Spring 5-30-2023

Towards generalizable machine learning models for computer-aided diagnosis in medicine

Yiyang Wang
DePaul University, ianwang7152@gmail.com

TOWARDS GENERALIZABLE MACHINE LEARNING MODELS FOR COMPUTER-AIDED DIAGNOSIS IN MEDICINE

BY

YIYANG WANG

A DISSERTATION SUBMITTED TO THE COLLEGE OF COMPUTING AND DIGITAL MEDIA OF DEPAUL UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY COMPUTER SCIENCE

DEPAUL UNIVERSITY
CHICAGO, ILLINOIS
June 2023

DePaul University
College of Computing and Digital Media

Dissertation Verification

This doctoral dissertation has been read and approved by the dissertation committee below according to the requirements of the Computer and Information Systems PhD program and DePaul University.

Name: Yiyang Wang

Title of dissertation:
Towards Generalizable Machine Learning Models for Computer-Aided Diagnosis in Medicine

Date of Dissertation Defense:
May 30, 2023

Daniela Stan Raicu
Dissertation Advisor*

Jacob Furst
1st Reader

Thiruvarangan Ramaraj
2nd Reader

Samuel G. Armato
3rd Reader

4th Reader (if applicable)

5th Reader (if applicable)

** A copy of this form has been signed, but may only be viewed after submission and approval of FERPA request letter.*

Abstract

Hidden stratification represents a phenomenon in which a training dataset contains unlabeled (hidden) subsets of cases that may affect machine learning model performance. Machine learning models that ignore the hidden stratification phenomenon--despite promising overall performance measured as accuracy and sensitivity--often fail at predicting the low prevalence cases, but those cases remain important. In the medical domain, patients with diseases are often less common than healthy patients, and a misdiagnosis of a patient with a disease can have significant clinical impacts. Therefore, to build a robust and trustworthy CAD system and a reliable treatment effect prediction model, we cannot only pursue machine learning models with high overall accuracy, but we also need to discover any hidden stratification in the data and evaluate the proposing machine learning models with respect to both overall performance and the performance on certain subsets (groups) of the data, such as the 'worst group'.

In this study, I investigated three approaches for data stratification: a novel algorithmic deep learning (DL) approach that learns similarities among cases and two schema completion approaches that utilize domain expert knowledge. I further proposed an innovative way to integrate the discovered latent groups into the loss functions of DL models to allow for better model generalizability under the domain shift scenario caused by the data heterogeneity.

My results on lung nodule Computed Tomography (CT) images and breast cancer histopathology images demonstrate that learning homogeneous groups within heterogeneous data significantly improves the performance of the computer-aided diagnosis (CAD) system, particularly for low-prevalence or worst-performing cases. This study emphasizes the importance of discovering and learning the latent stratification within the data, as it is a critical step towards building ML models that are generalizable and reliable. Ultimately, this discovery can have a profound impact on clinical decision-making, particularly for low-prevalence cases.

Acknowledgement

Firstly, I am deeply grateful to my advisor, Dr. Daniela Stan Raicu, and my co-advisor, Dr. Jacob Furst, for their invaluable encouragement and advice throughout my master's and Ph.D. studies. Their guidance has not only shaped the direction of my research but has also instilled in me a greater sense of confidence. Their unwavering support and belief in my abilities have transformed me into a better researcher and a more resilient individual. I truly appreciate their expertise, dedication, and continuous guidance, which have played a pivotal role in navigating the challenges of my academic journey.

I would also like to express my gratitude to the members of my dissertation committee, Dr. Thiruvarangan Ramaraj and Dr. Samuel Armato III. Their insightful feedback, constructive criticism, and valuable suggestions have greatly contributed to this research work.

Additionally, I want to express my deepest appreciation to my parents for their unconditional love. I am truly fortunate to have parents who have gone above and beyond to provide me with every opportunity for success.

Last but not least, I want to say thank you to all my friends. I am grateful for the countless moments we have shared. Your friendship and support have been invaluable to me.

Table of Contents

List of Figures

List of Tables

CHAPTER 1. Introduction

Hidden data stratification is a significant obstacle to build robust and trustworthy machine learning models. For example, in computer-aided diagnosis (CAD) tasks, despite extraordinary overall model performance reported in the literature, the heterogeneity in the visual appearance of medical images often causes machine learning models fail at predicting certain cases that could have critical clinical impact. In the context of treatment effect prediction, patient heterogeneity can cause contradictory conclusions across different prediction models. Therefore, discovering and addressing the hidden stratification phenomenon are important steps towards building generalizable and reliable machine learning models.

My dissertation work focuses on the following two research questions (RQs) in the context of medical diagnosis and treatment:

RQ1: Are there any latent data representations that can capture semantically meaningful groups with respect to different diagnoses?

RQ2: Is the integration of semantically meaningful groups into machine learning diagnostic models boosting the models' generalization capabilities?

Section 1.1 introduces the hidden or latent stratification concept, why hidden stratification causes model degradation, and briefly describe common techniques to discover hidden stratification. Section 1.2 presents an overview of the Group

Distributionally Robust Optimization (gDRO), a representative model optimization method aiming to mitigate the hidden stratification problem. Section 1.3 summarizes the machine learning and medical imaging contributions of this work.

1.1 Hidden Stratification

Hidden stratification represents a phenomenon in which a training dataset contains unlabeled (hidden) subsets of cases that may affect machine learning model performance [1]. The causes of hidden stratification can be a combination of poor label collection process, low prevalence of a hidden data group, spurious correlation between an object of interest and image background, and subtle discriminative features across hidden subgroups [1]. Figure 1.1 shows four hidden stratification examples from two public available datasets [2, 3].

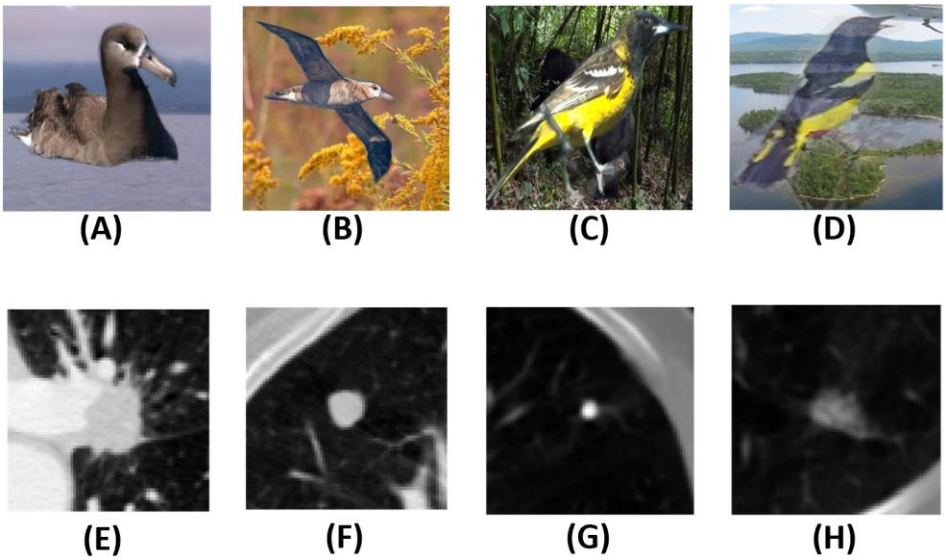

Figure 1.1 An illustration of hidden stratification. Images (A) and (B) were labeled as "Waterbird": (A) Waterbird appearing against a water background (majority of the cases), and (B) Waterbird appearing against a land background (minority cases). Image (C) and (D) were labeled as "Landbird": (C) Landbird at front of land background (majority of the case), and (D) Landbird at front of water background. Images (E) and (F) were labeled as "Malignant" lung nodules: (E) typical malignant nodule with ovoid shape and spiculation (majority of the cases), and (F) malignant nodule that looks like a typical benign nodule that has round shape and a sharp margin (minority of the cases). Images (G) and (H) were labeled as "Benign" lung nodules: (G) typical Benign nodule and (H) Benign nodule that looks like a malignant nodule.

Machine learning models that ignore the hidden stratification phenomenon--despite promising overall performance measured as accuracy and sensitivity--often fail at predicting the low prevalence, but those cases are still important. For example, for a waterbird classification task, most classification models can correctly classify waterbirds against water backgrounds, but these models frequently make mistakes on classifying waterbirds in front of land backgrounds

[2]. In this example, we refer to “waterbirds” and “landbirds” as *superclass* labels, and we call ‘waterbirds in front of water’, ‘waterbirds in front of land’, ‘landbirds in front of land’, and ‘landbirds in front of water’ as *subclass* labels. In most real-world scenarios, annotators only provide coarse-grained super-class labels without detailed subclass labels [1, 4, 5].

In the medical domain, *patients with diseases are often less common than healthy patients, and a misdiagnosis of a patient with a disease can have significant clinical impacts* [1]. Therefore, to build a robust and trustworthy computer-aided diagnosis system (CAD) and a reliable treatment effect prediction model, we cannot only pursue machine learning models with high overall accuracy, but we also need to discover any hidden stratification in the data and evaluate the proposing machine learning models with respect to both overall performance and the performance on certain subsets (groups) of the data, such as the ‘worst group’ performance.

Rayner et al. [1] summarized three methods for discovering hidden stratification: *schema completion*, *error auditing* and *algorithmic measurement*. *Schema completion* requires annotators to provide a more detailed set of sub-class labels. For example, before schema completion, the label of the image (B) in Figure 1.1 is "water bird", but after schema completion, the label is "water bird on land". *Error auditing* asks auditors examine consistently misclassified cases by machine learning models and observe potential hidden patterns in the data. For example, in

the context of chest X-ray image pathology detection, Rayner et al. [1] observed that pneumothorax cases without chest drains are prevalent among false negative cases. Therefore, they further labeled two sub-class labels for the pneumothorax superclass label: "chest drain" and "no chest drain". They observed that the classification performance on the "no chest drain" instances is significantly lower compared to other subclasses. *Algorithm measurement*, which is the focus in this study, uses unsupervised methods, such as clustering, to discover unlabeled subgroups in the data. Both schema completion and error auditing are time consuming, and the auditor's ability to recognize unlabeled sub-class labels or anomalous patterns is the only factor to decide the success of hidden stratification discovery. On the contrary, algorithm measurement reduces human efforts and identifies hidden patterns and subgroups in the data from learned data representations.

In this study, I propose to *build a generalizable hidden stratification discovery framework* that will be validated against two different scenarios with respect to the data acquisition modality: lung nodule computed tomography (CT) scans from the NIH/NCI Lung Image Database Consortium (LIDC) dataset [3] and breast histopathology images [6]. My *first hypothesis (H1)* is:

H1: Clustering-based stratification can reveal semantically meaningful latent groups within the data that correspond to the categorization perceived by the domain experts.

1.2 Distributionally Robust Optimization (DRO)

In the context of building machine learning models, the hidden stratification problem can be viewed as a type of domain shift or out-of-distribution (OOD) generalization problem in which subsets of the data distribution changed in the testing domain. Among all OOD generalization techniques, optimization methods directly guarantee the worst-case performance under distribution shift [7]. Distributionally Robust Optimization (DRO) is one of the optimization methods and its objective function has two terms: a learner and an adversary. The adversary maximizes the expected loss through shifting the test distribution from the training distribution as far as possible while the learner minimizes the adversarial expected loss [8]. Deviated from DRO, Sagawa et al. [2] proposed Group Distributionally Robust Optimization (gDRO) method that optimizes the *worst group* performance instead of the worst case performance and showed that a regularized gDRO method achieves better generalization results compared with DRO result. Sohoni et al. [4] implemented gDRO on multiple real world benchmark datasets and showed that gDRO can significantly reduce the model degradation caused by the hidden stratification. My *second hypothesize (H2)* is:

H2: Integrating stratification information into machine learning models can significantly improve the CAD generalization ability measured by both the overall and worst-group accuracy.

1.3 Contributions

My contributions are:

- It is necessary to address the hidden stratification problem as indicated by the different ERM model performance across stratification groups in a malignancy classification task. For the LIDC dataset, a high degree of overlap between the clustering-based stratification results and malignancy likelihood stratification provided by radiologists indicate that my algorithmic hidden stratification discovery method results are aligned with domain experts' annotations, which is in support of my first hypothesis.
- The integration of the stratification groups into malignancy classification models, and the systematic evaluation of the model performance with various model inputs, deep learning architectures and training strategies show that, under the domain shift scenario caused by the disease heterogeneity, the model performance is boosted by 5% as measured by the worst group accuracy when compared with

model without subgroup learning, which is in support of my second hypothesis 2.

CHAPTER 2. Related Work

This chapter reviews two main categories of approaches: disease subtype discovery (Section 2.1) and domain shift generalization (Section 2.2). As mentioned in Chapter 1, hidden stratification can be viewed as a type of domain shift or out-of-distribution (OOD) generalization problem in which subsets of the data distribution change in the testing domain. Disease subtype discovery techniques reveal the hidden stratification phenomena in medical data, and domain shift generalization methods mitigate model degradation under domain shift.

