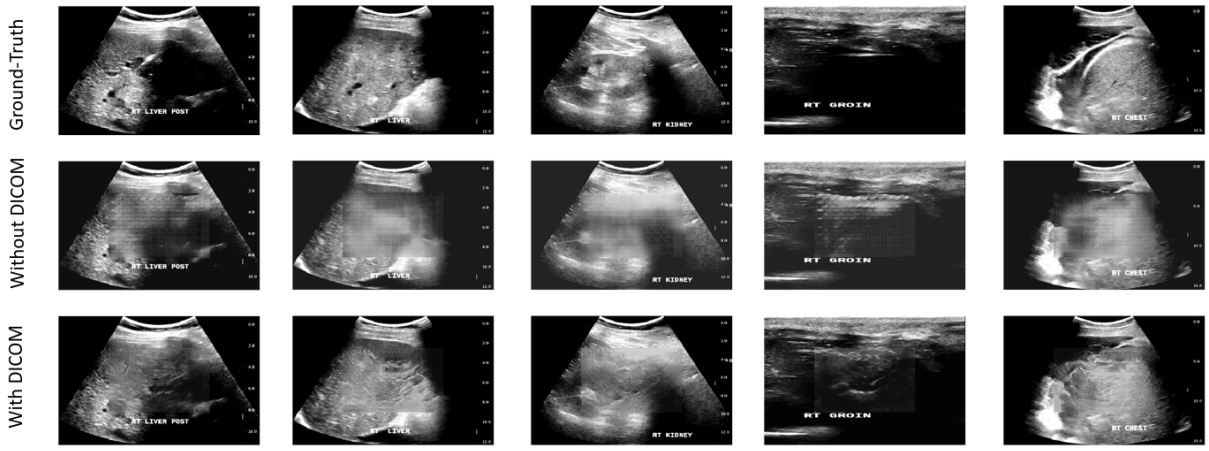


Figure 3-5: Example of semantic in-painting on various organs using the validation set.

3.7.2 Downstream Tasks

Next, we examined whether downstream segmentation and classification can benefit from the pre-trained encoder. The results are summarized in Table 3.4.

Backbone	**pre-training**	**Downstream Task**		
		Quality	Liver/Kidney	Thyroid
VGG16-BN	Baseline	0.558 ± 0.031	0.801 ± 0.011	0.856 ± 0.012
VGG16-BN	ImageNet	0.656 ± 0.034	0.824 ± 0.008	0.876 ± 0.009
VGG16-BN	DCM	0.652 ± 0.020	0.829 ± 0.008	0.859 ± 0.011
VGG16-BN	CE	0.629 ± 0.022	0.816 ± 0.009	0.858 ± 0.011
VGG16-BN	CE+DCM	**0.657 ± 0.031**	**0.832 ± 0.009**	**0.883 ± 0.010**
VGG16-BN	CE+DCM+F	0.645 ± 0.035	0.826 ± 0.009	0.879 ± 0.011
ResNet-50	Baseline	0.706 ± 0.029	0.765 ± 0.013	0.843 ± 0.011
ResNet-50	ImageNet	0.708 ± 0.035	0.811 ± 0.014	0.849 ± 0.011
ResNet-50	DCM	0.706 ± 0.028	0.753 ± 0.016	0.850 ± 0.011
ResNet-50	CE	0.715 ± 0.028	0.781 ± 0.011	0.849 ± 0.010
ResNet-50	CE+DCM	0.715 ± 0.024	0.807 ± 0.011	0.852 ± 0.011
ResNet-50	CE+DCM+F	**0.754 ± 0.024**	**0.814 ± 0.014**	**0.865 ± 0.010**

Table 3.4: Performance evaluation of downstream tasks undergoing pre-training. The reported value is the mean ± standard deviation on the held-out test set. The standard deviation is derived from bootstrapping 1000 times on the test set. Each time we sample 50% of the test data with replacement. The best performance is highlighted in boldface. The abbreviation of each pre-training configuration is specified in the section 3.6.2.

The performance with pre-training improved significantly across different tasks and configurations compared to the random-initialized baseline models. The pre-training from ImageNet and DICOM also works reasonably well; however, our method, context encoder with DICOM, consistently obtains the best results. The effect of freezing the encoder differs between the two backbones. When freezing the encoder, we are reusing the learned features directly, and the ResNet models benefit more from this approach; when unfreezing the encoders, we treat it as the self-supervised initialization, and the VGG16 gains more from this approach. The observation is consistent with the conclusion in Kolesnikov et al. [2019], i.e., the quality of the representation learned in self-supervised tasks deteriorates toward the final layers of the VGG network. In contrast, the skip connections in the ResNet architecture help prevent the degradation of the representation and this is the best performer when reusing the features up to the pre-logit layers.

To emphasize the impact of adding DICOM metadata, we further repeat the experiments using a smaller data regime. We compare three configurations: CE + DCM (+F), CE, and baseline model, but only using 5% of the data for all three tasks. The results are shown in Figure 3-6; note that we only freeze the encoder for the ResNet backbone given the previous conclusion. The box-plots show that adding the metadata improved the performance in all cases. The difference between CE + DCM (+F) and CE is statistically significant (p-value < 0.05) in all combinations, except for the quality score classification using VGG16 as the backbone (p-value $= 0.111$).

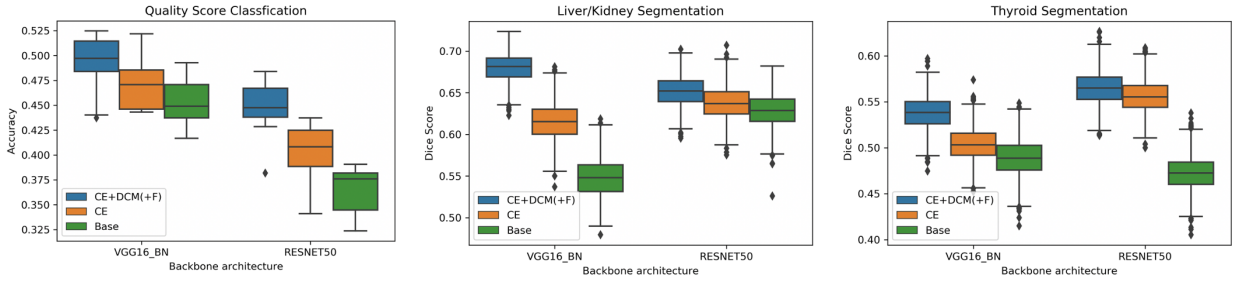


Figure 3-6: Box-plots between the three configurations **CE+DCM(+F)**, **CE** and **Baseline** across two backbone architectures and three downstream tasks. **CE+DCM(+F)** here denotes **CE+DCM+F** when we use ResNet as backbone and **CE+DCM** when using VGG16.

3.8 Discussion

In this study, we demonstrate that the performance of existing self-supervision techniques can be consistently boosted with DICOM metadata as weak labels. Compared to other pre-training data sources like ImageNet, which often comprises millions of entries, our methods achieve comparable performance with only around 10,000 images. Mahajan et al. [2018] suggests that, while increasing the size of the pre-training dataset may be beneficial, selecting a label space for the source task to match that of the target task is even more fruitful. In our experiments, the pre-training and the downstream dataset share a similar distribution; they both cover standard ultrasound examination views such as the abdominal and thyroid scans. Our results further emphasize the benefits of pre-training from data of the same domain.

Such a method can be particularly useful in limited data and low-resource settings where obtaining and training large-scale annotated data is not feasible, and that leads to reducing the gap toward building a generalized and robust medical imaging pre-training technique.

The choice of DICOM tags is also crucial to the success of the application. We only experiment with ultrasound images and two DICOM tags. Different image modality like computed tomography (CT) and magnetic resonance imaging (MRI) have their specific metadata and would require further investigation to identify the proper candidates. Potential targets such as voxel information (pixel spacing, Hounsfield units), study details (anatomic structure, patient orientation), or patient-level data (demographics, diagnosis) can provide meaningful semantic information for supervised learning. Although some tags such as study descriptions or study findings may be inconsistent among different acquisition devices or institutions, they can still be valuable with proper categorization by the clinical experts.

In our experiments, we focus on the advantages of DICOM metadata and only investigate one self-supervised method. However, the methodology used in this paper, adding the DICOM weak labels to the discriminator, can be generalized to other pretext tasks given that adversarial training is often used as an auxiliary loss to many existing models. For example, Chen et al. [2019d] extended the self-supervised rotation loss with a GAN-based structure. Exploration of different pretext tasks will be necessary for future studies.

3.9 Summary

In this chapter, we demonstrate the potential of using DICOM metadata from ultrasound images as weak labels to improve representation learning in a self-supervised schema. The method can significantly impact resource-limited regions by leveraging its ability to effectively utilize the pre-existing information, curtailing the need for additional annotations, which require high skill and are expensive. The method can be extended to other medical image modalities with DICOM tags, like CT or MRI.

Chapter 4

Contrastive Learning for Class Imbalance

4.1 Overview

Deep neural networks have been investigated in learning latent representations of medical images, yet most of the studies limit their approach to a single supervised convolutional neural network (CNN) method, which usually relies heavily on a large-scale annotated dataset for training. To learn image representations with less supervision involved, we develop deep Siamese CNN (SCNN) architecture that can be trained with only binary image pair information. We evaluate the learned image representations on a task of content-based medical image retrieval (CBMIR) using a publicly available multiclass diabetic retinopathy (DR) fundoscopic imaging dataset. The experimental results show that our proposed deep SCNN is comparable to the state-of-the-art single supervised CNN, and requires much less supervision for training.

4.2 Background

Effective feature extraction and data representation are critical factors to successful medical imaging tasks. Researchers usually adopt medical domain knowledge and ask for annotations from clinical experts. For example, using traditional image processing techniques such as filters or edge detection techniques to extract clinically relevant spatial features from images obtained by different image modalities, such as mammography [Tsochatzidis et al., 2017],

lung computed tomography (CT) [Dhara et al., 2017], and brain magnetic resonance imaging (MRI) [Jenitta and Ravindran, 2017]. The handcrafted features with supervised learning using expert-annotated labels work appropriately for specific scenarios. However, using predefined expert-derived features for data representation limits the chance to discover novel features. It is also very expensive to have clinicians and experts label the data manually, and such a labor-intensive annotation task limits the scalability of learning generalizable medical imaging representations.

To learn efficient data representations of medical images, researchers recently have used different deep learning approaches and applied these to various medical image machine learning tasks, such as image classification [Esteva et al., 2017, Gulshan et al., 2016], image segmentation [Havaei et al., 2017, Guo et al., 2017], or CBMIR [Litjens et al., 2017, Sun et al., 2017, Anavi et al., 2016, Liu et al., 2016, Shah et al., 2016].

CBMIR is a task that helps clinicians make decisions by retrieving similar cases and images from the electronic medical image database [Müller et al., 2004] (Figure 4-1). CBMIR requires expressive data representations for knowledge discovery and similar image identification in massive medical image databases, and has been explored by different algorithmic approaches [Kumar et al., 2013, Müller et al., 2004].

However, the previous works of CBMIR mainly focused on using shallow learning algorithms, or combining a single pre-trained CNN structure with other techniques [Litjens et al., 2017, Sun et al., 2017, Anavi et al., 2016, Liu et al., 2016, Shah et al., 2016], which relies heavily on manually annotated, high-quality ground truth labeling.

To mitigate these issues, we develop an SCNN-based method that learns fixed-length latent image representation from solely image pair information in order to reduce the dependency on using actual class labels annotated by human experts [Bromley et al., 1994]. We then evaluate the learned image representations on the task of CBMIR using a publicly available DR fundus image dataset. We compare the image representations learned by the proposed deep SCNN with the single pre-trained supervised CNN architecture [He et al., 2016].

The architecture of the deep SCNN is illustrated in Figure 4-2. The deep SCNN learns to differentiate an image pair by evaluating the similarity and relationship between the given

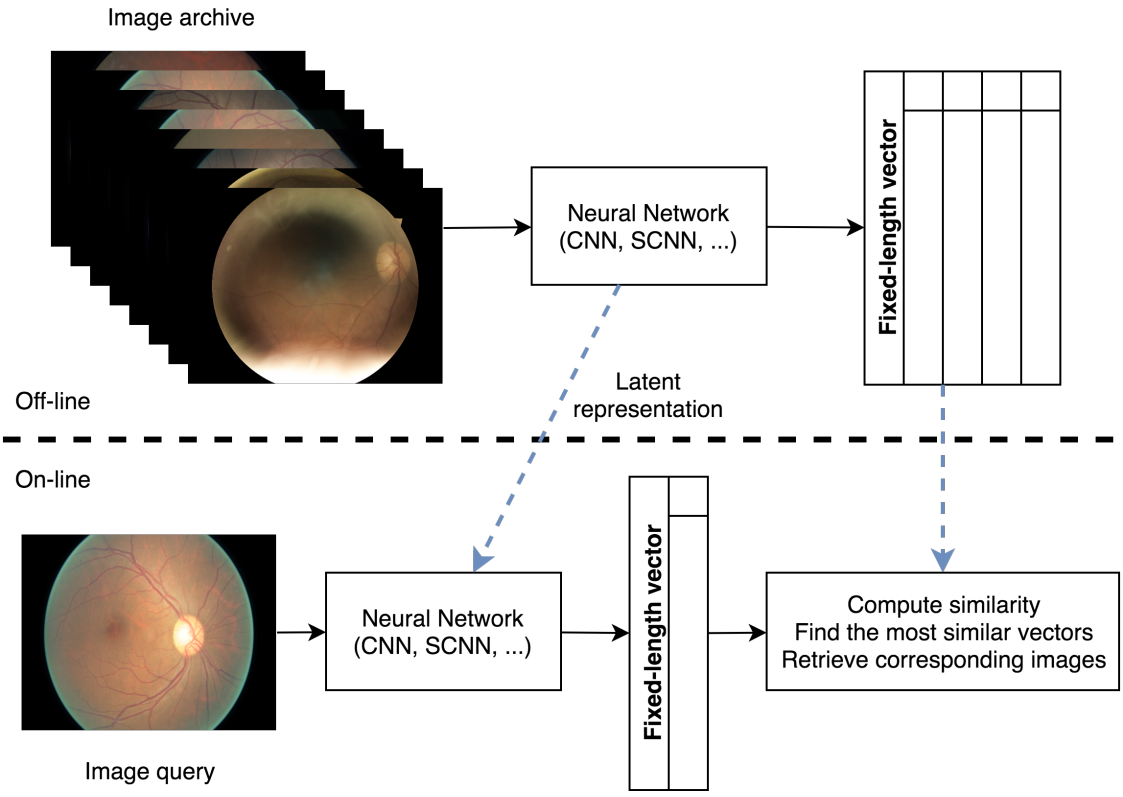


Figure 4-1: Overview of content-based medical image retrieval (CBMIR).

images. Each image in the image pair is fed into one of the identical CNN, and the contrastive loss is computed between two outputs of CNNs. The model is an end-to-end structure to obtain a latent representation of the image, which can be used for further CBMIR tasks.

The main contributions of this work are that we develop an end-to-end deep SCNN model for learning latent representations of medical images with minimal expert labeling efforts by reducing the multiclass problem to a binary class learning problem, and applying the model to the task of CBMIR using retina fundoscopic imaging as a proof of concept. Experimental results show that SCNN's performance is comparable to that of the state-of-the-art CBMIR method using a single supervised pre-trained CNN, but requires much less supervision for training.

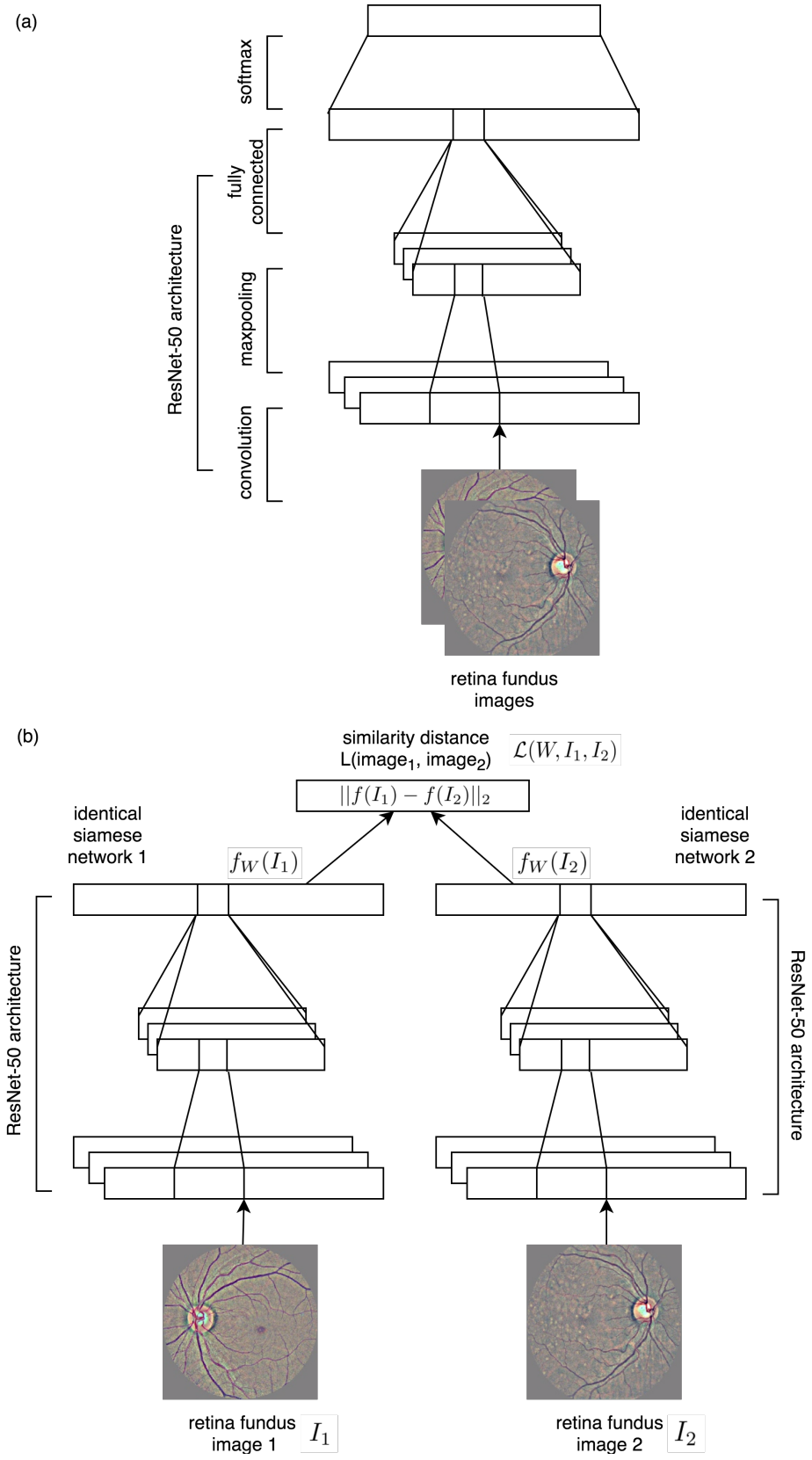


Figure 4-2: Architecture of the model used in the study. (a) Single CNN, and (b) deep Siamese CNNs.

4.3 Related Works

Recently, deep neural networks have been adopted in medical image learning tasks and yield state-of-the-art performance in many medical imaging problems [Litjens et al., 2017]. Using deep neural networks allows automatic feature extraction and general, expressive representation learning for different computer vision tasks [Bengio et al., 2013], including machine learning tasks for medical imaging. After Krizhevsky et al. [2012] yielded a breakthrough performance using deep CNN for the ImageNet challenge [Deng et al., 2009], supervised learning with CNN architecture has become a general structure for visual tasks. For medical images, researchers mainly use CNNs, stacked autoencoders [Cheng et al., 2016], and restricted Boltzmann machines [Brosch et al., 2013] for different tasks such as classification [Esteva et al., 2017, Gulshan et al., 2016], segmentation [Havaei et al., 2017, Guo et al., 2017], image generation and synthesis [Nie et al., 2017, Van Nguyen et al., 2015], image captioning [Moradi et al., 2016, Shin et al., 2015], and CBMIR [Sun et al., 2017, Anavi et al., 2016, Liu et al., 2016, Shah et al., 2016].

Deep learning is not yet widely adopted in CBMIR except for a few studies on lung CT [Sun et al., 2017], prostate MRI [Shah et al., 2016], and X-ray images [Anavi et al., 2016, Liu et al., 2016]. Sun et al. [2017] applied CNN with a residual network to retrieve lung CT images. Shah et al. [2016] adopted CNN with a hashing-forest method for prostate MRI image retrieval. Anavi et al. [2016] used a five-layered pre-trained CNN, extracted the image representation in the fully-connected layer, integrated textual metadata, and fed it into a support vector machine (SVM) classifier for distance measurement. Liu et al. [2016] combined three-layer CNN with Radon barcodes to retrieve images from 14,410 chest X-ray images. Different from the previous works, our approach elaborates the capability of deep SCNN to reduce the labeling effort by using only binary image pair information, rather than the exact multiclass labeling.

4.4 Methods

Our method utilizes end-to-end deep SCNN architecture to learn fixed-length representations from images with minimal expert labeling information [Bromley et al., 1994].

4.4.1 Deep Siamese Convolutional Neural Networks

Deep SCNN architecture is a neural network variant that can find the relationship and similarity between the input objects. It has multiple symmetric subnetworks tying the same parameters and weights and updating mirrored, and cojoining at the top by an energy function. Siamese neural networks are designed initially to solve the signature verification problem of image matching [Bromley et al., 1994]. It has also been used for one-shot image classification [Koch et al., 2015].

We construct deep SCNN for learning fixed-length representations using two identical CNNs sharing the same weights. In our experiment, each identical CNN is built using the ResNet-50 architecture with the ImageNet pre-trained weights [He et al., 2016]. We use 25% dropout for regularization to reduce overfitting and adopted batch normalization [Srivastava et al., 2014, Ioffe and Szegedy, 2015]. The rectified linear units (ReLU) nonlinearity is applied as the activation function for all layers, and we use the adaptive moment estimation (Adam) optimizer to control the learning rate [Kingma and Ba, 2014]. The similarity between images is calculated by Euclidean distance, and we define the loss function by computing the contrastive loss [Hadsell et al., 2006], which can be presented in the equation:

$$\mathcal{L}(W, I_1, I_2) = \mathbf{1}(L = 0)\frac{1}{2}D^2 + \mathbf{1}(L = 1)\frac{1}{2}[\max(0, margin - D)]^2 \tag{4.1}$$

where I_1 and I_2 are a pair of retina fundoscopic images fed into each of two identical CNNs. $\mathbf{1}(\cdot)$ is an indicator function to show whether two images have the same label, where $L = 0$ represents that the images have the same label and $L = 1$ represents the opposite. W is the shared parameter vector that neural networks will learn. $f(I_1)$ and $f(I_2)$ are the latent representation vectors of input I_1 and I_2, respectively. D is the Euclidean distance between $f(I_1)$ and $f(I_2)$, which is $||f(I_1) - f(I_2)||_2$.

Compared to the single supervised CNN, which uses multiclass information, the SCNN transforms the multiclass problem into a binary classification learning problem.

4.4.2 Baseline

In this study, we compare end-to-end deep SCNN with an end-to-end single supervised ResNet-50 architecture. We implement all neural networks with Keras.

4.4.3 Evaluation

We use two metrics to evaluate the performance of CBMIR, (1) mean reciprocal rank (MRR),

$$MRR = \frac{1}{Q}\Sigma_{i=1}^{Q}\frac{1}{rank_i} \tag{4.2}$$

where Q is the query size and $rank_i$ means that the rank of the real first-ranked item in the i-th query and (2) mean average precision (MAP),

$$MAP = \frac{1}{Q}\Sigma_{i=1}^{Q}AveP \tag{4.3}$$

where $AveP$ is the area under the precision-recall curve.

4.5 Experiments

We conduct experiments and train our model on a subset of the DR fundoscopic image dataset to demonstrate the capability of the SCNN architecture. We then analyze and evaluate the performance of learning representations and CBMIR between different approaches.

4.5.1 The Diabetic Retinopathy Fundus Image Dataset

As a severe complication of diabetic mellitus, DR is a common cause of blindness worldwide, especially in developed countries due to the high prevalence of diabetes mellitus. Screening and detection of early DR are therefore critical for disease prevention. The DR fundoscopic

image database* is collected, maintained, and released by EyePACS, a free platform for retinopathy screening, and released as the dataset for Kaggle competition. We use the whole training set of the Kaggle Diabetic Retinopathy Detection challenge with 35,125 fundoscopic images. Five clinical severity labels from normal/healthy to severe (labeled as 0, 1, 2, 3, and 4 in the dataset) were given by experts and used for the single CNN approach in the study.

4.5.2 Data Preprocessing and Augmentation

To remove variations caused by camera and lighting conditions of different fundoscopes, we rescale and normalize all images to the same radius, subtract the local average color and preserve the central 90% of images to minimize boundary effects, and resize the images to 224×224 pixels.

There are 25,809 images in the largest class (normal) and only 708 images in the smallest class (most severe DR) (Figure 4-3). The label distribution shows the severe class imbalance in our dataset.

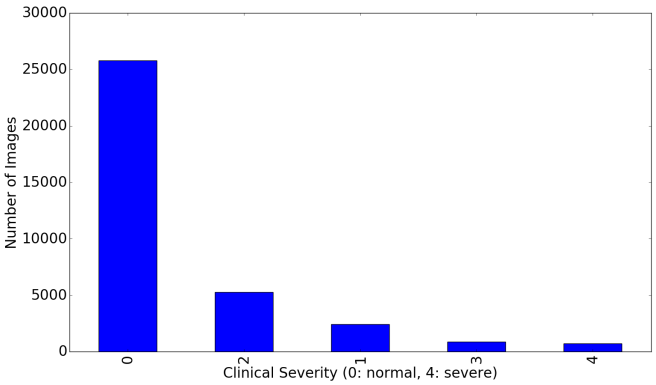


Figure 4-3: Expert-annotated label distribution.

To handle class imbalance, we augment the numbers of images of all classes to the same as the largest class by randomly selecting images from the minor classes and performing Krizhevsky style random offset cropping [Krizhevsky et al., 2012], random horizontal and vertical flipping, Gaussian blurring, and rotation between 0° and 360°. The original and augmented images are pooled together and split into 70% train and 30% test data based on

*https://www.kaggle.com/c/diabetic-retinopathy-detection

the stratification of class labels.

4.5.3 Learning Latent Representations

For both the single supervised CNN and deep SCNN architectures, we extract the last bottleneck layer as our latent image representation. Principal component analysis is first adopted to reduce the feature dimension to 50, t-Distributed Stochastic Neighbor Embedding (t-SNE) is then applied to further reduce the dimension to two [Maaten and Hinton, 2008] for visualizing the latent feature embeddings.

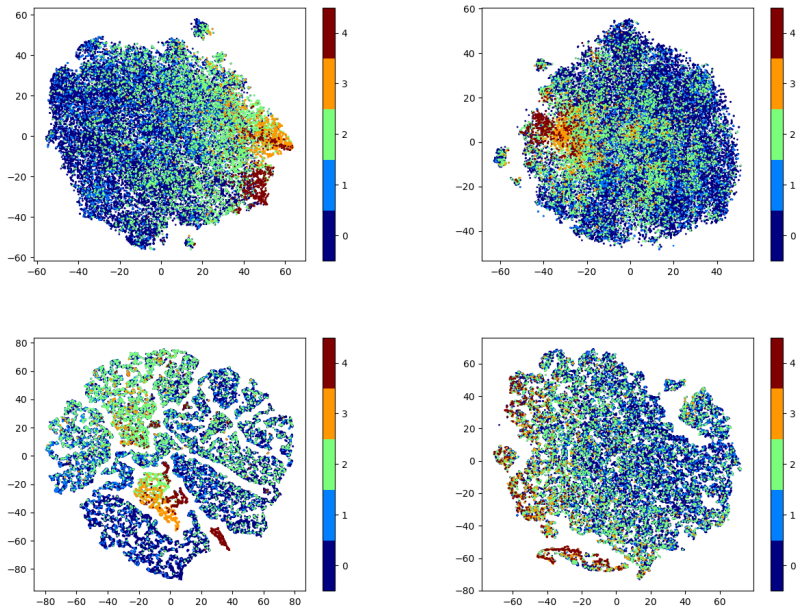


Figure 4-4: t-Distributed Stochastic Neighbor Embedding (t-SNE) visualizations for the distribution of learned retinal fundoscopic image representation embedding in the two-dimension vector space. (upper left) The embedding from the third-to-last layer of single CNN. (upper right) The embedding from the second-to-last layer of a single CNN. (lower left) The embedding from the last softmax layer of single CNN. (lower right) The embedding from the last layer of deep SCNN. Colors represent the actual expert-labeled severity of diabetic retinopathy (DR), where blue indicates normal/healthy cases and dark red represents severe DR cases. SCNN embedding shows us a color gradient change from low-grade DR to high-grade DR in the representation. We suppose that SCNN may learn the progressive changes of DR instead of just providing arbitrary borders between grades.

In Figure 4-4, we demonstrate the data distribution of image representations extracted from different layers of CNN and SCNN. A clear clinically interpretable severity transi-

tion from healthy cases (class 0) to severe disease (classes 3 and 4) is shown in the t-SNE visualization of baseline and proposed representations, which indicates that the learned representations are reliable.

The softmax (last) layer of CNN learns the tighter representation using multiclass information. However, the borders between different ground truth DR grading categories are arbitrary. The actual DR condition is progressive gradually instead of having a strict cutting-off boundary between each stage of severity. Compared to the softmax layer of CNN, the last layer of deep SCNN and the third-to-last and second-to-last layers of CNN learn the representations that capture the progressive changes in DR. Such representations are therefore more desirable to express the actual DR pathology.

4.5.4 Content-Based Medical Image Retrieval

In the experiment of CBMIR in DR, we compare the performance of our deep SCNN model with the corresponding single supervised pre-trained ResNet-50 architecture, and perform image retrieval on a few sample queries of DR fundoscopic images (Figure 4-5).

Table 4.1 shows that the deep SCNN architecture yields comparable performance in image retrieval using minimal expert labeling with only binary image pair information, compared to the single CNN model, which requires exact expert labeling of multiclass information. Considering the preferred representations that capture the progressive changes in DR, the representation learned by deep SCNN with binary labeling outperforms those learned by either third-to-last or second-to-last layers of single CNN.

Layer	MAP	MRR
CNN (third-last)	0.6209	0.7608
CNN (second-last)	0.6369	0.7691
CNN (softmax)	0.6673	0.7745
SCNN (last layer)	0.6492	0.7737

Table 4.1: Performance measurement of CBMIR using latent representations from single pre-trained CNN or deep Siamese convolutional neural networks (SCNN).

To evaluate the quality of CBMIR using deep SCNN, we make real queries and extract similar images for clinical qualitative evaluation. In Figure 4-5, we show three sample queries

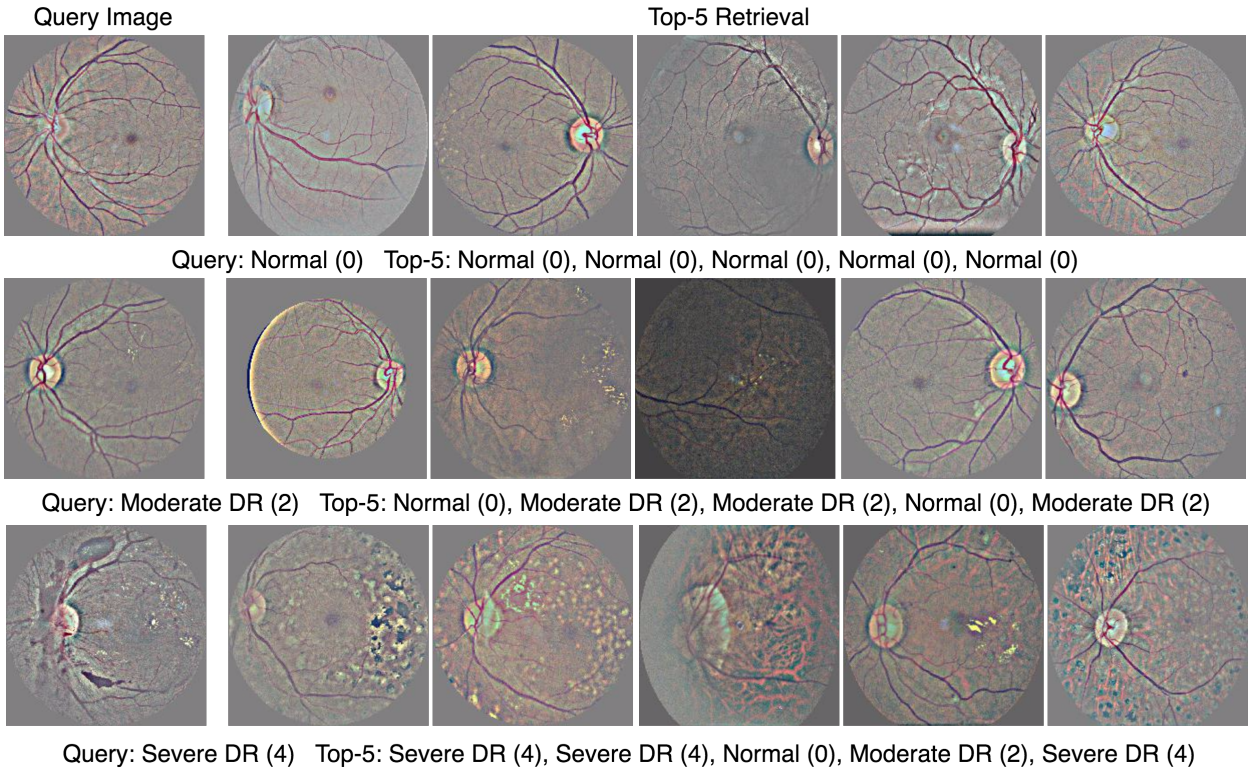


Figure 4-5: Three samples of the top-5 content-based medical image retrieval (CBMIR).

with different DR severity and the top five corresponding retrieved examples using deep SCNN.

The first query using a normal/healthy fundoscopic image yields fundoscopic images with exactly the same expert label. There are a few inconsistencies while using the image of moderate or severe DR as query inputs. However, in Figure 4-5 we are able to see that the inconsistencies result from either artifact of fundoscopic images or ambiguous diagnosis of DR severity. For example, the third retrieval of the query (third row in Figure 4-5) using severe DR images shows that the artifact in the original image leads to incorrect classification. The fourth retrieval of severe DR query is an image on the borderline between moderate and severe DR, which may also be a challenging case for some clinicians.

In general, the deep SCNN can identify and extract fundoscopic images that have the same expert-annotated label as the input query, or images with very similar patterns but having different severity labels.

4.6 Summary

In this chapter, we present a new strategy to learn latent representations of medical images by learning an end-to-end deep SCNN, which only requires binary image pair information. We experiment on the CBMIR task using a publicly available DR image dataset and demonstrated that the performance of deep SCNN is comparable to the commonly used single CNN architecture, which requires actual multiclass expert labeling that is expensive in the medical machine learning tasks. Future investigation may focus on performing experiments on different network architectures, other ranking metrics for evaluation such as recall on top-N, and applying the proposed method to different medical image datasets, such as chest X-ray imaging.

Chapter 5

Meta-learning for Class Imbalance

5.1 Overview

Class imbalance is a common problem in medical diagnosis, causing a standard classifier to be biased towards the common classes and to perform poorly on the rare classes. This is especially true for dermatology, a specialty with thousands of skin conditions but many of which have a low prevalence in the real world. Motivated by recent advances, we explore few-shot learning (FSL) methods and conventional class imbalance techniques for the skin condition recognition problem and propose an evaluation setup to fairly assess the real-world utility of such approaches. We find the performance of FSL methods does not reach that of conventional class imbalance techniques, but combining the two approaches using a novel ensemble improves model performance, especially for the rare skin condition classification. We conclude that ensembling can help address the class imbalance problem, yet real-world evaluation setups can further accelerate progress for benchmarking new methods.

5.2 Background

Skin disease is the fourth leading cause of non-fatal medical conditions burden worldwide [Seth et al., 2017]. Due to the global shortage of dermatologists, access to dermatology care is limited, leading to rising costs, poor patient outcomes, and health inequalities. Recent research endeavors have demonstrated that deep learning systems built to detect skin conditions from

either dermatoscopic or digital camera images can achieve expert-level performance in diagnosing certain skin conditions [Esteva et al., 2017, Liu et al., 2020b]. Despite the encouraging progress, such systems can only identify a few common skin conditions well, leaving a vast number of skin conditions still unaddressed in the real world (Figure 5-1 (a)).

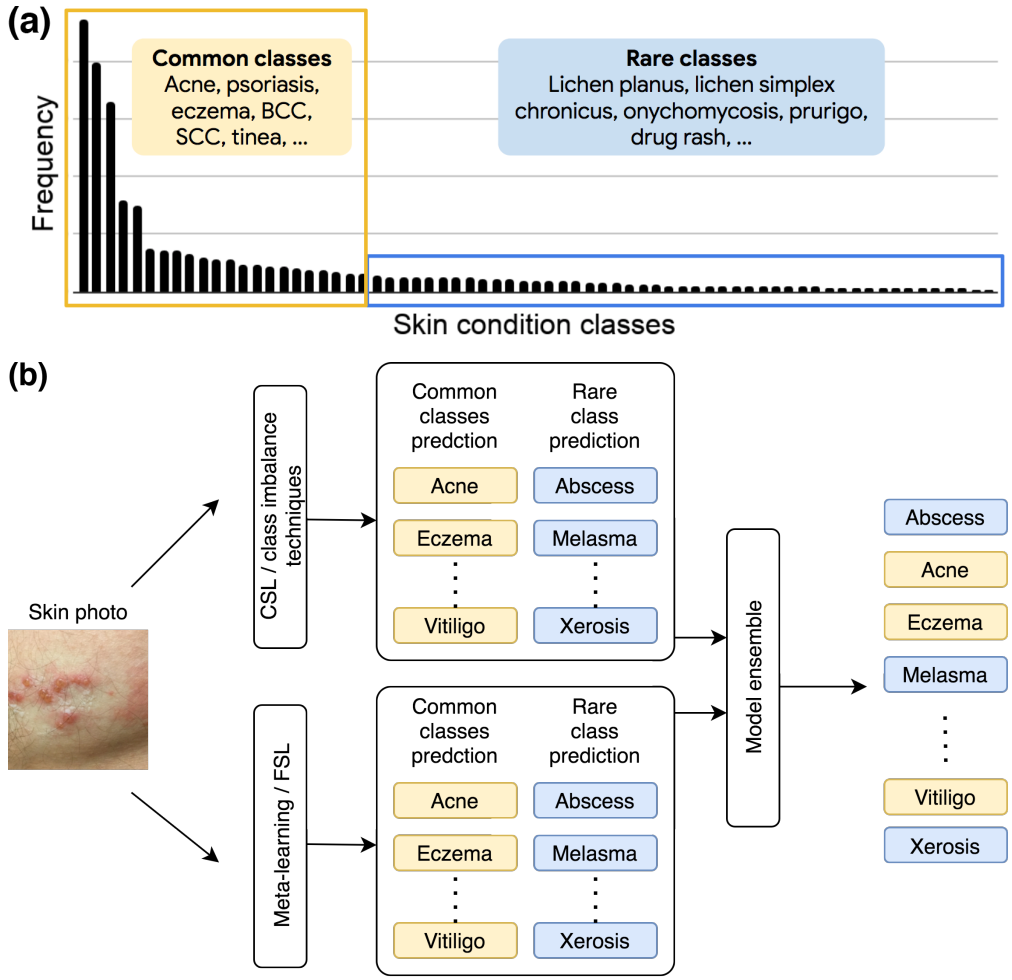


Figure 5-1: Class imbalance problem in dermatology at a glance (a) and proposed modeling framework (b). CSL: conventional supervised learning. FSL: few-shot learning.

As many skin conditions infrequently occur in the real world, datasets collected from a natural patient population are highly unbalanced, with few examples for many skin conditions. This makes it challenging to build models that can reliably detect rare skin conditions. Such limitations not only may diminish the clinical utility of these systems due to the low skin condition coverage, but also may cause harm to patients as they are often biased towards common diseases.

A wide variety of techniques have been proposed over the years to address the class imbalance problem in machine learning. These include relatively classic approaches like modified resampling strategies [Chawla et al., 2002], loss reweighting [Eban et al., 2017], focal loss (FL), and bias initialization (BI) [Lin et al., 2017], to more complex techniques such as using generative models to augment the rare classes with synthetic images [Ghorbani et al., 2020].

A related counterpart to the class imbalance problem is *FSL*, a learning scenario inspired by the human ability to learn from very few examples [Lake et al., 2011]. It can further be generalized to the meta-learning framework that aims to gain learning experiences from solving many meta tasks in order to achieve better performance for the tasks (in this case, the rare skin condition classification) with limited training examples [Vinyals et al., 2016].

The classic evaluation framework for FSL follows an "N-way-k-shot" setting: given k training examples for N classes as a random subset of all classes, test if the FSL techniques can correctly distinguish among the N classes. It is often not understood whether FSL will translate well beyond this contrived setting to the *real-world* problem, when the task requires recognizing the correct class among *all* possible classes (all-way classification) in the wild [Chen et al., 2019e, Triantafillou et al., 2020]. Furthermore, it is challenging to compare the classification performance between FSL and the conventional supervised training without a unified evaluation framework.

In this study, we investigate various FSL methods to tackle the class imbalance problem in skin condition classification (Figure 5-1 (b)). We modify the typical FSL evaluation setup to adapt to the real-world setting, and design studies to compare FSL methods to *conventional supervised learning* (CSL) with various classic techniques to address the class imbalance problem. We further explore the potential use of the combination of FSL and CSL with various class imbalance techniques, motivated by the hypothesis that a combination may provide independent predictions, capturing different aspects of the data. Specifically, our contributions are:

- We propose a real-world evaluation framework to compare the FSL methods against conventional methods (i.e., CSL with class imbalance techniques) for the class imbalance problem.

- We find that FSL methods don't outperform CSL with class imbalance techniques, yet an ensemble of these two types can outperform either alone, especially for the rare classes.

5.3 Related Works

5.3.1 Class Imbalance

Unbalanced class distribution is common in real-world datasets. For mild class imbalance, machine learning algorithms, such as support vector machines [Cortes and Vapnik, 1995], random forests [Breiman, 2001], and modern deep learning techniques can usually handle such cases well. However, special care is required to manage moderate or extreme imbalance situations (Figure 5-2), when minority classes typically constitute less than 20% or 1% of the training data, respectively.

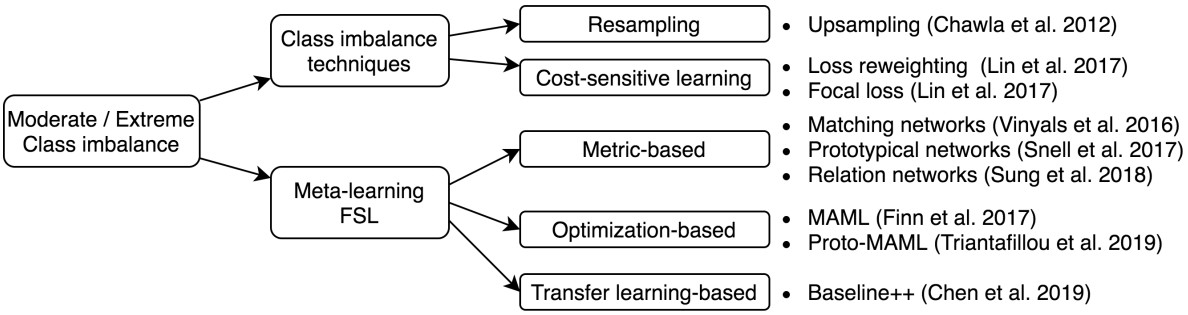


Figure 5-2: Categories of methods for tackling the class imbalance problem.

Conventional Techniques

Researchers have developed various resampling and cost-sensitive learning strategies to improve models under such skewed class distribution settings. Resampling aims to balance the class distribution in the training data by either downsampling common classes or upsampling rare classes. Downsampling may increase variance; thus, upsampling is more commonly used [Chawla et al., 2002], yet upsampling procedures for extremely unbalanced data may be computationally expensive [Weng et al., 2019a].

In contrast, cost-sensitive learning penalizes algorithms by increasing the cost of classification mistakes in rare classes. It can be implemented in various ways, such as reweighting the losses of specific classes, introducing bias as a prior into the classification loss at each class [Lin et al., 2017], or less used approaches such as global objectives [Eban et al., 2017]. To prevent the easy examples in the common classes from dominating the gradients during classifier training, the FL function was developed [Lin et al., 2017], not only to over-weight rare classes but also to emphasize the hard examples during training.

Few-shot Learning

FSL can also be utilized to improve the classification performance for the classes with very few training examples. The FSL algorithms can be categorized into several types: metric-based, optimization-based, and transfer learning-based approaches.

Metric-based FSL learns a representation by learning to compare training examples. Koch et al. [2015] developed the Siamese neural network to compare two examples at the same time with two identical twin networks sharing the same weights. The matching network utilizes cosine similarity as the distance metric and adopts the long-short term memory (LSTM) network for generating embeddings [Vinyals et al., 2016]. Prototypical network, on the other hand, uses the Euclidean distance and convolutional networks to determine the embedding of the class prototypes [Snell et al., 2017]. Relation networks further concatenate the prototype of each class with query examples, and use the relation module, which is parameterized by additional layers, to predict the class probability given the query and score example similarity [Sung et al., 2018].

Optimization-based FSL aims to learn a set of parameters that allows a meta-learner to quickly adapt to new tasks [Finn et al., 2017]. Ravi and Larochelle [2017] designed an LSTM-based meta-learner to replace the stochastic gradient descent algorithm in order to learn a better optimizer. One can also learn a better model initialization for faster task adaptation with fewer gradient updates (Model-Agnostic Meta-Learning, MAML) [Finn et al., 2017]. Others integrated both metric and optimization-based FSL to further improve performance [Triantafillou et al., 2020].

Finally, transfer learning tackles the few-shot classification by fine-tuning a model pre-

trained on a much larger dataset [Chen et al., 2019e, Tian et al., 2020]. Chen et al. [2019e] adopted the training-then-fine-tuning process for few-shot classification without meta-learning, and named the method as "Baseline++". Recently, Tian et al. [2020] demonstrated that transfer learning under the meta-learning framework yielded strong performance for the few-shot classification problem.

Despite the progress in FSL, such techniques are typically evaluated in a contrived setting and little is known about how they work in real-world all-way classification problems. In this study, we therefore propose an evaluation method that allows us to use FSL algorithms for the all-way classification problem, and compare them with the CSL-based methods.

5.3.2 Machine Learning and FSL in Dermatology

Artificial intelligence in dermatology has been a rapidly growing topic of interest in recent years [Cruz-Roa et al., 2013, Esteva et al., 2017, Liu et al., 2020b, Yang et al., 2018, Prabhu et al., 2019, Li et al., 2020, Mahajan et al., 2020, Guo et al., 2020, Le et al., 2020]. For example, Esteva et al. [2017] applied deep learning to clinical skin photos for two binary skin cancer classification tasks. Liu et al. [2020b] developed a CSL-based system that identifies 26 common skin conditions with performance superior to general practitioners and on par with dermatologists.

However, it is challenging to extend such systems to support rare skin conditions due to the limited training examples. Previous efforts have explored various approaches, such as adopting domain knowledge to learn a better representation [Yang et al., 2018], developing or modifying the meta-learning-based methods [Prabhu et al., 2019, Li et al., 2020, Mahajan et al., 2020], and comparing different approaches to tackle the extreme class imbalance problem in dermatology [Guo et al., 2020, Le et al., 2020]. Yet there is no study investigating both the FSL and CSL-based class imbalance techniques and comparing them under the real-world all-way classification setting.

Our work focuses on the all-way classification under an extreme long-tailed data distribution that often occurs in real-world medical imaging tasks. Under a unified evaluation framework, we study comparisons between FSL and CSL-based class imbalance techniques. We further propose ensembling FSL with conventional methods and demonstrate gains from

combining both strategies.

5.4 Methods

In this work, we investigate the model performance on the skin condition classification problem across CSL baseline, CSL with class imbalance techniques, FSL techniques, and different ensembles between these approaches.

5.5 Class Imbalance Methods

5.5.1 CSL-based class imbalance techniques

The following list describes the details of class imbalance techniques adopted for the CSL.

- Upsampling: on top of the CSL baseline with the cross entropy (CE) loss, the uniform sampling is performed across all classes during training.

- Bias initialization (BI) [Lin et al., 2017]: the final layer of the network, i.e., classification heads, is initialized with a bias of the log values of the number of training examples in the class.

- Inverse frequency weighting (IFW): the CE loss is further weighted by the inverse frequency of the number of training examples in the class.

- Focal loss (FL) [Lin et al., 2017]: the CE loss is extended to the α-balanced CE loss that accounts for the class imbalance by multiplying by a weighting factor $\alpha \in [0, 1]$: $\mathcal{L}_{\alpha\text{CE}} = -\sum_i \alpha_i \log(p_i)$, where $\alpha_i = \alpha$ for correct prediction, and $\alpha_i = 1 - \alpha$ otherwise. It also attempts to down-weight easy samples and focus on hard samples by introducing a modulating factor $(1 - p_i)^\gamma$ with a focusing parameter $\gamma \geq 0$. The objective can be expressed as $\mathcal{L}_{\text{focal}} = -\sum_i \alpha_i (1 - p_i)^\gamma \log(p_i)$.

- Prior correction: predictions from a CSL baseline model are further divided by maximum-likelihood estimates for class priors based on the training set. This method can be seen

as a specific use case of the post-hoc correction method introduced by Latinne et al. [2001] to account for label shift and calibrate the model predictions to a domain with the uniform class distribution.

We further combine BI/IFW, as well as FL/IFW for the combined class imbalance techniques.

5.5.2 FSL algorithms

The methods for FSL can be categorized into batch training and episodic training methods. For batch learning, we explore the following:

- k-nearest neighbor (k-NN): the k-NN method classifies the given test example based on the cosine distance between the test example and the class centroids in the vector space.
- Baseline++ [Chen et al., 2019e]: Baseline++ is a fine-tune model that uses the support set in testing episodes to train the final layer on top of the embeddings.

For the episodic training, we select the following meta-learners from the metric-based and optimization-based algorithms:

- Metric-based algorithms
  - Matching networks [Vinyals et al., 2016]: matching networks use the average of class example embeddings, which is computed by bilateral LSTM, as class prototypes, and classify query examples based on the cosine distance-weighted linear combination of the support labels.
  - Prototypical networks [Snell et al., 2017]: prototypical networks use the average of embeddings learned from a convolutional neural network as class prototypes, and classify query examples by computing the Euclidean distance between the embeddings of prototypes and queries.
  - Relation networks [Sung et al., 2018]: relation networks also average the embeddings to create prototypes, yet concatenate the prototype of each class with query

examples, and use the relation module, which is parameterized by additional layers, to predict the class probability given the query.

- Optimization-based algorithms
  - Model Agnostic Meta-Learning (MAML) [Finn et al., 2017]: MAML is an optimization-based method that attempts to learn a set of optimal initialization parameters to fast learn on different downstream tasks with a small number of gradient steps. We use the first-order approximation MAML in our study.
  - Proto-MAML [Triantafillou et al., 2020]: Proto-MAML combines the ideas of prototypical networks and MAML by reinterpreting the prototypical network as a linear layer on top of the learned embedding [Snell et al., 2017]. Then the linear layer with prototypical network-equivalent weights and bias can be used to initialize the last layer of MAML during training.

5.5.3 FSL Task and Evaluation Setup

We use an evaluation setup that allows us to compare all methods for the real-world classification problem, which entails properly addressing all skin conditions in our datasets. In this section, we explain how we modify the evaluation setup to adapt FSL to this real-world scenario and introduce the learning algorithms and metrics used for evaluation.

Task Formulation and Standard Evaluation

FSL follows the meta-learning setting, consisting of training, validation, and testing phases: training for learning a classifier (meta-learner), validation for hyperparameter tuning, and testing for evaluating the learned classifier. Data used in each of the three phases must be split into support and query sets. For the standard FSL evaluation, the classes used in these three phases should be all disjoint (Figure 5-3 (a)).

Given a dataset with M classes, the training can be batch or episodic. For batch training, we train a model using all examples from a N random classes ($N < M$) in the train set; for episodic training, we train a classifier under an N-way-k-shot learning scenario: in each

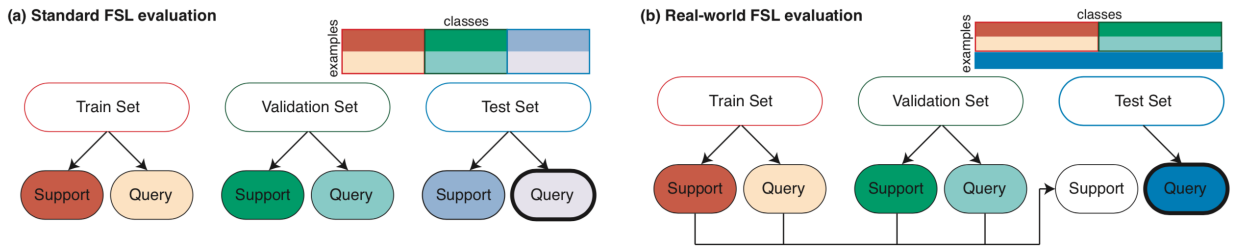


Figure 5-3: Differences between the standard (a) and the real-world (b) few-shot learning (FSL) evaluation settings. For the real-world classification problem, we need to modify the evaluation process of FSL methods in order to adapt them to evaluate the whole test set, and also for all classes. We report metrics over test query sets illustrated with bold box.

training episode, we choose k random samples from the N random classes ($N < M$) in the train set to form a support set to learn a model, and then evaluate on a query set that includes other samples in the same N classes in the training set. The validation and test sets are used to evaluate the model during and after training, respectively.

Real-world Evaluation for FSL

The standard N-way-k-shot framework doesn't fit the real-world classification problem, where we need to discriminate all classes simultaneously. Therefore, we introduce the following changes to the classic FSL evaluation setup (Figure 5-3 (b)):

(1) The dataset is split into development (train $\cup$ validation) and test sets in advance, where both sets may contain samples from all possible classes.

(2) Different from the classical setup where the classes in all three splits are disjoint, in the real-world setup, only training and validation have disjoint classes, and both the support set and the query set for those classes come from the development set. During testing, the support set that is used to train the final model comes from the development set, whereas the query set is exactly the same as the test set and includes disjoint samples from *all classes*.

5.5.4 Modeling

The following FSL algorithms are used in the study (Figure 5-2): k-NN and Baseline++ [Chen et al., 2019e], matching networks [Vinyals et al., 2016], prototypical networks [Snell et al., 2017], relation networks [Sung et al., 2018], MAML [Finn et al., 2017], and Proto-MAML [Tri-

antafillou et al., 2020]. We also implement the following CSL-based class imbalance techniques: upsampling with uniform sampling based on the ground truth class during training, BI with the log of frequency in the training set [Lin et al., 2017], IFW, FL [Lin et al., 2017], and the combination of BI/IFW and FL/IFW.

5.5.5 Metrics

We report the balanced accuracy (a.k.a. normalized sensitivity, or macro recall) separately for the common, rare, and all classes for the real-world evaluation. The balanced accuracy is used to account for the class imbalance issue to avoid overweighting the common classes. We also report top-1 all-way accuracy in the Appendix. We further report the 95% binomial confidence interval, derived from the mean μ and standard deviation σ of accuracies of E episodes of FSL or E runs of the CSL-based model, which can be expressed as $\mu \pm 1.96\frac{\sigma}{\sqrt{E}}$.

5.6 Experiments

5.6.1 Data

We use the clinical skin images dataset collected by a teledermatology service serving 17 different clinical sites [Liu et al., 2020b]. We divide the data into different subsets as illustrated in Figure 5-3 according to the temporal order (75% in the development set and 25% in the test set). Specifically, the test set is chosen for the more recent patient visits, to mimic the real-world setting as using earlier data for model training and deploying the model to serve future patients. The development set is further partitioned into train and validation sets based on per skin condition stratified sampling to ensure enough samples for training and validation. The statistics of the data splits are shown in Table 5.1.

Split	Train	Validation	Test
Patient number	9249	302	2755
Image number	11403	526	3136

Table 5.1: Statistics of the skin dataset in the study.

There are a total of 317 skin conditions in the dataset. We define the common classes based on the selection criteria used in [Liu et al., 2020b] (more than 100 cases in the development set), and the rare classes as those with (1) more than 20 cases in the development set, and (2) more than 5 cases in the test set. The criteria were established in order to ensure sufficient examples for training and evaluation. For other extremely rare classes (i.e., classes with <20 samples in the development set or <5 samples in the test set), we group them into a single aggregated category "Other" that belongs to a common class due to its sample size. In summary, we have 27 common classes and 42 rare classes (Figure 5-1 (a), Table 5.2).

5.6.2 Training Frameworks

The input for any model is an image of a size of 448×448, and the output is the corresponding skin condition prediction. The Inception-V4 backbone is used for the CSL baseline and CSL with class imbalance techniques. Inception-V4 is evaluated as the most performant architecture in an internal neural architecture search experiment.

For FSL, we adopt the implementation in the meta-dataset work with the ResNet-18 backbone, which is one of the best performant network architectures in various FSL studies [Triantafillou et al., 2020, Tian et al., 2020]. We train each model using the Adam optimizer for 75000 steps with exponential decay. We follow Triantafillou et al. [2020] for the learning rate and weight decay setups. We use a batch size of 64 and standard operations like random brightness, saturation, hue, contrast normalization, flip, rotation, and bounding box cropping for data augmentation.

5.7 Results

We perform experiments to understand (1) whether FSL can be viably applied to the skin condition classification task, (2) how FSL compares against current CSL-based class imbalance techniques, especially on the rare classes, (3) whether combining CSL-based techniques and FSL can further improve the performance across the different skin condition classes.

	#Class	Included classes
Common	27	Acne, Actinic Keratosis, Allergic Contact Dermatitis, Alopecia Areata, Androgenetic Alopecia, Basal Cell Carcinoma, Cyst, Eczema, Folliculitis, Hidradenitis, Lentigo, Melanocytic Nevus, Melanoma, Other, Post Inflammatory Hyperpigmentation, Psoriasis, Squamous cell carcinoma/squamous cell carcinoma in situ (SCC/SCCIS), seborrheic keratosis/irritated seborrheic keratosis (SK/ISK), Scar Condition, Seborrheic Dermatitis, Skin Tag, Stasis Dermatitis, Tinea, Tinea Versicolor, Urticaria, Verruca vulgaris, Vitiligo
Rare	42	Abscess, Acanthosis nigricans, Acne keloidalis, Amyloidosis of skin, Central centrifugal cicatricial alopecia, Condyloma acuminatum, Confluent and reticulate papillomatosis, Cutaneous lupus, Dermatofibroma, Dissecting cellulitis of scalp, Drug Rash, Erythema nodosum, Folliculitis decalvans, Granuloma annulare, Hemangioma, Herpes Simplex, Herpes Zoster, Idiopathic guttate hypomelanosis, Inflicted skin lesions, Insect Bite, Intertrigo, Irritant Contact Dermatitis, Keratosis pilaris, Lichen Simplex Chronicus (LSC), Lichen planus/lichenoid eruption, Lichen sclerosus, Lipoma, Melasma, Milia, Molluscum Contagiosum, Onychomycosis, Paronychia, Perioral Dermatitis, Photodermatitis, Pigmented purpuric eruption, Pityriasis rosea, Prurigo nodularis, Pyogenic granuloma, Rosacea, Scabies, Telogen effluvium, Xerosis
Other (one category)	248	Accessory nipple, Acquired digital fibrokeratoma, Acute generalised exanthematous pustulosis, Adnexal neoplasm, Alopecia mucinosa, Alopecia neurotica, Angiofibroma, Angiokeratoma of skin, Angiosarcoma of skin, Apocrine cystadenoma, Arterial ulcer, Beau's lines, Becker's nevus, Benign neoplasm of nail apparatus, Brachioradial pruritus, Bullosis diabeticorum, Bullous Pemphigoid, Burn of skin, Cafe au lait macule, Calcinosis cutis, Candida, Canker sore, Cellulitis, Central centrifugal cicatricial alopecia, Chicken pox exanthem, Chilblain, Chondrodermatitis nodularis, Clavus, Comedone, Cutaneous capillary malformation, Cutaneous metastasis, Cutaneous neurofibroma, Cutaneous sarcoidosis, Cutaneous T Cell Lymphoma, Cutis verticis gyrata, Deep fungal infection, Dental fistula, Dermatitis herpetiformis, Dermatofibrosarcoma protuberans, Dermatomyositis, Diabetic dermopathy, Diabetic ulcer, Ecthyma, Epidermal nevus, Erosive pustular dermatosis, Erythema ab igne, Erythema annulare centrifugum, Erythema dyschromicum perstans, Erythema migrans, Erythema multiforme, Erythrasma, Erythromelalgia, Fordyce spots, Foreign body reaction of the skin, Fox-Fordyce disease, Frontal fibrosing alopecia, Ganglion cyst, Granulomatous cheilitis, Grover's disease, Hailey Hailey disease, Hairy tongue, Hand foot and mouth disease, Hematoma of skin, Hemosiderin pigmentation of skin, Hirsutism, Hyperhidrosis, Hypersensitivity, Ichthyosis, Impetigo, Infected eczema, Inflammatory linear verrucous epidermal nevus, Ingrown hair, Kaposi's sarcoma of skin, Keratoderma, Keratolysis exfoliativa, Knuckle pads, Leukocytoclastic Vasculitis, Leukonychia, Leukoplakia of skin, Lichen nitidus, Lichen planopilaris, Lichen striatus, Lichenoid myxedema, Lipodermatosclerosis, Livedo reticularis, Livedoid vasculopathy, Longitudinal melanonychia, Mucocele, Mucocutaneous venous malformation, Nail dystrophy due to trauma, Necrobiosis lipoidica, Nevus comedonicus, Nevus lipomatosus cutaneous superficialis, Nevus sebaceous, Nevus spilus, Nodular vasculitis, Notalgia paresthetica, O/E - ecchymoses present, Onychocryptosis, Onycholysis, Onychomadesis, Onychoschizia, Oral fibroma, Paronychia, Pearly penile papules, Pemphigoid gestationis, Pemphigus vulgaris, Perforating dermatosis, Perleche, Pilonidal cyst, Pincer nail deformity, Pitted keratolysis, Pityriasis alba, Pityriasis amiantacea, Pityriasis lichenoides, Pityriasis rubra pilaris, Poikiloderma, Porokeratosis, Porphyria cutanea tarda, Post-Inflammatory hypopigmentation, Pressure ulcer, Pseudopelade, Pterygium of nail, Pruritic urticarial papules and plaques of pregnancy, Purpura, Pyoderma Gangrenosum, Remove from labeling tool, Retention hyperkeratosis, Rheumatoid nodule, Sebaceous hyperplasia, SJS/TEN, Skin and soft tissue atypical mycobacterial infection, Skin striae, Sweet syndrome, Syphilis, Tattoo, Telangiectasia disorder, Trachyonychia, Traumatic bulla, Traumatic ulcer, Trichotillomania, Varicose veins of lower extremity, Venous Stasis Ulcer, Viral Exanthem, Xanthoma, Yellow nail syndrome, Zoon's balanitis

Table 5.2: Detailed categorization for common and rare skin conditions in the study.