2.1 Disease Subtype Discovery

In most disease subtype discovery studies, these algorithms are based on clustering models or their variations with different input features or partition methods. The goal of disease subtype discovery is to divide patient populations into distinct and relatively homogeneous subgroups (subtypes) [9]. Since the type of data plays an important role in selecting appropriate disease discovery models, the following review is organized according to different data categories.

2.1.1 Image Data

Several studies have investigated the brain lesion subtype discovery based on magnetic resonance imaging (MRI). Wen et al. [10] first utilized a non-negative

matrix factorization algorithm [11] to extract multi-scale, biologically interpretable features and then implemented ensemble of support vector machines (SVMs) to create a nonlinear polytope that separates healthy and patient group. Each face of the polytope represents one subtype. Four subtypes of Alzheimer's disease (AD) were identified, and it was observed that the clinical characteristics of these subtypes were similar to their neuroanatomical patterns [10]. Ezzati et al. [12] first selected important region of interests (ROIs) utilizing principal factor analysis (PCA) and then conducted latent class analysis (LCA) with ROIs as input. They discovered four amnestic mild cognitive impairment (aMCI) subgroups and characterized them in global atrophy, hippocampus, atrophy and cognitive performance scores. Chen et al. [13] extracted intensity, shape and texture features from segmented tumor regions and leveraged an autoencoder approach to reduce the feature dimensions and to learn the most representative features in latent space. They implemented Gaussian Mixture Model (GMM) clustering and found three brain tumor subtypes that have different tumor size, heterogeneity and elongation. Yang et al. [14] extracted the latent representation from a generative adversarial network (GAN) architecture and implemented a semi-supervised clustering approach to divide AD patients into two subgroups that have different cortical atrophies and focal atrophy patterns.

Researchers also use MRIs for breast lesion subtype discovery. Wu et al. [15] extracted quantitative image features that describe tumor volume, parenchymal

enhancement and tumor surrounding parenchymal enhancement. They implemented consensus clustering, a combination of k-medoids clustering and bootstraps, to partition the breast MRI images into different sub-groups and characterized each subtype with intra-tumor heterogeneity and background parenchymal enhancement (BPE). Fan et al. [16] et al. extracted texture, intensity features and morphological features from breast tumor dynamic contrast-enhanced MRI data and used these features as input to a survival analysis model to divide patients into subgroups with different survival rates.

2.1.2 Genome Data

Schulz et al [17] used the Cancer Genome Atlas (TCGA) project dataset [18] to discover cancer subtypes. They calculated Shapley Additive Explanations (SHAP) value [19] for each instance from a super-class classification task (cancer vs. not cancer), and then inputted SHAP values to an agglomerative clustering model. In this study, they used six cancer tissue's immune model-based subtypes as the ground truth and reported the adjusted mutual information between the subtype ground truth and discovered subtypes from clustering to measure the subtype recovery ability. Using the same TCGA dataset, Arslanturk et al. [20] addressed the data heterogeneous problem through integrating clustering results from qualitative data and quantitative data. For qualitative data, they implemented Partition Around Medoids (PAM) clustering with Jaccard Index as distance

measurement and for quantitative data, they performed K-means clustering with Euclidean distance as distance measurement. They averaged the similarity matrix obtained from each cluster and implemented a new clustering algorithm using the averaged similarity matrix. They discovered four patient subgroups with different survival rates and linked each subgroup with a specific gene biomarker.

Using microRNA expression data, Vasudevan et al. [21] proposed a max-flow/min-cut graph clustering approach to detect four glioblastoma multiforme (GBM) subtypes: mesenchymal, classical, proneural, and neural, and they found that implementing the max-flow/min-cut graph clustering achieves better clustering accuracy when compared with K-means, nonnegtive matrix factorization (NMF), and iCluster. Utilizing a similar gene expression dataset, Anderson et al. [22] embedded the prior subtype knowledge in a graph structure and incorporated it to a deep neural network prediction loss function. They showed that utilizing prior subtype information can reduce the variability of breast cancer subtype predictions measured as an 5% increase of overlapped identified important genes across multiple model training.

2.1.3 Electronic Health Record (EHR) Data

This line of research uses patient electronic health record (HER) data and most studies extract patient representative vectors from time series models such as Long-Short Term Memory (LSTM) and memory networks (MN). Extracted patient

representative vectors are then treated as input to an unsupervised learning model to find patient subgroups.

Using the Parkinson Progression Marker Initiative (PPMI) study data, Zhang et al. [23] extracted embeddings from hidden layers of a LSTM model with sequential demographics, biospecimen, imaging and clinical features as input and then inputted embeddings to a t-Stochastic Neighbor Embedding (t-SNE) algorithm with a dynamic time warping (DTW) distance to find patient subtypes. They characterized three patient subtypes with different mean values of demographics, such as age, duration, and education, and with different mean values of clinical features such as Hoehn and Yahr Scale and MDS-Unified Parkinson's Disease Rating Scale (MDS-UPDRS). Similarly, to find acute kidney injury (AKI) patient subgroups, Xu et al. [24] extracted patient representation vectors from a memory networks (MN) trained on the Medical Information Mart for Intensive Care III (MIMIC- III) dataset and inputted the embeddings to a t-SNE algorithm. They identified three patient subgroups and differentiate these subgroups with age and clinical features such as Serum Creatinine (SCr) and Glomerular Filtration Rate Test (eGFR).

Instead of extracting patient embeddings from time sequence models, Xu et al. [25] chose the most important clinical features from a Gradient Booting Decision Tree (GBDT) and inputted these important features to a hierarchical clustering

model. They found three depression patient subgroups that are differentiated in age, the frequency of comorbidities and the amount of taken medications.

2.1.4 Subtype Discovery Evaluation Methods

Typical methods for evaluating the subtype discovery aim to capture three important aspects: *cluster (subtype) quality*, *subtype recovery examination*, and *cluster semantic meaning examination*.

Cluster quality measurement metrics include Silhouette Coefficient [17, 21], Bouldin Index [17] and the Calinski-Harabaz Index [17]. For *Subtype recovery examination*, when subtype ground-truth is available, discovered subtypes were compared with actual subtypes through visual inspection in the PCA space [17], t-SNE space [23], and through quantitative metrics such as adjusted mutual information [17]. In situations where subtype ground-truth is not available, Wen et al. and Yang et al. [10, 14] conducted a study where they generated simulation data from healthy control samples. The purpose was to examine whether their clustering approach could accurately identify the appropriate number of clusters and corresponding simulated neuroanatomical patterns. In Wu et al. [15], researchers implemented a consensus clustering [26] on training and validation data separately and tested if discovered subgroups from training and from validation were similar using in-group proportion (IGP) statistic [27]. Wu et al. [15] also used discovered subtypes as prediction labels and built a classifier with an accuracy of 90.6% to

predict subclass labels. *Cluster semantic meaning examination* requires domain knowledge, such as comparing the survivor rate of each cluster [13, 15, 20] or conduct statistical analysis for semantic meaningful features across different clusters [12, 15, 23-25].

2.2 Domain Shift Generalization

Traditional machine learning (ML) methods assume that training and test data are identically and independently distributed (i.i.d.). However, contrary to the i.i.d. assumption, we often see distribution shift or out-of-distribution situation in which the testing data distribution and the training data distribution are different. ML algorithms' performance drops significantly under distribution shift, therefore, how to solve the out-of-distribution (OOD) problem has become a crucial research direction in the ML community. In this section, I will systematically discuss state-of-the-art OOD generalization methods and applications to the medical domain.

Similar to Shen et al. [7], I review OOD generalization literature in four categories: Disentangle Representation Learning that aims to learn meaningful latent factors from the data, Domain Generalization that aims to learn transferrable and robust representations to different domains, Causal and Invariant Learning, and Optimization Methods for OOD Generalization. Under each category, I will introduce basic concepts, most representative works, and applications to the medical domain.

2.2.1 Disentangled Representation Learning

Let us consider an example from the medical imaging field and a task of detecting lung nodules in computerized tomography (CT) scans. The dataset may contain scans with different disease manifestations, showing different nodule characteristics under the same malignancy category, and acquired using various devices in different hospitals. In this specific task, we aim to extract a representation that is equivariant to the lung nodule location in a CT scan and invariant to the disease manifestation shift and acquisition shift. In this task, equivariance means the representation will change if the lung nodule location changed while invariance means the representation will not change when the disease manifestation and acquisition methods changed. In order to separate the equivariant and invariant variables, we need a disentangled representation learning (DRL) process that is able to decompose the input data into disentangled factors (also named as generating factors in DRL literature), where each factor corresponds to an important data generating factor [28]. In the lung nodule detection example, variables of interest are nodule shape, location, texture etc. Among these variables, nodule location is an equivariant variable and nodule shape and texture are invariant variables. In this section, we will review four different disentanglement architectures: Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), Normalizing Flows and Content-Style Disentanglement.

2.2.1.1 Variational Autoencoders (VAEs)

One typical DRL architecture is based on Variational Autoencoders (VAEs) [29], which is derived from Autoencoders (AEs) neural network architecture [30]. AEs consists of an Encoder that transforms high-dimensional data into low-dimensional representations and a Decoder that maps the representation from the bottleneck layer back to a reconstruction of the input. When compared with a standard AEs, VAEs have a regularized bottleneck layer, which enforces the distributions returned by the encoder to be close to Gaussian distribution. Higgins et al. [31] added an additional regularization hyperparameter β ($\beta > 1$) into the VAEs objective function (β-VAEs) in order to extract the most informative and dissimilar latent factors [32]. Increasing β value helps generate better disentangled representations but under the sacrifice of image reconstruction fidelity. Kim and Mnih [33] introduced the FactorVAE method as a means to achieve a more favorable balance between disentanglement and reconstruction quality in β-VAE. This approach encourages the representations' distribution to be factorial, thereby ensuring independence across dimensions.

Aforementioned VAEs approaches all assume latent factors are independent, while in most scenarios, factors with semantics are causally correlated. CausalVAE [34] is a representative work that included a linear Structural Causal Model (SCM) in VAEs to transform independent latent factors into causal representations. Shen

et al. [35] extended the work of CausalVAE through incorporating a nonlinear SCM with a bidirectional generative model.

To handle sequential data, Zhu et al. [36] proposed a sequential variational autoencoder model that enforces the latent variable to be disentangled into a static representation and a dynamic representation. In Zhu et al. [36], disentangled representations were used for representation swapping and video generation.

In medical field, sequential VAEs were utilized for disease decomposition [37-40]. Couronne et al. [37] proposed a generic deep longitudinal model to separate the variance caused by Alzheimer's progression from the inter-patient variability. A similar application in Alzheimer's disease characterizations is Yang et al.'s work [41] that they successfully disentangled latent variables into time-variant and time-invariant components across the entire region of interest (ROI) in the brain. They achieved this by incorporating a disease prediction loss into the image reconstruction loss function. Besides disease progression decomposition, VAEs have also been applied in medical image classification task. Gyawali et al. [42] proposed to first obtain the data embedding from a VAE model in an unsupervised learning manner and then follow by a self-ensembling network. Gyawali et al. [42] evaluated their model on a public available X-rays dataset [43] for thoracic disease study and showed an improved performance over other classification models.

2.2.1.2 Generative Adversarial Network (GANs)

The second line of DRL research is based on Generative Adversarial Networks (GANs) that consist of a generator and a discriminator. The generator network is responsible for generating new samples based on a noise variable. Throughout the training process, the generator is trained in an adversarial manner, competing against a discriminator that seeks to differentiate between genuine data samples and samples generated by the generator [44].

GANs learn disentangled representations through adding regularization terms, and InfoGAN [45] is a representative work that uses GANs for DRL. InfoGAN includes an additional term in the loss function aiming at maximizing the mutual information between the input noise variables and structured semantic features of the data distribution. While InfoGAN is designed for extracting interpretable and disentangled latent variables, Mukherjee et al. [46] demonstrated that the cluster structure is not retained in the InfoGAN latent space. Therefore, they trained the GAN with an inverse-mapping network and a clustering-specific loss. Another limitation of InfoGAN is it transforms a latent vector from source domain to target domain directly and the input latent space must follow the same training data probability density, causing some degrees of unavoidable entanglement. Karras et al. [47] devised a StyleGAN architecture that utilizes a generator to modify the image's style at each convolutional layer, effectively exploring an intermediate

space between the source domain and target domain. This adjustment is achieved through the manipulation of the latent code.

In addition, StyleGAN is able to separate the fine-grained and coarse-grained features, a property that makes StyleGAN a good candidate for DRL [48]. Based on the StyleGAN architecture, Nie et al. [48] expands the study to include a semi-supervised setting and demonstrated that utilizing merely 0.25% to 2.5% of labeled data is adequate for achieving notable disentanglement in high-resolution images.