5.7.1 Standard FSL Evaluation

In Figure 5-4 and Table 5.3, we first benchmark and examine the feasibility of adopting the standard FSL evaluation (5-way-5-shot classification) to the teledermatology dataset.

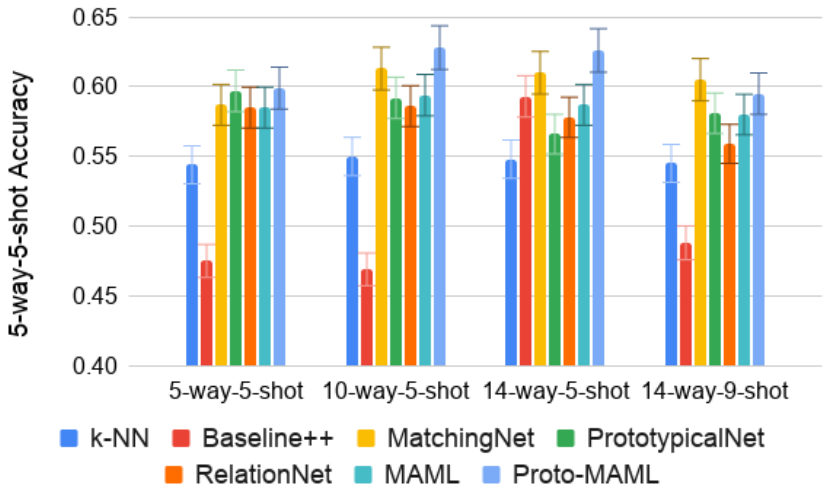


Figure 5-4: Standard FSL evaluation (5-way-5-shot accuracy with 95% confidence interval) on the teledermatology dataset. We investigate the change of N and k values during training. The x-axis is the FSL setting during training.

We find that increasing N yields a trend of better performance. For this particular dermatology use case, a higher N is a better option to achieve the best performance under the standard FSL evaluation (Figure 5-4; N-way-5-shot bars where $N = 5, 10, 14$). However, increasing k is not helpful, possibly because of overfitting (Figure 5-4; 14-way-k-shot bars where $k = 5, 9$). This finding is consistent with the previous literature [Snell et al., 2017]. Among all methods, Proto-MAML and prototypical networks are comparable and consistently outperform the others for the teledermatology dataset. This dataset has shown similar properties as other FSL datasets under the standard FSL evaluation.

5.7.2 Real-world Evaluation

FSL-based Approaches

In the real-world setting, we report all-way (69-way) performance (Figure 5-5, Table 5.4) on the test set (i.e., query set of the testing phase; see Figure 5-3 (b)).

Method	Train Eval	5-way-5-shot 5-way-5-shot	10-way-5-shot 5-way-5-shot	14-way-5-shot 5-way-5-shot	14-way-9-shot 5-way-5-shot
k-NN		0.544±0.010	**0.55±0.010**	0.548±0.010	0.545±0.010
Baseline++		0.475±0.010	0.469±0.010	**0.593±0.010**	0.488±0.010
Matching networks		**0.597±0.010**	0.592±0.010	0.566±0.010	0.581±0.010
Prototypical networks		0.587±0.010	**0.613±0.010**	0.610±0.010	0.605±0.010
Relation networks		0.585±0.010	**0.586±0.010**	0.578±0.010	0.559±0.010
MAML		0.585±0.010	**0.594±0.010**	0.587±0.010	0.580±0.010
Proto-MAML		0.599±0.010	**0.628±0.010**	0.626±0.010	0.595±0.010

Table 5.3: Standard few-shot learning (FSL) evaluation for the skin dataset with the change of N and k value (5-way-5-shot accuracy, with 95% confidence interval). "14-way-9-shot" indicates that we use 14-way-9-shot for the training in meta-learning, but use 5-way-5-shot for all evaluations. The boldface values indicate the best setting for each FSL algorithm. We find that higher N in general help improve the performance, yet the 10-way-5-shot training is best for standard FSL using the teledermatology dataset to prevent overfitting/underfitting while performing meta-learning. Proto-MAML and prototypical networks outperform other methods.

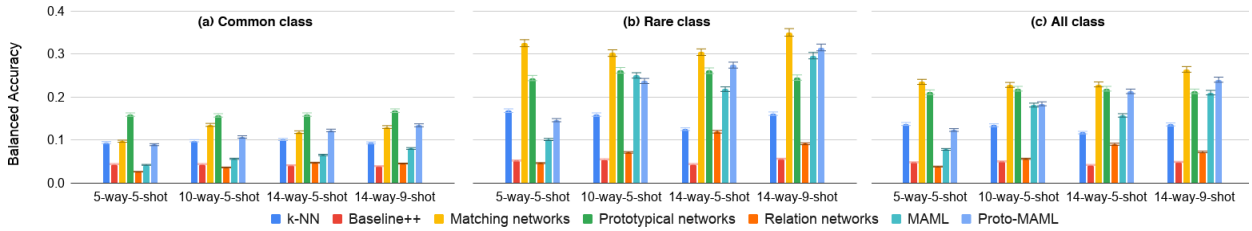


Figure 5-5: Real-world few-shot learning (FSL) evaluation. We report the balanced accuracy for (a) common class (b) rare class (c) all class prediction using different N-way-k-shot settings.

Method	5-way-5-shot	10-way-5-shot	14-way-5-shot	14-way-9-shot
k-NN	0.137 (0.094/0.168)	0.134 (0.098/0.159)	0.117 (0.101/0.125)	0.136 (0.093/0.161)
Baseline++	0.048 (0.044/0.052)	0.050 (0.044/0.055)	0.042 (0.041/0.044)	0.049 (0.039/0.056)
Matching networks	**0.235 (0.097/0.325)**	**0.228 (0.135/0.302)**	**0.229 (0.118/0.304)**	**0.264 (0.130/0.350)**
Prototypical networks	0.211 (0.159/0.244)	0.219 (0.157/0.262)	0.219 (0.159/0.261)	0.213 (0.168/0.245)
Relation networks	0.038 (0.026/0.046)	0.056 (0.036/0.071)	0.090 (0.047/0.119)	0.072 (0.045/0.091)
MAML	0.078 (0.042/0.101)	0.181 (0.056/0.250)	0.157 (0.065/0.218)	0.210 (0.080/0.296)
Proto-MAML	0.123 (0.089/0.146)	0.184 (0.107/0.237)	0.213 (0.122/0.274)	0.240 (0.134/0.308)

Table 5.4: Real-world few-shot learning (FSL) evaluation. We report the performance of models trained by different N-way-k-shot settings. The value outside the parentheses is the balanced accuracy of *all* classes, and the values inside the parentheses are the balanced accuracy of common/rare classes, respectively. The boldface values indicate the best algorithm for each training setup. Unlike the standard FSL evaluation, higher N and k yield a better meta-learner in the real-world FSL evaluation. The performance of classifying rare classes is consistently better than classifying the common classes. Matching networks and Proto-MAML using 14-way-9-shot training outperforms the other methods under the real-world evaluation. We use them for the method comparison and model ensemble.

The balanced accuracy is chosen as the metric to avoid overweighting the common classes. We conduct the N-way-k-shot experiments again to find the optimal training setting under the real-world setup. Unlike the standard FSL evaluation findings, we find that the optimal N and k values are not entirely consistent across methods. The best performant method in the real-world FSL evaluation is also matching networks. In Figure 5-5 (a, b) and Table 5.4, we also find that the FSL methods consistently perform better on the rare classes than the common classes.

In brief, we find that the conclusion from the standard FSL evaluation that higher N-way training yields a better meta-learner is inconsistent with the real-world evaluation results. Thus, more in-depth exploration is required in the real world to identify the optimal N-way setting. We also confirm our hypothesis that the FSL models can be beneficial for rare class prediction but may not be for common classes. For the common class prediction, relying on the CSL methods may be the better option in this case. We later use the top performant models based on validation, which is the matching network model trained in a 14-way-9-shot setting, for the model ensemble experiments below.

CSL-based Class Imbalance Techniques

Next, we compare the real-world all-way classification performance between FSL methods and CSL-based class imbalance techniques under the single model (non-ensemble) setting. In Figure 5-6 and Table 5.5, we find that the best FSL model (matching network) is only slightly better than the CSL baseline for the rare class prediction, possibly because of less training data. Yet the CSL-based class imbalance techniques yield better balance between common and rare classes, as demonstrated by all-class balanced accuracy. Moreover, the CSL-based model integrating FL and IFW (FL/IFW) yields even better performance on the rare class classification. Such results suggest that FSL is not beneficial when used independently in this real-world setting. One possible explanation could be that the given FSL models are not using all the data from the common classes during training, hindering the learning of good low-level visual features, whereas CSL-based models can take advantage of all the data.

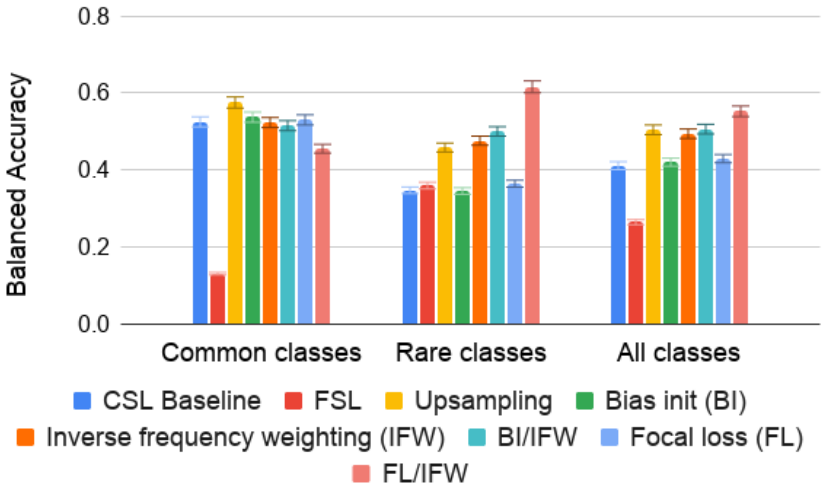


Figure 5-6: Comparison between FSL and CSL-based class imbalance techniques on the all-way skin condition classification.

	All-way accuracy	Balanced accuracy	Balanced acc. (common class)	Balanced. acc. (rare class)
FSL	0.147	0.264	0.130	0.359
CSL baseline	0.648	0.411	0.525	0.347
Upsampling	0.610	0.504	**0.575**	0.458
BI	**0.661**	0.420	0.537	0.345
IFW	0.621	0.494	0.523	0.476
BI/IFW	0.620	0.506	0.515	0.500
FL	0.657	0.429	0.530	0.364
FL/IFW	0.395	0.552	0.455	0.616
Prior Correction	0.450	**0.580**	0.518	**0.620**

Table 5.5: Comparison between few-shot learning (FSL), conventional supervised learning (CSL) baseline and other class imbalance techniques in the all-way classification problem under a single model (non-ensemble) setting. The boldface values indicate the best method for each evaluation metric. Note that all-way accuracy is a biased metric towards common classes, as the test set is dominated by common classes. Also note that we include here an additional class-imbalance method "Prior Correction". Prior correction achieves the greatest balanced accuracy, an observation theoretically supported by its interpretation as a label-shift correction and the fact that balanced accuracy is equivalent to overall accuracy in a domain with the uniform class distribution. The single FSL model (matching networks) is not beneficial while using it independently for all-way classification. Instead, the single CSL model with upsampling strategy outperforms others in the common class classification, and the single CSL model with focal loss (FL) / inverse frequency weighting (IFW) yields the best performance for the rare class classification.

Model Ensemble

To better utilize the FSL and CSL-based class imbalance techniques, we experiment with ensembling the trained models. We approximate the joint ensemble output probability by computing the geometric mean of the probabilities across M selected models:

$$\mathrm{P}_c = \left(\prod_{m=1}^{M} \mathrm{P}_c^{(m)} \right)^{\frac{1}{m}}$$

where $\mathrm{P}_c^{(m)}$ is the probability the m^{th} model in the ensemble gives to class c. We compute the corresponding ensemble logits by first normalizing each model's logits ($f_c^{(m)}$) using LogSumExp($\cdot$) (note this normalization does not alter the probabilities) before aggregating the logits from different models:

$$\bar{f}_c = \frac{1}{M} \sum_{m=1}^{M} \left(f_c^{(m)} - \log \left(\sum_{i=1}^{C} \exp \left(f_i^{(m)} \right) \right) \right)$$

where $\bar{f}_c$ is the normalized logit for the ensemble. We apply Softmax($\cdot$) to the normalized logits to compute the final ensemble probabilities.

To utilize the CSL-based methods in common class prediction, and FSL in rare class prediction, we further use the prediction from CSL-based methods if the ensemble prediction falls into the group of common classes, and use the prediction from FSL if it is from the group of rare classes while ensembling FSL with CSL-based models. Based on our hypothesis and the results of the validation set, we use the best performant FSL model for rare classes, which is a matching network, and the best CSL-based class imbalance model for common classes, which is the upsampling method, as basic components for the ensemble. In Figure 5-7 (a) and Table 5.6, we find that FSL-only ensembles do not perform well, which is consistent with the observations from the previous single model setting.

In contrast, ensembling FSL with CSL-based methods leads to a slight decrease in the common classes yet some improvement over rare classes and all classes in terms of balanced accuracy (Figure 5-7); ensembling the matching network model with the CSL model with upsampling technique (Figure 5-7 (a) orange) tends to strike a better balance between common

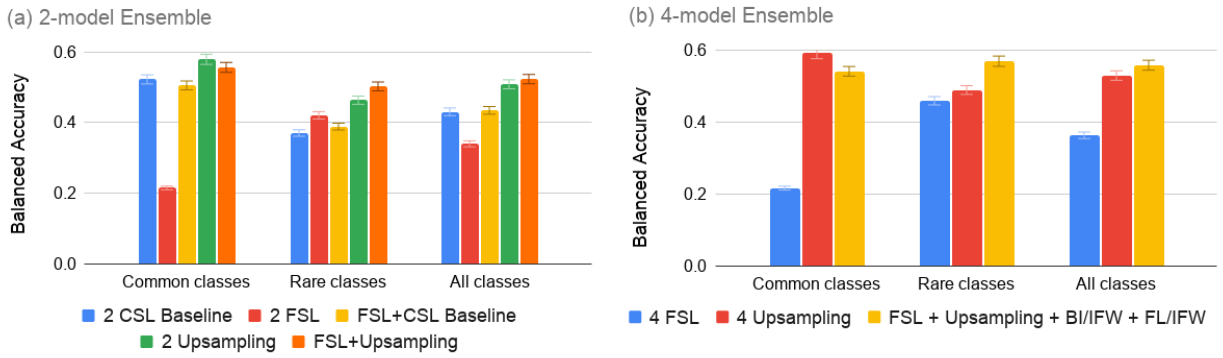


Figure 5-7: Model ensemble. (a) 2-model ensemble to show the improvement after using FSL. (b) 4-model ensemble to demonstrate the improvement after ensembling models with different mechanisms.

#Model	Combination	All-way accuracy	Balanced accuracy	Balanced acc. (common class)	Balanced acc. (rare class)
2	2 FSL	0.233	0.340	0.215	0.421
	2 CSL Baseline	0.663	0.431	0.523	0.371
	FSL+CSL Baseline	0.633	0.435	0.506	0.389
	2 Upsampling	0.617	0.509	0.58	0.464
	FSL+Upsampling	0.582	0.524	0.557	0.503
4	4 FSL	0.24	0.363	0.216	0.459
	4 Upsampling	0.623	0.529	0.591	0.489
	FSL+Upsampling+BI/IFW+FL/IFW	0.582	0.558	0.541	0.569

Table 5.6: Comparison between ensemble few-shot learning (FSL), conventional supervised learning (CSL) baseline, and CSL-based class imbalance techniques in the all-way classification problem under ensemble setting. Note that all-way accuracy is a biased metric towards common classes, as the test set is dominated by common classes. In all-way classification, the balanced accuracy on all classes improves after ensembling with the FSL model (2 CSL baseline versus FSL+CSL baseline, and 2 upsampling versus FSL+upsampling). Similarly, combining models with different learning mechanisms helps improve the balanced accuracy across all classes (4 FSL versus 4 upsampling versus 4 different models).

and rare classes. In Figure 5-7 (b), we find that making the ensemble more heterogeneous by using models with different mechanisms yields an even better trade-off between common and rare classes along with a more notable performance increase, especially for rare classes. These findings confirm our hypothesis that the ensemble of FSL and CSL can benefit such a real-world unbalanced skin condition classification problem.

5.7.3 Qualitative Analysis

In Figure 5-8, we present eight examples from rare classes alongside predictions of different models.

We show that the FSL and CSL-based class imbalance techniques predict the rare classes more accurately while the CSL baseline tends to be biased toward the common classes such as acne, eczema, and psoriasis. For example, eczema is a very common skin condition with various presentations, but it can be visually similar to drug rash or other skin lesions. While the baseline tends to over-predict common classes, the FSL can correctly distinguish the drug rash from eczema or inflicted skin lesions.

5.8 Summary

In this work, we develop a real-world evaluation framework to fairly assess and compare the performance between FSL and CSL-based methods on the all-way skin condition classification problem, where extreme class imbalance exists. We find that FSL methods do not outperform class imbalance techniques, yet when ensemble with CSL-based class imbalance techniques, they lead to improved model performance, especially for the rare classes. We also observe that ensembling the models with different strengths and more heterogeneity, such as upsampling and FSL, yields promising results.

To further improve the FSL methods, researchers have, in recent times, proposed more robust representation learning strategies via feature reuse [Raghu et al., 2020], or self-supervised learning [Tian et al., 2020]. We believe that evaluating the FSL methods on real-world benchmarks and use cases like the one outlined in this work can significantly accelerate progress in the development of FSL methods with real-world utility.

Figure 5-8: Case study of rare classes. FSL and class imbalance techniques are better at identifying rare classes than the baseline model.

We also need to account for potential model failures for future deployment and real-world impact. For example, false-positive cases may lead to wastage of medical resources, while false negatives may lead to delayed treatment, especially serious for malignant cases. The ensemble of both CSL and FSL approaches may mitigate these issues, yet to assess the generalizability of the results one needs to evaluate our proposed methods on other datasets and domains, which is an interesting future direction.

Chapter 6

Multimodal Multitask Learning for Heterogeneous Data

6.1 Overview

Instead of using metadata as input signals for learning representations of medical imaging, in this chapter, metadata become the information we want to predict using machine learning models. Metadata are general characteristics of the data in a well-curated and condensed format, and have been proven to be useful for decision making, knowledge discovery, and also heterogeneous data organization of a biobank. Among all data types in the biobank, pathology is the key component and also serves as the gold standard of diagnosis. To maximize the utility of a biobank and allow the rapid progress of biomedical science, it is essential to organize the data with well-populated pathology metadata. However, manual annotation of such information is tedious and time-consuming. In this study, we develop a multimodal multitask learning framework to predict four major slide-level metadata of pathology images. The framework learns generalizable representations across tissue slides, pathology reports, and case-level structured data. We demonstrate improved performance across all four tasks with the proposed method compared to a single modal single task baseline on two test sets, one external test set from a distinct data source, The Cancer Genome Atlas (TCGA), and one internal held-out test set (TTH). In the test sets, the performance improvements on the averaged area under the receiver operating characteristic curve across the four tasks are

16.48% and 9.05% on TCGA and TTH, respectively. Such a pathology metadata prediction system may be adopted to mitigate the effort of expert annotation and ultimately accelerate data-driven research by better utilization of the pathology biobank.

6.2 Background

Metadata have been proven to be useful for decision making and knowledge discovery [Yee et al., 2003, Stephens, 2004]. They provide compact and domain-specific data representations for machine learning in different knowledge domains [Linkert et al., 2010, Malet et al., 1999, Johnson et al., 2015]. In the previous chapter, we also learned that it is possible to utilize metadata and learn better representations during model pre-training (Chapter 3). In the biomedical domain, metadata are critical for both data archiving and other downstream data-driven applications [Posch et al., 2016, Weng et al., 2017b]. One of the prominent applications is the development of biobanks [Coppola et al., 2019].

A biobank is a repository that stores biological tissue samples for research usage, and the essence of developing a biobank is to curate an organized dataset with well-populated metadata. Biobanks have been playing a critical role in numerous scientific and clinical breakthroughs. For example, the development of one of the effective breast cancer immunotherapies, Herceptin, has greatly relied on metadata of a well-organized biobank to curate a cohort for the initial validation [Coppola et al., 2019]. The Cancer Genome Atlas (TCGA) and UK Biobank are also two large-scale repositories that support enormous studies and clinical trials to accelerate biomedical research [Sudlow et al., 2015]. Biobanks usually contain a large amount of data from different modalities in heterogeneous formats, such as genome sequences, pathology images, and clinical reports. To utilize these datasets, having well-populated metadata is necessary.

Pathology has been recognized as the gold standard for diagnosis across different medical specialties [Kumar et al., 2014, Nagpal et al., 2019]. In a biobank, pathology metadata along with the long-term follow-up survival data are the most valuable for modeling disease progression and patient outcomes. These data are also arguably of greater value due to the more extended follow-up periods. However, they are also often the least organized data.

This is primarily due to the variation in the data archiving process across institutes and the tedious manual processing of biological tissues in the modern pathology workflow. Although an ideal approach for collecting accurate metadata is to ask pathologists for annotation while signing out the cases, it is infeasible for the already developed biobanks that have abundant longitudinal survival data. Therefore, to maximize the utilization of the biobank, automated pathology metadata annotation is necessary [Guo et al., 2016].

Through advances in machine learning, modern computational techniques have shown promising results in automated prediction tasks. For pathology, researchers have shown the capability of pathologist-level performance on various tasks, such as tumor detection [Litjens et al., 2017, Coudray et al., 2018], survival outcome prediction [Nagpal et al., 2019], and even augmentation of the workflow through real-time feedback [Chen et al., 2019b]. However, most works are based on the imaging modality alone without considering data from other modalities, such as free text reports. Those works mainly focus on a single diagnosis task on a few tissue types, different from a problem like biobank data curation that requires a well-performing model across multiple tasks on a wide range of specimens. The heterogeneous, limited, and unbalanced data also makes the automated pathology metadata prediction even more challenging. It is also a common issue for machine learning and its downstream applications [Lin et al., 2017].

In this study, we investigate a multimodal multitask learning approach to jointly predict multiple slide-level metadata simultaneously from a shared representation across image, text, and structured categorical variables in a limited and unbalanced sample size regime. Multitask learning (MTL) leverages multiple prediction tasks to mitigate the issue of limited sample size since different tasks may share similar representations. Multimodal learning utilizes data of different modalities to learn a shared representation. Specifically, we incorporate case-level text from pathology reports with slide-level tissue images, because each of them holds different information that links to different metadata. Figure 6-1 and Table 6.2 illustrate the multi-level information in a single case. We further conduct ablation analysis to investigate the importance and utility of different modalities.

To this end, we make the following contributions in this study:

- We develop a multimodal multitask learning framework using images, free texts, and

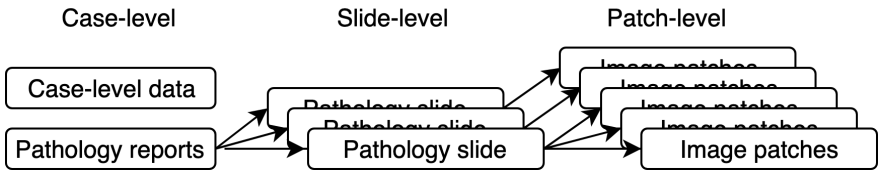


Figure 6-1: Illustration of the information across the case, slide, and patch levels from a single patient.

structured data for pathology metadata prediction with limited and unbalanced data.

- Our multimodal multitask framework outperforms the baseline single modal single task framework across all four pathology metadata prediction tasks.
- We observe the synergistic effect that adding multimodal information on top of a multitask framework outperforms either multimodal or multitask alone.

6.3 Related Works

In this work, we develop a single model that can tackle multiple problems simultaneously by learning a shared generalizable representation with multimodal limited, unbalanced pathology data. Based on the existing works, we introduce challenges in multimodal learning and MTL, and justify the selected strategies for overcoming these challenges in the proposed framework.

6.3.1 Multimodal Learning

The goal of multimodal learning is to learn a data representation capturing shared and independent information from data across different modalities. There have been successes of multimodal learning in many different fields [Ngiam et al., 2011, Baltrušaitis et al., 2018]. In the biomedical domain, there is a strong need for such an approach because patient data usually come with multiple modalities such as waveforms, claims data, free texts, images, and genome sequences. Researchers have utilized information from different modalities to approach various biomedical problems, such as tissue pattern recognition [Schlegl et al., 2015], report generation [Liu et al., 2019a], medical language translation [Weng et al., 2019b], and

clinical event prediction [Suresh et al., 2017]. However, the key challenge of multimodal learning is the heterogeneity across different modalities due to vastly different statistical properties and varying levels of noise. In this study, we explore different strategies to effectively merge the heterogeneous modalities.

There are two major approaches for multimodal learning: shared representation learning and cross-modal coordinated representation learning [Ngiam et al., 2011]. Shared representation learning, i.e., learning a common embedding space, enforces the model to have a single latent representation from multimodal data. On the other hand, cross-modal coordinated representation learning, or embedding alignment, uses an additional step to align representations from different modalities with the assumption that the geometric structure of the representations is similar [Chung et al., 2018].

For shared representation learning, early fusion and late fusion are the two most prominent approaches [Baltrušaitis et al., 2018]. Early fusion helps capture low-level interactions between modalities, which is ideal when there are dependencies between features in different modalities. Late fusion instead builds a meta-classifier to preserve more single-modal information since this approach doesn't model the interactions between modalities at low-level.

Since our goal is to learn a general representation from heterogeneous pathology images, texts, and structured data, we adopt the early fusion strategy to capture low-level interactions between modalities and keep the generalizability as much as possible. We also investigate the model performance of using different operations for merging modalities.

6.3.2 Multitask Learning

MTL, also known as joint learning or learning with auxiliary tasks, is a machine learning scenario that uses training signals from other related tasks to solve the more complex tasks simultaneously [Caruana, 1993, Ruder, 2017]. MTL has been widely used in different domains such as natural language processing, speech, and computer vision [Ruder, 2017]. It has also been applied to biomedical problems. One of the earliest applications is the pneumonia risk stratification task using the lab value prediction as an auxiliary MTL task [Caruana et al., 1996]. Despite the strong theoretical support for the utility of MTL, the performance gain of adopting MTL in the biomedical domain is not guaranteed [Caruana et al., 1996, Nori

et al., 2015].

We argue that MTL can help leverage supervision signals from other tasks by training a model to predict correlated pathology metadata simultaneously from a data representation. A natural question that arises from this design is the ways to do parameter sharing. There are two main strategies for MTL to propagate signals back to the data representation, via hard parameter sharing or soft parameter sharing. Hard sharing strategy has a single pathway from input to a data representation, and following the representation are task-specific heads with independent parameters. The approach enables lower layer parameters to be shared while parameters in each head are task-specific. This enables learning a generalizable shared representation while also optimizing for the downstream tasks [Caruana, 1993]. In the soft sharing strategy, each task has its pathway from input to output. The parameters for different pathways are soft-shared by imposing a joint regularization. This allows a certain degree of similarity across representations for different task pathways without forcing them to be identical [Duong et al., 2015].

Besides parameter sharing, another critical aspect in MTL is the weighting of loss across different tasks. A prominent recent approach is multi-objective optimization, which integrates the interactions between tasks into the loss fucntion [Kendall et al., 2018]. In this study, we incorporate the hard parameter sharing strategy with gradient-based multi-objective optimization to learn a better representation that can be shared across tasks.

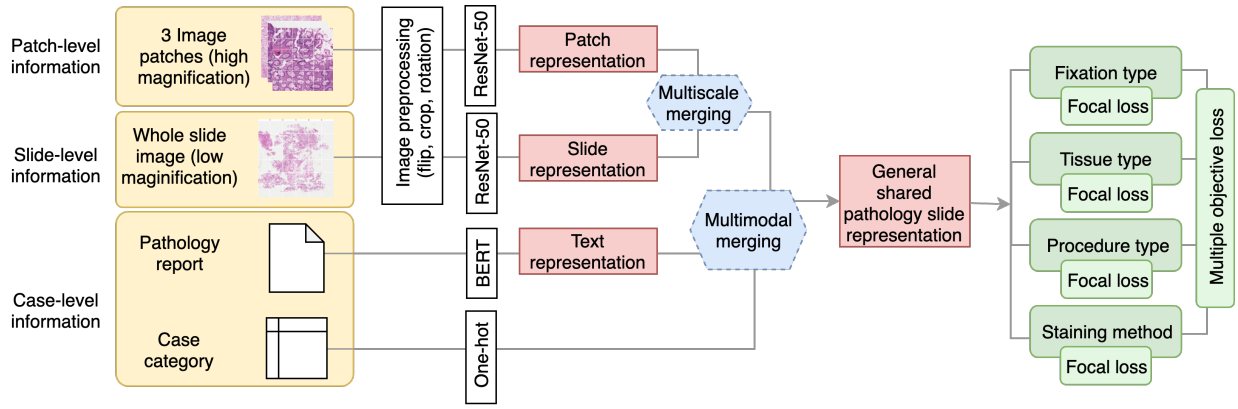


Figure 6-2: Study overview and the proposed multimodal multitask learning framework. Color scheme: input data (yellow), data encoder (white), multimodal merging operations (blue), representations (red), tasks (green), learning objectives (light green).

6.4 Methods

We propose a framework for joint prediction of pathology metadata while leveraging multimodal data, including images at different scales, texts from pathology reports, and the case-level structured data (Figure 6-2). We focus on four metadata commonly used for constructing research cohorts from the pathology samples in biobanks. They are tissue type, fixation type, procedure type, and staining method of a slide. For example, to construct the research cohort for the lymph node metastasis detection, we need to identify samples of lymph node specimen (tissue type) using hematoxylin and eosin staining (H&E) (staining method) obtained by biopsy (procedure type) and fixed by formalin-fixed paraffin-embedded fixation (FFPE) (fixation type) [Liu et al., 2017]. Figure 6-3 shows samples of pathology images with their corresponding metadata.

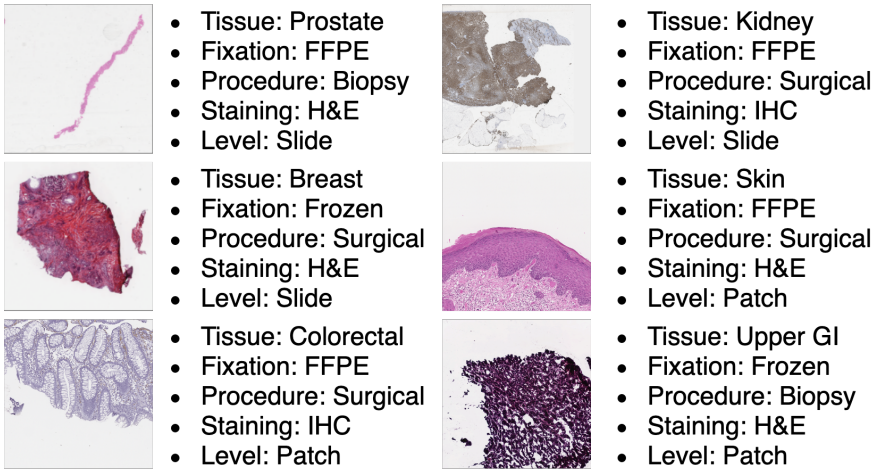


Figure 6-3: Examples of pathology slide- and patch-level images with their metadata.

The framework consists of two main parts; a multitask output and a multimodal input. For the output, the model predicts four metadata tasks considering the multi-objective loss. For the input, we adopt ResNet [He et al., 2016] and Bidirectional Encoder Representations from Transformers (BERT) [Devlin et al., 2019] to extract features from slide images and free texts, respectively. For the case-level structured data such as primary cancer sites, we encode them in the one-hot scheme. The intention of using multi-level data is to incorporate prior knowledge into the model for narrowing the prediction search space. For example, using

primary cancer sites as the input helps focus the model on fewer tissue types that might occur in such a case, yet without the issue of data leakage from the pathology perspective.

6.4.1 Multitask Learning

We use MTL with hard parameter sharing to learn the shared representation of pathology data from joint supervision across different tasks (Figure 6-2). The proposed MTL loss function consists of two parts, multi-objective loss, and task-specific losses.

For the multi-objective loss, to utilize the interactions between supervision from different tasks, we extend a Gaussian likelihood-based multitask loss accounting for all tasks simultaneously to mitigate the task-specific noise. The approach models the estimated noise of task t as a trainable parameter σ_t. Different from treating all tasks equally without scaling the gradient of each task, the multi-objective loss has the potential to prevent overfitting to any specific task by considering the geometric average of all task-specific losses. This is different from the commonly seen arithmetic mean approach and addresses the issue of vastly different scales across task-specific losses.

In more detail, for the metadata prediction task t, we adopt a Gaussian likelihood to model the label y_{it} and the predicted output from the neural network f_t with training input x_i and weights w. The output noise is modeled as a Gaussian distribution with zero mean and a standard deviation of σ_t. To find the stationary points of the Gaussian log likelihood function $\sum_{it} \log(\frac{1}{\sqrt{2\pi}\sigma_t} \exp(-\frac{(y_{it}-f_t(x_i;w))^2}{2\sigma_t^2}))$, we set the partial derivative of the likelihood function with respect to σ^2 to zero and obtain the update equation for σ^2. By replacing σ^2 in the likelihood function, we obtain the loss function:

$$\mathcal{L}_{\text{multi}} = \sum_t \log(\sum_i (y_{it} - f_t(x_i; w))^2) \tag{6.1}$$

The loss function is a geometric average of the task-specific mean squared error, different from the commonly-seen arithmetic mean. The details are as follows:

Given the metadata prediction task t, we make the assumption that the task predicted output y_{it} for the training input x_i returned by the neural network f_t under the weight set w, the noise scalar (error) of the output is zero-mean and normally distributed with standard

deviation σ_t. Thus, we want to find the stationary points of the Gaussian log likelihood:

$$\sum_{it} \log(\frac{1}{\sqrt{2\pi}\sigma_t} \exp(-\frac{(y_{it} - f_t(x_i; w))^2}{2\sigma_t^2})) \tag{6.2}$$

We set the partial derivative of the Gaussian log likelihood with respect to the variance σ^2 to zero and obtain the update equation for σ^2:

$$\frac{\partial}{\partial \sigma_t^2} \sum_i \log(\frac{1}{\sqrt{2\pi}\sigma_t} \exp(-\frac{(y_{it} - f_t(x_i; w))^2}{2\sigma_t^2})) = 0 \tag{6.3}$$

$$\frac{\partial}{\partial \sigma_t^2} \sum_i (-\frac{\log \sigma_t^2}{2} - \frac{(y_{it} - f_t(x_i; w))^2}{2\sigma_t^2}) = 0 \tag{6.4}$$

$$\sum_i (-\frac{1}{2\sigma_t^2} + \frac{(y_{it} - f_t(x_i; w))^2}{2\sigma_t^4}) = 0 \tag{6.5}$$

$$\sum_i (-1 + \frac{(y_{it} - f_t(x_i; w))^2}{\sigma_t^2}) = 0 \tag{6.6}$$

$$\sigma_t^2 = \frac{\sum_i (y_{it} - f_t(x_i; w))^2}{N}, \tag{6.7}$$

where N is the size of training data. Then we replace the σ_t^2 in the log likelihood function and get the following simplified version:

$$\sum_t (-\frac{N}{2} \log(\frac{1}{N} \sum_i (y_{it} - f_t(x_i; w))^2) - \frac{N}{2}) \tag{6.8}$$

$$- \sum_t \log(\sum_i (y_{it} - f_t(x_i; w))^2). \tag{6.9}$$

To maximize the likelihood, we convert the above to the format of loss function and minimize it:

$$\mathcal{L}_{\text{multi}} = \sum_t \log(\sum_i (y_{it} - f_t(x_i; w))^2), \tag{6.10}$$

The above loss function can be optionally exponentiated as:

$$\mathcal{L}_{\text{multi}} = \prod_t \sum_i (y_{it} - f_t(x_i; w))^2 \tag{6.11}$$

where we can see the loss is a geometric average instead of the arithmetic mean.

To mitigate the class imbalance issue, we apply focal loss (FL) for task-specific objectives [Lin et al., 2017]. In a binary classification task, we frequently use the cross entropy (CE) loss: $\mathcal{L}_{\text{CE}} = -\sum_i \log(p_i)$, where $p_i = p$ if the label is correctly predicted, else $p_i = 1-p$. $p \in [0, 1]$ is the estimated probability for correct class prediction. The CE loss can be further extended to the α-balanced CE loss that considers the class imbalance by multiplying a weighting factor $\alpha \in [0, 1]$: $\mathcal{L}_{\alpha\text{CE}} = -\sum_i \alpha_i \log(p_i)$, where $\alpha_i = \alpha$ for correct prediction, else $\alpha_i = 1 - \alpha$. However, both CE and α-balanced CE loss can not differentiate easy and hard samples well. Thus FL is introduced to reshape the α-balanced CE loss to down-weight easy samples and focus on hard samples by introducing a modulating factor $(1 - p_i)^\gamma$ with a focusing parameter $\gamma \geq 0$. The objective can be expressed as

$$\mathcal{L}_{\text{focal}} = -\sum_i \alpha_i (1 - p_i)^\gamma \log(p_i). \tag{6.12}$$

where M is a multiplier. When $\gamma > 0$, the loss contribution of easy samples will be discounted, else the loss function will turn into CE loss.

The final loss function in our MTL framework is the combination of the task-specific losses and the multi-objective loss:

$$\mathcal{L} = \mathcal{L}_{\text{multi}} + \sum_t \mathcal{L}_{\text{focal}t}, \tag{6.13}$$

where t is the task index.

6.4.2 Multimodal Learning

To develop a joint representation across modalities, we explore two early fusion methods. We investigate a widely-used vector concatenation and the compact bilinear pooling (CBP)

that captures interactions between modalities more expressively [Fukui et al., 2016].

Vector concatenation is well-known as a solid approach for merging modalities. For the vector concatenation method, the shared representation used for the downstream tasks is derived as the following: $\mathcal{V}_{\text{shared}} = [\mathcal{V}_{\text{image}}; \mathcal{V}_{\text{text}}; \mathcal{V}_{\text{structured}}]$, where $\mathcal{V}$ is the vector representation of the modality.

Bilinear models take the outer product of two representations to form the joint representation. It is an alternative approach to learning the interaction between two vector representations but with a substantial computational cost. Fukui et al. [2016] proposed CBP to mitigate the issue of the computationally heavy outer product by adopting Fourier transformation tricks to operate in the transformed space. Besides, CBP also utilizes a Count Sketch projection (CSP) function Ψ to project the high-dimensional outer product vector to a lower-dimensional space as a more compact representation.

In details, the inputs of CSP function are the outer product vector $v \in \mathbb{R}^n$ and two randomly uniformly initialized constant vectors $s \in \{-1, 1\}^n$ and $h \in \{1, ..., d\}^n$. The function outputs a latent representation vector $\mathcal{V}_{\text{shared}} \in \mathbb{R}^d$, where $n \gg d$. The j-th element of $\mathcal{V}_{\text{shared}}$, $\mathcal{V}_{\text{shared}}(j)$, is defined as $\sum_{i \in \{i|h_i=j\}} s_i \times v_i$. A CSP function of an outer product between two vectors $\mathbf{X}, \mathbf{Y}$ is equivalent to the convolution between the CSP applied to $\mathbf{X}$ and $\mathbf{Y}$. Precisely, $\mathcal{V}_{\text{shared}} = \Psi(\mathbf{X} \otimes \mathbf{Y}, s, h) = \Psi(\mathbf{X}, s, h) * \Psi(\mathbf{Y}, s, h)$, where $\otimes$ is the outer product, $*$ is the convolution operation. This can again be rewritten in a format using fast Fourier transformation (FFT) and inverse Fourier transformation (FFT^{-1}), such that $i * j = \text{FFT}^{-1}(\text{FFT}(i) \odot \text{FFT}(j))$, where $\odot$ is the element-wise product. Thus, the original Ψ function for the dimensionality reduction of outer product can be operated using Fourier transformation as following: $\mathcal{V}_{\text{shared}} = \text{FFT}^{-1}(\text{FFT}(\Psi(\mathbf{X}, s, h)) \odot \text{FFT}(\Psi(\mathbf{Y}, s, h)))$, where $\mathbf{X}, \mathbf{Y}$ are input vectors and s, h are randomly assigned vectors mentioned above. These transformation tricks have the benefits of reduced computation and memory usage, enabling operations on high-dimensional vectors.

6.4.3 Multiscale Imaging

We also explore multimodal learning in the context of multiscale imaging. This is inspired by pathologists' workflow in examining images at different image magnifications to get both the

context and the details. We use the whole slide images in low magnification and three high magnification image patches randomly cropped from the tissue area in the slide image. We apply vector concatenation and CBP for multiscale learning. Using the vector concatenation, the image representation $\mathcal{V}_{\text{image}}$ will be: $\mathcal{V}_{\text{image}} = [\mathcal{V}_{\text{slide}}; \mathcal{V}_{\text{patch}_1}; \mathcal{V}_{\text{patch}_2}; \mathcal{V}_{\text{patch}_3}]$. If using CBP, $\mathcal{V}_{\text{image}} = \text{CBP}([\mathcal{V}_{\text{slide}}, [\mathcal{V}_{\text{patch}_1}; \mathcal{V}_{\text{patch}_2}; \mathcal{V}_{\text{patch}_3}])$, where $\text{CBP}(\cdot)$ is the CBP operation described above. For CBP, we follow the parameter setting in [Fukui et al., 2016] with a 16000-dimension output representation.

6.4.4 Natural Language Representation

To extract representations from the free text pathology reports, we use an attention-based BERT model as the encoder [Devlin et al., 2019]. The implementation of the BERT encoder contains 12 Transformer blocks and 12 self-attention heads that output a 768-dimension representation. The encoder is pre-trained on large English corpora consisting of Wikipedia and BookCorpus. The last two Transformer blocks are set to be trainable to be fine-tuned for our tasks. After integrating the pre-trained BERT into the proposed multimodal merging and multitask framework, two Transformer blocks are fine-tuned with the backpropagated gradient from the overall loss (Figure 6-2).

6.5 Data

We collect two datasets for this study, a dataset from a tertiary teaching hospital (TTH) and a publicly available TCGA dataset. The metadata of two datasets has been annotated by board-certified pathologists. The TTH dataset is split into three subsets of 80% (18,413 slides / 4,972 cases), 10% (2,570 slides / 1,324 cases), 10% (2,570 slides / 1,324 cases) for training, validation and testing, respectively. We adopt iterative stratified sampling to ensure that the proportions of classes are nearly equally distributed in three subsets. A random subset of the TCGA dataset (10,440 slides / 7,084 cases) is annotated and used as an independent hold-out test set to evaluate the generalization of our method to the unseen data source.

All samples are categorized into two fixation types, 14 tissue types (with one type as "others"), two procedure types, and two staining methods, as shown in Table 6.1. Extremely

rare tissue types are recategorized into "others" unless they are the top-8 tissue types in the TCGA dataset.

Task	Dataset	TTH						TCGA	
	Split	Train		Val		Test		Test	
	Count/%	N	%	N	%	N	%	N	%
	#Case	4972		1324		1432		7084	
	#Slide	18413		2570		2835		10440	
Fixation	FFPE	17790	96.6	2409	93.7	2660	93.8	4734	45.3
	Frozen	623	3.4	161	6.3	175	6.2	5706	54.7
Tissue	LN	2561	13.9	324	12.6	394	13.9	0.0	0.0
	Uterus/cervix	1697	9.2	269	10.5	263	9.3	701	6.7
	Breast	2029	11.0	298	11.6	301	10.6	1010	9.7
	Other	2131	11.6	242	9.4	264	9.3	3284	31.5
	Skin	1966	10.7	235	9.1	290	10.2	178	1.7
	Prostate	2048	11.1	279	10.9	305	10.8	465	4.5
	Colorectal	1739	9.4	241	9.4	282	9.9	707	6.7
	H&N	1610	8.7	221	8.6	243	8.6	403	3.9
	Thyroid	1025	5.6	161	6.3	151	5.3	447	4.3
	UGI	960	5.2	147	5.7	174	6.1	589	5.6
	Ovary	581	3.1	89	3.5	104	3.7	495	4.7
	Kidney	29	0.2	29	1.1	29	1.0	1130	10.8
	Lung	37	0.2	35	1.4	35	1.2	1031	9.9
Procedure	Surgical	10853	58.9	1435	55.8	1651	58.2	9551	91.5
	Biopsy	7560	41.1	1135	44.2	1184	41.7	889	8.5
Staining	H&E	15086	81.9	1787	69.5	1963	69.2	10425	99.9
	IHC	3327	18.1	783	30.5	872	30.8	15	0.01

Table 6.1: Datasets statistics and label distributions.

6.6 Experiments

6.6.1 Data Preprocessing and Resampling

In MTL, the class imbalance issue is aggravated because of the number of tasks. In this study, there are 112 possible label combinations across four tasks (2 fixation types $\times$ 14 tissue types $\times$ 2 procedure types $\times$ 2 staining methods), Table 6.1. The ratio of cases between

the most frequent combination and the least frequent combination grows exponentially with respect to the number of tasks. For example, the number of cases in the most common combination (FFPE, LN, Surgical, H&E) is 16,415 times more than in the least frequent combination (Frozen, Lung, Biopsy, IHC). To address this issue for model development, we upsample rare combinations to 500 cases per combination and downsample combinations with abundant data to 1000 cases per combination in the training set. There is no change to the class distribution for validation and test sets.

We use images with two different scales in this study, low magnification whole slide images and high magnification image patches from slides. For whole slide images, due to the difficulty of fitting gigapixel images into the memory [Liu et al., 2017], all the images are retrieved at 0.3125× and then rescaled to 512×512 pixels by bilinear interpolation regardless of the aspect ratio. The slide-level features such as spatial relationship and colors of the tissues are still preserved through the rescaling. From the pathology perspective, such preprocessing does not affect the decision of metadata annotation because the spatial relationship is preserved. We use the high magnification image patches for incorporating fine-grained features. The images are retrieved from the pathology slide images at 5× magnification and randomly cropped out three patches within the tissue area at 299 × 299 pixels.

For the free-text pathology reports, we set a fixed text sequence length of 64 for the BERT encoder input. The average sequence length of the pathology reports in our datasets is around 100, yet the report lengths are skewed since most biopsy reports are concise. Also, the most critical information in the pathology reports is usually shown at the beginning, such as the diagnosis of the case. Thus, we set the maximal sequence length at around the 60% quantile of all report lengths for text encoding.

We also select the primary cancer site as structured data input to capture more related information at the case level. Even though the primary cancer site seems to be similar to the tissue type, they are different from the pathology perspective. The primary cancer site is the case-level information yet tissue type is slide-level information (Figure 6-1, Table 6.2). The tissue type of a slide can be completely different from the case-level primary cancer site even though they are from the same patient. For example, a patient with the primary cancer site of the breast may have multiple pathology slides. Some of the slides can be lymph nodes

or skin tissues due to metastasis and invasion.

Visit		Generated data	
Time	Reason	Case-level	Slide-level
1	Breast mass, arrange biopsy	Clinical note×1	
2	Biopsy	Pathology report×1 (benign)	Slide images×2
3	Followup	Clinical note×1	
4	Another mass, arrange biopsy	Clinical note×1	
5	Biopsy	Pathology report×1 (high grade IDC)	Slide images×3
6	Followup, arrange surgery	Clinical note×1	
7	Surgery	Pathology report×1 (IDC, stage 2a, T2N1M0) surgery report×1	Slide images×10 (both breast and lymph node tissues, with frozen and FFPE fixations, H&E and IHC staining)

Table 6.2: An example of patient visits from complaining about a breast mass to surgery. The pathology data at the case-level and slide-level provide different information about the disease.

We use a one-hot representation to encode the case-level structured input.

6.6.2 Neural Network and Baseline

We use the standard ResNet-50 model for learning the slide- and the patch-level image representations [He et al., 2016]. Text representation is learned from the BERT encoder [Devlin et al., 2019]. Case-level structured data are featurized with a one-layer fully connected network.

We apply the FL with $\gamma = 2.0, \alpha = 0.5$, and use the Adam optimizer with $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$ and set a clipping normalization value of 0.5 for optimization. Exponential decay learning rate scheduling with an initial learning rate of 10^{-3}, with a decay rate of 0.9 for every 200 training steps, is used. We train the model with five epochs with a batch size of 32.

The baseline configuration for comparison is the single modal single task framework with multiscale image (slide + patch) inputs. We use both slide and patch-level information as the baseline since it is the most comprehensive way to read the slide for pathologists.

6.6.3 Evaluation

We evaluate the models' performance on metadata prediction with two standard metrics, the macro average of the area under the receiver operator characteristic curve (AUC-ROC). The AUC-ROC measures the probability of the model ranking a randomly chosen positive sample higher than a randomly chosen negative sample. Macro AUC-ROC and take an equally weighted average of AUC-ROCs across all classes, respectively. An equal-weighted average is desired because rare cases are treated equally as common cases. Other commonly used metrics such as accuracy, and micro average AUC-ROC are not suitable for this study due to the highly unbalanced test set. These metrics are computed per sample instead of per class; therefore, a model can achieve high accuracy simply by predicting well on the majority classes while ignoring minority classes.

6.7 Results and Discussions

6.7.1 Metadata Prediction with Multimodal Multitask Learning

We start with comparing the proposed multimodal multitask framework and the baseline single modal single task benchmark in Table 6.3. The proposed method, multimodal multitask (MM), achieves higher AUC-ROCs compared to the baseline, single modal single task (SS), on three out of four tasks and on-par on one task, on the hold-out independent TCGA test set (Table 6.3, Δ(MM, SS) = 35% (tissue), 2.33% (fixation), 0% (procedure), 1.15% (staining)). On the TTH test set, MM achieves higher AUC-ROCs on two tasks, on-par on one task, and slightly lower on one task (Table 6.3, Δ(MM, SS) = 23.94% (tissue), -1.01% (fixation), 4.41% (procedure), 0% (staining)). Predicting tissue type is a task with significant improvement by using a multimodal multitask framework.

We find that the main contribution of the performance improvement comes from the mul-

timodal component instead of the multitask component (Table 6.3, Δ(MS, SS) = 23.58% (tissue), 1.40% (fixation), -0.76% (procedure), -2.92% (staining) versus Δ(SM, SS) = 2.82% (tissue), -0.65% (fixation), -1.47% (procedure), 0.60% (staining)). However, combining multitask with multimodality often leads to further increases in performance (Table 6.3, Δ(MM, MS) = 7.32% (tissue), -0.60% (fixation), 2.99% (procedure), 3.60% (staining)). It allows the multitask framework to utilize the inductive bias from one task to learn a more generalized representation that improves or keeps the performance of other tasks. Such effect is consistent with the argument that shared representation learned from multitask helps generalize across tasks, especially when the major information sources are slightly different but correlated between tasks. In our case, image mainly provides the evidence for fixation and staining prediction, whereas texts are informative for tissue type prediction; incorporating both multimodal and multitask frameworks helps learn a better representation that shares more underlying patterns in pathology.

Among the four prediction tasks, tissue type is the most challenging task for pathologists. To classify tissue type from images, a pathologist will need to review the tissue morphology in both low and high magnification while incorporating prior knowledge about the case from the pathology report. On the other hand, the other three tasks are relatively simple and can be learned by a layman with appropriate training. Furthermore, the tissue type prediction task has 14 classes while the other tasks are binary classification problems. Due to the intrinsic difficulty of the task, the baseline only reaches an AUC-ROC of 0.60 on the TCGA test set, while the multimodal multitask framework significantly improves the performance to an AUC-ROC of 0.81 (Table 6.3, TCGA MM), and up to 0.92 after further optimization (Table 6.6).

We observe that the majority of the performance gain came from the incorporation of multimodality instead of multitask. This is because the critical information included in the reports are pathology findings, which are highly relevant to tissue type, and partially related to the procedure since different procedures may provide different findings mentioned in the reports. The reports and case-level information are less informative for fixation and staining prediction since these two metadata are directly related to slide images and are less emphasized in the reports. However, comparable results are still observed since MTL

can utilize the tissue type prediction as an auxiliary task to keep or improve the model prediction power for other tasks. Therefore, we argue that the additional modalities, free text, and structured data, provide useful information about the specific task of tissue type prediction. This is also intuitively reasonable because understanding the pathology reports provides a strong prior for a model to predict tissue type among fewer options consistent with the report.

Dataset	Cfg	Tissue	Fixation	Procedure	Staining
TCGA (external)	SS	0.60 (0.51, 0.67)	0.86 (0.84, 0.88)	0.67 (0.60, 0.72)	0.87 (0.16, 1.00)
	SM	0.60 (0.57, 0.62)	0.84 (0.81, 0.86)	0.65 (0.58, 0.71)	0.87 (0.18, 1.00)
	MS	0.79 (0.72, 0.85)	**0.91 (0.90, 0.93)**	0.65 (0.58, 0.71)	0.84 (0.22, 1.00)
	MM	**0.81 (0.79, 0.83)**	0.88 (0.85, 0.90)	**0.67 (0.61, 0.72)**	**0.88 (0.24, 1.00)**
TTH	SS	0.71 (0.60, 0.81)	0.99 (0.98, 1.00)	0.68 (0.60, 0.75)	0.84 (0.77, 0.91)
	SM	0.75 (0.69, 0.80)	**1.00 (0.98, 1.00)**	0.67 (0.58, 0.75)	**0.85 (0.78, 0.91)**
	MS	0.82 (0.75, 0.89)	0.96 (0.87, 1.00)	0.69 (0.60, 0.77)	0.82 (0.75, 0.89)
	MM	**0.88 (0.85, 0.91)**	0.98 (0.94, 1.00)	**0.71 (0.63, 0.79)**	0.84 (0.77, 0.90)

Table 6.3: Quantitative evaluation across different configurations on the TCGA and TTH test sets. TCGA is an independent test set with different data distributions from the TTH dataset. We report the values of macro AUC-ROC with 95% confidence intervals. Abbreviations: configurations (Cfg), single modal single task (SS), single modal multitask (SM), multimodal single task (MS), multimodal multitask (MM).

6.7.2 Utility of the Multitask Framework Alone

Although the multimodal multitask learning approach shows improved performance over the baseline, limited improvements are seen if we consider MTL alone without multimodal inputs (Table 6.3, Δ(SM, SS) = 2.82% (tissue), -0.65% (fixation), -1.47% (procedure), 0.60% (staining), where the performance of the two approaches are on-par considering the confidence interval). This is likely due to insufficient information in the imaging modality alone

for some tasks, such as predicting tissue type from the image alone, which is regarded as a nontrivial task by pathologists. Although MTL alone does not lead to a significant improvement over the baseline, it demonstrates on-par performance using only one model with four task heads instead of four independent models. This indicates that the multitask framework yields a more generalized representation for different tasks while it greatly reduces the required computation resource for model development and speeds up model iterations and inferences.

6.7.3 Comparison of Multimodality Strategies

For the multimodality module, we also investigated different merging strategies for integrating image and text information under the multitask scenario. Table 6.4 shows the performance comparison of vector concatenation and CBP. We observe higher performance with vector concatenation in most tasks except for the tissue type prediction problem. In Fukui et al. [2016], CBP worked reasonably well for visual-text question answering problems, where image and text modalities both have a good correlation to the targeted task. Similarly, in this study, for tissue type prediction, which requires both image and text modalities, CBP shows an improved performance by leveraging both modalities. On the other hand, for tasks where a single modality is sufficient and other modalities are not expected to contain task-related information, CBP yields inferior performance. For example, text reports usually do not contain information about fixation and staining (AUC-ROC of 0.59 using only text modality on TCGA). Therefore, we observe worse performance with CBP relative to the concatenation method (Table 6.4).

Dataset	MM-Strategy	Tissue	Fixation	Procedure	Staining
TCGA (external)	Concat	0.81 (0.79, 0.83)	**0.88 (0.85, 0.90)**	**0.67 (0.61, 0.72)**	**0.88 (0.24, 1.00)**
	CBP	**0.86 (0.85, 0.87)**	0.52 (0.48, 0.55)	0.59 (0.53, 0.64)	0.40 (0.06, 0.79)
TTH	Concat	0.88 (0.85, 0.91)	**0.98 (0.94, 1.00)**	**0.71 (0.63, 0.79)**	**0.84 (0.77, 0.90)**
	CBP	**0.89 (0.85, 0.92)**	0.52 (0.32, 0.71)	0.50 (0.41, 0.60)	0.46 (0.36, 0.55)

Table 6.4: Performance comparison between different representation merging methods on the testing sets, TCGA and TTH. We report the values of macro AUC-ROC with 95% confidence intervals.

6.7.4 Ablation Analysis to Understand the Importance of Different Modalities

To explore the effect of each modality, we conduct ablation analysis by removing text and structured data one at a time while keeping the whole slide image modality (Table 6.5). The whole slide image modality is not removed because it is the base component for the pathology metadata prediction.

Dataset	Modality	Tissue	Fixation	Procedure	Staining
TCGA (external)	All	**0.81 (0.79, 0.83)**	**0.88 (0.85, 0.90)**	**0.67 (0.61, 0.72)**	**0.88 (0.24, 1.00)**
	All w/o text	0.66 (0.64, 0.68)	0.84 (0.82, 0.86)	0.66 (0.60, 0.72)	0.86 (0.17, 1.00)
	All w/o structured	0.76 (0.74, 0.78)	0.85 (0.83, 0.88)	0.60 (0.54, 0.66)	0.87 (0.30, 1.00)
TTH	All	**0.88 (0.85, 0.91)**	**0.98 (0.94, 1.00)**	**0.71 (0.63, 0.79)**	**0.84 (0.77, 0.90)**
	All w/o text	0.78 (0.74, 0.83)	0.98 (0.94, 1.00)	0.69 (0.60, 0.77)	0.84 (0.77, 0.90)
	All w/o structured	0.85 (0.82, 0.89)	0.98 (0.94, 1.00)	0.70 (0.61, 0.77)	0.83 (0.76, 0.90)

Table 6.5: Ablation analysis by removing input modality one at a time for model development. We report the values of macro AUC-ROC with 95% confidence intervals on the testing sets, TCGA and TTH.

Among the two additional modalities, we find that removing texts decreased the performance the most, especially on tissue type and procedure type metadata prediction (Table 6.5, Δ(All w/o text, All) = -14.94% (tissue), -2.28% (fixation), -2.16% (procedure), -1.14% (staining)). Case-level structured data is also predictive for some tasks but not as informative as texts (Table 6.5, Δ(All w/o structured, All) = -4.79% (tissue), -2.84% (fixation), -5.93% (procedure), -1.17% (staining)). The observed trend is consistent with the understanding that the pathology reports contain information closely related to diagnosis and tissue type but not fixation or staining information.

6.7.5 Additional Informative Modality Might Not Be Helpful

For research on multimodal modeling, a common understanding is that adding informative modalities is helpful for prediction. However, we observe inferior performance after adding an informative modality. Specifically, we explore incorporating high magnification image patches to improve the performance across all four tasks. With only image modality as the input, patch information helps the model perform better in three out of four tasks (Table 6.6,

Δ(Image, Image w/o patch) = 0.86% (tissue), 5.15% (fixation), -2.26% (procedure), 0.6% (staining)). However, adding patches yields inferior performance for the tissue and procedure type prediction when text and structured data are also used (Table 6.6, Δ(All, All w/o patch) = -10.77% (tissue), 2.27% (fixation), -11.69% (procedure), 1.16% (staining)).