In the medical domain, GANs were mainly used to decompose an abnormal image into patient-specific normal scans and scans with disease regions. Tang et al. [49] proposed a deep disentangled generative model (DGM) that consists of three branches: one for synthetic normal scan generation using GAN; one for disease separation using an encoder-decoder structure and the last one for training enhancement on noisy data using a similar encoder-decoder structure. Tang et al. [49] trained the DGM model and evaluated the model on a NIH Clinical Center Chest X-ray dataset [50]. Qualitatively, Tang et al. [49] showed that normal scans generated by DGM are more visually "radiorealistic" and disease region residual maps are more meaningful and interpretable than other competing methods. Quantitatively, in a lung opacity detection task, they showed an average 5% increase of precision by using DGM when compared with results obtained from other state-of-art methods. Similarly, Xia et al. [51] proposed a GAN structure

model to disentangle the pathology information from the healthy regions on images from three Magnetic Resonance Imaging (MRI) datasets. In an image retrieval task, Kobayashi et al. [52] proposed to use decomposed healthy MRI scans, decomposed abnormal scans or original abnormal scans as three different query image types. For image classification, Ben-Cohen et al. [53] mixed the disentangled class specified and unspecified representation into data augmentation process and increased the liver lesion classification accuracy by 7.4% over the baseline models. Michela et al. [54] disentangled the Contrast Agent effects, which is crucial for lesion classification purposes from all the other image components while performing the breast lesion classification.

2.2.1.3 Normalizing Flows (NFs)

Despite the impressive performance in learning distributions of images, VAEs and GANs have two main drawbacks: 1) in VAEs and GANs, we randomly draw new sample from the latent space and map them to the relevant domain through a decoder. Nevertheless, since the distribution of the latent space does not match that of the training data, both VAEs and GANs cannot precisely evaluate the probability density of novel points; 2) the training process using VAEs and GANs can be challenging due to vanishing gradients, training instability, mode collapse and posterior collapse [55]. Mode collapse occurs when the generator, during each iteration, excessively optimizes for a specific discriminator, preventing the

discriminator from effectively escaping this trap and learning a more diverse set of patterns [56]. Posterior collapse occurs when the variational posterior distribution closely aligns with the prior for a specific subset of latent variables, causing the generative model to disregard the influence of these latent variables [57].

A normalizing flow (NF) utilizes a series of invertible and differentiable mappings to transform a simple probability distribution into a more complex distribution [55]. Esser et al. [58] added a flow-based invertible network to a pretrained autoencoder model to translate hidden representations onto semantic concepts that are comprehensible to the user. Examples of semantic concepts are ‘digit’ and ‘color’ in MNIST dataset [59] and ‘beardiness’ and ‘smiling’ in CelebA dataset [60]. Instead of training on embeddings from a pre-trained network as in Esser et al. [58], Sankar et al. [61] trained a flow-based generative model directly on a brain tumor MRI dataset and factorized the latent space into anomaly representations, slice location and other semantic concepts. Wang et al. [62] employed a normalizing-flow-based approach to conduct counterfactual inference on a structural causal model, enabling the harmonization of diverse medical data sources. Wang et al. [62] trained their model on two source Alzheimer's Disease Neuroimaging Initiative (ADNI) sites data, evaluated the model on two different target ADNI sites and showed an increased classification result with data harmonization.

2.2.1.4 Content-Style Disentanglement (CSD)

VAEs, GANs and NFs decomposes an input image into a single vector of latent variable representation while content-style disentanglement (CSD) generates two vectors: a latent variable representation vector associating with the image appearance such as color; and a tensor latent variable vector representing the image content in terms of objects [63]. For example, a house is an image content, but the image itself can be a photorealistic style or Van Gogh style. Gatys et al. [63] firstly proposed to separate the domain invariant image content and domain specific style information during the CNN training process. Huang et al. [64] implemented the concept of CSD to transfer an image from a source domain to a target domain without seeing any examples of corresponding image pairs.

In the medical domain, CSD was used for image to image translation [65], data harmonization [66], segmentation [67] and classification [68]. Li et al. [65] disentangled domain features into domain-shared structural features and domain-independent appearance features. They utilized this disentanglement to synthesize invasive and harmful Fluorescein Fundus Angiography (FFA) images using non-invasive Fluorescein Fundus (FF) images. Zuo et al. [66] learned a globally disentangled latent space that encompasses both anatomical and contrast information, enabling harmonization. Chartsias et al. [67] factorized 2D cardiac images into spatial anatomical factors and non-spatial modality factors. Spatial

anatomical factors maintain pixel-level correspondences with the input, such as factors related to myocardium, left and right ventricles locations in an MRI image, and can be used for multi-class segmentation. On the other hand, none-spatial modality factors contain modality-specific information, such as the distribution of intensities of the spatial regions. Chartsias et al. [67] showed that with only very limited number of labels, the CSD model significantly improved the segmentation performance when compared with other state-of-art semi-supervised segmentation models. To improve the interpretability of brain MRI classification, Bass et al. [68] disentangled class relevant features from irrelevant confounds.

2.2.2 Domain Generalization (DG)

A domain is referred as data sampled from one distribution [69]. Domain Generalization (DG) aims to train a model on different but related domains that will produce generalizable results on unseen domains [70]. For example, images acquired from different scanners / medical centers often have different intensity distribution, contrast, and noise levels [69]. In this section, I will review domain-invariant representation learning methods, and different DG training strategies. Domain-invariant representation learning methods are related to DRL, but instead of distengling features into domain shared or domain-specific representations, domain-invariant representation learning aims to learn the domain invariant representations directly with a supervised learning manner.

2.2.2.1 Domain-Invariant Representation Learning

Domain-invariant representation learning assumes there are representations that are transferrable and robust on different domains. The goal of domain-invariant representation learning is to train a model with minimized representation discrepancy between different source domains, thus, the trained model is generalizable to the unseen domain [70]. Similar to [7], I divide the domain-invariant representation learning methods into domain adversarial learning, feature alignment and normalization, and kernel-based methods.

Domain Adversarial Learning. Based on the basic generative adversarial networks (GANs) architecture, Ganin and Lempitsky [71] and Ganin et al. [72] included a domain-discriminator that distinguishes the source and target domains. During the training process, the model confuses the domain discriminator to learn the domain invariant representations. Li et al. [73] expanded the capabilities of adversarial autoencoders by utilizing the Maximum Mean Discrepancy (MMD) to align distributions across various domains. They further aligned the distribution with an arbitrary prior distribution through the application of adversarial feature learning. Li et al. [73] make the assumption that the conditional distribution of labels given image features remains unchanged across domains. However, this assumption may not always be valid in practical applications. To address this limitation, Li et al. [74] introduced a conditional invariant adversarial network

designed to learn domain-invariant representations by considering the joint distribution of image invariant representations and image labels. Domain adversarial learning is a commonly employed technique in image-to-image translation tasks. Zhu et al. [75] addressed the challenge of unpaired image-to-image mapping by integrating two mapping functions into the GAN architecture. One mapping function translates a source image to the target image, while the other mapping function converts the transformed image back to the source domain. The training process is governed by the forward cycle loss and the backward cycle loss, ensuring consistency in both directions. Instead translating a source image directly into a target image, Gong et al. [76] translated a source image to a sequence of intermediate images between the source and target domains and proved that using intermediate images is better for down-stream tasks, such as segmentation and classification.

In medical domain, domain adversarial learning was mainly used for segmentation. Li et al. [69] tackled a cross-domain medical image segmentation challenge by introducing a semantic discriminator. This discriminator ensures a comparable image-to-label mapping between the source and target domains, effectively addressing the problem of cross-domain medical image segmentation. Li et al. [69] trained and evaluated their model on three different brain MRI datasets and showed the highest dice score results compared with results using other cross-domain image segmentation approaches. Chen et al. [77] applied adversarial

learning to enhance the image and feature alignment between two domains (MRI and CT) and showed an improved cardiac substructure segmentation and abdominal multi-organ segmentation results.

Feature Alignment and Normalization. Feature alignment and normalization methods aim to learn domain invariant representations through aligning features across different source domains. Based on a siamese network architecture, Motiian et al. [78] semantically aligned samples from different domains, such as images from Amazon, Webcam, and DSLR in Office dataset [79], by minimizing the distance between same class samples while at the same time maximizing the distance between samples with different class labels and domains. Other feature alignment methods focus on minimizing the distance between feature distributions with different distance measurements, such as Wasserstein distance [80], maximum mean discrepancy distance (MMD) [81], and the second order correlation [82].

In a glaucoma detection task, Zhou et al. [83] proposed a data augmentation-based (DA) feature alignment (DAFA) method to enhance the out-of-distribution (OOD) generalization of a single fundus image dataset. Instead of implementing feature alignment between two source datasets [78], DAFA performs the feature alignment from a single source dataset but between two augmented views. Zhou et al. [83] trained the model on one private dataset, evaluated the model on images

from other six datasets, and showed the highest AUC (greater than 0.1 increase) compared with other state-of-art classification algorithms.

Kernel-Based Methods. Kernel-based methods use specific kernel functions to transform the input to a high-dimensional feature space. Pan et al. [84] introduced the concept of Transfer Component Analysis (TCA), which aims to learn transfer components across domains within a reproducing kernel Hilbert space. Grubinger et al. [85] extended the formulation of TCA to multiple source and target domains. In addition to TCA, domain-invariant component analysis (DICA) [86] is another notable approach. DICA is a kernel-based optimization algorithm that learns an invariant transformation by minimizing the variance across domains while preserving the functional relationship between input and output variables. In the supervised setting, DICA relies on the inverse of a covariance operator, which can be computationally expensive and prone to instability in practical applications [87]. To overcome this limitation, Gan et al. [87] proposed a solution by incorporating a centered kernel alignment that incorporates attribute labeling information. They demonstrated that utilizing DICA with the centered kernel alignment enables the learning of representations that are both category-invariant and attribute-discriminative. In the medical field, Opbroek et al. [88] explored kernel learning as a means to mitigate differences between training and test data. They conducted experiments on brain tissue, white matter lesion, and hippocampus segmentation tasks.

2.2.3 Training Strategy

Some studies focus on different training strategies to achieve domain generalization. We will review four categories: meta learning, ensemble learning, unsupervised, and semi-supervised DG [7].

2.2.3.1 Meta Learning

Meta learning involves training a model on multiple tasks with the aim of enabling the model to tackle new learning tasks using only a small number of training samples. Finn et al. [89] introduced a model- and task-agnostic algorithm designed for meta-learning schemes. Li et al. [90] then implemented the model agnostic concept on domain generalization problem. To mimic real train-test domain shifts, Li et al. [90] partitioned the original source domains into meta-train domains and meta-test domains. This approach allowed for the simultaneous optimization of the loss functions for both meta-train and meta-test domains. Balaji et al. [91] added a regularization function to the meta learning framework and showed that the notion of domain generalization can be explicitly encoded in the regularization function. Du et al. [92] employed a meta variational information bottleneck (MetaVIB) to learn domain-invariant representations. The MetaVIB approach gradually reduces the domain gaps throughout the meta training process.

2.2.3.2 Ensemble Learning

Ensemble learning incorporates models for different domains to achieve generalization. Mancini et al. [93] suggested utilizing multiple domain-specific classifiers during the training phase and estimating the probabilities of a target sample belonging to each source domain. Segu et al. [94] trained multiple domain-dependent classifiers, collected independent domain's statistics through BN layers, and then mapped statistics to a domain invariant feature space.

2.2.3.3 Unsupervised and Semi-supervised Domain Generalization

Unsupervised domain generalization (UDG) aims to learn generalizable representations across different domains in an unsupervised manner and thus reduce the dependence on labeled data [95]. Zhang et al. [95] proposed a contrastive learning algorithm that forces the model to ignore domain-related features based on the instance similarity between different domains. During the training, the algorithm selects instances from different domains but ignore the instances from similar domains. Regarding semi-supervised learning, Liao et al. [96], presented an approach that combines Wasserstein generative adversarial network with gradient penalty (WGAN-GP) based adversarial learning and pseudo label-based semi-supervised learning. This method leverages both a fully labeled source domain dataset and a completely unlabeled source domain dataset simultaneously. In medical domain, Perone et al. [97] showed that by using a small amount of unlabeled brain MRI images from multiple domains, they can significantly increase

the gray matter segmentation performance. In a similar vein, Zhang et al. [98] introduced a semi-supervised domain generalization approach to address fundus image segmentation and chest X-ray diagnosis tasks. Their method involves utilizing one labeled source domain, provided by a medical center with ample expert effort, along with multiple unlabeled source domains gathered from different centers.