Dataset	Modality	Tissue	Fixation	Procedure	Staining
TCGA (external)	Image	**0.58 (0.56, 0.61)**	**0.86 (0.84, 0.88)**	0.67 (0.62, 0.73)	0.86 (0.15, 1.00)
	Image w/o patch	0.57 (0.55, 0.59)	0.78 (0.75, 0.81)	**0.72 (0.67, 0.77)**	**0.88 (0.40, 1.00)**
	All	0.81 (0.79, 0.83)	**0.88 (0.85, 0.90)**	0.67 (0.61, 0.72)	**0.88 (0.24, 1.00)**
	All w/o patch	**0.92 (0.91, 0.93)**	0.84 (0.82, 0.86)	**0.77 (0.72, 0.81)**	0.87 (0.37, 1.00)
TTH	Image	**0.75 (0.69, 0.80)**	**0.99 (0.98, 1.00)**	**0.68 (0.60, 0.76)**	**0.85 (0.78, 0.91)**
	Image w/o patch	0.71 (0.66, 0.76)	0.98 (0.94, 1.00)	0.66 (0.57, 0.74)	0.82 (0.73, 0.89)
	All	0.88 (0.85, 0.91)	0.98 (0.94, 1.00)	0.71 (0.63, 0.79)	**0.84 (0.77, 0.90)**
	All w/o patch	**0.95 (0.92, 0.96)**	**0.98 (0.95, 1.00)**	**0.77 (0.69, 0.84)**	0.83 (0.76, 0.90)

Table 6.6: Ablation analysis on multiscale image information on the testing sets, TCGA and TTH. We report the values of macro AUC-ROC with 95% confidence intervals.

We argue that the inferior performance is due to the similarity in information between patch images and texts/structured data. Patch image is expected to be noisier than the other two modalities primarily due to image noise and patch sampling noise. The patch image only contains a narrow specific region. Therefore, the noise might come from capturing tissue images with similar visual features that are common across different tissue types. Since we only pick up three random patches, the patches may not be representative enough for a specific tissue type. This finding also leads to a potential future direction of developing a method to identify the label-specific region for patch generation without extensive annotation.

In Table 6.6, we identify that patch integration doesn't work well while other modalities (text and structured data) are used. We consult the board-certified pathologists to ensure our findings and interpretation are reasonable. We demonstrate some examples in Figure 6-4 that are misclassified by the model using patch information, yet correctly classified by the model without the patch.

As we mentioned, the patch information may not be representative enough through the generation process, which is commonly used for most machine learning tasks in pathology. For example, we expect to identify the breast tissue by seeing the breast epithelial or tubular structures. However, the patch may focus solely on fat, muscular or connective tissues

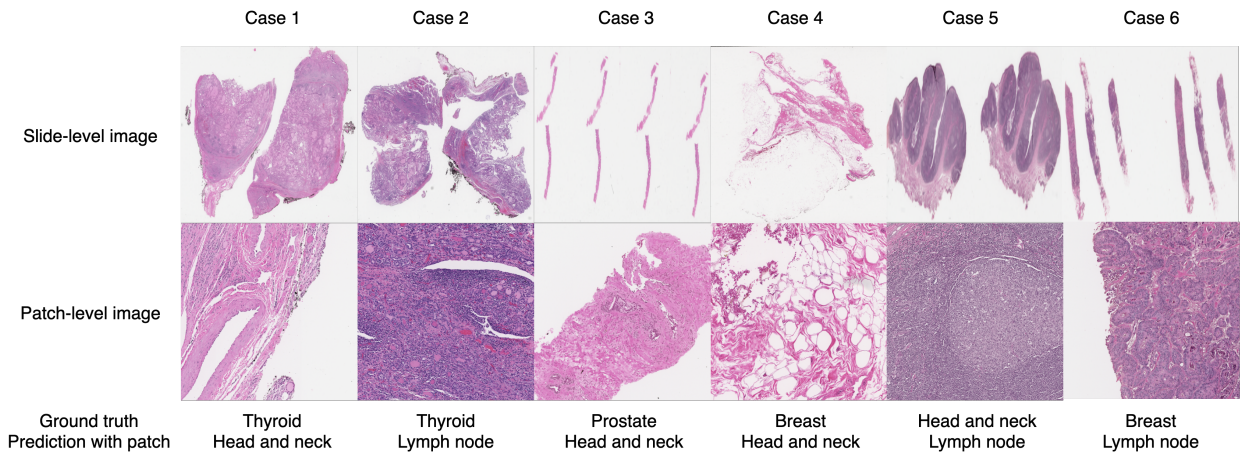


Figure 6-4: Examples of the cases in which the models with patch integration fail.

(e.g., collagen fibers), which are general across many tissue types and therefore tend to be misclassified to the head and neck tissue, which has all these features in the submucosal layer (Figure 6-4 Case 1, 3, 4).

The case with the patches that are cell-abundant may also tend to be misclassified. For instance, the dense regions of the thyroid tissue are similar to those in the lymph node tissue (Figure 6-4 Case 2). Even more, there are some regions in the head and neck is gathered the lymphatic cells (Figure 6-4 Case 5). Including such correct but misleading patches may bias the result toward the incorrect prediction, i.e., from the head and neck tissue to lymph node tissue.

Finally, the patch of cancer metastasis is also a source of misclassification between primary cancer organs and metastatic sites, e.g., breast and lymph nodes (Figure 6-4 Case 6). Even though the patch information is helpful while other modalities such as reports are not used, further patch processing is required to obtain many representative patches for better integration.

6.7.6 Relations between Multiscale Imaging and Prediction Tasks

Both slide- and patch-level images are essential for metadata prediction from the pathology perspective.

The tissue type prediction requires patch images to observe cellular and regional tissue

information. This is hard at the whole slide level at low magnification. Since the same tissue type specimen can be processed in various ways, it can eventually result in different shapes at low magnification and therefore the whole slide image can't be a good pattern for identifying the tissue type. High magnification images that demonstrate cell morphology and tissue structure are essential if there are no additional modalities used.

For the fixation type prediction, it is difficult to identify it at the whole slide level but relatively easier in the patch since the intracellular matrix in frozen sections is usually not preserved well and fragmented at high magnification due to the fast but less delicate tissue fixation process. Also, the frozen section staining process is not as robust as the FFPE fixation and therefore the images usually have less contrast. However, the contrast issue can be found not only because of the fixation type but also other issues, such as the stain normalization problem. Therefore, patch-level images are still required for better fixation type prediction.

Identifying the procedure type is challenging at both slide and patch levels due to the definition ambiguity of the procedure. The biopsy can be a needle core biopsy, incisional biopsy, or excisional biopsy. The latter two biopsy types can be very similar to surgical resection, and therefore hard to be discriminated from each other. This is also challenging for an annotation since every pathologist and specialist has a different interpretation of the procedure. For example, we usually use incisional/excisional biopsy for skin surgery rather than calling it major resection. Instead of image modality, we may rely on other inputs such as free text reports to improve model performance.

Staining type prediction has fewer issues but it highly depends on color normalization across hospitals and labs. Even for the same hospitals, preservation and timing are also critical for the staining quality.

6.8 Summary

Due to the need and challenges of acquiring well-curated metadata for large-scale biobanking, in this study, we explore the potential for using a machine learning approach for slide-level metadata prediction. We develop a multimodal multitask model to leverage information

across different modalities for the prediction of several important metadata jointly. The results show that our framework outperforms a single modality single task baseline. It also shows better performance when generalizing to the independent TCGA test set from a different source. This generalization is essential because it provides a better estimation of the actual performance on other future unseen datasets. We expect this model to be helpful in increasing the utilization of existing biobank archives, and the developed framework to be a valuable reference for future multimodal multitask learning research in the biomedical space.

Chapter 7

Robust Training for Dataset Shift

7.1 Overview

Dataset shift is a challenging issue in machine learning for medical settings. There are usually some semantic differences in medical imaging, such as contrast, brightness, noise level, across hospitals, or even devices within the same hospital. Previously, transfer learning and data augmentation were commonly used to mitigate the dataset shift problem for model generalizability; however, the model robustness often does not transfer, as demonstrated in our experiments. To tackle this problem, we adopt adversarial learning techniques to train a high-quality model with good generalizability and robustness under the dataset shift setting. Extensive experiments on a synthetic dataset that mimics the dataset shift across hospital settings and lung pathology classification tasks using real-world chest X-ray (CXR) datasets demonstrate that the proposed approach is effective, and the robustly-trained models obtain much larger adversarial accuracy and certified accuracy against input perturbations compared to nominal (non-robust) models.

7.2 Background

In machine learning, models are usually learned under the assumption that training and testing data are from the same distribution. However, the assumption often does not hold in the settings of machine learning for healthcare. It is not uncommon that there are variations

in device setups and manual operations within a dataset from the same hospital, not to mention there are often more significant discrepancies between different hospitals [Gong et al., 2017]. Oftentimes, different hospitals may have diverse patient populations, treatment plans, and protocols for using medical devices, which results in different dataset distributions. More concretely, for the current well-known benchmarks such as the CXR datasets, there exist many different sources: the NIH ChestX-ray14 dataset includes mainly standard routine examinations [Wang et al., 2017]. In contrast, the Stanford CheXpert dataset has both outpatient and inpatient images [Irvin et al., 2019] and in MIMIC-CXR, there are more patients in critical conditions [Johnson et al., 2019]. Even in the same clinical dataset, there may exist a data distribution shift problem due to a change in recording devices [Gong et al., 2017]. Such dataset shift problems result in challenges of adopting the model trained on one dataset to the other new, unseen datasets, and is known as the model generalization problem. Additional challenges occur when we further consider the dataset shift problem accompanied by the aforementioned intrinsic variations. Therefore, how to train a machine learning model that can obtain good generalizability and the properties of robustness as well as safety and trustworthiness in the scenario of dataset shift becomes a critical issue [Challen et al., 2019, Subbaswamy and Saria, 2020], especially for a decision-critical domain such as healthcare [Pooch et al., 2020, Quionero-Candela et al., 2009].

To improve the safety of adopting machine learning in healthcare, in this work, we learn models whose performance on generalization and robustness can both be kept while transferring to a dataset with different data distribution. i.e., generalizability and robustness preserving w.r.t. dataset shift. We investigate different robust training mechanisms, from random and semantic data augmentations to adversarial and certified robust training schemes to improve the model robustness. We use six CXR lung pathology classification tasks, which can be deployed in real-world clinical practice to improve residents' sensitivity to diagnosis [Hwang et al., 2019], as a use case, and the synthetic dataset based on the MNIST benchmark for a sanity check, to demonstrate the potential of utilizing the robust training for the dataset shift problem in machine learning for healthcare.

In this work, we start from the real-world need of mitigating the dataset shift and investigate how different machine learning approaches work under different levels of dataset

shift, in order to understand how to improve the model robustness with suitable techniques. Specifically, we make the following contributions and generalizable insights:

- We first apply and compare different techniques, including image data transformation, adversarial and certified robust training methods, for the real-world dataset shift problem in medical imaging.
- We examine the proposed methods on (1) a synthetic dataset based on MNIST data which mimics the dataset shift between data from two hospitals, and (2) on two real-world CXR imaging datasets.
- We demonstrate that the robust training techniques provide us better-transferred model robustness under dataset shift with intrinsic variation, while standard training and image transformation outperform the robust training under dataset shift without intrinsic variation in the real-world CXR datasets but not in the synthetic dataset.

7.3 Related Works

7.3.1 Dataset Shift in Machine Learning for Healthcare

Dataset shift is a challenging model generalizability problem where the data distribution is different in the training and testing phase [Quionero-Candela et al., 2009]. The common causes of the dataset shift are sample selection bias and non-stationary environments. Conventional approaches to deal with the dataset shift problem include removing features that contribute to the covariate shift or performing importance reweighting [Shimodaira, 2000, Sugiyama et al., 2007]. In the field of machine learning for healthcare, the dataset shift problem has been explored in several different studies. For example, Gong et al. [2017] and Nestor et al. [2019] investigated the non-stationary electronic health record, where the dataset shift problem happened in the temporal dimension. In the field of medical imaging, AlBadawy et al. [2018] performed the brain tumor segmentation task and showed that the model performance significantly degraded when the model was trained on data from a different dataset (different distribution); while they tried to resolve this problem by training

the model also on multi-institution data, the model performance was still poor. Meanwhile, Pooch et al. [2020] also showed that such damage to model generalizability existed in the CXR pathology classification problem. Similarly, Zech et al. [2018] trained and validated the convolutional neural network (CNN)-based CXR pneumonia detection classifier, and found that the model performed significantly worse when applied to a new dataset due to the significant dataset shift. Recently, Janizek et al. [2020] found that the adversarial learning approach from domain adaptation [Louppe et al., 2017] helped to improve model predictive performance when there was a dataset shift problem on an external dataset. Motivated by their works, in this chapter, we investigate the dataset shift problem on the CXR lung pathology classification task through the lens of adversarial learning, which originated from the field of neural network adversarial robustness. In particular, we study the model performance and robustness under different levels of dataset shift with additional challenges on inevitable dataset intrinsic variations.

7.3.2 Robustness and robust learning

Adversarial attacks have become a critical topic in machine learning due to the security vulnerability of deep learning systems against fabricated adversarial samples [Szegedy et al., 2014]. There are two major attack categories, one is white-box attacks, where an attacker has full knowledge about the machine learning model; while the other is black-box attacks, where an attacker can only query the probability or label on the machine learning model without access to the model internal architecture and parameters. Popular methods for the white-box attack are gradient-based methods, including fast gradient sign method (FGSM) [Goodfellow et al., 2015], projected gradient descent (PGD) attack [Madry et al., 2018], Carlini and Wagner (C&W) attacks [Carlini and Wagner, 2017]; while the attack algorithms for black-box attacks are often based on gradients estimation through finite differences or genetic algorithms. On the other hand, researchers explored various techniques to improve the model's robustness against the different attacks, including the adversarial learning approaches based on FGSM adversary [Goodfellow et al., 2015], PGD adversary [Madry et al., 2018, Wong et al., 2020] or use adversarial risk as an objective [Uesato et al., 2018]. Another recently prevalent approach for adversarial learning is through formal verification methods [Katz et al., 2017,

Sinha et al., 2018, Wong and Kolter, 2018, Weng et al., 2018a], and such approach is known as certified training/defense or verification-based learning. Formal verification (also known as robustness certification) can provide formal guarantees, a.k.a. robustness certificate, on any given input for a machine learning model such the prediction results are consistent [Katz et al., 2017, Sinha et al., 2018, Wong and Kolter, 2018, Weng et al., 2018a, Zhang et al., 2018, Boopathy et al., 2019].

We note that recent literature has shown that adversarial attacks are possible on medical deep learning systems [Finlayson et al., 2019]. Although the adversarial perturbations are not necessarily realistic in the medical imaging application, as the natural variations (such as random noise, semantic perturbations) are more common, we can think of the adversarial perturbations as the "worst-case" perturbations covering the natural variations and therefore applicable for us to develop effective adversarial learning or the robust learning algorithms to enhance model robustness. Specifically, in this work, we apply robust training methods (both adversarial learning and verification-based learning) to the dataset shift problem in medical imaging and compare them with other non-robust training approaches.

7.4 Methods

To investigate the model transfer robustness of the different machine learning techniques, we explore the approaches using the image data transformation and robust training methods including adversarial and certified training techniques.

7.4.1 Approach 1 - Image data transformation

The conventional method to train deep learning models robust to dataset shift is through training on transformed data. As the variations of random noise and semantic transformation (e.g., rotations, brightness and contrast changes, etc.) are more likely to occur in a medical imaging dataset, we train models with semantic transformation. We note that since our focus is the CXR data, which has a natural orientation, it would be more realistic to carefully design the transformation. For example, a CXR image with a vertical flip won't be realistic in a real-world setting. The setup we consider includes realistic changes on the rotation ($\pm 15^{\circ}$),

translation (±15%), scaling (±15%), shearing (±7.5°), brightness, contrast and saturation (±25%), horizontal and vertical flip with the probability of 0.25, and adding Gaussian noise with σ between 0 and 0.025. Note that the data transformation approach using simple image transformations like rotation and translation has been proven to be helpful to break the robustness of the standard training model if they are not trained with transformed images [Engstrom et al., 2017]. Therefore, we enhance our standard training models by training models with such transformed images.

7.4.2 Approach 2 - Robust Training

Instead of adding data transformation (or perturbation) before training in Approach 1, we consider and incorporate the perturbation during the training in Approach 2.

Adversarial Training

The idea of adversarial training attempts to learn networks f_θ that are robust to the threat model $\Delta = \{\delta : ||\delta||_\infty \leq \epsilon\}$ where $\epsilon > 0$. The formulation is a robust optimization problem [Madry et al., 2018]:

$$\min_{\theta} \max_{\delta \in \Delta} \mathcal{L}(f_\theta(x + \delta), y), \tag{7.1}$$

where (x, y) are data points following training distribution $\mathcal{D}_{\text{train}}$, and $\mathcal{L}$ is the loss function. For adversarial training, we can use the FGSM [Goodfellow et al., 2015] or the PGD attack. FGSM approximates the inner maximization with the closed form as:

$$\delta^* = \epsilon \cdot \text{sign}(\nabla_x \mathcal{L}(f(x), y)). \tag{7.2}$$

However, the FGSM has a relatively inaccurate approximation for perturbations. In PGD training, several smaller FGSM steps with the size of α are performed to reach a better approximation [Madry et al., 2018]. To accelerate the training, FREE adversarial training, the method that claims to have the same computational cost as conventional natural training, is introduced by taking the FSGM steps with the size of $\alpha = \epsilon$, then updating the model weights by minibatch replays without resetting the perturbations between minibatches [Shafahi et al.,

2019]. Recent work shows that using FGSM with random initialization [Wong et al., 2020], known as FAST-FGSM, is as effective as the PGD-based training but has a much lower computational cost. Therefore, in our study, we use FAST-FGSM to train our robust model.

Certified Robust Training

Certified Robust Training, also known as verification-based training or certified defense, has attracted lots of interest in this field recently. Along this line, as the first seminal work, Xiao et al. [2019] proposed a co-design training algorithm based on verification and showed that the exact verification method could be accelerated by imposing weight sparsity and activation stability (so-called rectified linear unit (ReLU) stability) [Tjeng and Tedrake, 2019]. On the other hand, Kolter and Wong [2018], Raghunathan et al. [2018], Gowal et al. [2019], Mirman et al. [2018] aim to train a more *robust* or *verifiable* model targeted on different tightness-level verifiers to trade-off computational efficiency [Kolter and Wong, 2018, Raghunathan et al., 2018, Weng et al., 2018a, Singh et al., 2019, Zhang et al., 2018, Boopathy et al., 2019]. By incorporating the robustness verification bounds into the training process, the learned model yields strengthened robustness with the certificate.

Nevertheless, current verification-based training methods are multiple times slower than standard (non-robust) training per training step. Interval bound propagation (IBP) [Gowal et al., 2019], instead, is faster than the other verifiers at the price of a weaker robustness certificate [Weng et al., 2018a, Singh et al., 2019, Zhang et al., 2018, Boopathy et al., 2019]. Intuitively, IBP finds a box bounding each layer, which can result in very loose bounds for general networks, as demonstrated in Kolter and Wong [2018], Gehr et al. [2018]. However, the IBP verifier can be strengthened by training the model with IBP bounds. Considering the computation efficiency, we thus adopt IBP training in our study on the dataset shift problem. The way that IBP works is to bound each layer in a neural network with a fixed upper and lower bound. These bounds are then propagated at each layer of the network using the previous layer's bounds. Specifically, given the layer-wise bounds at the i-th layer

where $l^i \leq \mathbf{z^i} \leq u^i$, the next layer's bounds are found as:

$$u^{i+1} = W_+^{i+1}\sigma(u^i) + W_-^{i+1}\sigma(l^i) + b^{i+1} \tag{7.3}$$

$$l^{i+1} = W_+^{i+1}\sigma(l^i) + W_-^{i+1}\sigma(u^i) + b^{i+1} \tag{7.4}$$

where W_+ and W_- denote the positive and negative components of the weight matrix W respectively with other entries being zeros otherwise, b as bias parameters, and σ is the monotonically-increasing activation function. Lower bounds are found similarly.

7.4.3 Similarity between Datasets

To evaluate the similarity between training and testing datasets, we compute the eigenvector score between them, which has been adopted in natural language processing research to evaluate the distance between two language embeddings [Søgaard et al., 2018, Weng et al., 2019b]. The central idea is that the eigenvalues and eigenvectors can be used to capture the graph structure. A higher eigenvector score indicates that the given two embedding spaces are less similar. Based on the derivation of the Laplacian matrices eigenvalues, the eigenvector score can be computed as follows:

- Derive the nearest neighbor graphs, G_1, G_2, from the learned embedding spaces, then compute $L_1 = D_1 - A_1$ and $L_2 = D_2 - A_2$, where L_i, D_i, A_i are the Laplacian matrices, degree matrices, and adjacency matrices of G_i, respectively.

- Search for the smallest value of k for each graph such that the sum of the largest k Laplacian eigenvalues is smaller than 90% of the summation of all Laplacian eigenvalues.

- Select the smallest k across two graphs and compute the squared differences, which is the eigenvector score, between the largest k eigenvalues in two Laplacian matrices.

7.5 Experiments

7.5.1 Datasets

Synthetic MNIST-SHIFT Dataset

To mimic the dataset shift problem between hospitals, we first create the MNIST-SHIFT dataset that is based on the MNIST dataset. We first split the original MNIST into two equal subsets (*source 1* and *source 2*, shown in Table 7.1, 7.2, 7.3), and then we further split each source into training and testing sets in the ratio of 6 : 1, and then perform data transformation on source 2. We have created two transformation setups, transformation 1 and 2, where the transformation 1 (denoted as *Transform 1* in Table 7.1, 7.2, 7.3) includes image rotation of $\pm 5°$, translation and scaling of $\pm 5\%$, shearing of $2.5°$, the jittering of brightness, contrast and saturation with $\pm 10\%$ range. For transformation 2 (denoted as *Transform 2* in Table 7.1, 7.2, 7.3), it uses transformation 1 with an additional 25% chance of random horizontal flip. Transformation 2 on MNIST-SHIFT shares a similar image transformation process with the one we adopt for the CXR imaging, which also further deviates the data distribution between two datasets, and therefore makes the MNIST-SHIFT an appropriate toy dataset for mimicking the dataset shift while transferring a model trained on one hospital to the other hospital setting. Note that *Transform 0* in Table 7.1, 7.2, 7.3 represents no data transformation, i.e., it uses the original MNIST data.

Medical Imaging—Chest X-ray

Two CXR datasets, CHEXPERT and MIMIC-CXR, are used for the dataset shift problem. The CHEXPERT collects the CXR studies from Stanford Hospital inpatient and outpatient services across 16 years, and includes 191,229 anterior-posterior (AP) and posterior-anterior (PA) views CXR images [Irvin et al., 2019]. The MIMIC-CXR contains the CXR studies from Beth Israel Deaconess Medical Center Emergency Department between 2011 and 2016 [Johnson et al., 2019]. It includes 242,306 AP and PA view CXR images. Images with other views, such as lateral and oblique views, are removed. The labels of both CHEXPERT and MIMIC-CXR are derived from the free text reports by the same labeling algorithm,

NegBio and CheXpert [Peng et al., 2018, Irvin et al., 2019]. Both datasets includes 13 CXR lung pathology labels, which are "Enlarged Cardiomediastinum", "Cardiomegaly", "Lung Opacity", "Lung Lesion", "Edema", "Consolidation", "Pneumonia", "Atelectasis", "Pneumothorax", "Pleural Effusion", "Pleural Other", "Fracture", and "Support Devices". We choose 6 of 13 labels, which include the detection of "Cardiomegaly", "Lung Opacity", "Edema", "Pneumonia", "Atelectasis", and "Pleural Effusion", as binary classification tasks for our experiments.

7.5.2 Network Architecture and Training

For the MNIST-shift experiments, we use two layers of CNN with the kernel size of 5 and the ReLU as an activation function. Stochastic gradient descent (SGD) with a learning rate schedule starting from 0.001 and decaying 90% per 5 steps, and a momentum of 0.9 is used for optimization. We set $\alpha = 0.375$ and 20 attack iterations for FGSM, while for IBP training, we select the best hyper-parameters from training epochs $= \{20, 50, 100\}$, learning rate $= \{0.001, 0.0005\}$, 2 MLP architecture (2 layer, 3 layer) with 64 hidden nodes per layer and 2 CNN architecture (2 layer, 3 layer) with 5x5 convolutional kernel.

For the CXR experiments, we use ResNet50 as the backbone for image encoding [He et al., 2016], and add the classification head with a 13-dimension linear layer and a sigmoid activation function. We use an SGD optimizer with a learning rate initializing at 0.001, weight decay of 10^{-4}, and momentum of 0.9 to optimize the binary cross entropy loss. Early stopping is used based on the loss of the validation subset. We also perform image data transformation based on the corresponding experiments (random transformation, semantic transformation), and normalize all images according to the specific mean and standard deviation values of each dataset. For the adversarial robust training, we set $\epsilon = 10^{-2}$ and $\alpha = 10^{-2}$ for FGSM and Fast-FGSM. Additionally for Fast-FGSM, we set the minibatch replay of 8, and 5 restarts, 10 attack iterations for the PGD attack.

All experiments are conducted in a Python 3.7.6 environment, PyTorch 1.7.1, Torchvision 0.8.2, and trained on a server with 4 NVIDIA TITAN RTX GPUs.

7.5.3 Evaluation

In MNIST-SHIFT experiments, we compute the certified accuracy of the classification. Instead, we compute the area under the receiver operating characteristic curve (AUC-ROC) for the lung pathology classification problem using CXRs. The reason for using the AUC-ROC rather than accuracy is that the CXR datasets have skewed, unbalanced class distribution; also we care more about true positive and false negative samples while making clinical diagnoses.

7.6 Results and Discussions

7.6.1 Results on Synthetic MNIST-SHIFT Dataset

Standard training, two image data transformation methods, and two robust training methods, including adversarial training with FAST-FGSM and certified training with IBP are performed on the synthetic MNIST-SHIFT dataset. In Figure 7-1, we show examples of the synthetic MNIST-SHIFT dataset with different transformation methods.

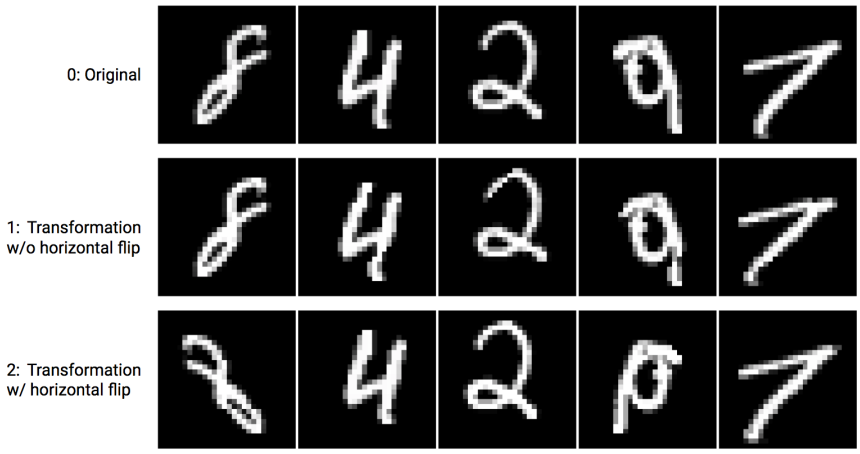


Figure 7-1: Examples of different transformations in the synthetic MNIST-SHIFT dataset.

Table 7.1 and 7.2 demonstrate the comparison between the performance of standard training and FAST-FGSM, where model 1 and 2 represent two MNIST-SHIFT subsets that mimic data from two different hospitals. In particular, the models are trained using a state-of-the-art efficient adversarial training method (FAST-FGSM) in Table 7.1-7.2 and certified

robust training method (IBP) with $\epsilon_{\text{train}} = \{0, 0.05\}$ in Table 7.3. The model with $\epsilon_{\text{train}} = 0$ means it's a standard (non-robust) model, while the model with $\epsilon_{\text{train}} = 0.05$ is a robustly-trained model. We evaluate the models using both IBP verifier (Table 7.1, 7.3) and FGSM attack (Table 7.2). In the tables, the left 2 columns (Source 1, Internal and Source 1, Transfer) denote that the models are trained on the source 1 dataset and evaluated on the internal setting (i.e., on source 1) or transfer setting (i.e., on source 2). Similarly, the right 2 columns (Source 2, Internal and Source 2, Transfer) denote that the models are trained on the source 2 dataset, and evaluated on internal setting (i.e., on source 2) or transfer setting (i.e., on source 1). We report both the result of internal and transfer accuracy in % on 3 trials (3 random seeds) and 3 data transformation schemes: {0: no transformation, 1: no random horizontal flip, 2: with 20% random horizontal flip}. Higher accuracy is better.

In Table 7.1, the results indicate that standard training (transform = 0, $\epsilon_{\text{train}} = 0$) and data transformation methods (transform = 1 or 2, $\epsilon_{\text{train}} = 0$) yield comparable performance while transferring the model trained on one dataset to the other dataset, where there is merely a gap between internal and simple transfer accuracy. However, the performance of models using standard training and data transformation methods decreases significantly when the dataset shift is more significant during the testing time via test data perturbation with the FGSM attack ($\epsilon_{\text{test}} > 0$). We also find that the image data transformation methods are less robust than standard training under the larger testing data perturbation. Adopting FAST-FGSM ($\epsilon_{\text{train}} = 0.05$) provides the model transfer robustness to mitigate such damage against the larger intrinsic variation during the testing time.

Nevertheless, the FAST-FGSM is still vulnerable if we further use a stronger IBP verifier (Table 7.2). Instead, adopting the certified training allows us to enhance the robustness and make models tolerant to more significant data perturbations ($\epsilon_{\text{train}} \geq 0.05$ in Table 7.3).

Therefore, we conclude that standard training and data transformation methods work for dataset shift without intrinsic variation, but don't provide transfer robustness with a larger data deviation of datasets between training and testing time. The robust training techniques can be helpful to mitigate this issue, and we demonstrate that the certified training method, IBP, yields stronger robustness than the state-of-the-art adversarial training technique, FAST-FGSM.