2.2.4 Causal Learning

Causal learning assumes that when other variables are altered, the conditional distribution of the target variable given the direct cause variable remains unchanged. An example of a target variable is “intensive treatment”, indicating whether a patient needs intensive treatment, and a direct cause of “intensive treatment” is the level of pre-treatment overall fitness [99]. Under the assumption that direct causal variables are stable across domains, if we can find the direct causes of the target variable, we can achieve OOD generalization. Peters et al. [100] leveraged this assumption and proposed Invariant Causal Prediction (ICP). Pfister et al. [101] relaxed the assumption in Peters et al. [100] that all environments are known, and proposed to detect causal relations using sequential non-stationary data. Heinze-Deml et al. [102] further extended the ICP into a non-linear model. The existence of hidden confounders violates the invariance assumption in ICP and a traditional strategy to deal with hidden confounders is to introduce instrument

variables. Considering a situation that we want to investigate the effect of acute myocardial infarction (AMI) treatment on mortality, the distance to the nearest hospital with cardiac catheterization is correlated with the mortality but does not have a causal relationship with it, except via AMI [103]. In this situation, the distance to the nearest hospital is an instrument variable and researchers have proposed various methods to use instrument variables, such as two stage least squares (2SLS) [104] and bivariate probit with correlated errors [105] etc. When utilizing instrumental variables, it is typically necessary for them to have no direct impact on the hidden confounding variable or the outcome variable. However, Rothenhäusler et al. [106] relaxed this requirement by introducing anchor variables, which can directly influence both the hidden confounding variable and the outcome variable. Oberst et al. [107] further relaxed the assumption in Rothenhäusler et al. [106] that anchor variables can be observed and proposed to use the noisy proxy of anchor variables.

In medical domain, Castro et al. [108] systematically discussed the importance of establishing the causal relationship between images and their annotations and provided recommendations for medical image analysis, including establishing the predictive causal direction, identifying any evidence of mismatch between datasets and determining whether the data collection was biased etc. Amsterdam et al. [99] addressed the bias in treatment effect estimation arising from colliders and proposed a method that utilizes the similarity between the final layer of a CNN and

linear regression to mitigate the collider effect in a lung cancer survival prediction task.

2.2.5 Optimization for OOD Generalization

Optimization methods for OOD generalization focus on ensuring the worst-case performance in the presence of distribution shifts [7]. When compared with disentangled representation learning and domain generalization methods, optimization methods are often model and data structure agnostic [7].

2.2.5.1 Distributionally Robust Optimization (DRO)

Similar with GAN, DRO objective function has two terms: a learner and an adversary. The adversary maximizes the expected loss through shifting the test distribution from the training distribution as far as possible but within a pre-specified range. The learner minimizes the adversarial expected loss [8]. The key component of DRO studies is to choose the class of uncertainty sets for testing distribution under certain constraints, where f-divergence [109] and Wasserstein distance constraints [110] are the most commonly used ones. DRO with f-divergences lets the uncertainty set for test distribution be an f-divergence ball from a training distribution [8, 111]. Despite DRO with f-divergences is a common choice in the literature [112], Hu et al. [8] revealed that DRO ends up optimizing the given training distribution instead of testing distribution in

classification tasks due to the overly flexible uncertainty set distributions and the losses used in classification. Compared with f-divergence, Wasserstein distance constrain is more flexible. One representative application of DRO with Wasserstein distance is Sinha et al's work [113], they addressed the distribution shift problem caused by adversarial input perturbations and achieved better classification performance compared to empirical risk minimization (ERM). Sagawa et al. [2] found that applying DRO naively to neural networks without regularizations often fails, and proposed to impose l_2 penalty [114] or early stopping to add penalty as model complexity increase. In addition, Sagawa et al. [2] firstly optimized the worst group performance (Group-DRO) instead of worst case performance. They showed that a regularized Group-DRO method achieves better generalization results compared with results using ERM and DRO. Sohoni et al. [4] combined a subclass discovery process with Group-DRO and demonstrated that Group-DRO boosted worst-case subclass accuracy by up to 22% on benchmark image classification datasets.

2.2.5.2 Invariant-Based Optimization

DRO methods directly optimize the training process to achieve the worst-case performance, however, in practice, datasets are frequently assembled without source labels available [115]. For example, when we try to collect different animal images online, we cannot always acquire the camera setting information, which

makes the domain labels unclear. Invariance-based optimization aims to identify invariant properties within the data and leverages multiple environments to discover such invariance for the purpose of OOD generalization. According to Arjovsky et al. [116], an environment is defined as an instance of the variable E that impacts the prior distribution of variables. Liu et al. [115] introduced the Heterogeneous Risk Minimization (HRM) framework, which facilitates the simultaneous learning of latent heterogeneity and invariant relationships among the data. Through searching features that are invariant across environments, Chang et al. [117] identified a subset of input features that are causally related to the prediction. Similarly, Koyama et al. [118] found the invariant representations through a Maximal Invariant Predictor (MIP).

2.2.6 OOD Generalization Evaluation Methods

According to Ye et al. [119], the test accuracy in a single environment is misleading in OOD situation and we need to consider test accuracy in multiple environments. In this review, we assess three evaluation metrics: average accuracy, worst-case accuracy, and the standard deviation of accuracies. Average accuracy measures the overall performance among different testing distributions. A significant disadvantage of average accuracy is it treats all testing distributions the same without considering the frequency it occurs. As suggested in Shen et al. [7], it is possible to assign weights to the accuracy based on the disparity between each

testing distribution and the training distribution, for example, we can assign a smaller weight to an accuracy generated from a testing data that has a relatively larger Kolmogorov-Smirnov distance to the training data. Worst-case accuracy measures the worst-case accuracy across all testing distributions. Worst-case accuracy has been widely employed in the literature on DRO and plays a critical role in high-stakes applications like computer-aided diagnosis and financial security. On the other hand, the standard deviation of accuracies quantifies the performance variation across different distributions. This metric serves as a measure of the algorithm's robustness and stability.

CHAPTER 3. Methodology

This chapter introduces hidden stratification discovery and subgroup learning methodologies. Section 3.1 describes a hidden stratification discovery method using an algorithmic approach. The subclass labels generated in the hidden stratification discovery process will be used for further gDRO model training and for model evaluation. Section 3.2 presents different training strategies with various model inputs, architectures, and loss functions for malignancy classification tasks. Section 3.3 explains an innovative training strategy, Classifier Retraining on Representative Independent Splits (CRRIS), which combines ERM and gDRO during training. Figure 3.1 provides an overview of the experimental design in this study.

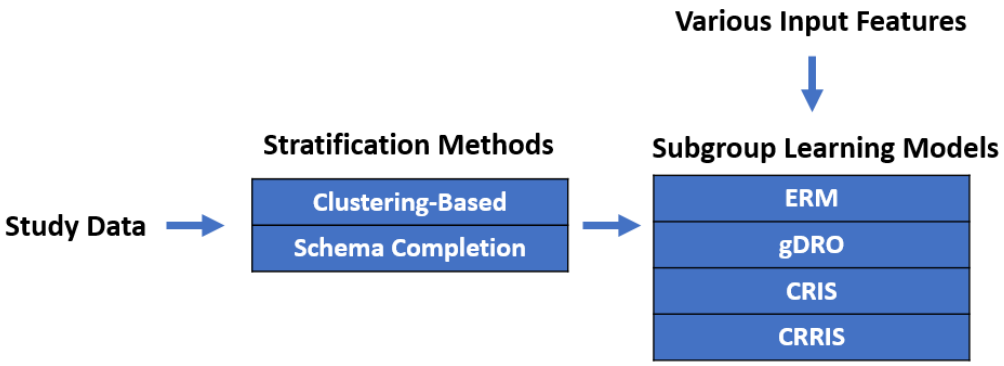

Figure 3.1 Methodology Overview. This study investigates clustering-based and schema completion hidden stratification methods and explores subgroup learning models with different input features.

3.1 Hidden Stratification Discovery

This section will only describe the algorithmic or clustering-based approach to hidden stratification discovery methodology. This is because schema completion requires domain knowledge, and the subgroup labels depend on the study dataset. Clustering-based method consists of deep image feature extraction (Section 3.1.1), feature reduction (Section 3.1.2) and clustering (Section 3.1.3) as illustrated in Figure 3.2.

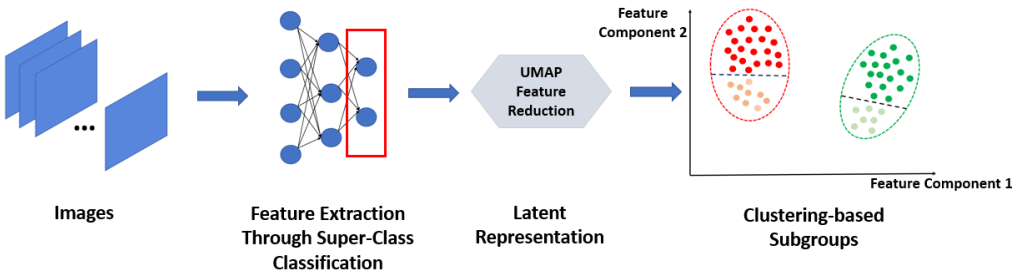

Figure 3.2 An illustration of clustering-based hidden stratification discovery. It consists of deep image feature extraction, feature reduction, and clustering.

3.1.1 Deep Image Feature Extraction

The image feature extraction process consists of two steps: First, I trained a CNN classifier to classify malignancy (malignant vs. benign) using input study images. Second, I saved the output of the last convolutional layer as our image features during the malignancy prediction process. Due to the small number of training images, I used a ResNet CNN network [120] pre-trained on ImageNet

[121]. Since lower-level convolutional layers produce lower-level features such as lines and corners, while higher-level convolutional layers provide high-level features that better describe the content of an input image, we expect these deep image features extracted from the last convolutional layer to enable the discovery of latent malignant and benign subtypes. After the feature extraction process, each cropped nodule image is associated with a vector of 512-dimensional CNN features.

3.1.2 Uniform Manifold Approximation and Projection (UMAP)

Uniform Manifold Approximation and Projection (UMAP) algorithm [122] was used as a feature reduction technique in this study. Compared with other feature reduction techniques, *UMAP has the advantage of preserving the cluster relationship of data points in the high-dimensional space.* Mathematically, assume there are n neighboring data points of each data point in the high-dimensional space (n is a hyperparameter in UMAP), given a random data point x_1, we can define a high-dimensional similarity score (SH) between x_1 and another random data point x_2 in the high-dimensional space as:

$$SH(x_1, x_2) = e^{-\frac{\left(dH_{(x_1,x_2)} - dH_{(x_1,x_{1_{nearest}})}\right)}{\delta}} \quad (1)$$

where $dH_{(x_1,x_2)}$ represents distance between x_1 and x_2 in the high dimensional space

$dH_{(x_1, x_{1_{nearest}})}$ represents the distance between x_1 and its nearest neighbor in the high dimensional space

δ is a hyperparameter that forces the sum of all pairwise SH equals to $log_2(n)$

To make the SH symmetrical, we define a symmetric high-dimensional similarity score SH' between x_1 and x_2 as:

$$SH(x_1, x_2)' = SH(x_1, x_2) + SH(x_2, x_1) - SH(x_1, x_2) \times SH(x_2, x_1) \quad (2)$$

where $SH(x_2, x_1)$ and $SH(x_2, x_1)$ are calculated with Equation (1)

Then, UMAP algorithm calculates low-dimensional similarity scores SL between $x_1$and x_2 utilizing a fixed, symmetrical curve derived from *t*-distribution:

$$SL(x_1, x_2) = \frac{1}{1 + \alpha dL_{(x_1, x_2)}{}^{2\beta}} \quad (3)$$

where $dL_{(x_1, x_2)}$ is the distance between data point x_1 and x_2 in the low-dimensional space

α and β control how tightly the low dimensional points can be packed together

UMAP dynamically changes the location of data points in the low dimensional space in order to reproduce the high-dimensional cluster. Given a random data point x_1, UMAP algorithm randomly picks a data point x_2 from its neighboring data points. The chance of picking data point x_2 is proportional to $SH(x_1, x_2)'$ using Equation (2). At the same time, UMAP algorithm randomly picks a data point x_3 that does not belong to the neighboring data points. The goal is to find the best new

position of x_1 in the low-dimensional space that minimizes the $\mathrm{SH}(\mathrm{x}_1, \mathrm{x}_2)$ while maximizes the $\mathrm{SH}(\mathrm{x}_1, \mathrm{x}_3)$. UMAP algorithm repeat this process for all data points in the low dimensional space.

In this study, the input of the UMAP algorithm is the 512-dimensional CNN features extracted in Section 3.1.1, and the output is 2-dimensional features that were used as input for the clustering model. As a comparison, I also implemented principal component analysis (PCA) [123] as another approach for feature reduction. Similarly, the input of PCA is the 512-dimensional CNN features, and the output is 2-dimensional feature embeddings that were used as input for the clustering model.

3.1.3 Gaussian Mixture Clustering

Gaussian mixture model assumes all data points were collected from a mixture of Gaussian distributions with unknown parameters and these unknown parameters can be estimated with an Expectation-Maximization (EM) Algorithm. In the expectation phase, for each data point i, the algorithm computes its probability p_{ic} that it is belongs to a certain cluster c:

$$p_{ic} = \frac{\pi_c \, \mathcal{N}(x_i; \mu_c; cov_c)}{\sum_{c'}^{C} \pi_{c'} \, \mathcal{N}(x_i; \mu_{c'}; cov_{c'})} \tag{4}$$

C represents the total number of clusters

π_c represents the total number of data points in cluster c

μ_c represents the mean value of data points in cluster c

cov_c represents the covariance of data points in cluster c

In the maximization phase, for each cluster c, the algorithm first calculates the weight m_c :

$$m_c = \sum_i^{\pi_c} p_{ic} \tag{5}$$

π_c represents the total number of data points in cluster c

Then the algorithm updates the size, mean, and covariance of each cluster with:

$$\pi_c = \frac{m_c}{m} \tag{6}$$

m represents the sum of the weights across all clusters

$$\mu_c = \frac{1}{m_c} \sum_i^{\pi_c} p_{ic} x_i \tag{7}$$

$$cov_c = \frac{1}{m_c} \sum_i^{\pi_c} p_{ic} (x_i - \mu_c)^T (x_i - \mu_c) \tag{8}$$

At each step, the algorithm increases the log likelihood of the model (Equation 9) until convergence.

$$log\, P(X) = \sum_i^C log[\sum_{c'}^C \pi_{c'}\, \mathcal{N}(x_i; \mu_{c'}; cov_{c'})] \tag{9}$$

Compared to other clustering algorithms such as K-means, Gaussian Mixture Clustering has the following advantages: First, it allows for more flexible cluster

shapes by modeling each cluster as a mixture of Gaussian distributions. Second, Gaussian Mixture Clustering captures the covariance structure between variables, allowing it to model complex relationships in the data. It can capture correlations and dependencies between different features. Third, it is robust to outliers, the probabilistic modeling of GMM assigns lower probabilities to outliers, reducing their influence on the clustering process. My initial clustering results indicate that I can well separate malignant and benign nodules (one pure malignant cluster and one pure benign cluster), therefore, to discover the hidden sub-classes of malignant and benign nodules, I performed Gaussian Mixture clustering on benign and malignant nodules separately. I chose the number of clusters with the highest Silhouette Coefficient [124] since a higher Silhouette Coefficient indicates clusters are well apart from each other and clusters are clearly distinguished. Mathematically, Silhouette Coefficients (SC) is defined as:

$$SC = \frac{b - a}{max(a, b)} \tag{10}$$

where a represents average intra-cluster distance

b represents average inter-cluster distance

3.2 Subgroup Learning Models and Loss Functions

This section introduces different model architectures and loss functions for subgroup learning. Section 3.2.1 introduces a transfer learning approach and a

convolutional neural network (CNN) architecture that takes images as input. Section 3.2.2 and Section 3.2.3 describe two classification architecture that takes numerical features as input and a combination of images and numerical features as input respectively. Section 3.2.4 explains empirical risk minimization (ERM) loss function and group distributionally optimization (gDRO) function. Section 3.2.5 provides details on the data splits and model evaluation methods.

3.2.1 Transfer Learning and ResNet18

Transfer learning method addresses the limitation of small amounts of training data by first pre-training a deep learning model on a publicly available large dataset, then fine tuning the model on our own dataset [125]. In this study, I used a pre-trained ResNet18 convolutional neural network on ImageNet [121] for both the deep image feature extraction (Section 3.1.1) and for subgroup learning with images as model inputs. Table 3.1 shows the convolutional layer structure of ResNet18. The number of freezing convolutional layers, the number of fully connected layers, and the dimension of output layers are hyperparameters and I will provide more details in Chapter 5.

Table 3.1 Convolutional Layer Structure of ResNet18

Layer Name	# filters *(filter size)	Output size
Convolutional 1	64*(7*7), stride 2	112*112 * 64
Max Pooling	1*(3*3), stride 2	
Convolutional 2 and 3	64*(3*3), stride 2	56*56 *64
Convolutional 4 and 5	128*(3*3), stride 2	28*28*128
Convolutional 6 and 7	256*(3*3), stride 2	14*14*256
Convolutional 8 and 9	512*(3*3), stride 2	7*7*512
Average Pooling	1* (7*7)	1*512

3.2.2 Fully Connected Network

When the model input is numerical features, I used a fully connected network. The number of hidden layers, the number of dimensions of each layer, and the dimensions of the output layers are hyperparameters in this study.

3.2.3 Composite Convolutional Neural Network (CompNet)

Composite convolutional neural network (CompNet) proposed by Qiu et al. [126] is specially designed for combining images and numerical features, such as designed features and semantic features, into the neural network. Figure 3.3 illustrates the CompNet architecture.

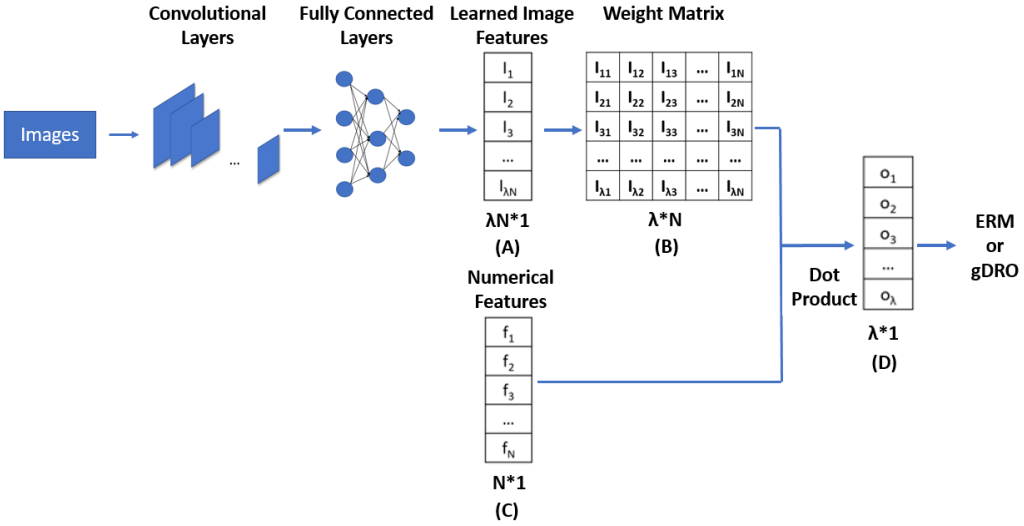

Figure 3.3 CompNet Architecture. λ represents the number of classification classes; N is the total number of numerical features. (A) Learned Image feature vector with a dimension λ*N; (B) Weight Matrix with λ *N dimensions generated from (A); (C) Input numerical feature vector; (D) Dot product of weight matrix and numerical feature vector.

Instead of simply flatting and concatenating image pixel values with other numerical features as input of a CNN [127-129], CompNet learns weights of each numerical feature from input images and uses these weights to refine the model during the training. There are three important steps in CompNet: the first step is to set the dimension of last fully connected layer as λ *N, where λ is the number of classification labels and N is the number of numerical features, thus the output of fully connected layers is a λ*N by 1 feature vector; the second step is to resize the fully connected layers output to a λ by N weight matrix; the third step is to conduct dot product between the λ by N weight matrix and the input N * 1 numerical

features and the dot product result will be the input of the loss calculation process. In this study, I used the same dimensions of convolutional and fully connected layers as described in Section 3.3.1.

3.2.4 Empirical Risk Minimization (ERM) and Group Distributionally Optimization (gDRO)

Empirical Risk Minimization (ERM) aims to minimize the *average loss* of all samples. Given a training dataset $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ with n data points, where x_i is the observation and y_i is the label, a classification model F, and a loss function l. ERM loss is defined as:

$$loss_{ERM}(D) = \frac{1}{n}\sum_{i=1}^{n} l(F(x_i), y_i) \tag{11}$$

Instead of treating each sample equally, group Distributionally Robust Optimization (gDRO) divides training dataset D into K disjoint sub-datasets $D_1, D_2, \dots, D_K$, and focus on the sub-dataset that has the maximum ERM loss. During the training, a model with gDRO loss minimizes the loss of the worst group. gDRO loss is defined as:

$$loss_{gDRO}(D) = \sum_{k=1}^{K} q_k^{(t)} loss_{ERM}(D_k) \tag{12}$$

Where $q_k^{(t)}$ represents weight for subgroup k at iteration t.

In this study, I trained different classification models (Figure 3.3) with both ERM and gDRO losses and compared their results.

3.2.5 Data Splits and Model Evaluation Methods

To obtain generalizable classification results, I randomly shuffled the training (80%), validation (10%) and testing data (10%) 30 times using stratification method based on nodule malignancy ratings. For each study dataset, I used the same 30 training sets to build machine learning models, same 30 validation sets to tune the hyper parameters and same 30 testing sets to report classification results (mean accuracy and 95% confidence interval across 30 trials). I conducted a two-tail T-test to compare significance between two sets of results.

In addition to overall model accuracy, I evaluated the performance of classification models at each subclass level and used the accuracy of the worst-performing subgroup as a metric to measure the model's generalization ability. The worst-performing subgroup is defined as the subgroup with the lowest accuracy across all subgroups from an ERM model.

3.3 Classifier Retraining for Model Robustness Improvement

One limitation of the base gDRO model introduced in Section 3.2.4 is it requires subclass labels of all the training data. Nguyen et al. [130] proposed a *classifier retraining on independent splits (CRIS)* method that only requires a

subset of subclass labels and achieved higher worst group accuracy by implementing CRIS when compared with the base gDRO model results. Section 3.3.1 describes CRIS and Section 3.3.2. introduces classifier retraining on representative independent splits (CRRIS), an innovative training method derived from CRIS. Both CRIS and CRRIS require images as input.

3.3.1 Classifier Retraining on Independent Splits (CRIS)

Compared to traditional data split methods in Section 3.2.5, Nguyen et al. [130] further *randomly* split the training data into two subsets ("Independent Splits"), trained a CNN model with ERM loss on the first training subset, transferred the weights of the trained CNN model to another newly initiated CNN model, and trained fully connected layers of the new CNN model with gDRO loss on the second training subset ("Classifier Retraining"). The motivation behind CRIS training strategy is to separate the image feature extraction process and classification process during the training, and use a model with ERM loss to extract features and a model with gDRO loss for classification. Different with the base gDRO model that requires the subgroup labels of entire training dataset, CRIS only needs subclass labels of a subset of training data that is used for gDRO model training. Figure 3.4 illustrates CRIS method.

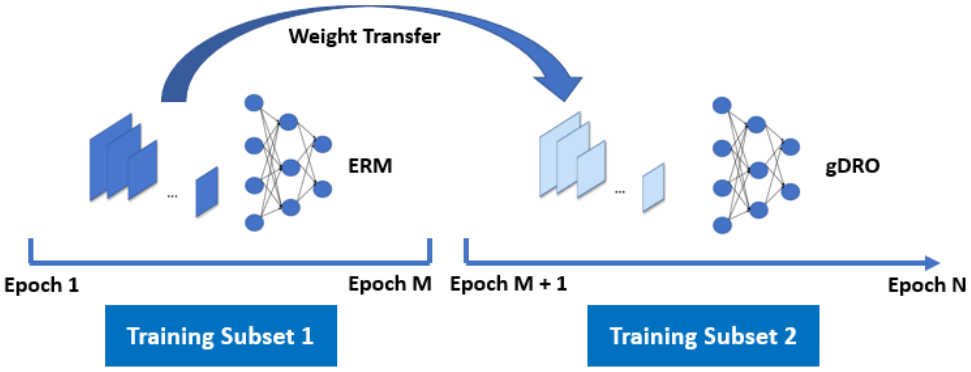

Figure 3.4 An illustration of Classifier Retraining on Independent Splits (CRIS). N represents the total number of epochs, M represents the total number of training epoch for the first CNN model with ERM loss. Both M and N are hyperparameters.