			Source 1, Internal		Source 1, Transfer		Source 2, Internal		Source 2, Transfer	
			mean	std	mean	std	mean	std	mean	std
Transform	ϵ_{test}	ϵ_{train}								
0	0	0	98.61	0.09	98.75	0.17	98.68	0.14	98.51	0.16
		0.05	98.83	0.04	98.87	0.08	98.84	0.11	98.72	0.13
	0.01	0	96.73	0.59	97.05	0.45	96.09	1.12	95.95	1.28
		0.05	98.51	0.13	98.57	0.17	98.46	0.06	98.29	0.24
	0.05	0	77.23	4.84	77.68	4.15	70.32	12.88	70.73	12.63
		0.05	96.26	0.14	96.56	0.47	96.62	0.16	96.29	0.42
	0.1	0	40.84	13.2	41.34	13.84	43.79	18.45	43.45	17.76
		0.05	89.51	1.87	90.03	2.12	90.11	1.66	89.95	1.13
	0.2	0	8.91	5.22	8.92	5.15	10.79	7.04	11.07	6.67
		0.05	61.09	5.6	61.31	5.27	59.49	10.71	59.77	10.55
	0.3	0	4.25	2.64	4.37	2.56	4.81	3.83	5.02	3.76
		0.05	29.29	4.75	30.06	4.47	31.03	12.91	30.68	12.27
1	0	0	96.97	0.31	96.9	0.38	96.79	0.16	96.51	0.31
		0.05	97.29	0.3	97.4	0.26	97.15	0.2	97.17	0.37
	0.01	0	93.57	1.61	93.27	1.4	92.05	2.25	91.49	2.05
		0.05	96.91	0.63	96.86	0.18	96.79	0.17	96.49	0.23
	0.05	0	63.06	5.59	62.77	5.28	57.07	13.66	56.75	13.87
		0.05	92.18	0.6	92.41	0.23	92.23	0.18	92.19	0.34
	0.1	0	28.54	10.95	28.65	11.47	35.91	13.95	36	14.81
		0.05	80.24	4.72	80.45	4.32	80.51	1.99	80.54	1.94
	0.2	0	6.93	4.31	6.83	3.94	6.83	4.03	7.28	4.18
		0.05	44.72	7.07	45.22	5.66	43.47	11.37	43.51	10.79
	0.3	0	4.06	2.59	3.57	2.08	3.36	2.64	3.44	2.41
		0.05	19.45	3.01	20.21	3.61	21.12	9.83	20.77	9.47
2	0	0	81.32	1.16	81.17	0.84	80.87	0.84	80.54	1.07
		0.05	81.65	0.54	81.57	0.65	81.41	0.3	81.3	0.42
	0.01	0	77.35	1.91	76.52	2.22	75.67	1.84	75.19	1.46
		0.05	81.03	0.76	80.69	1.05	81.03	0.62	79.89	0.56
	0.05	0	48.74	4.33	48.67	4.13	44.73	11.28	44.23	10.98
		0.05	75.67	0.37	75.39	1.1	75.19	0.96	75.81	1.3
	0.1	0	21.35	7.57	21.53	8.43	25.31	7.93	24.57	7.61
		0.05	64.45	3.68	64.41	3.35	64.23	2.58	64.11	1.71
	0.2	0	4.74	2.88	4.29	2.67	4.63	1.75	4.64	1.71
		0.05	34.41	6	34.83	4.64	34.36	9.76	34.2	8.97
	0.3	0	2.26	1.34	2.18	1.22	2.64	1.48	2.63	1.47
		0.05	14.59	1.7	15.26	2.43	16.62	7.82	16.55	7.9

Table 7.1: The models are trained using a state-of-the-art efficient adversarial training method (FAST-FGSM) with $\epsilon_{\text{train}} = \{0, 0.05\}$ and evaluated on ϵ_{test} using **FGSM attack**.

			Source 1, Internal		Source 1, Transfer		Source 2, Internal		Source 2, Transfer	
			mean	std	mean	std	mean	std	mean	std
Transform	ϵ_{test}	ϵ_{train}								
0	0	0	98.61	0.09	98.75	0.17	98.68	0.14	98.51	0.16
		0.05	98.83	0.04	98.87	0.08	98.84	0.11	98.72	0.13
	0.01	0	45.68	13.6	46	13.37	36.37	12.7	36.36	12.17
		0.05	89.59	0.55	89.93	0.71	88.66	2.46	88.52	2.04
	0.05	0	0	0	0	0	0	0	0	0
		0.05	0	0	0	0	0	0	0	0
	0.1	0	0	0	0	0	0	0	0	0
		0.05	0	0	0	0	0	0	0	0
	0.2	0	0	0	0	0	0	0	0	0
		0.05	0	0	0	0	0	0	0	0
	0.3	0	0	0	0	0	0	0	0	0
		0.05	0	0	0	0	0	0	0	0
1	0	0	96.92	0.17	96.62	0.29	96.53	0.34	96.52	0.39
		0.05	97.35	0.27	97.61	0.14	97.3	0.2	97.12	0.29
	0.01	0	32.04	9.86	32.4	9.7	24.3	9.03	24.25	8.66
		0.05	79.84	0.56	79.64	1.59	77.93	4.02	77.41	4
	0.05	0	0	0	0	0	0	0	0	0
		0.05	0	0	0	0	0	0	0	0
	0.1	0	0	0	0	0	0	0	0	0
		0.05	0	0	0	0	0	0	0	0
	0.2	0	0	0	0	0	0	0	0	0
		0.05	0	0	0	0	0	0	0	0
	0.3	0	0	0	0	0	0	0	0	0
		0.05	0	0	0	0	0	0	0	0
2	0	0	81.43	0.72	80.75	0.76	80.59	0.6	80.25	0.48
		0.05	82.02	0.5	81.55	0.51	81.53	0.71	81.29	1.36
	0.01	0	24.57	7.87	25.33	7.73	18.55	6.69	18.74	6.64
		0.05	64.06	0.03	63.83	0.88	62.63	3.47	61.49	3.09
	0.05	0	0	0	0	0	0	0	0	0
		0.05	0	0	0	0	0	0	0	0
	0.1	0	0	0	0	0	0	0	0	0
		0.05	0	0	0	0	0	0	0	0
	0.2	0	0	0	0	0	0	0	0	0
		0.05	0	0	0	0	0	0	0	0
	0.3	0	0	0	0	0	0	0	0	0
		0.05	0	0	0	0	0	0	0	0

Table 7.2: The models are trained using state-of-the-art efficient adversarial training method (FAST-FGSM) with $\epsilon_{\text{train}} = \{0, 0.05\}$ and *verified* on ϵ_{test} using **IBP verifier**.

			Source 1, Internal		Source 1, Transfer		Source 2, Internal		Source 2, Transfer	
			mean	std	mean	std	mean	std	mean	std
Transform	ϵ_{test}	ϵ_{train}								
0	0	0	98.09	0.04	98.11	0.36	98.06	0.14	97.87	0.21
		0.05	97.15	0.16	97.04	0.24	97.32	0.02	97.35	0.14
	0.01	0	72	3.43	72.24	3.17	70.96	3.65	70.85	3.7
		0.05	96.34	0.24	96.38	0.19	96.7	0.12	96.6	0.12
	0.05	0	0	0	0.01	0.01	0.03	0.02	0	0
		0.05	91.58	0.39	92.02	0.46	92.14	0.6	91.81	0.69
	0.1	0	0	0	0	0	0	0	0	0
		0.05	73.74	4.09	73.83	3.93	38.13	6.5	37.93	7.05
	0.2	0	0	0	0	0	0	0	0	0
		0.05	0	0	0	0	0.01	0.01	0	0
	0.3	0	0	0	0	0	0	0	0	0
		0.05	0	0	0	0	0	0	0	0
1	0	0	95.85	0.36	95.59	0.42	95.76	0.14	95.55	0.34
		0.05	93.58	0.8	93.49	0.53	93.99	0.47	93.85	0.51
	0.01	0	54.48	3.89	54.35	3.28	52.99	4.5	52.55	2.99
		0.05	92.43	0.69	92.2	0.99	92.9	0.4	92.67	0.39
	0.05	0	0	0	0.01	0.01	0.01	0.02	0.01	0.01
		0.05	83.09	0.52	83.54	0.39	83.69	1.32	83.31	1.14
	0.1	0	0	0	0	0	0	0	0	0
		0.05	55.11	7.44	55.98	7.28	22.43	4.22	22.4	4.42
	0.2	0	0	0	0	0	0	0	0	0
		0.05	0	0	0	0	0	0	0	0
	0.3	0	0	0	0	0	0	0	0	0
		0.05	0	0	0	0	0	0	0	0
2	0	0	80.25	0.2	79.97	0.61	79.73	0.65	79.93	0.34
		0.05	79.49	0.13	79.77	0.34	79.26	0.51	79.03	0.58
	0.01	0	43.94	3.22	44.15	2.56	43.09	3	41.99	1.95
		0.05	76.91	0.6	77.27	1.14	77.48	0.13	77.33	0.57
	0.05	0	0	0	0.01	0.01	0.01	0.02	0	0
		0.05	67.13	0.74	66.76	1.32	66.91	1.54	66.39	1.56
	0.1	0	0	0	0	0	0	0	0	0
		0.05	43.14	6.14	43.81	5.47	17.44	3.25	17.54	3.55
	0.2	0	0	0	0	0	0	0	0	0
		0.05	0	0	0.01	0.01	0	0	0	0
	0.3	0	0	0	0	0	0	0	0	0
		0.05	0	0	0	0	0	0	0	0

Table 7.3: The models are trained via IBP certified training with $\epsilon_{\text{train}} = \{0, 0.05\}$ and *verified* on ϵ_{test} using **IBP verifier**.

7.6.2 Results on Real Chest X-ray Datasets

In the section, we move from the synthetic dataset to the real-world medical imaging dataset, CheXpert and MIMIC-CXR CXR imaging datasets, for six lung pathology classification problems [Hwang et al., 2019]. Standard training, both random and semantic data transformation methods, and the adversarial training with FGSM and Fast-FGSM are conducted.

Image transformation

We find that semantic image transformation improves the performance while dataset shift without intrinsic variation but is not robust to dataset shift with intrinsic variation. Figure 7-2 shows examples of random and semantic image data transformations using the parameters described in the method section. The examples using random transformation can be less realistic, while those generated by semantic transformation are similar to the real-world CXR images.

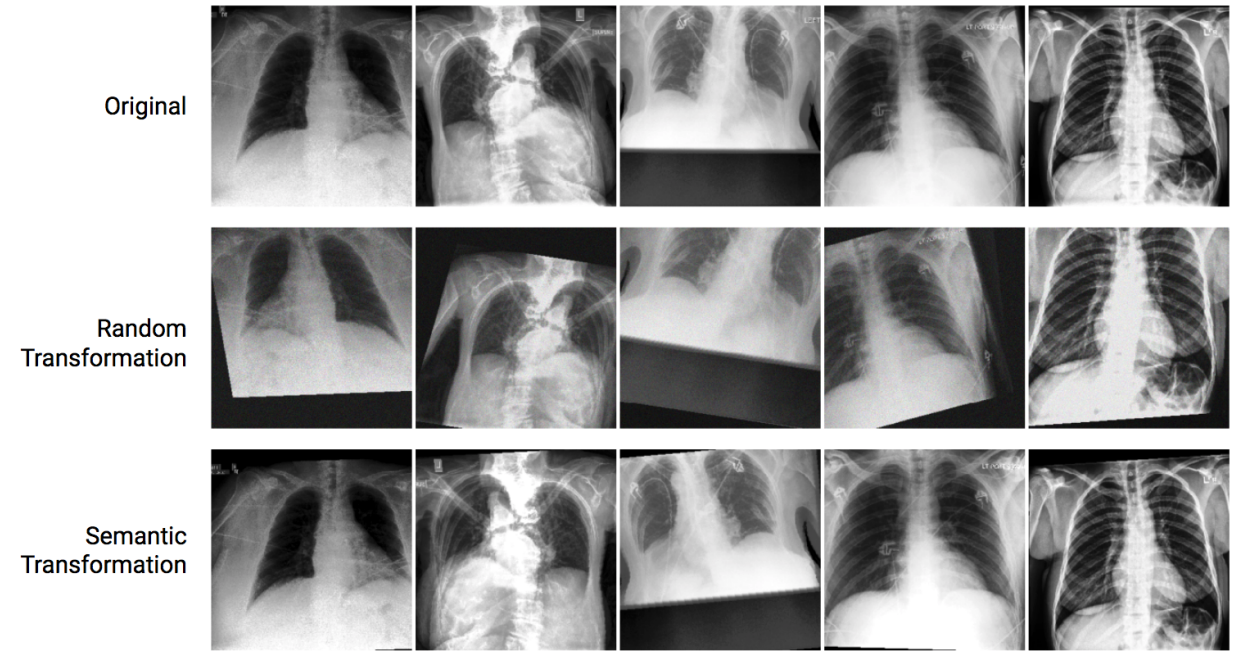


Figure 7-2: Examples from the original training dataset (CheXpert as an example), random transformation, and semantic transformation.

In Table 7.4, we find that the models trained on CheXpert usually yield better dataset shift without intrinsic variation, i.e., less performance drop, while using semantic transforma-

tion. Yet the models trained on MIMIC-CXR often perform well on dataset shift without intrinsic variation under standard training, followed by semantic transformation. Robust training techniques don't provide a comparable performance under such dataset shift without intrinsic variation across six lung pathology classification tasks. The findings align with the results we observed in the synthetic MNIST-SHIFT dataset.

Task	Method	Source	Internal	Transfer	Source	Internal	Transfer
Cardiomegaly	Standard	CheXpert	0.820	0.726 (-11.5%)	MIMIC-CXR	**0.770**	0.820 (+6.5%)
	Random Trans.	CheXpert	0.823	0.720 (-12.5%)	MIMIC-CXR	0.741	0.804 (+8.5%)
	Semantic Trans.	CheXpert	**0.844**	**0.730** (-13.5%)	MIMIC-CXR	0.761	**0.822** (+8.0%)
	FGSM	CheXpert	0.814	0.719 (-11.7%)	MIMIC-CXR	0.649	0.602 (+8.6%)
	FAST-FGSM	CheXpert	0.779	0.654 (-16.0%)	MIMIC-CXR	0.730	0.793 (+9.4%)
Lung Opacity	Standard	CheXpert	0.873	0.611 (-30.0%)	MIMIC-CXR	**0.667**	0.801 (+20.1%)
	Random Trans.	CheXpert	**0.887**	0.615 (-30.7%)	MIMIC-CXR	0.651	0.744 (+14.3%)
	Semantic Trans.	CheXpert	0.879	**0.616** (-29.9%)	MIMIC-CXR	0.665	0.793 (+19.2%)
	FGSM	CheXpert	0.852	0.595 (-30.2%)	MIMIC-CXR	0.596	0.834 (+39.9%)
	FAST-FGSM	CheXpert	0.810	0.546 (-32.6%)	MIMIC-CXR	0.627	**0.838** (+33.7%)
Edema	Standard	CheXpert	0.873	0.783 (-10.3%)	MIMIC-CXR	**0.817**	0.870 (+6.5%)
	Random Trans.	CheXpert	**0.899**	0.798 (-11.2%)	MIMIC-CXR	0.790	**0.879** (+11.3%)
	Semantic Trans.	CheXpert	0.884	**0.800** (-9.5%)	MIMIC-CXR	0.810	**0.879** (+8.5%)
	FGSM	CheXpert	0.869	0.761 (-12.4%)	MIMIC-CXR	0.774	0.845 (+9.2%)
	FAST-FGSM	CheXpert	0.834	0.718 (-13.9%)	MIMIC-CXR	0.790	0.847 (+7.2%)
Pneumonia	Standard	CheXpert	**0.813**	0.628 (-22.8%)	MIMIC-CXR	**0.656**	0.686 (+4.6%)
	Random Trans.	CheXpert	0.764	0.639 (-16.4%)	MIMIC-CXR	0.634	0.722 (+13.9%)
	Semantic Trans.	CheXpert	0.744	**0.650** (-12.6%)	MIMIC-CXR	0.631	0.697 (+10.5%)
	FGSM	CheXpert	0.722	0.591 (-18.1%)	MIMIC-CXR	0.530	0.803 (+51.5%)
	FAST-FGSM	CheXpert	0.530	0.559 (+5.5%)	MIMIC-CXR	0.593	**0.831** (+40.1%)
Atelectasis	Standard	CheXpert	0.729	0.675 (-7.4%)	MIMIC-CXR	**0.727**	**0.817** (+12.4%)
	Random Trans.	CheXpert	**0.766**	**0.701** (-8.5%)	MIMIC-CXR	0.710	0.815 (+14.8%)
	Semantic Trans.	CheXpert	0.755	0.692 (-8.3%)	MIMIC-CXR	0.721	0.814 (+12.9%)
	FGSM	CheXpert	0.722	0.657 (-9.0%)	MIMIC-CXR	0.674	0.784 (+16.3%)
	FAST-FGSM	CheXpert	0.713	0.605 (-15.1%)	MIMIC-CXR	0.699	0.800 (+14.4%)
Pleural Effusion	Standard	CheXpert	0.908	0.822 (-9.5%)	MIMIC-CXR	**0.862**	**0.904** (+4.9%)
	Random Trans.	CheXpert	**0.911**	0.832 (-8.7%)	MIMIC-CXR	0.831	0.876 (+5.4%)
	Semantic Trans.	CheXpert	0.908	**0.834** (-8.1%)	MIMIC-CXR	0.851	0.878 (+3.2%)
	FGSM	CheXpert	0.885	0.795 (-10.2%)	MIMIC-CXR	0.789	0.828 (+4.9%)
	FAST-FGSM	CheXpert	0.797	0.720 (-9.7%)	MIMIC-CXR	0.821	0.869 (+5.8%)

Table 7.4: Transferring chest X-ray pathology classifier to the other dataset with dataset shift. We report the AUC-ROC of the binary classification tasks. The percentage values in the parentheses indicate the performance drop from testing on the source dataset to the transfer dataset.

Robust training

We find that robust training holds performance under dataset shift with intrinsic variation. To deal with a larger dataset shift, we further perform model transfer where dataset shift

with intrinsic variation exists via injecting Gaussian noise in the testing data. We add Gaussian noise with a standard deviation of 0.05 and 0.1 into the testing dataset, where the noise-injected datasets are still human-readable without significant visual changes that can hide the targeted lung pathology (Figure 7-3).

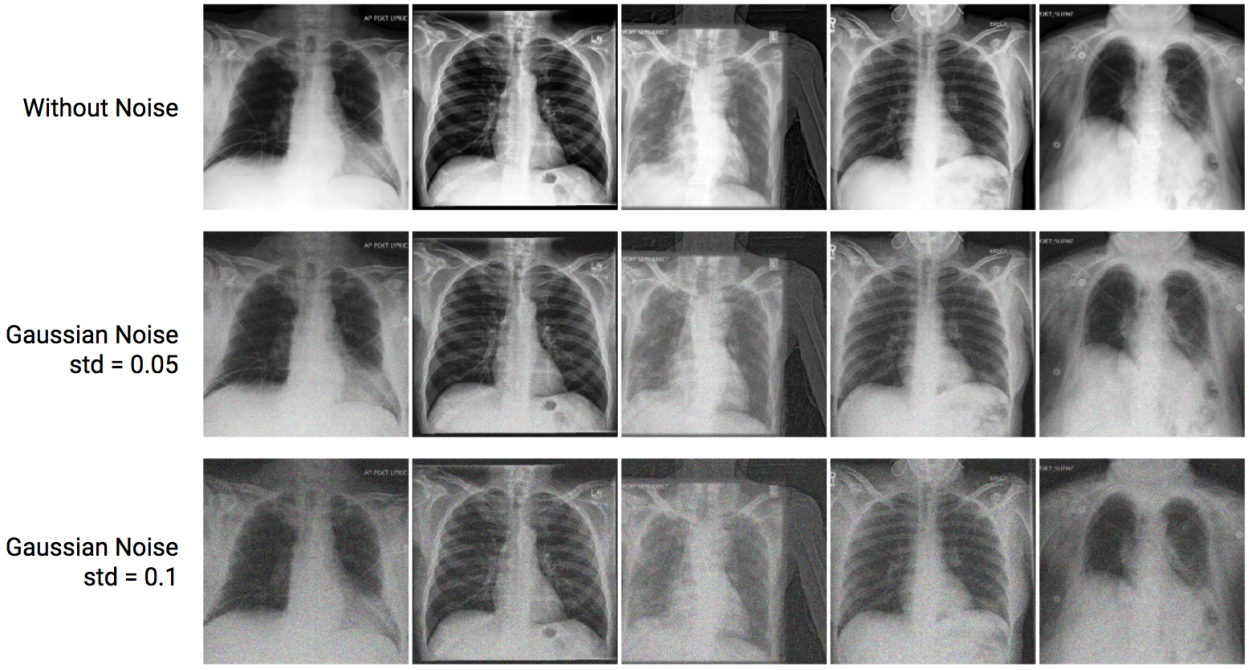


Figure 7-3: Examples from the original testing dataset (CheXpert as an example), testing examples with Gaussian noise with standard deviation of 0.05 and 0.1.

By computing the eigenvector score between source and target datasets [Søgaard et al., 2018], we also show that adding Gaussian noise with larger standard deviation (i.e., higher intrinsic variation) yields a larger eigenvector score, i.e., the image embeddings of the two given datasets are less similar, and therefore a larger shift between datasets (Table 7.5).

		CheXpert		
	Gaussian noise	0	0.05	0.1
MIMIC-CXR	0	1.52	1.527	1.93
	0.05	1.753	-	-
	0.1	2.362	-	-

Table 7.5: Eigenvector score between the training and testing dataset pairs. A higher score indicates the dissimilarity between datasets. The value of noise level is the standard deviation of Gaussian noise.

In Table 7.6, we demonstrate the results of experiments on dataset shift with intrinsic variation. We show that both standard training and image data transformation methods do not provide good transfer robustness while testing time data perturbation. Furthermore, even though the random and semantic transformation methods work well under dataset shift without intrinsic variation, we find that these image data transformation-based methods yield a larger performance drop than standard training when the noise exists at testing time. In other words, the transformation-based methods are less robust to the model transfer with a larger noise level than standard training, which we also observe in the MNIST-SHIFT experiments.

Instead, robust training methods (FGSM and FAST-FGSM) hold the transfer robustness without losing their performance when there is a larger dataset shift via noise injection. Such findings are consistent with the results in the MNIST-SHIFT synthetic dataset. We also find that FGSM performs better on the CHEXPERT-trained models, and FAST-FGSM yields better results on the MIMIC-CXR-trained models.

Task	Method	Source	Transfer Without Noise	Transfer Gaussian 0.05	Transfer Gaussian 0.1	Source	Transfer Without Noise	Transfer Gaussian 0.05	Transfer Gaussian 0.1
Cardiomegaly	Standard	CheXpert	0.726	0.714 (-1.7%)	0.699 (-3.7%)	MIMIC-CXR	0.820	**0.838** (+2.2%)	0.814 (-0.7%)
	Random Trans.	CheXpert	0.720	0.707 (-1.8%)	0.659 (-8.5%)	MIMIC-CXR	0.804	0.789 (-1.7%)	0.707 (-12.1%)
	Semantic Trans.	CheXpert	**0.730**	0.703 (-3.7%)	0.697 (-4.5%)	MIMIC-CXR	0.822	0.796 (-3.2%)	0.731 (-11.1%)
	FGSM	CheXpert	0.719	**0.720** (+0.1%)	**0.721** (+0.3%)	MIMIC-CXR	0.793	0.792 (-0.1%)	0.791 (-0.3%)
	FAST-FGSM	CheXpert	0.654	0.661 (+1.1%)	0.668 (+2.1%)	MIMIC-CXR	**0.823**	0.819 (-0.5%)	**0.816** (-0.9%)
Lung Opacity	Standard	CheXpert	0.611	0.581 (-4.9%)	0.541 (-11.5%)	MIMIC-CXR	0.801	0.712 (-11.1$)	0.681 (-15.0%)
	Random Trans.	CheXpert	0.615	**0.616** (+0.2%)	0.542 (-11.9%)	MIMIC-CXR	0.744	0.580 (-22.0%)	0.453 (-39.1%)
	Semantic Trans.	CheXpert	**0.616**	0.600 (-2.6%)	0.532 (-13.6%)	MIMIC-CXR	0.793	0.643 (-18.9%)	0.520 (-34.4%)
	FGSM	CheXpert	0.595	0.596 (+0.2%)	**0.598** (+0.5%)	MIMIC-CXR	0.834	0.833 (-0.1%)	0.831 (-0.4%)
	FAST-FGSM	CheXpert	0.546	0.547 (+0.2%)	0.546 (0)	MIMIC-CXR	**0.838**	**0.839** (+0.1%)	**0.838** (0)
Edema	Standard	CheXpert	0.783	0.721 (-7.9%)	0.634 (-19.0%)	MIMIC-CXR	0.870	0.841 (-3.3%)	0.815 (-6.3%)
	Random Trans.	CheXpert	0.798	**0.776** (-2.8%)	0.574 (-28.1%)	MIMIC-CXR	**0.879**	**0.867** (-1.4%)	0.828 (-5.8%)
	Semantic Trans.	CheXpert	**0.800**	0.752 (-6.0%)	0.675 (-15.6%)	MIMIC-CXR	**0.879**	0.829 (-5.7%)	0.807 (-8.2%)
	FGSM	CheXpert	0.761	0.761 (0)	**0.762** (+0.1%)	MIMIC-CXR	0.845	0.844 (-0.1%)	0.845 (0)
	FAST-FGSM	CheXpert	0.718	0.723 (+0.7%)	0.719 (+0.1%)	MIMIC-CXR	0.847	0.846 (-0.1%)	**0.846** (-0.1%)
Pneumonia	Standard	CheXpert	0.628	0.621 (-1.1%)	0.608 (-3.2%)	MIMIC-CXR	0.686	0.666 (-2.9%)	0.787 (+14.7%)
	Random Trans.	CheXpert	0.639	0.649 (+1.6%)	0.548 (-14.2%)	MIMIC-CXR	0.722	0.535 (-25.9%)	0.593 (-17.9%)
	Semantic Trans.	CheXpert	**0.650**	**0.654** (+0.6%)	0.588 (-9.5%)	MIMIC-CXR	0.697	0.582 (-16.5%)	0.68 (-2.4%)
	FGSM	CheXpert	0.591	0.591 (0)	**0.593 (+0.3%)**	MIMIC-CXR	0.803	0.805 (+0.2%)	0.806 (+0.4%)
	FAST-FGSM	CheXpert	0.559	0.556 (-0.5%)	0.545 (-2.5%)	MIMIC-CXR	**0.831**	**0.828** (-0.4%)	**0.828** (-0.4%)
Atelectasis	Standard	CheXpert	0.675	0.623 (-7.7%)	0.552 (-18.2%)	MIMIC-CXR	**0.817**	0.795 (-2.7%)	0.701 (-14.2%)
	Random Trans.	CheXpert	**0.701**	0.605 (-13.7%)	0.471 (-32.8%)	MIMIC-CXR	0.815	**0.810** (-0.6%)	0.694 (-14.8%)
	Semantic Trans.	CheXpert	0.692	0.592 (-14.5%)	0.536 (-22.5%)	MIMIC-CXR	0.814	0.798 (-2.0%)	0.649 (-20.3%)
	FGSM	CheXpert	0.657	**0.658** (+0.2%)	**0.658** (+0.2%)	MIMIC-CXR	0.784	0.785 (+0.1%)	0.782 (-0.3%)
	FAST-FGSM	CheXpert	0.605	0.604 (-0.2%)	0.609 (+0.7%)	MIMIC-CXR	0.8	0.798 (-0.3%)	**0.791** (-1.1%)
Pleural Effusion	Standard	CheXpert	0.822	0.810 (-1.5%)	0.763 (-7.2%)	MIMIC-CXR	**0.904**	**0.890** (-1.5%)	0.83 (-8.2%)
	Random Trans.	CheXpert	0.832	**0.827** (-0.6%)	0.704 (-15.4%)	MIMIC-CXR	0.876	0.865 (-1.3%)	0.798 (-8.9%)
	Semantic Trans.	CheXpert	**0.834**	0.816 (-2.2%)	0.746 (-10.6%)	MIMIC-CXR	0.878	0.858 (-2.3%)	0.805 (-8.3%)
	FGSM	CheXpert	0.795	0.795 (0)	**0.796** (+0.1%)	MIMIC-CXR	0.828	0.829 (+0.1%)	0.828 (0)
	FAST-FGSM	CheXpert	0.72	0.723 (+0.4%)	0.725 (0.7%)	MIMIC-CXR	0.869	0.864 (-0.6%)	**0.862** (-0.8%)

Table 7.6: Transferring chest X-ray pathology classifier to the other dataset with Gaussian noise injection and dataset shift. We report the AUC-ROC of the binary classification tasks. The percentage values in the parentheses indicate the performance drop from testing on the clean transfer dataset to testing on the transfer dataset with Gaussian noise.

7.7 Summary

We demonstrate that the robust training techniques preserve model robustness under the larger dataset shift with intrinsic variation. We show that the finding is consistent across the synthetic MNIST-SHIFT dataset and two real-world CXR datasets, CHEXPERT and MIMIC-CXR. However, the standard training and image transformation methods might outperform under dataset shift without intrinsic variation.

We observe that the models trained on MIMIC-CXR usually yield better performance across six CXR lung pathology classification tasks while testing on the CHEXPERT dataset. Such results may arise from the data heterogeneity due to the difference between the severity of the disease and the quality of images in the two datasets. The CHEXPERT dataset includes radiographs from the inpatient and outpatient services while the MIMIC-CXR contains CXR images from an emergency department in a tertiary medical center. In other words, MIMIC-CXR images can be more complicated (i.e., a more challenging machine learning task) and have lower image quality due to the nature of taking medical imaging in the emergency service, and therefore leads to lower performance while transferring the CHEXPERT model to the MIMIC-CXR dataset, and better performance vice versa. The finding also gives us an insight that having much more diverse training data can be potentially helpful for better transfer capability of machine learning models.

Even though image data transformation is a standard technique in medical imaging tasks and general deep learning, these methods are not robust to larger dataset shift with a larger internal variation during testing time. With theoretical analysis, Eghbal-zadeh et al. [2020] also demonstrated that the general-purpose data augmentations should be applied carefully since these methods may not take the specific characteristics of the task and data into consideration. Further investigation of robust data augmentations and image transformations that preserve model robustness will be a future direction. For example, designing a threat model that can defend against semantic perturbations such as color shifting and lighting condition is essential for real-world medical imaging machine learning tasks [Mohapatra et al., 2020].

Regarding the dataset shift across the CXR datasets, we consider not only the dataset shift without intrinsic variation, which directly transfers the model to the other dataset,

but also consider the dataset shift with intrinsic variation that further perturbs the testing dataset to make a larger dataset shift. In the MNIST-SHIFT, we generate the adversarial examples ($\epsilon_{\text{test}} > 0$) for attack. Instead, we inject Gaussian noise into the testing dataset while conducting the experiments on CXR datasets since this is a possible dataset shift condition in the real-world medical setting, compared with the adversarial examples, which may not necessarily be realistic [Finlayson et al., 2019]. For example, we can think of the CXR images with Gaussian noise as images with relatively low-quality 7-3, or images taken under portable devices, which also yield more significant noise levels.

Limitations

Some limitations in the study can shed light on future research directions. First, we find that the certified robust training method such as IBP can defend against strong attackers, yet it's challenging to adapt them for a very large network. Further investigation is needed to efficiently compute the tractable verification for much larger network architecture, or even for the complicated transfer learning and meta-learning settings [Shafahi et al., 2020, Wang et al., 2021]. We also only investigate methods using a synthetic task and six CXR lung pathology classification tasks. To make the conclusion more generalizable, we may extend the approaches to other medical imaging datasets for more tasks, such as dermatological disease classification using the International Skin Imaging Collaboration (ISIC) and SD-198 datasets [Weng et al., 2020b]. In the study, we use accuracy and AUC-ROC to evaluate the model performance, which is widely accepted for the robust training purpose. However, we may also consider more clinically relevant performance metrics, such as sensitivity and specificity, and other class-imbalance sensitive metrics for real-world considerations (e.g., macro precision, recall and F1-score) for better clinical interpretation of the results. Finally, in the future, we will focus more on the interpretability and explainability of the model, to ensure that the robust models are also interpretable and acceptable by healthcare providers for a potential real-world clinical use in the future.

Chapter 8

Conclusions and Discussions

In this dissertation, we systematically explore machine learning frameworks for limited data, data imbalance, and heterogeneous data, using cross-domain learning, self-supervised learning, contrastive learning, meta-learning, multitask learning, and robust learning. We present six studies with different medical applications, such as clinical language translation, pathology metadata prediction, ultrasound image classification and segmentation, diabetic retinopathy image retrieval, skin diagnosis classification, and lung pathology classification under dataset shift, to demonstrate how we approach the limited and heterogeneous medical data, and learn data representations from them. The studies are not exhaustive but also indicate that no single machine learning technique will be the best approach for all problems. Our findings also provide insights and caveats for applying machine learning methods to medical data and motivate future research directions of machine learning with low-resource and high-dimensional data.

In Chapter 2, we introduce the setup of learning cross-domain representations for a clinical natural language processing (NLP) task with limited, unpaired resources. We focus on the limited data problem in clinical language translation, a scenario of low-resource language translation. We use the completely unsupervised embedding spaces alignment method with identical anchors to approach the problem and conduct the statistical machine translation (SMT) under the limited, unparalleled data condition. We compare the proposed method with the current commonly used method, the dictionary-based algorithm, with the newly designed clinically interpretable criteria of clinical correctness and readability, to understand

whether the method yields better quality of translation. We show that our framework yields the best performance on both word- and sentence-level translation. Such a fully-unsupervised strategy overcomes the limited annotation problems, the SMT helps learn usable language models with limited data, and the designed clinically meaningful evaluation reduces biases from inappropriate evaluators, which are critical in clinical machine learning.

The proposed method's advantage is that it doesn't require large and paired training data for learning model parameters, which is ideal for the limited data and annotation setup. The method simply uses the trick of linear algebra with statistical learning-based language modeling instead of state-of-the-art deep learning-based approaches, which are data-hungry. However, the limitation is that we cannot handle word sense disambiguation using this one-to-one word mapping approach. To resolve this issue, contextualized natural language processing techniques such as Embeddings from Language Models (ELMo) and Transformer-based models are required [Aldarmaki and Diab, 2019, Peters et al., 2018, Vaswani et al., 2017]. We may consider embedding a medical knowledge graph to learn the concept-level embedding rather than simple word embedding for capturing better clinical language meaning [Liu et al., 2020a].

We may also adapt ideas from methods for the text style transfer task [Jin et al., 2022a]. Without parallel data for training, Jin et al. [2019] developed an iterative matching-translation-refinement approach with a standard sequence-to-sequence (Seq2Seq) neural network to generate a pseudo-parallel corpus, and then applied the unpaired, limited data scheme. They further developed the TitleStylist method, which combines the text summarization and reconstruction tasks into a Seq2Seq-based multitask learning framework for the unparalleled small data problem [Jin et al., 2020]. The latter approach may be a direction for summarizing long, complicated medical notes into a shorter, easily understandable summary for patients. However, both approaches can be computationally heavy with high time complexity. For future research directions, we may consider focusing more on concept translation and also solving the word sense disambiguation problem using modified neural network-based models.

In chapter 3, we use the self-supervised learning method, context encoder, which adopts the semantic in-painting technique to tackle the limited annotation problem. The technique

aims to reconstruct a central missing part of the input using the encoder-decoder architecture with adversarial learning that improves the quality of representations via discriminating between the actual and the predicted missing part. Additionally, we integrate the metadata as a multimodal input to further improve the quality of learned representations. We use the ultrasound image and its corresponding Digital Imaging and Communications in Medicine (DICOM) metadata for the experiment. We then transfer the learned ultrasound image representations to different downstream ultrasound imaging tasks—ultrasound quality classification and liver/thyroid nodule segmentation. As a pre-training method, we compare the self-supervised learning method with other approaches such as randomized initialization and ImageNet pre-training. We find that pre-training with the context encoder and multimodal information helps us learn the representation that yields better downstream task performance.

Yet the caveat of the study is the choice of metadata. That is, the selection of DICOM tags in this study is critical. Further investigation is required to know whether other metadata, such as voxel information, study details, or patient demographics, may or may not provide additional semantic information for better representation learning. We can, of course, use the prior medical knowledge to make a selection, yet it is also possible to use machine learning to choose an optimal set of metadata. Other state-of-the-art self-supervised learning techniques for pre-training may also help learn better representations. For example, the methods using negative examples for contrastive learning, such as SimCLR [Chen et al., 2020b] and MoCo (Momentum Contrast) [Chen et al., 2020c], or the techniques that don't use negative examples but add the projection prediction layer, such as BYOL (Bootstrap Your Own Latent) [Grill et al., 2020], and SimSiam [Chen and He, 2021], may be considered for more effective learning. Since we have multimodal data for training, it is also possible to adopt conditional image generation methods such as CLIP (Contrastive Language-Image Pre-training) to guide output generation conditioning on external multimodal information, like free text or metadata [Radford et al., 2021].

In chapter 4, we explore the utility of the Siamese convolutional neural network (SCNN), a contrastive learning-based method, in order to mitigate the challenge of learning generalizable representations from a limited and unbalanced medical dataset. Contrastive learning

aims to maximize the mutual information between the examples with the same label or similar patterns, and repel the representations which belong to the examples from different labels or concepts. We utilize the SCNN architecture, a contrastive learning algorithm for few-shot learning (FSL) [Bromley et al., 1994, Koch et al., 2015], to learn representations from a highly unbalanced fundoscopic image dataset. Then we examine the quality of representations via the downstream tasks of diabetic retinopathy image retrieval and stage prediction. We compare the quality of the representations learned from the SCNN against the representations learned from different layers of the standard supervised convolutional neural networks (CNN).

SCNN is known to have the strength of dealing with small data problems, providing robustness to class imbalance, and outputting the distance metric so we can easily compute the similarity between data points [Bromley et al., 1994, Koch et al., 2015]. With this network architecture, we can also learn representations without exact multiclass labels, but only use a binary label of whether the two input images have the same/different label. Yet the trade-off of using such methods is that we need more training time since the SCNN requires quadratic pairs to learn from instead of point-wise learning. The state-of-the-art contrastive learning methods mentioned above, such as SimCLR, MoCo, BYOL, and SimSiam, actually share the same idea of SCNN—they all use identical dual networks to learn representations. For future direction, we may adopt the data augmentation idea from these new contrastive learning algorithms to define the contrast and further reduce the dependency of using labels, i.e., we don't even need the binary same/different label. However, further investigation of these methods is needed since their behaviors are mostly empirical but without theoretical evidence.

In chapter 5, we focus on utilizing meta-learning/FSL for limited and extremely unbalanced medical data, and use skin diagnosis classification as a use case. We examine different meta-learning methods, specifically FSL algorithms, which include batch learning with fine-tuning, episodic metric-based, and optimization-based few-shot learning algorithms to learn representations from limited and unbalanced data. We compare the model performance of skin disease classification between these FSL methods and conventional supervised learning (CSL), i.e., direct classification, approaches. We also consider the model ensemble of CSL

and FSL methods since the former technique is superior in predicting common classes while the latter one may be helpful for rare class prediction—this is critical for skin disease classification since the skin disease distribution is long-tailed not only in the dataset but also in the real-world.

We find the model performance of solely using FSL is not superior to those trained by CSL with conventional class imbalance techniques. Yet, the model ensemble generally improves model performance, especially for rare class predictions. We also notice that the standard evaluation for FSL in general, i.e., n-way-k-shot accuracy, is not realistic. Thus, we develop a real-world evaluation method to test FSL on all data points in the test set at once, which is comparable with the standard supervised learning evaluation. However, for the real impact of the developed method, both model ensemble and the real-world FSL evaluation, further investigation on more datasets is required, especially on general domain data.

Adopting the self-supervised loss to the meta-learning/FSL framework but applying the new scheme to real-world benchmarks can also be an interesting research direction [Liu et al., 2021]. We also consider further improving the performance of meta-learning/FSL methods via more advanced techniques, such as the ANIL (Almost No Inner Loop) algorithm that focuses on feature reuse rather than rapid adaptation and learning [Raghu et al., 2020]. With a large dataset for pre-training, we may also simply rely on self-supervised learning to learn generalizable representations, i.e., learning good reusable embeddings can be better than complicated meta-learning algorithms [Tian et al., 2020].

For skin diagnosis classification, we may need to investigate the model bias/fairness problem since the dataset is highly skewed, and we should also consider skin tone, which is an attribute that may lead to a safety issue if we don't appropriately address it. To approach the problem, we may bring the skin tone information into the classification model by calculating the skin tone for unaffected skin first, then take the predicted skin tone score as an input into the final model [Kinyanjui et al., 2020].

In chapter 6, we apply the multimodal multitask framework to utilize multiple data modalities, including image, free text, and structured data, with the multi-objective loss that provides inductive bias and captures interactions between multiple tasks, to learn shared representations from the heterogeneous pathology biobank dataset for pathology metadata

prediction. The models created by our multimodal multitask framework outperform those models using the single modal single task (most of the state-of-the-art pathology machine learning models adapt this framework), single modal multitask, and multimodal single task frameworks. Pathologists' interpretation of the results also provides insights and caveats about the method.

We find that the complicated multimodal fusion method, such as compact bilinear pooling (CBP), may not be better than simple vector concatenation for some tasks. The selection of the multimodal fusion algorithm depends on the discrepancy between modalities. For modalities with huge differences, for example, image and metadata, simple vector concatenation can be helpful. Yet for fusing similar modalities, such as whole images and patches, CBP may be a better option to preserve information. We may select and design different machine learning approaches for different modality fusion problems. For example, Chen et al. [2020a] considered the patch-level histopathological image as a graph, and applied the graph convolution network (GCN) to learn hidden representations of the patches. The researcher combined the whole slide image, patches, and genomic profile using standard CNN, GCN, and feed-forward network with attention, followed by the Kronecker product to learn the final representation. Chen et al. [2021a] further adopted the co-attention transformer architecture to integrate genomic information into the pathology slide level survival prediction problem. Dealing with heterogeneous multimodal data is still an active field of machine learning, and there is no consensus on how best to approach it. We still need to wisely choose appropriate approaches based on the understanding of the data modality and domain knowledge.

In chapter 7, we use robust training techniques to learn better data representations that are robust enough to tolerate the dataset shift in the heterogeneous data setting, e.g., inputs with perturbation, transformation, or noise, which is also common in the raw medical data, and also between data sources. Comparing models trained with standard training, we find that models trained with robust training techniques yield better performance when a dataset shift exists. Extensive experiments on a synthetic dataset that mimics the dataset shift across hospital settings, and the lung pathology classification task using two real-world chest X-ray (CXR) datasets demonstrate that our approach can be effective, and the robustly-trained models obtain much larger adversarial accuracy and certified accuracy against input

perturbations compared to nominal (non-robust) models.

Even though the certified robust training methods such as interval bound propagation (IBP) are tolerant to distributional shifts, it is challenging to adapt them for a very large network, which is a standard for current machine learning in medicine. We need to explore solutions to compute the tractable verification for larger networks or much more complicated learning scenarios such as transfer learning and meta-learning [Shafahi et al., 2020, Wang et al., 2021]. Extending from the robust training for dataset shift problem, we can also combine the idea of robust training with clinically interpretable metrics for model explainability since it can be a byproduct of the adversarial robustness measurement [Etmann et al., 2019]. We may also need to apply the proposed methods to different tasks to make the conclusion of this work generalizable.

In summary, we are glad to contribute to the progression of machine learning for medicine by providing insights on learning better data representations with limited and heterogeneous medical data. For future research directions, we would like to highlight four general but major limitations and challenges that we don't focus on and address in the thesis—interpretability, generalizability, bias and fairness, and deployment.

Interpretability

First, we mainly use qualitative approaches for model analysis and interpretation in these case studies. Recently, interpretable deep learning methods allow model designers to interrogate, understand, debug, and even improve the systems by analyzing and interpreting the behavior of black-box systems quantitatively [Jin et al., 2022b, **?**]. Quantitative methods such as Shapley Additive Explanations (SHAP) also allow end-users to evaluate the model's decision-making much more objectively [Lundberg and Lee, 2017]. Thus, we may explore an approach to learn clinically interpretable models under the heterogeneous, limited, and unbalanced data set using an algorithm to compute the robustness estimate of neural network classifiers [Weng et al., 2018b].

Generalization

Our model generalizability is still limited. As mentioned in the previous discussion, there is no one solution for all different problems in machine learning for medicine. Thus, we still need to investigate more approaches to learn better representations to transfer knowledge between data distributions.

In the medical domain, data distributional shift not only exists between data sources, but also in the temporal aspects such as changes in medical practice, disease prevalence, and patient populations [Gong et al., 2017]. How to preserve the model performance under such a distributional shift becomes the most critical issue for model generalizability. Some strategies at the model level, such as model refitting, probability calibration, model updating, model selection, or feature-level approaches, can be potentially helpful [Guo et al., 2021]. Domain generalization and unsupervised domain adaptation, which aim to develop much more robust models against unseen, out-of-distribution data, can be a critical, potential solution for this challenge but require more investigation [Gulrajani and Lopez-Paz, 2021, Guo et al., 2022]. Other methods that consider uncertainty and the confidence level of predictions may also help deal with data shift/drift, erroneous data, and missingness [Shashikumar et al., 2021].

We may also consider using the conditional computing-based method, such as mixture-of-experts [Jacobs et al., 1991], with the state-of-the-art vision encoder such as vision transformer [Riquelme et al., 2021], for pre-training that yields better data representations. Developing unified, massive pre-trained models through a multimodal data and/or multitask setup, such as the ExT5 model that uses multitask objectives for self-supervised learning, is also a promising approach for model generalization [Bugliarello et al., 2021, Hendricks et al., 2021, Aribandi et al., 2022]. However, we should also remember that the model might be eventually deployed on edge devices instead of a centralized, cloud computing center. Other smaller machine learning models can be further investigated instead of pursuing massive pre-training, which is usually less efficient. For example, more efficient fine-tuning methods that only updates a small number of parameters using an adapter instead of full fine-tuning for parameters [He et al., 2022], sparse training, network pruning [Liu et al., 2019c], quantization [Polino et al., 2018], and model distillation [Hinton et al., 2014], can all be effective

ways to reduce the model scale potentially.

Algorithmic bias and fairness

As an emerging problem, algorithmic bias and fairness have also become a critical issue for machine learning modeling in healthcare, not only for the reason of ethical concerns such as discrimination based on gender, race, or political beliefs, but also for potential consequences such as misdiagnosis, health disparities, and mistrust [Mehrabi et al., 2021, Manrai et al., 2016, Boag et al., 2018]. For instance, a biased skin disease diagnosis model may give an unfair prediction because of dataset selection/sampling bias, especially due to skin tone and race factors [Kinyanjui et al., 2020]. To mitigate such issues, we can approach them from two directions. From the data perspective, we may develop suitable metrics to approximate the quantity of bias in the data, then use the metrics as features for downstream tasks [Boag et al., 2018]. We can also collect more data for minority representation to reduce the bias from datasets. Yet we should also remember the potential harmful bias and toxicity in the training data. Overfitting these data can also be a critical issue while using the trained model. From the machine learning technique perspective, we may adopt fair methods that consider some specific data attributes, such as race, sex, and disability, to satisfy the definitions of fairness [Mehrabi et al., 2021]. Furthermore, we may consider both fairness and model robustness simultaneously with a causal framework in order to transfer model fairness and develop much more reliable machine learning models for future real-world deployment [Schrouff et al., 2022].

Toward Real-world Deployment

Finally, even though our experimental setups consider the real-world setting, all these studies are still at the research project stage and have not yet reached clinical practice. As model developers, we first need to collaborate with clinicians and domain experts to reach a consensus of actual problems that have real clinical needs closely [Saleh et al., 2020]. With clinicians' insights, we may overcome common barriers to real-world deployment, such as inadequate attention to health system needs, logistical constraints, and end-user acceptability.

We should also keep in mind considering choosing suitable evaluation metrics during

model development. Researchers usually use the area under the receiver operator curve (AUC-ROC) for binary prediction tasks. Yet, the metric has unclear meaning from a clinical perspective [Pinker, 2018], which can lead to misinterpretation. However, it is better to convince the real-world users to adopt machine learning methods by including metrics such as F1 score, precision, recall, or other clinically understandable measurements that are widely used in the medical world like sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV).

Once the methods are developed, clinical validation, regulation, real-world deployment, workflow integration, and monitoring of the real clinical effect are all essential in order to ensure that the methods actually bring expected benefits to the target audience [Liu et al., 2019b, Chen et al., 2019c]. Admittedly, there is a substantial critical gap while bridging machine learning and the clinical world due to the lack of well-developed mechanisms to implement effective machine learning-based solutions. To achieve this goal, we should seriously think about having clinical artificial intelligence departments in hospitals [Cosgriff et al., 2020], connecting different stakeholders such as research institutes, hospitals, government, and maybe big tech companies, and adopting implementation science, which translates knowledge gained through methodology innovations into improvements in clinical care delivery. For example, the "Ecosystem as a Service" (EaaS) approach builds a collaboration network to bridge researchers and clinicians and support translation, innovation, and prospective evaluations of machine learning models in the real-world [Ishii-Rousseau et al., 2022]. The Massachusetts Institute of Technology Critical Data (MIT-CD) consortium* is an EaaS initiative that offers a sustainable and cost-effective network to accelerate the real-world deployment of machine learning for medicine.

Improving medicine requires changes from different perspectives. This thesis tries to make contributions to helping clinical decision making by demonstrating different machine learning-based approaches to tackle medical data, and providing insights and caveats for using them in practice. We hope that the insights from the thesis can be stepping stones for future research and clinical adoption and eventually make a positive real-world impact.

*https://criticaldata.mit.edu/consortium/

Bibliography

Michael David Abràmoff, Yiyue Lou, Ali Erginay, Warren Clarida, Ryan Amelon, James C Folk, and Meindert Niemeijer. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Investigative Ophthalmology & Visual Science*, 57(13):5200–5206, 2016.

Ehab A AlBadawy, Ashirbani Saha, and Maciej A Mazurowski. Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing. *Medical Physics*, 45(3):1150–1158, 2018.

Hanan Aldarmaki and Mona Diab. Context-aware cross-lingual mapping. *NAACL-HLT*, 2019.

Emily Alsentzer and Anne Kim. Extractive summarization of ehr discharge notes. *arXiv*, 2018.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *Clinical Natural Language Processing (Clinical NLP) Workshop at NAACL*, 2019.

Yaron Anavi, Ilya Kogan, Elad Gelbart, Ofer Geva, and Hayit Greenspan. Visualizing and enhancing a deep learning framework using patients age and gender for chest X-ray image retrieval. *SPIE Medical Imaging*, 2016.

Vamsi Aribandi, Yi Tay, Tal Schuster, Jinfeng Rao, Huaixiu Steven Zheng, Sanket Vaibhav Mehta, Honglei Zhuang, Vinh Q Tran, Dara Bahri, Jianmo Ni, et al. Ext5: Towards extreme multi-task scaling for transfer learning. *ICLR*, 2022.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. *AAAI*, 2018a.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. *ACL*, 2018b.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. Unsupervised neural machine translation. *ICLR*, 2018c.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *ICLR*, 2014.

Pierre Baldi. Autoencoders, unsupervised learning, and deep architectures. *Proceedings of ICML workshop on Unsupervised and Transfer Learning*, pages 37–49, 2012.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE TPAMI*, 41(2):423–443, 2018.

Antonio Valerio Miceli Barone. Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders. *RepL4NLP*, 2016.

Inci M Baytas, Cao Xiao, Xi Zhang, Fei Wang, Anil K Jain, and Jiayu Zhou. Patient subtyping via time-aware lstm networks. *KDD*, 2017.

Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22): 2199–2210, 2017.

Yoshua Bengio. Deep learning of representations for unsupervised and transfer learning. *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, pages 17–36, 2012.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *TPAMI*, 35(8):1798–1828, 2013.

Or Biran, Samuel Brody, and Noémie Elhadad. Putting it simply: A context-aware approach to lexical simplification. *ACL*, 2011.

Willie Boag, Harini Suresh, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Racial disparities and mistrust in end-of-life care. *MLHC*, 2018.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *TACL*, 2017.

Akhilan Boopathy, Tsui-Wei Weng, Pin-Yu Chen, Sijia Liu, and Luca Daniel. CNN-cert: An efficient framework for certifying robustness of convolutional neural networks. *AAAI*, 2019.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. *ICLR*, 2019.

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a "Siamese" time delay neural network. *NIPS*, 1994.

Tom Brosch, Roger Tam, Alzheimer's Disease Neuroimaging Initiative, et al. Manifold learning of brain MRIs by deep learning. *MICCAI*, 2013.

Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language BERTs. *TACL*, 9:978–994, 2021.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. *2017 IEEE Symposium on Security and Privacy*, 2017.

Rich Caruana, Shumeet Baluja, and Tom Mitchell. Using the future to" sort out" the present: Rankprop and multitask learning for medical risk evaluation. *NIPS*, 1996.

Richard A Caruana. Multitask connectionist learning. *Connectionist Models Summer School*, 1993.

Robert Challen, Joshua Denny, Martin Pitt, Luke Gompels, Tom Edwards, and Krasimira Tsaneva-Atanasova. Artificial intelligence, bias and clinical safety. *BMJ Quality & Safety*, 28(3):231–237, 2019.

Dustin Charles, Meghan Gabriel, and Michael F Furukawa. Adoption of electronic health record systems among us non-federal acute care hospitals: 2008-2012. *ONC Data Brief*, 9:1–9, 2013.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

Zhengping Che, David Kale, Wenzhe Li, Mohammad Taha Bahadori, and Yan Liu. Deep computational phenotyping. *KDD*, 2015.

Jinying Chen, Abhyuday Jagannatha, Samah Fodeh, and Hong Yu. Ranking medical terms to support expansion of lay language resources for patient comprehension of electronic health record notes: Adapted distant supervision approach. *JMIR Medical Informatics*, 5 (4):e42, 2017.

Jinying Chen, Emily Druhl, Balaji Polepalli Ramesh, Thomas Houston, Cynthia Brandt, Donna Zulman, Varsha Vimalananda, Samir Malkani, and Hong Yu. A natural language processing system that links medical terms in electronic health record notes to lay definitions: System development using physician reviews. *JMIR*, 20(1):e26, 2018.

Mickaël Chen, Thierry Artières, and Ludovic Denoyer. Unsupervised object segmentation by redrawing. *NeurIPS*, 2019a.

Po-Hsuan Cameron Chen, Krishna Gadepalli, Robert MacDonald, Yun Liu, Shiro Kadowaki, Kunal Nagpal, Timo Kohlberger, Jeffrey Dean, Greg S Corrado, Jason D Hipp, et al. An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis. *Nature Medicine*, 25(9):1453, 2019b.

Po-Hsuan Cameron Chen, Yun Liu, and Lily Peng. How to develop machine learning models for healthcare. *Nature Materials*, 18(5):410, 2019c.

Richard J Chen, Ming Y Lu, Jingwen Wang, Drew FK Williamson, Scott J Rodig, Neal I Lindeman, and Faisal Mahmood. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Transactions on Medical Imaging*, 2020a.

Richard J Chen, Ming Y Lu, Wei-Hung Weng, Tiffany Y Chen, Drew FK Williamson, Trevor Manz, Maha Shady, and Faisal Mahmood. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. *ICCV*, 2021a.

Richard J Chen, Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Jana Lipkova, Muhammad Shaban, Maha Shady, Mane Williams, Bumjin Joo, Zahra Noor, et al. Pan-cancer integrative histology-genomic analysis via interpretable multimodal deep learning. *arXiv*, 2021b.

Ting Chen, Xiaohua Zhai, Marvin Ritter, Mario Lucic, and Neil Houlsby. Self-supervised GANs via auxiliary rotation loss. *CVPR*, 2019d.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *ICML*, 2020b.

Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. *ICLR*, 2019e.

Xinlei Chen and Kaiming He. Exploring simple Siamese representation learning. *CVPR*, 2021.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv*, 2020c.

Jie-Zhi Cheng, Dong Ni, Yi-Hong Chou, Jing Qin, Chui-Mei Tiu, Yeun-Chung Chang, Chiun-Sheng Huang, Dinggang Shen, and Chung-Ming Chen. Computer-aided diagnosis with deep learning architecture: applications to breast lesions in us images and pulmonary nodules in ct scans. *Scientific Reports*, 6:24454, 2016.

Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor AI: Predicting clinical events via recurrent neural networks. *MLHC*, 2016a.

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. *NIPS*, 2016b.

Youngduck Choi, Chill Yi-I Chiu, and David Sontag. Learning low-dimensional representations of medical concepts. *AMIA CRI*, 2016:41, 2016c.

Yu-An Chung and Wei-Hung Weng. Learning deep representations of medical images using siamese cnns with application to content-based image retrieval. *NIPS Workshop on Machine Learning for Health (ML4H)*, 2017.

Yu-An Chung, Wei-Hung Weng, Schrasing Tong, and James Glass. Unsupervised cross-modal alignment of speech and text embedding spaces. *NeurIPS*, 2018.

Yu-An Chung, Wei-Hung Weng, Schrasing Tong, and James Glass. Towards unsupervised speech-to-text translation. *ICASSP*, 2019.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. *ICLR*, 2018.

Luigi Coppola, Alessandra Cianflone, Anna Maria Grimaldi, Mariarosaria Incoronato, Paolo Bevilacqua, Francesco Messina, Simona Baselice, Andrea Soricelli, Peppino Mirabelli, and Marco Salvatore. Biobanking in health care: evolution and future directions. *J Transl Med*, 2019.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3): 273–297, 1995.

Christopher V Cosgriff, David J Stone, Gary Weissman, Romain Pirracchio, and Leo Anthony Celi. The clinical artificial intelligence department: a prerequisite for success. *BMJ Health & Care Informatics*, 27(1), 2020.

Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre L Moreira, Narges Razavian, and Aristotelis Tsirigos. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature Medicine*, 24(10):1559, 2018.

Angel Alfonso Cruz-Roa, John Edison Arevalo Ovalle, Anant Madabhushi, and Fabio Augusto González Osorio. A deep learning architecture for image representation, visual interpretability and automated basal-cell carcinoma cancer detection. *MICCAI*, 2013.

Sandeep Dalal, Vadiraj Hombal, Wei-Hung Weng, Gabe Mankovich, Thusitha Mabotuwana, Christopher S Hall, Joseph Fuller, Bruce E Lehnert, and Martin L Gunn. Determining follow-up imaging study using radiology reports. *Journal of Digital Imaging*, 33(1):121–130, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. *CVPR*, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT*, 2019.

Ashis Kumar Dhara, Sudipta Mukhopadhyay, Anirvan Dutta, Mandeep Garg, and Niranjan Khandelwal. Content-based image retrieval system for pulmonary nodules: Assisting radiologists in self-learning and diagnosis of lung cancer. *Journal of Digital Imaging*, 30(1): 63–77, 2017.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. Improving zero-shot learning by mitigating the hubness problem. *ICLR Workshop*, 2015.

Jeff Donahue and Karen Simonyan. Large scale adversarial representation learning. *NeurIPS*, 2019.

Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *ICLR*, 2017.

Finale Doshi-Velez, Yaorong Ge, and Isaac Kohane. Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis. *Pediatrics*, 133(1):e54–e63, 2014.

Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. *ACL-IJCNLP*, 2015.

Elad Eban, Mariano Schain, Alan Mackey, Ariel Gordon, Ryan Rifkin, and Gal Elidan. Scalable learning of non-decomposable objectives. *Artificial Intelligence and Statistics*, pages 832–840, 2017.

Hamid Eghbal-zadeh, Khaled Koutini, Verena Haunschmid, Paul Primus, Michal Lewandowski, Werner Zellinger, and Gerhard Widmer. Adversarial robustness in data augmentation. *Towards Trustworthy ML: Rethinking Security and Privacy for ML ICLR 2020 Workshop*, 2020.

Noemie Elhadad and Komal Sutaria. Mining a lexicon of technical terms and lay equivalents. *BioNLP*, 2007.

Logan Engstrom, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. A rotation and a translation suffice: Fooling CNNs with simple transformations. *arXiv*, 2017.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. Why does unsupervised pre-training help deep learning? *JMLR*, 11 (Feb):625–660, 2010.

Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017.

Christian Etmann, Sebastian Lunz, Peter Maass, and Carola-Bibiane Schönlieb. On the connection between adversarial robustness and saliency map interpretability. *ICML*, 2019.

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. A comparison of features for automatic readability assessment. *COLING*, 2010.

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

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *ICML*, 2017.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *EMNLP*, 2016.

Joseph Futoma, Jonathan Morris, and Joseph Lucas. A comparison of models for predicting early hospital readmissions. *JBI*, 56:229–238, 2015.

Jevgenij Gamper and Nasir Rajpoot. Multiple instance captioning: Learning representations from histopathology textbooks and articles. *CVPR*, 2021.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 17(59):1–35, 2016.

Romane Gauriau, Christopher Bridge, Lina Chen, Felipe Kitamura, Neil A Tenenholtz, John E Kirsch, Katherine P Andriole, Mark H Michalski, and Bernardo C Bizzo. Using DICOM metadata for radiological image series categorization: a feasibility study on large clinical brain MRI datasets. *Journal of Digital Imaging*, pages 1–16.

T. Gehr, M. Mirman, D. Drachsler-Cohen, P. Tsankov, S. Chaudhuri, and M. Vechev. Ai2: Safety and robustness certification of neural networks with abstract interpretation. *IEEE Symposium on Security & Privacy (SP)*, pages 948–963, 2018.

Marzyeh Ghassemi, Tristan Naumann, Finale Doshi-Velez, Nicole Brimmer, Rohit Joshi, Anna Rumshisky, and Peter Szolovits. Unfolding physiological state: Mortality modelling in intensive care units. *KDD*, 2014.

Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L Beam, Irene Y Chen, and Rajesh Ranganath. A review of challenges and opportunities in machine learning for health. *AMIA Summits on Translational Science Proceedings*, 2020:191, 2020.

Amirata Ghorbani, Vivek Natarajan, David Coz, and Yuan Liu. DermGAN: Synthetic generation of clinical skin images with pathology. *NeurIPS Workshop on Machine Learning for Health (ML4H)*, 2020.

Traber Davis Giardina and Hardeep Singh. Should patients get direct access to their laboratory test results?: An answer with many questions. *JAMA*, 306(22):2502–2503, 2011.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *ICLR*, 2018.

Uma M Girkar, Ryo Uchimido, Li-wei H Lehman, Peter Szolovits, Leo Celi, and Wei-Hung Weng. Predicting blood pressure response to fluid bolus therapy using attention-based neural networks for clinical interpretability. *NeurIPS Workshop on Machine Learning for Health (ML4H)*, 2018.