3.3.2 Classifier Retraining on Representative Independent Splits (CRRIS)

In this study, instead of randomly splitting the training data into two subsets as in Nguyen et al. [130], I innovatively split the training data into *representative* and *atypical* instances based on the instance representative and named this method *Classifier Retraining on Representative Independent Splits (CRRIS).* An instance is a representative case if the Euclidean distance between its image feature and the image feature of the cluster center is within the smallest p% among the Euclidean distances between image features of every instance within the cluster and the image feature of the cluster center. p is a hyperparameter, where a higher p implies more instances are representative and fewer instances are atypical instances. I used the same image features described in Section 3.1.1 and cluster centers generated from

Gaussian Mixture Clustering described in Section 3.1.3. Figure 3.5 illustrates the process of choosing representative and atypical instances. My hypothesis is that, *compared to the CRIS method that trains an ERM model on one random split and then retrains a model with gDRO loss on another random split, training the ERM model on typical instances or atypical instances produces more generalizable features that can further improve the model's generalization ability measured as the worst group accuracy.*

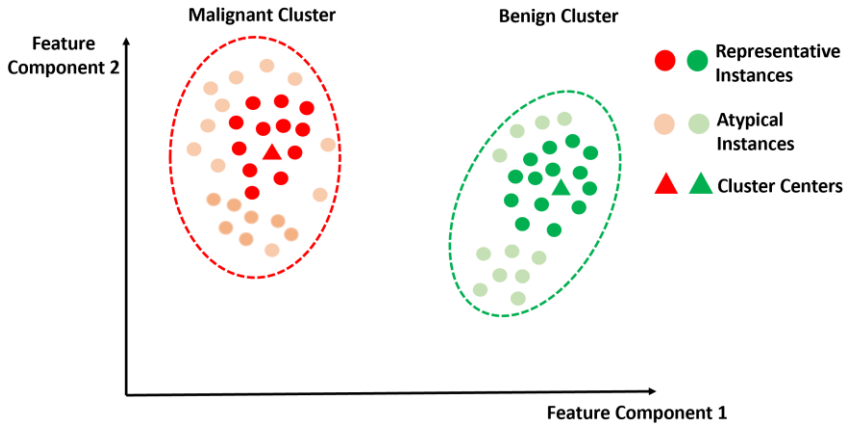

Figure 3.5 An illustration of choosing representative and atypical instances. Triangles represent two cluster centers. Dots in darker color and shallow color represent representative and atypical instances respectively. Representative instances are determined by the Euclidean distance between image features of each instance within the cluster and the image features of the cluster center.

Figure 3.6 and Figure 3.7 illustrate two CRRIS models with different training orders. For the first CRRIS model (Figure 3.6), an ERM model is firstly trained on atypical instances and then retrain the model with gDRO loss on typical instances. For the second CRRIS model (Figure 3.7), the order of the atypical and typical instances reversed that an ERM model is firstly trained on typical instances and then retrain the model with gDRO loss on atypical instances.

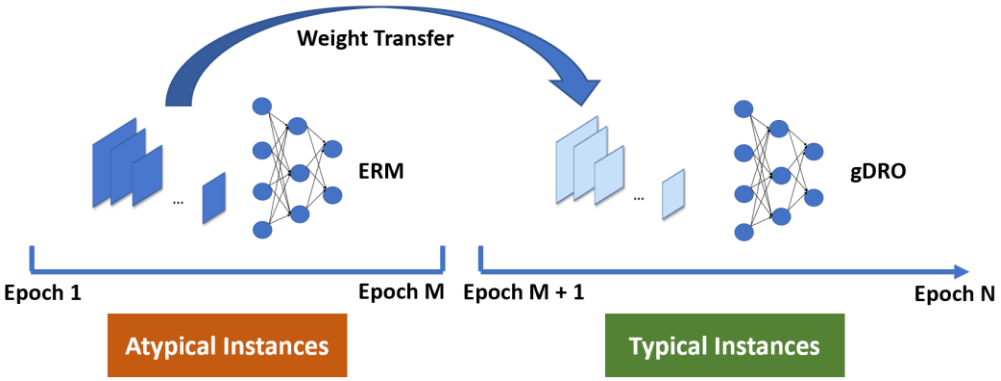

Figure 3.6 An illustration of Classifier Retraining on Representative Independent Splits (CRRIS) with an ERM model trained on atypical instances. In this situation, we first train a CNN model with ERM loss on atypical instances and then retrain the same CNN model with gDRO loss on typical instances. N represents the total number of epochs, M represents the total number of training epoch for the first CNN model with ERM loss. Both M and N are hyperparameters.

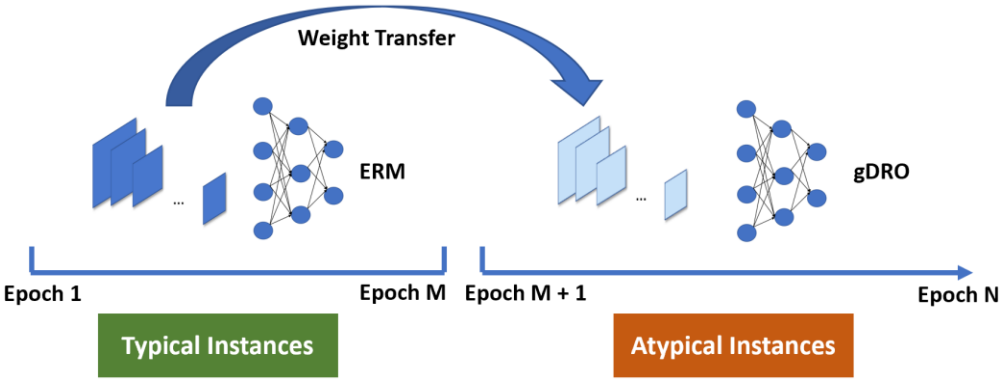

Figure 3.7 An illustration of Classifier Retraining on Representative Independent Splits (CRRIS) with an ERM model trained on typical instances. In this situation, we first train a CNN model with ERM loss on typical instances and then retrain the same CNN model with gDRO loss on atypical instances. N represents the total number of epochs, M represents the total number of training epoch for the first CNN model with ERM loss. Both M and N are hyperparameters.

CHAPTER 4. Applications to Lung Cancer

4.1 The Lung Image Database Consortium (LIDC) dataset

The Lung Image Database Consortium (LIDC) dataset [3, 131] contains 2,680 distinct nodules found in Computed Tomography (CT) scans from 1,010 patients. Radiologists manually identify, delineate, and semantically characterize nodules that are three millimeters or larger across nine semantic characteristics, which are calcification, internal structure, lobulation, malignancy, margin, sphericity, spiculation, subtlety, and radiographic texture (Table 4.1). The radiologists assign ordinal or nominal ratings to the nodules based on these features.

Table 4.1 Semantic features in LIDC datasets.

Ratings / Features	1	2	3	4	5	6
Calcification	Popcorn	Laminated	Solid	Non-central	Central	Absent
Internal Structure	Soft tissue	Fluid	Fat	Air	-	-
Lobulation	Marked	-	-	-	-	None
Malignancy	Highly unlikely	Moderately unlikely	Indeter minate	Moderately Suspicious	Highly Suspicious	Absent
Margin	Poorly defined	-				Sharp
Sphericity	Linear	-	Ovoid	-	Round	-
Spiculation	None	-	-	-	Marked	-
Subtlety	Extremely subtle	Moderately subtle	Fairly Subtle	Moderately Obvious	Obvious	-
Radiographic Texture	Non-solid	-	Part Solid	-	Solid	-

In this study, I implemented the following data preprocessing steps: First, I cropped nodules into images of size 71 x 71, which is the size of the largest nodule in the dataset. The cropped images were centered on the nodule's center based on the radiologists' delineated nodule boundaries. Second, for each semantic characteristic, I took the mode of the four radiologists' ratings as the final rating. If a nodule did not have a mode rating, I used the average of the ratings. If the average rating was a decimal number, I rounded it down to the nearest integer. For example, if the average rating of a nodule was 3.7, I used 3 as the final rating of the nodule. Third, I assigned malignancy classification labels based on the mode or average value of malignancy ratings. Nodules with malignancy ratings of 1 (highly unlikely) and 2 (moderately unlikely) were labeled as 'Benign', while nodules with malignancy ratings of 4 (moderately suspicious) and 5 (highly suspicious) were labeled as 'Malignant'. Since the goal was to identify hidden sub-class labels (subtypes) for malignancy and benign nodules, we focused the analysis on the set of nodules that had a higher likelihood of being 'malignant' and 'benign' and removed nodules with malignancy rating 3 (indeterminate). After data preprocessing, we were left with 1,605 nodules, of which 699 were malignant and 906 were benign. In addition to the semantic features listed in Table 4.1, previous studies have generated 64 designed features [132] that describe the intensity, texture, and shape of the nodules.

4.2 Hidden Stratification Discovery on LIDC dataset

Section 4.2.1 presents the results of hidden stratification discovery using an algorithmic measurement approach, while sections 4.2.2 and 4.2.3 demonstrate the results using schema completion.

4.2.1 Clustering-Based Hidden Stratification Discovery

My results demonstrate that the use of UMAP and Gaussian Mixture Modeling-based clustering in the feature space, extracted via deep learning algorithms, enables the capture of the likelihood of malignancy in different groups/subtypes, as perceived by domain experts.

Figure 4.1 (A) shows different silhouette coefficients for various numbers of clusters. The highest silhouette coefficients were obtained when the number of clusters was 2. Figure 4.1 (B) visualizes the two output clusters from the Gaussian Mixture Clustering model in UMAP space. *We can see a clear distinction between malignant and benign superclass labels in UMAP space*. This result indicates that the extracted deep image features reflect the image semantic information. Additionally, this result is consistent with the finding in the literature that UMAP has the advantage of preserving the cluster relationship of data points in the high-dimensional space when compared with other feature reduction techniques, such as t-SNE and PCA [122].

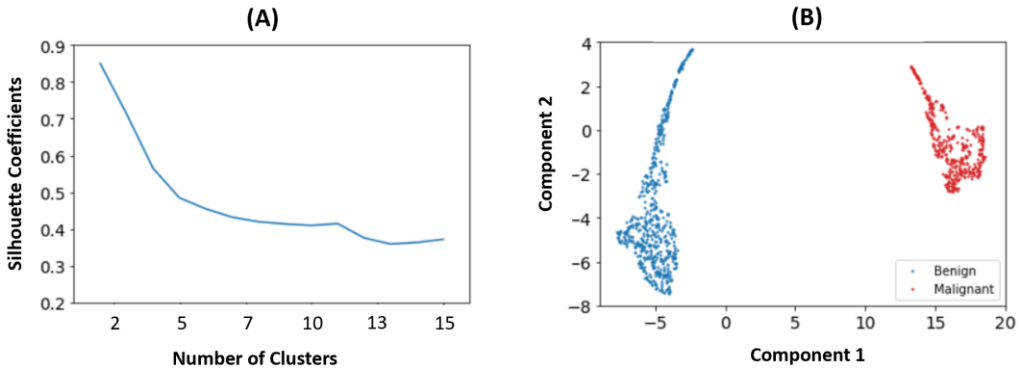

Figure 4.1 Clustering results on all LIDC data points. (A) Silhouette coefficients with various number of clusters; (B) A visualization of two clusters in UMAP space. Blue dots represent benign nodules and red dots represent malignant nodules. Component 1 and Component 2 represent two dimensions reduced from high dimensional image feature space.

The clustering results presented in Figure 4.2 delineate only the super-classes. However, since the goal of this study is to discover hidden subclasses, I further implemented Gaussian Mixture Clustering on benign and malignant nodules separately. The aim was to identify any hidden sub-malignant and sub-benign clusters. Figure 4.2 (A) and Figure 4.2 (B) show that for both benign and malignant nodules, the optimal number of clusters (subgroups) is two. From Figure 4.2 (C) and Figure 4.2 (D), it is evident that there is a clear separation between the two malignant sub-clusters and the two benign sub-clusters. This result implies that there are at least two hidden subgroups within both the benign and malignant

categories. Additionally, UMAP successfully preserves the cluster relationships of data points in the image feature space.

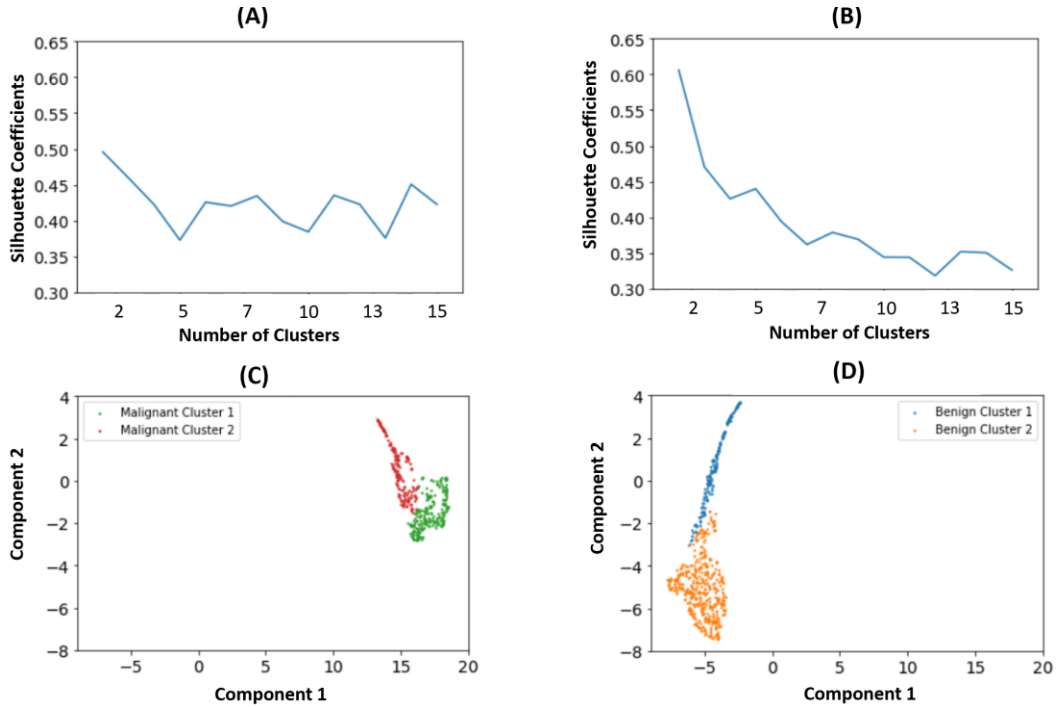

Figure 4.2 Clustering results on benign and malignant lung nodules individually. (A) and (B) show silhouette coefficients with various number of clusters using malignant and benign nodules respectively; (C) and (D) are visualizations of two sub-clusters of malignant and benign nodules in UMAP space. Component 1 and Component 2 represent two dimensions reduced from high dimensional image feature space.