Jen J Gong, Tristan Naumann, Peter Szolovits, and John V Guttag. Predicting clinical outcomes across changing electronic health record systems. *KDD*, 2017.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NIPS*, 2014.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *ICLR*, 2015.

Sven Gowal, Krishnamurthy Dj Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. Scalable verified training for provably robust image classification. *ICCV*, 2019.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *NeurIPS*, 2020.

Mark Oliver Gueld, Michael Kohnen, Daniel Keysers, Henning Schubert, Berthold B Wein, Joerg Bredno, and Thomas Martin Lehmann. Quality of DICOM header information for image categorization. *Medical Imaging 2002: PACS and Integrated Medical Information Systems: Design and Evaluation*, 4685:280–287, 2002.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *ICLR*, 2021.

Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22):2402–2410, 2016.

Huazhang Guo, Joe Birsa, Navid Farahani, Douglas J Hartman, Anthony Piccoli, Matthew O'Leary, Jeffrey McHugh, Mark Nyman, Curtis Stratman, Vanja Kvarnstrom, et al. Digital pathology and anatomic pathology laboratory information system integration to support digital pathology sign-out. *J Pathol Inform*, 7, 2016.

Lin Lawrence Guo, Stephen R Pfohl, Jason Fries, Jose Posada, Scott Lanyon Fleming, Catherine Aftandilian, Nigam Shah, and Lillian Sung. Systematic review of approaches to preserve machine learning performance in the presence of temporal dataset shift in clinical medicine. *Applied Clinical Informatics*, 12(04):808–815, 2021.

Lin Lawrence Guo, Stephen R Pfohl, Jason Fries, Alistair EW Johnson, Jose Posada, Catherine Aftandilian, Nigam Shah, and Lillian Sung. Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine. *Scientific reports*, 12(1):1–10, 2022.

Yanrong Guo, Yaozong Gao, and Dinggang Shen. Deformable MR prostate segmentation via deep feature learning and sparse patch matching. 2017.

Yunhui Guo, Noel C Codella, Leonid Karlinsky, James V Codella, John R Smith, Kate Saenko, Tajana Rosing, and Rogerio Feris. A broader study of cross-domain few-shot learning. *ECCV*, 2020.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. *CVPR*, 2006.

Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron C. Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. Brain tumor segmentation with deep neural networks. *Medical Image Analysis*, 35:18–31, 2017.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. *ICLR*, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, 2016.

Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking ImageNet pre-training. *ICCV*, 2019.

Kenneth Heafield. Kenlm: Faster and smaller language model queries. *WMT*, 2011.

Lisa Anne Hendricks, John Mellor, Rosalia Schneider, Jean-Baptiste Alayrac, and Aida Nematzadeh. Decoupling the role of data, attention, and losses in multimodal transformers. *TACL*, 9:570–585, 2021.

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *NIPS Deep Learning Workshop*, 2014.

Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

Steven Horng, David A Sontag, Yoni Halpern, Yacine Jernite, Nathan I Shapiro, and Larry A Nathanson. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. *PLoS one*, 12(4):e0174708, 2017.

Tzu-Ming Harry Hsu, Wei-Hung Weng, Willie Boag, Matthew McDermott, and Peter Szolovits. Unsupervised multimodal representation learning across medical images and reports. *NeurIPS Workshop on Machine Learning for Health (ML4H)*, 2018.

Szu-Yen Hu, Shuhang Wang, Wei-Hung Weng, JingChao Wang, XiaoHong Wang, Arinc Ozturk, Quan Li, Viksit Kumar, and Anthony E Samir. Self-supervised pretraining with dicom metadata in ultrasound imaging. *MLHC*, 2020a.

Szu-Yeu Hu, Andrew Beers, Ken Chang, Kathi Höbel, J Peter Campbell, Deniz Erdogumus, Stratis Ioannidis, Jennifer Dy, Michael F Chiang, Jayashree Kalpathy-Cramer, et al. Deep feature transfer between localization and segmentation tasks. *arXiv*, 2018.

Szu-Yeu Hu, Wei-Hung Weng, Shao-Lun Lu, Yueh-Hung Cheng, Furen Xiao, Feng-Ming Hsu, and Jen-Tang Lu. Multimodal volume-aware detection and segmentation for brain metastases radiosurgery. *Workshop on Artificial Intelligence in Radiation Therapy*, 2019.

Szu-Yeu Hu, Shuhang Wang, Wei-Hung Weng, JingChao Wang, XiaoHong Wang, Arinc Ozturk, Qian Li, Viksit Kumar, and Anthony E Samir. Weakly supervised context encoder using dicom metadata in ultrasound imaging. *ICLR AI4AH*, 2020b.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. ClinicalBERT: Modeling clinical notes and predicting hospital readmission. *CHIL*, 2020.

Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liou, Yen-Yu Lin, and Ming-Hsuan Yang. Adversarial learning for semi-supervised semantic segmentation. *BMVC*, 2018.

Eui Jin Hwang, Ju Gang Nam, Woo Hyeon Lim, Sae Jin Park, Yun Soo Jeong, Ji Hee Kang, Eun Kyoung Hong, Taek Min Kim, Jin Mo Goo, Sunggyun Park, et al. Deep learning for chest radiograph diagnosis in the emergency department. *Radiology*, 293(3):573–580, 2019.

Vladimir Iglovikov and Alexey Shvets. Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation. *arXiv*, 2018.

Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating deep network training by reducing internal covariate shift. *ICML*, 2015.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *AAAI*, 2019.

Julian Euma Ishii-Rousseau, Shion Seino, Daniel K Ebner, Maryam Vareth, Ming Jack Po, and Leo Anthony Celi. The "Ecosystem as a Service (EaaS)" approach to advance clinical artificial intelligence (cAI). *PLoS Digital Health*, 1(2):e0000011, 2022.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

Joseph D Janizek, Gabriel Erion, Alex J DeGrave, and Su-In Lee. An adversarial approach for the robust classification of pneumonia from chest radiographs. *CHIL*, 2020.

A. Jenitta and R. Samson Ravindran. Image retrieval based on local mesh vector co-occurrence pattern for medical diagnosis from mri brain images. *Journal of Medical Systems*, 41(10):157, 2017.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, Lisa Orii, and Peter Szolovits. Hooks in the headline: Learning to generate headlines with controlled styles. *ACL*, 2020.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205, 2022a.

Di Jin, Elena Sergeeva, Wei-Hung Weng, Geeticka Chauhan, and Peter Szolovits. Explainable deep learning in healthcare: A methodological survey from an attribution view. *WIREs Mechanisms of Disease*, page e1548, 2022b.

Zhijing Jin, Di Jin, Jonas Mueller, Nicholas Matthews, and Enrico Santus. IMaT: Unsupervised text attribute transfer via iterative matching and translation. *arXiv*, 2019.

Alistair EW Johnson, Tom Pollard, Lu Shen, Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

Justin Johnson, Lamberto Ballan, and Li Fei-Fei. Love thy neighbors: Image annotation by exploiting image metadata. *ICCV*, 2015.

Justin M Johnson and Taghi M Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54, 2019.

Rohit Joshi and Peter Szolovits. Prognostic physiology: modeling patient severity in intensive care units using radial domain folding. *AMIA*, 2012.

Sasikiran Kandula, Dorothy Curtis, and Qing Zeng-Treitler. A semantic and syntactic text simplification tool for health content. *AMIA*, 2010.

Guy Katz, Clark Barrett, David L Dill, Kyle Julian, and Mykel J Kochenderfer. Reluplex: An efficient SMT solver for verifying deep neural networks. *International Conference on Computer Aided Verification*, 2017.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. *CVPR*, 2018.

Alla Keselman, Catherine Arnott Smith, Guy Divita, Hyeoneui Kim, Allen Browne, Gondy Leroy, and Qing Zeng-Treitler. Consumer health concepts that do not map to the umls: Where do they fit? *JAMIA*, 15(4):496–505, 2008.

Kaung Khin, Philipp Burckhardt, and Rema Padman. A deep learning architecture for de-identification of patient notes: Implementation and evaluation. *arXiv preprint arXiv:1810.01570*, 2018.

Jin-Hwa Kim, Kyoung Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. *ICLR*, 2017.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. *NeurIPS*, 2018.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2014.

Newton M Kinyanjui, Timothy Odonga, Celia Cintas, Noel CF Codella, Rameswar Panda, Prasanna Sattigeri, and Kush R Varshney. Fairness of classifiers across skin tones in dermatology. *MICCAI*, 2020.

Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. *ICML Deep Learning Workshop*, 2, 2015.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. *NAACL-HLT*, 2003.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. *ACL Interactive Poster and Demonstration Sessions*, 2007.

Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. *CVPR*, 2019.

J Zico Kolter and Eric Wong. Provable defenses against adversarial examples via the convex outer adversarial polytope. *ICML*, 2018.

Matthieu Komorowski, Leo A Celi, Omar Badawi, Anthony C Gordon, and A Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11):1716, 2018.

Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? *CVPR*, 2019.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *NIPS*, 2012.

Ashnil Kumar, Jinman Kim, Weidong Cai, Michael Fulham, and Dagan Feng. Content-based medical image retrieval: a survey of applications to multidimensional and multimodality data. *Journal of Digital Imaging*, 26(6):1025–1039, 2013.

Vinay Kumar, Abul K Abbas, Nelson Fausto, and Jon C Aster. *Robbins and Cotran pathologic basis of disease*. Elsevier/Saunders, 2014.

Brenden Lake, Ruslan Salakhutdinov, Jason Gross, and Joshua Tenenbaum. One shot learning of simple visual concepts. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 33(33), 2011.

John Lalor, Hao Wu, Li Chen, Kathleen Mazor, and Hong Yu. Comprehenotes, an instrument to assess patient reading comprehension of electronic health record notes: Development and validation. *JMIR*, 20(4):e139, 2018.

Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. *ICLR*, 2018a.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Phrase-based & neural unsupervised machine translation. *EMNLP*, 2018b.

Patrice Latinne, Marco Saerens, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities may significantly improve classification accuracy: evidence from a multi-class problem in remote sensing. *ICML*, 2001.

Duyen NT Le, Hieu X Le, Lua T Ngo, and Hoan T Ngo. Transfer learning with class-weighted and focal loss function for automatic skin cancer classification. *arXiv*, 2020.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 2019.

Eric P Lehman, Rahul G Krishnan, Xiaopeng Zhao, Roger G Mark, and H Lehman Li-wei. Representation learning approaches to detect false arrhythmia alarms from ecg dynamics. *MLHC*, 2018.

Xiaomeng Li, Lequan Yu, Chi-Wing Fu, and Pheng-Ann Heng. Difficulty-aware meta-learning for rare disease diagnosis. *MICCAI*, 2020.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *ICCV*, 2017.

Melissa Linkert, Curtis T Rueden, Chris Allan, Jean-Marie Burel, Will Moore, Andrew Patterson, Brian Loranger, Josh Moore, Carlos Neves, Donald MacDonald, et al. Metadata matters: access to image data in the real world. *JCB*, 189(5):777–782, 2010.

Zachary C Lipton, David C Kale, Charles Elkan, and Randall Wetzel. Learning to diagnose with lstm recurrent neural networks. *ICLR*, 2016.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

Chen Liu, Yanwei Fu, Chengming Xu, Siqian Yang, Jilin Li, Chengjie Wang, and Li Zhang. Learning a few-shot embedding model with contrastive learning. *AAAI*, 2021.

Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. Clinically accurate chest X-ray report generation. *MLHC*, 2019a.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. K-BERT: Enabling language representation with knowledge graph. *AAAI*, 2020a.

Xinran Liu, Hamid R. Tizhoosh, and Jonathan Kofman. Generating binary tags for fast medical image retrieval based on convolutional nets and radon transform. *IJCNN*, 2016.

Yuan Liu, Ayush Jain, Clara Eng, David H Way, Kang Lee, Peggy Bui, Kimberly Kanada, Guilherme de Oliveira Marinho, Jessica Gallegos, Sara Gabriele, et al. A deep learning system for differential diagnosis of skin diseases. *Nature Medicine*, pages 1–9, 2020b.

Yun Liu, Krishna Gadepalli, Mohammad Norouzi, George E Dahl, Timo Kohlberger, Aleksey Boyko, Subhashini Venugopalan, Aleksei Timofeev, Philip Q Nelson, Greg S Corrado, et al. Detecting cancer metastases on gigapixel pathology images. *arXiv*, 2017.

Yun Liu, Po-Hsuan Cameron Chen, Jonathan Krause, and Lily Peng. How to read articles that use machine learning: users' guides to the medical literature. *JAMA*, 322(18):1806–1816, 2019b.

Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. Rethinking the value of network pruning. *ICLR*, 2019c.

Gilles Louppe, Michael Kagan, and Kyle Cranmer. Learning to pivot with adversarial networks. *NIPS*, 2017.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *NIPS*, 2017.

Mario Lučić, Marvin Ritter, Michael Tschannen, Xiaohua Zhai, Olivier Frederic Bachem, and Sylvain Gelly. High-fidelity image generation with fewer labels. *ICML*, 2019.

Ruibin Ma, Po-Hsuan Cameron Chen, Gang Li, Wei-Hung Weng, Angela Lin, Krishna Gadepalli, and Yuannan Cai. Human-centric metric for accelerating pathology reports annotation. *NeurIPS Workshop on Machine Learning for Health (ML4H)*, 2019.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *JMLR*, 9(Nov): 2579–2605, 2008.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *ICLR*, 2018.

Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. *ECCV*, 2018.

Kushagra Mahajan, Monika Sharma, and Lovekesh Vig. Meta-DermDiagnosis: Few-shot skin disease identification using meta-learning. *CVPR Workshops*, 2020.

Gary Malet, Felix Munoz, Richard Appleyard, and William Hersh. A model for enhancing internet medical document retrieval with "medical core metadata". *JAMIA*, 6(2):163–172, 1999.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. *ACL System Demonstrations*, 2014.

Arjun K Manrai, Birgit H Funke, Heidi L Rehm, Morten S Olesen, Bradley A Maron, Peter Szolovits, David M Margulies, Joseph Loscalzo, and Isaac S Kohane. Genetic misdiagnoses and the potential for health disparities. *New England Journal of Medicine*, 375(7):655–665, 2016.

Matthew BA McDermott, Tzu Ming Harry Hsu, Wei-Hung Weng, Marzyeh Ghassemi, and Peter Szolovits. CheXpert++: Approximating the chexpert labeler for speed, differentiability, and probabilistic output. *MLHC*, 2020.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54 (6):1–35, 2021.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *NIPS*, 2013.

Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific Reports*, 6:26094, 2016.

Matthew Mirman, Timon Gehr, and Martin Vechev. Differentiable abstract interpretation for provably robust neural networks. *ICML*, 2018.

Takeru Miyato and Masanori Koyama. cGANs with projection discriminator. *ICLR*, 2018.

Pooya Mobadersany, Safoora Yousefi, Mohamed Amgad, David A Gutman, Jill S Barnholtz-Sloan, José E Velázquez Vega, Daniel J Brat, and Lee AD Cooper. Predicting cancer outcomes from histology and genomics using convolutional networks. *PNAS*, 115(13): E2970–E2979, 2018.

Jeet Mohapatra, Tsui-Wei Weng, Pin-Yu Chen, Sijia Liu, and Luca Daniel. Towards verifying robustness of neural networks against a family of semantic perturbations. *CVPR*, 2020.

Mehdi Moradi, Yufan Guo, Yaniv Gur, Mohammadreza Negahdar, and Tanveer Syeda-Mahmood. A cross-modality neural network transform for semi-automatic medical image annotation. *MICCAI*, 2016.

Henning Müller, Nicolas Michoux, David Bandon, and Antoine Geissbuhler. A review of content-based image retrieval systems in medical applications—clinical benefits and future directions. *International Journal of Medical Informatics*, 73(1):1–23, 2004.

Kunal Nagpal, Davis Foote, Yun Liu, Po-Hsuan Cameron Chen, Ellery Wulczyn, Fraser Tan, Niels Olson, Jenny L Smith, Arash Mohtashamian, James H Wren, et al. Development and validation of a deep learning algorithm for improving gleason scoring of prostate cancer. *npj Digital Medicine*, 2(1):48, 2019.

Bret Nestor, Matthew BA McDermott, Willie Boag, Gabriela Berner, Tristan Naumann, Michael C Hughes, Anna Goldenberg, and Marzyeh Ghassemi. Feature robustness in non-stationary health records: caveats to deployable model performance in common clinical machine learning tasks. *MLHC*, 2019.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. *ICML*, 2011.

Dong Nie, Roger Trullo, Jun Lian, Caroline Petitjean, Su Ruan, Qian Wang, and Dinggang Shen. Medical image synthesis with context-aware generative adversarial networks. *MICCAI*, 2017.

Nozomi Nori, Hisashi Kashima, Kazuto Yamashita, Hiroshi Ikai, and Yuichi Imanaka. Simultaneous modeling of multiple diseases for mortality prediction in acute hospital care. *KDD*, 2015.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. *ECCV*, 2016.

Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier GANs. *ICML*, 2017.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *NeurIPS*, 2018.

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. *CVPR*, 2016.

Lina Pedraza, Carlos Vargas, Fabián Narváez, Oscar Durán, Emma Muñoz, and Eduardo Romero. An open access thyroid ultrasound image database. *10th International Symposium on Medical Information Processing and Analysis*, 9287:92870W, 2015.

Yifan Peng, Xiaosong Wang, Le Lu, Mohammadhadi Bagheri, Ronald Summers, and Zhiyong Lu. NegBio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Summits on Translational Science Proceedings*, 2018.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *NAACL-HLT*, 2018.

Edieal Pinker. Reporting accuracy of rare event classifiers. *npj Digital Medicine*, 1(1):1–2, 2018.

Rimma Pivovarov, Adler J Perotte, Edouard Grave, John Angiolillo, Chris H Wiggins, and Noémie Elhadad. Learning probabilistic phenotypes from heterogeneous ehr data. *JBI*, 58:156–165, 2015.

Ramesh Polepalli, Thomas Houston, Cynthia Brandt, Hua Fang, and Hong Yu. Improving patients' electronic health record comprehension with noteaid. *Studies in Health Technology and Informatics*, 192:714–718, 2013.

Antonio Polino, Razvan Pascanu, and Dan Alistarh. Model compression via distillation and quantization. *ICLR*, 2018.

Eduardo HP Pooch, Pedro Ballester, and Rodrigo C Barros. Can we trust deep learning based diagnosis? the impact of domain shift in chest radiograph classification. *International Workshop on Thoracic Image Analysis*, pages 74–83, 2020.

Ryan Poplin, Avinash V Varadarajan, Katy Blumer, Yun Liu, Michael V McConnell, Greg S Corrado, Lily Peng, and Dale R Webster. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering*, 2(3):158, 2018.

Lisa Posch, Maryam Panahiazar, Michel Dumontier, and Olivier Gevaert. Predicting structured metadata from unstructured metadata. *Database*, 2016.

Viraj Prabhu, Anitha Kannan, Murali Ravuri, Manish Chaplain, David Sontag, and Xavier Amatriain. Few-shot learning for dermatological disease diagnosis. *MLHC*, 2019.

Sampo Pyysalo, Filip Ginter, Hans Moen, Tapio Salakoski, and Sophia Ananiadou. *Distributional semantics resources for biomedical text processing*. 2013.

Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. The MIT Press, 2009.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *ICLR*, 2016.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *ICML*, 2021.

Aniruddh Raghu, Matthieu Komorowski, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Continuous state-space models for optimal sepsis treatment-a deep reinforcement learning approach. *MLHC*, 2017.

Aniruddh Raghu, Maithra Raghu, Samy Bengio, and Oriol Vinyals. Rapid learning or feature reuse? towards understanding the effectiveness of MAML. *ICLR*, 2020.

Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. *NeurIPS*, 2019.

Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. *ICLR*, 2018.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. CheXnet: Radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv*, 2017.

Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. *ICLR*, 2017.

Bhanu Pratap Singh Rawat, Wei-Hung Weng, So Yeon Min, Preethi Raghavan, and Peter Szolovits. Entity-enriched neural models for clinical question answering. *ACL BioNLP Workshop*, 2020.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. *KDD*, 2016.

Salah Rifai, Pascal Vincent, Xavier Muller, Xavier Glorot, and Yoshua Bengio. Contractive auto-encoders: Explicit invariance during feature extraction. *ICML*, 2011.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *NeurIPS*, 2021.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *MICCAI*, 2015.

Stephen Ross and Chen-Tan Lin. The effects of promoting patient access to medical records: A review. *JAMIA*, 10(2):129–138, 2003.

Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv*, 2017.

Shems Saleh, William Boag, Lauren Erdman, and Tristan Naumann. Clinical collabsheets: 53 questions to guide a clinical collaboration. *MLHC*, 2020.

Thomas Schlegl, Sebastian M Waldstein, Wolf-Dieter Vogl, Ursula Schmidt-Erfurth, and Georg Langs. Predicting semantic descriptions from medical images with convolutional neural networks. *IPMI*, 2015.

Jessica Schrouff, Natalie Harris, Oluwasanmi Koyejo, Ibrahim Alabdulmohsin, Eva Schnider, Krista Opsahl-Ong, Alex Brown, Subhrajit Roy, Diana Mincu, Christina Chen, et al. Maintaining fairness across distribution shift: do we have viable solutions for real-world applications? *arXiv*, 2022.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. *ACL*, 2016.

Divya Seth, Khatiya Cheldize, Danielle Brown, and Esther E Freeman. Global burden of skin disease: inequities and innovations. *Current Dermatology Reports*, 6(3):204–210, 2017.

Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! *NeurIPS*, 2019.

Ali Shafahi, Parsa Saadatpanah, Chen Zhu, Amin Ghiasi, Christoph Studer, David Jacobs, and Tom Goldstein. Adversarially robust transfer learning. *ICLR*, 2020.

Amit Shah, Sailesh Conjeti, Nassir Navab, and Amin Katouzian. Deeply learnt hashing forests for content based image retrieval in prostate mr images. *SPIE Medical Imaging*, 2016.

Supreeth P Shashikumar, Gabriel Wardi, Atul Malhotra, and Shamim Nemati. Artificial intelligence sepsis prediction algorithm learns to say "i don't know". *npj Digital Medicine*, 4(1):1–9, 2021.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

Hoo-Chang Shin, Le Lu, Lauren Kim, Ari Seff, Jianhua Yao, and Ronald M Summers. Interleaved text/image deep mining on a very large-scale radiology database. *CVPR*, 2015.

Yuqi Si, Jingqi Wang, Hua Xu, and Kirk Roberts. Enhancing clinical concept extraction with contextual embeddings. *Journal of the American Medical Informatics Association*, 26(11):1297–1304, 2019.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015.

Gagandeep Singh, Timon Gehr, Matthew Mirman, Markus Püschel, and Martin Vechev. An abstract domain for certifying neural networks. *POPL*, 2019.

Aman Sinha, Hongseok Namkoong, Riccardo Volpi, and John Duchi. Certifying some distributional robustness with principled adversarial training. *ICLR*, 2018.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *NIPS*, 2017.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. On the limitations of unsupervised bilingual dictionary induction. *ACL*, 2018.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 15(1): 1929–1958, 2014.

R Todd Stephens. Utilizing metadata as a knowledge communication tool. *IPCC*, 2004.

Adarsh Subbaswamy and Suchi Saria. From development to deployment: dataset shift, causality, and shift-stable models in health ai. *Biostatistics*, 21(2):345–352, 2020.

Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015.

Rebecca Sudore, Kristine Yaffe, Suzanne Satterfield, Tamara Harris, Kala Mehta, Eleanor Simonsick, Anne Newman, Caterina Rosano, Ronica Rooks, Susan Rubin, et al. Limited literacy and mortality in the elderly: The health, aging, and body composition study. *JGIM*, 21(8):806–812, 2006.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *JMLR*, 8(5), 2007.

Qinpei Sun, Yuanyuan Yang, Jianyong Sun, Zhiming Yang, and Jianguo Zhang. Using deep learning for content-based medical image retrieval. *SPIE Medical Imaging*, 2017.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. *CVPR*, 2018.

Harini Suresh, Nathan Hunt, Alistair Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Clinical intervention prediction and understanding with deep neural networks. *MLHC*, 2017.

Ilya Sutskever, Oriol Vinyals, and Quoc Le. Sequence to sequence learning with neural networks. *NIPS*, 2014.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *ICLR*, 2014.

Peter Szolovits. *Artificial intelligence in medicine*. Westview Press Boulder, CO, 1982.

Peter Szolovits and Stephen G Pauker. Categorical and probabilistic reasoning in medical diagnosis. *Artificial Intelligence*, 11(1-2):115–144, 1978.

Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? *ECCV*, 2020.

Vincent Tjeng and Russ Tedrake. Verifying neural networks with mixed integer programming. *ICLR*, 2019.

Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44, 2019.

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *ICLR*, 2018.

Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, et al. Meta-Dataset: A dataset of datasets for learning to learn from few examples. *ICLR*, 2020.

Lazaros Tsochatzidis, Konstantinos Zagoris, Nikolaos Arikidis, Anna Karahaliou, Lena Costaridou, and Ioannis Pratikakis. Computer-aided diagnosis of mammographic masses based on a supervised content-based image retrieval approach. *Pattern Recognition*, 71: 106–117, 2017.

Jonathan Uesato, Brendan O'donoghue, Pushmeet Kohli, and Aaron Oord. Adversarial risk and the dangers of evaluating against weak attacks. *ICML*, 2018.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *JAMIA*, 18(5):552–556, 2011.

Hien Van Nguyen, Kevin Zhou, and Raviteja Vemulapalli. Cross-domain synthesis of medical images using efficient location-sensitive deep network. *MICCAI*, 2015.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NIPS*, 2017.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. *ICML*, 2008.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *JMLR*, 11(Dec):3371–3408, 2010.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *NIPS*, 2016.

Vinod Vydiswaran, Qiaozhu Mei, David Hanauer, and Kai Zheng. Mining consumer health vocabulary from community-generated text. *AMIA*, 2014.

Ren Wang, Kaidi Xu, Sijia Liu, Pin-Yu Chen, Tsui-Wei Weng, Chuang Gan, and Meng Wang. On fast adversarial robustness adaptation in model-agnostic meta-learning. *ICLR*, 2021.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *CVPR*, 2017.

Yanshan Wang, Sijia Liu, Naveed Afzal, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Paul Kingsbury, and Hongfang Liu. A comparison of word embeddings for the biomedical natural language processing. *JBI*, 87:12–20, 2018.

Tsui-Wei Weng, Huan Zhang, Hongge Chen, Zhao Song, Cho-Jui Hsieh, Duane Boning, Inderjit S Dhillon, and Luca Daniel. Towards fast computation of certified robustness for ReLU networks. *ICML*, 2018a.

Tsui-Wei Weng, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, Dong Su, Yupeng Gao, Cho-Jui Hsieh, and Luca Daniel. Evaluating the robustness of neural networks: An extreme value theory approach. *ICLR*, 2018b.

Wei-Hung Weng. Machine learning for clinical predictive analytics. *Leveraging Data Science for Global Health*, pages 199–217, 2020.

Wei-Hung Weng and Peter Szolovits. Mapping unparalleled clinical professional and consumer languages with embedding alignment. *KDD MLMH Workshop*, 2018.

Wei-Hung Weng and Peter Szolovits. Representation learning for electronic health records. *arXiv*, 2019.

Wei-Hung Weng and Tsui-Wei Weng. Preserving model robustness for dataset shift in medical imaging. *In preperation*, 2021.

Wei-Hung Weng, Mingwu Gao, Ze He, Susu Yan, and Peter Szolovits. Representation and reinforcement learning for personalized glycemic control in septic patients. *NIPS Workshop on Machine Learning for Health (ML4H)*, 2017a.

Wei-Hung Weng, Kavishwar Wagholikar, Alexa McCray, Peter Szolovits, and Henry Chueh. Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. *BMC MIDM*, 17(1):155, 2017b.

Wei-Hung Weng, Yuannan Cai, Angela Lin, Fraser Tan, and Po-Hsuan Cameron Chen. Multimodal multitask representation learning for pathology biobank metadata prediction. *NeurIPS Workshop on Machine Learning for Health (ML4H)*, 2019a.

Wei-Hung Weng, Yu-An Chung, and Peter Szolovits. Unsupervised clinical language translation. *KDD*, 2019b.

Wei-Hung Weng, Yu-An Chung, and Schrasing Tong. Clinical text summarization with syntax-based negation and semantic concept identification. *arXiv*, 2020a.

Wei-Hung Weng, Jonathan Deaton, Vivek Natarajan, Gamaleldin F Elsayed, and Yuan Liu. Addressing the real-world class imbalance problem in dermatology. *NeurIPS Workshop on Machine Learning for Health (ML4H)*, 2020b.

Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. *ICML*, 2018.

Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. *ICLR*, 2020.

Cao Xiao, Tengfei Ma, Adji B Dieng, David M Blei, and Fei Wang. Readmission prediction via deep contextual embedding of clinical concepts. *PLoS one*, 13(4):e0195024, 2018.

Kai Y. Xiao, Vincent Tjeng, Nur Muhammad Shafiullah, and Aleksander Madry. Training for faster adversarial robustness verification via inducing ReLU stability. *ICLR*, 2019.

Qizhe Xie, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Self-training with noisy student improves ImageNet classification. *CVPR*, 2020.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. Normalized word embedding and orthogonal transform for bilingual word translation. *NAACL-HLT*, 2015.

Chaochao Yan, Jiawen Yao, Ruoyu Li, Zheng Xu, and Junzhou Huang. Weakly supervised deep learning for thoracic disease classification and localization on chest x-rays. *ACM-BCB*, 2018.

Jufeng Yang, Xiaoxiao Sun, Jie Liang, and Paul L Rosin. Clinical skin lesion diagnosis using representations inspired by dermatologist criteria. *CVPR*, 2018.

Ka-Ping Yee, Kirsten Swearingen, Kevin Li, and Marti Hearst. Faceted metadata for image search and browsing. *CHI*, 2003.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. *EMNLP*, 2017.

John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11): e1002683, 2018.

Qing Zeng and Tony Tse. Exploring and developing consumer health vocabularies. *JAMIA*, 13(1):24–29, 2006.

Qing Zeng-Treitler, Sergey Goryachev, Hyeoneui Kim, Alla Keselman, and Douglas Rosendale. Making texts in electronic health records comprehensible to consumers: A prototype translator. *AMIA*, 2007.

Huan Zhang, Tsui-Wei Weng, Pin-Yu Chen, Cho-Jui Hsieh, and Luca Daniel. Efficient neural network robustness certification with general activation functions. *NeurIPS*, 2018.

Henghui Zhu, Ioannis Ch Paschalidis, and Amir Tahmasebi. Clinical concept extraction with contextual word embedding. *NeurIPS Workshop on Machine Learning for Health (ML4H)*, 2018.

Rita Zielstorff. Controlled vocabularies for consumer health. *JBI*, 36(4-5):326–333, 2003.