Upon further analysis of the results with respect to the domain expert annotations, it is observed that the malignant and benign sub-clusters exhibit a high correlation with the malignancy ratings provided by radiologists. For instance, one malignant sub-cluster comprises mostly nodules with a malignancy rating of 5

(“highly suspicious”), while another malignant sub-cluster consists mainly of nodules with a malignancy rating of 4 (“moderately suspicious”). Given this correlation between the sub-clusters and the likelihood of malignancy as determined by the domain expert, each sub-cluster can be labeled using the frequency of the malignancy likelihood. These labels are as follows: "Predominantly Moderately Likely Benign" (the majority of nodules with a malignancy rating of 2), "Predominantly Most Likely Benign" (the majority of nodules with a malignancy rating of 1), "Predominantly Moderately Likely Malignant" (the majority of nodules with a malignancy rating of 4), and "Predominantly Most Likely Malignant" (the majority of nodules with a malignancy rating of 5).

To obtain stable clustering labels, I repeated the clustering process (described in Section 3.1.3) 30 times and took the mode of clustering labels across the 30 trials as the final subclass labels. Table 4.2 shows the rating distribution of the generated benign subclasses. It can be observed that 81.33% of the nodules in the Predominantly Moderately Likely Benign cluster were labeled as having a malignancy rating of 2 (“moderately unlikely”). Table 4.3 shows the rating distribution of the generated malignant subclasses. It is observed that 76.57% of the nodules in the Predominantly Moderately Likely Malignant cluster were labeled as having a malignancy rating of 4 (“moderately suspicious”), while 66.52% of the

nodules in the Predominantly Most Likely Malignant cluster were labeled as having a malignancy rating of 5 ("highly suspicious").

Table 4.2 Lung nodule malignancy rating distributions of derived benign subclasses using UMAP embeddings

Malignancy / Cluster Name	Rating 2	Rating 1	Total
Predominantly Moderately Likely Benign	514 (81.33%)	118 (18.67%)	632 (100%)
Predominantly Most Likely Benign	48 (18.11%)	217 (81.89%)	265 (100%)

Table 4.3 Lung nodule malignancy rating distributions of derived malignant subclasses using UMAP embeddings

Malignancy / Cluster Name	Rating 4	Rating 5	Total
Predominantly Moderately Likely Malignant	281 (76.57%)	86 (23.43%)	367 (100%)
Predominantly Most Likely Malignant	75 (33.48%)	149 (66.52%)	224 (100%)

I repeated the same clustering process using PCA embeddings. Table 4.4 and Table 4.5 show the rating distribution of the generated benign subclasses and malignant subclasses, respectively. From Table 4.4, we can see that when using PCA embeddings, we cannot separate the two sub-benign clusters, as both clusters are labeled as 'Predominantly Moderately Likely Benign'. Compared with the malignancy rating distributions obtained using UMAP (Table 4.2), we observe less overlap between the clustered stratification and the malignancy likelihood stratification (Table 4.4). This result is consistent with the finding in the literature

that UMAP has an advantage over PCA in preserving the cluster relationship of data points in high-dimensional space [122].

Table 4.4 Lung nodule malignancy rating distributions of derived benign subclasses using PCA embeddings

Malignancy / **Cluster Name**	**Rating 2**	**Rating 1**	**Total**
Predominantly Moderately Likely Benign	**186 (54.23%)**	157 (45.77%)	343 (100%)
Predominantly Moderately Likely Benign	**376 (67.87%)**	178 (32.13%)	554 (100%)

Table 4.5 Lung nodule malignancy rating distributions of derived malignant subclasses using PCA embeddings

Malignancy / **Cluster Name**	**Rating 4**	**Rating 5**	**Total**
Predominantly Moderately Likely Malignant	272 (72.92%)	101 (27.08%)	373 (100%)
Predominantly Most Likely Malignant	84 (38.53%)	134 (61.47%)	218 (100%)

4.2.2 Spiculation-Malignancy-Based Hidden Stratification Discovery

In addition to generating subclass labels using clustering results, the second and third stratification methods are based on the semantic features provided by radiologists. Previous studies have shown a high positive correlation between malignancy and spiculation [133, 134]. My hypothesis is that machine learning models with ERM loss will more accurately classify spiculated malignant nodules and unspiculated benign nodules, indicating a positive correlation between malignancy and spiculation. Conversely, machine learning models are more likely

to misclassify unspiculated malignant nodules and spiculated benign nodules, indicating a negative correlation between malignancy and spiculation. Based on this hypothesis, the spiculation-malignancy-based subclass labels are: *Unspiculated Benign*, *Spiculated Benign*, *Spiculated Malignant*, and *Unspiculated Malignant*.

Table 4.6 presents the malignant and benign nodule counts in the LIDC dataset. We observe that 79.29% of unspiculated nodules are benign (with malignancy rating 1 or 2) and 76.94% of spiculated nodules are malignant (with malignancy rating 4 or 5).

Table 4.6 Malignancy and spiculation lung nodule counts. Most benign nodules are unspiculated and most malignant nodules are spiculated

Malignancy / **Spiculation**	**Benign (1 or 2)**	**Malignant (4 or 5)**	**Total**
Unspiculated (1)	781(79.29%)	204 (20.71%)	985 (100%)
Spiculated (5)	116 (23.06%)	387 (76.94%)	503 (100%)

4.2.3 Malignancy-Likelihood-Based Hidden Stratification Discovery

By utilizing malignancy ratings directly, we identified malignancy-likelihood-based subclasses as follows: *'Most likely benign'* (rating 1), *'moderately likely benign'* (rating 2), *'moderately likely malignant'* (rating 4), and *'most likely malignant'* (rating 5). Table 4.7 presents the distribution of malignancy ratings, revealing that the majority of nodules in the LIDC dataset are classified as 'moderately likely benign' or 'moderately likely malignant'.

Table 4.7 Malignancy rating distribution.Most nodules are moderately likely benign or malignant

Malignancy / Likelihood	Benign (1 or 2)	Malignant (4 or 5)
Moderately Likely	562 (62.65%)	356 (60.24%)
Most Likely	335 (37.35%)	235 (39.76%)
Total	897 (100%)	591(100%)

4.3 Subtype Learning Results on LIDC Dataset

This section presents the results of binary lung nodule malignancy classification (malignant vs. benign) on the testing dataset using various input features, loss functions, and different sets of subclass labels. Section 4.3.1 illustrates the different input features used for the LIDC dataset, while Sections 4.3.2 to 4.3.4 present the clustering-based, spiculation-malignancy-based, and malignancy-likelihood-based classification results, respectively. The subclass labels were used as one of the inputs to the gDRO model and were also used to calculate the worst group accuracy for model evaluation.

4.3.1 Lung Nodule Malignancy Classification with Different Features and Loss Functions

Figure 4.3 shows experiment designs of training different lung nodule malignancy classification models. Model input include *original cropped nodule images*, *designed features*, *a combination of cropped images* and *designed features*

or *a combination of images and semantic features*. When the model input is cropped nodule images, I utilized a pretrained ResNet CNN architecture [120] (Section 3.2.1); when the input is designed features, I used a fully connected network (Section 3.2.2); and when the input is a combination of designed features and images, or a combination of semantic features and images, I used CompNet architecture [126] (Section 3.2.3). For loss functions, I compared classification results using ERM that ignores the hidden stratification and gDRO that address the hidden stratification problem (Section 3.2.4). *The hypothesis is gDRO model will generate higher worst group accuracy when compared with ERM model.*

Input	Model	Loss Function	Training and Prediction Labels
Images	ResNet	ERM	Superclass Labels
		gDRO	
Designed Features	Fully Connected Networks	ERM	
		gDRO	
Combined Images and Designed Features	CompNet	ERM	
		gDRO	
Combined Images and Semantic Features	CompNet	ERM	
		gDRO	

Figure 4.3 Different model inputs, model architectures and loss functions for lung nodule malignancy classification.

4.3.2 Classification Results with Clustering-Based Subclasses

Comparison between ERM and gDRO models with different model input. Using clustering-based subclass labels (Section 4.2.1), Table 4.8 to Table 4.11 present lung nodule malignancy classification results on testing data with different

model inputs: *images* (Table 4.8), *designed features* (Table 4.9), *a combination of images and designed features* (Table 4.10) and *a combination of images and semantic features* (Table 4.11). The worst performance group, defined as the group with the minimum accuracy using ERM loss, is *Predominantly Moderately Likely Malignant.* Numbers in each cell represent mean accuracy values and numbers in parentheses represent 95% confidence interval across 30 trials. Numbers in bold indicate there is a significant difference between ERM and gDRO models.

Table 4.8 Lung nodule malignancy classification results on testing data using clustering-based subclasses with images as model input

Loss Function / **Accuracy**	**ERM**	**gDRO**
Overall	0.849 (0.842, 0.856)	0.845 (0.839, 0.850)
Predominantly Most Likely Benign	0.986 (0.980, 0.991)	0.981 (0.972, 0.991)
Predominantly Moderately Likely Benign	**0.836 (0.818, 0.855)**	0.803 (0.785, 0.821)
Predominantly Moderately Likely Malignant	0.694 (0.664, 0.724)	**0.735 (0.710, 0.761)**
Predominantly Most Likely Malignant	0.980 (0.971, 0.990)	0.983 (0.975, 0.991)

Table 4.9 Lung nodule malignancy classification results on testing data using clustering-based subclasses with designed features as input

Loss Function / Accuracy	ERM	gDRO
Overall	0.859 (0.852, 0.867)	0.851 (0.841, 0.861)
Predominantly Most Likely Benign	0.981 (0.968, 0.995)	0.965 (0.950, 0.980)
Predominantly Moderately Likely Benign	**0.880 (0.868, 0.891)**	0.827 (0.809, 0.846)
Predominantly Moderately Likely Malignant	0.661 (0.642, 0.680)	**0.728 (0.709, 0.746)**
Predominantly Most Likely Malignant	0.986 (0.979, 0.994)	0.989 (0.984, 0.995)

Table 4.10 Lung nodule malignancy classification results on testing data using clustering-based subclasses with a combination of images and designed features as input.

Loss Function / Accuracy	ERM	gDRO
Overall	0.850 (0.843, 0.857)	0.842 (0.835, 0.848)
Predominantly Most Likely Benign	**0.980 (0.971, 0.989)**	0.965 (0.956, 0.974)
Predominantly Moderately Likely Benign	**0.839 (0.827, 0.851)**	0.812 (0.798, 0.827)
Predominantly Moderately Likely Malignant	0.692 (0.667, 0.716)	**0.726 (0.706, 0.745)**
Predominantly Most Likely Malignant	**0.989 (0.983, 0.995)**	0.972 (0.962, 0.983)

Table 4.11 Lung nodule malignancy classification results on testing data using clustering-based subclasses with a combination of images and semantic features as input

Loss Function / Accuracy	ERM	gDRO
Overall	0.872 (0.865, 0.878)	0.865 (0.857, 0.873)
Predominantly Most Likely Benign	**0.994 (0.990, 0.998)**	0.982 (0.973, 0.992)
Predominantly Moderately Likely Benign	**0.860 (0.847, 0.872)**	0.820 (0.800, 0.840)
Predominantly Moderately Likely Malignant	0.735 (0.717, 0.753)	**0.784 (0.764, 0.803)**
Predominantly Most Likely Malignant	0.988 (0.982, 0.995)	0.991 (0.986, 0.996)

We can observe that, across all types of model inputs, implementing gDRO loss *significantly improves the worst group overall accuracy while maintaining the overall accuracy at the same level*. When comparing the classification results for different input features, we can see that the highest ERM overall accuracy (0.872), the highest gDRO overall accuracy (0.865), the highest ERM worst group accuracy (0.735), and the highest gDRO worst group accuracy (0.784) are achieved using a combination of images and semantic features (Table 4.11).

Comparison between ERM, gDRO and CRIS models. Since the CRIS model requires images as input, CRIS classification results will only be compared with ERM and gDRO models with images as input. Table 4.3.2.5 and Table 4.3.2.6 show the classification results comparison between the ERM model and the CRIS model, and between the gDRO model and the CRIS model, respectively. Numbers

in each cell represent mean accuracy values, and numbers in parentheses represent 95% confidence intervals. Numbers in bold indicate a significant difference between the compared models.

From Table 4.12, we can see that implementing CRIS significantly improves the worst group accuracy, Predominantly Moderately Likely Malignant, but with a decreased overall accuracy when compared with the ERM model. This result is consistent with gDRO results in the literature that most optimization models have the ability to increase the worst group performance but normally at the sacrifice of the overall accuracy. From Table 4.13, we observe that implementing CRIS decreased the overall accuracy and generated a similar worst group accuracy when compared with the gDRO model.

Table 4.12 Lung nodule malignancy classification results comparison between ERM and CRIS models on testing data using clustering-based subclasses

Loss Function / **Accuracy**	**ERM**	**CRIS**
Overall	**0.849 (0.842, 0.856)**	0.835 (0.829, 0.841)
Predominantly Most Likely Benign	0.986 (0.980, 0.991)	0.980 (0.973, 0.987)
Predominantly Moderately Likely Benign	**0.836 (0.818, 0.855)**	0.778 (0.764, 0.792)
Predominantly Moderately Likely Malignant	0.694 (0.664, 0.724)	**0.740 (0.721, 0.760)**
Predominantly Most Likely Malignant	0.980 (0.971, 0.990)	0.985 (0.975, 0.994)

Table 4.13 Lung nodule malignancy classification results comparison between gDRO and CRIS models on testing data using clustering-based subclasses.

Loss Function / Accuracy	gDRO	CRIS
Overall	**0.845 (0.839, 0.850)**	0.835 (0.829, 0.841)
Predominantly Most Likely Benign	0.981 (0.972, 0.991)	0.980 (0.973, 0.987)
Predominantly Moderately Likely Benign	**0.803 (0.785, 0.821)**	0.778 (0.764, 0.792)
Predominantly Moderately Likely Malignant	0.735 (0.710, 0.761)	0.740 (0.721, 0.760)
Predominantly Most Likely Malignant	0.983 (0.975, 0.991)	0.985 (0.975, 0.994)

Comparison between CRIS and CRRIS models. Table 4.14 and Table 4.15 compare the classification results between CRIS and two CRRIS models. One CRRIS model was trained with an ERM model on typical instances and a gDRO model on atypical instances (Table 4.14), while the other CRRIS model was trained with an ERM model on atypical instances and a gDRO model on typical instances (Table 4.15).

The results in Table 4.14 indicate that there is no significant difference between CRIS and the CRRIS model trained on typical instances in terms of overall accuracy and worst group accuracy. However, Table 4.15 shows that the CRRIS model trained on atypical instances outperforms CRIS in terms of worst group accuracy.

Table 4.14 Lung nodule malignancy classification results comparison between CRIS and CRRIS model with an ERM trained on typical instances on testing data using clustering-based subclasses.

Loss Function / Accuracy	CRIS	CRRIS (ERM trained on typical instances)
Overall	0.835 (0.829, 0.841)	0.840 (0.830, 0.850)
Predominantly Most Likely Benign	0.980 (0.973, 0.987)	0.988 (0.982, 0.995)
Predominantly Moderately Likely Benign	0.778 (0.764, 0.792)	**0.799 (0.783, 0.815)**
Predominantly Moderately Likely Malignant	0.740 (0.721, 0.760)	0.737 (0.713, 0.760)
Predominantly Most Likely Malignant	**0.985 (0.975, 0.994)**	0.969 (0.954, 0.983)

Table 4.15 Lung nodule malignancy classification results comparison between CRIS and CRRIS model with an ERM trained on atypical instances on testing data using clustering-based subclasses

Loss Function / Accuracy	CRIS	CRRIS (ERM trained on atypical instances)
Overall	**0.835 (0.829, 0.841)**	0.824 (0.816, 0.832)
Predominantly Most Likely Benign	**0.980 (0.973, 0.987)**	0.952 (0.938, 0.965)
Predominantly Moderately Likely Benign	**0.778 (0.764, 0.792)**	0.753 (0.734, 0.772)
Predominantly Moderately Likely Malignant	0.740 (0.721, 0.760)	**0.754 (0.729, 0.779)**
Predominantly Most Likely Malignant	0.985 (0.975, 0.994)	**0.993 (0.985, 1.000)**

Comparison between ERM, gDRO and CRRIS with ERM trained on atypical instances. Since classification results presented in Table 4.15 implies that CRRIS with ERM trained on atypical instances outperforms CRIS measured by the worst group accuracy, I further compared the classification results between ERM

and CRRIS (Table 4.16) and between gDRO with CRRIS (Table 4.17). We can observe that implementing CRRIS with ERM trained on atypical instances and gDRO trained on typical instances significantly improved the worst group accuracy when compared with base ERM and base gDRO model but with a tradeoff of overall accuracy.

Table 4.16 Lung nodule malignancy classification results comparison between ERM and CRRIS model with an ERM trained on atypical instances on testing data using clustering-based subclasses.

Loss Function / **Accuracy**	**ERM**	**CRRIS (ERM trained on atypical instances)**
Overall	**0.849 (0.842, 0.856)**	0.824 (0.816, 0.832)
Predominantly Most Likely Benign	**0.986 (0.980, 0.991)**	0.952 (0.938, 0.965)
Predominantly Moderately Likely Benign	**0.836 (0.818, 0.855)**	0.753 (0.734, 0.772)
Predominantly Moderately Likely Malignant	0.694 (0.664, 0.724)	**0.754 (0.729, 0.779)**
Predominantly Most Likely Malignant	0.980 (0.971, 0.990)	**0.993 (0.985, 1.000)**

Table 4.17 Lung nodule malignancy classification results comparison between gDRO and CRRIS model with an ERM trained on atypical instances on testing data using clustering-based subclasses.

Loss Function / Accuracy	gDRO	CRRIS (ERM trained on atypical instances)
Overall	**0.845 (0.839, 0.850)**	0.824 (0.816, 0.832)
Predominantly Most Likely Benign	**0.981 (0.972, 0.991)**	0.952 (0.938, 0.965)
Predominantly Moderately Likely Benign	**0.803 (0.785, 0.821)**	0.753 (0.734, 0.772)
Predominantly Moderately Likely Malignant	0.735 (0.710, 0.761)	**0.754 (0.729, 0.779)**
Predominantly Most Likely Malignant	0.983 (0.975, 0.991)	**0.993 (0.985, 1.000)**

Classification results summary using clustering-based subclass labels. First, we observe that implementing the gDRO loss significantly increases the worst group accuracy compared to the ERM model, regardless of the model input (Tables 4.8–11). Second, CRRIS with ERM trained on atypical instances and gDRO trained on typical instances outperforms the ERM model (Table 4.16), the gDRO model (Table 4.17), and the CRIS model (Table 4.15) in terms of worst group accuracy, but at the cost of overall accuracy. This tradeoff is also observed in the literature.

4.3.3 Classification Results Using Spiculation-Malignancy-Based Subclasses

Comparison between ERM and gDRO models with different model input. Tables 4.18 to 4.21 present classification results with different model inputs utilizing spiculation-malignancy-based subclasses. The model inputs include

images (Table 4.18), designed features (Table 4.19), a combination of images and designed features (Table 4.20), and a combination of images and semantic features (Table 4.21). The numbers in each cell represent mean accuracy values, and the numbers in parentheses represent a 95% confidence interval. Numbers in bold indicate a significant difference between the compared models. The worst performance group, defined as the group with the minimum accuracy using ERM loss, is *Unspiculated Malignant* except for one situation when the model input is a combination of images and semantic features; the worst performance group is spiculated benign (Table 4.21).

From Table 4.18 and 4.19, we can observe that implementing gDRO significantly increased the worst group accuracy compared with the ERM model. However, when the input is images, implementing the gDRO loss results in a tradeoff between the worst group accuracy and the overall accuracy (Table 4.18). Table 4.20 and 4.21 show that there are no significant differences between ERM and gDRO when the input features are a combination of images and designed features (Table 4.20) or a combination of images and semantic features (Table 4.21).

Table 4.18 Lung nodule malignancy classification comparison between ERM and gDRO on testing data with images as model input using spiculation-malignancy based subclasses

Loss Function / Accuracy	ERM	gDRO
Overall Accuracy	**0.850 (0.842, 0.858)**	0.838 (0.829, 0.846)
Unspiculated Benign	**0.905 (0.893, 0.916)**	0.862 (0.840, 0.883)
Spiculated Benign	**0.775 (0.741, 0.808)**	0.700 (0.661, 0.739)
Spiculated Malignant	0.846 (0.827, 0.866)	**0.875 (0.853, 0.897)**
Unspiculated Malignant	0.701 (0.676, 0.727)	**0.766 (0.740, 0.791)**

Table 4.19 Lung nodule malignancy classification comparison between ERM and gDRO on testing data with designed features as model input using spiculation-malignancy based subclasses

Loss Function / Accuracy	ERM	gDRO
Overall Accuracy	0.856 (0.848, 0.864)	0.852 (0.841, 0.863)
Unspiculated Benign	**0.923 (0.913, 0.932)**	0.886 (0.866, 0.906)
Spiculated Benign	**0.815 (0.792, 0.838)**	0.756 (0.720, 0.792)
Spiculated Malignant	0.850 (0.833, 0.866)	**0.879 (0.867, 0.892)**
Unspiculated Malignant	0.642 (0.616, 0.667)	**0.729 (0.705, 0.753)**

Table 4.20 Lung nodule classification comparison between ERM and gDRO on testing data with a combination of images and designed features as model input using spiculation-malignancy based subclasses

Loss Function / Accuracy	ERM	gDRO
Overall Accuracy	0.850 (0.842, 0.858)	0.840 (0.823, 0.856)
Unspiculated Benign	0.893 (0.882, 0.904)	0.877 (0.858, 0.897)
Spiculated Benign	0.774 (0.750, 0.799)	0.745 (0.711, 0.779)
Spiculated Malignant	0.861 (0.850, 0.872)	0.859 (0.836, 0.883)
Unspiculated Malignant	0.713 (0.686, 0.741)	0.718 (0.696, 0.740)

Table 4.21 Lung nodule malignancy classification comparison between ERM and gDRO on testing data with a combination of images and semantic features as model input using spiculation-malignancy based subclasses

Loss Function / Accuracy	ERM	gDRO
Overall Accuracy	0.871 (0.864, 0.878)	0.877 (0.869, 0.884)
Unspiculated Benign	0.915 (0.905, 0.924)	0.915 (0.906, 0.925)
Spiculated Benign	0.731 (0.695, 0.767)	0.761 (0.722, 0.801)
Spiculated Malignant	0.889 (0.878, 0.901)	0.891 (0.879, 0.904)
Unspiculated Malignant	0.757 (0.735, 0.779)	0.772 (0.747, 0.797)

Comparison between ERM, gDRO and CRIS models. Table 4.22 and Table 4.23 show a comparison of classification results between the ERM model and the CRIS model, and between the gDRO model and the CRIS model, respectively. The numbers in each cell represent mean accuracy values, and the numbers in

parentheses represent the 95% confidence interval across 30 trials. Numbers in bold indicate a significant difference between the compared models.

From Table 4.22, we can see that implementing CRIS significantly improves the worst group accuracy for Unspiculated Malignant, but with a decreased overall accuracy when compared with the ERM model. Table 4.23 shows that there is no significant difference between the gDRO and CRIS models in terms of overall accuracy and worst group accuracy.

Table 4.22 Lung nodule malignancy classification results comparison between ERM and CRIS models on testing data using spiculation-malignancy based subclasses.

Loss Function / **Accuracy**	**ERM**	**CRIS**
Overall	**0.850 (0.842, 0.858)**	0.834 (0.823, 0.844)
Unspiculated Benign	**0.905 (0.893, 0.916)**	0.854 (0.838, 0.870)
Spiculated Benign	**0.775 (0.741, 0.808)**	0.724 (0.687, 0.760)
Spiculated Malignant	0.846 (0.827, 0.866)	**0.869 (0.854, 0.884)**
Unspiculated Malignant	0.701 (0.676, 0.727)	**0.759 (0.732, 0.786)**

Table 4.23 Lung nodule malignancy classification results comparison between gDRO and CRIS models on testing data using spiculation-malignancy based subclasses.

Loss Function / Accuracy	gDRO	CRIS
Overall Accuracy	0.838 (0.829, 0.846)	0.834 (0.823, 0.844)
Unspiculated Benign	**0.862 (0.840, 0.883)**	0.854 (0.838, 0.870)
Spiculated Benign	0.700 (0.661, 0.739)	**0.724 (0.687, 0.760)**
Spiculated Malignant	0.875 (0.853, 0.897)	0.869 (0.854, 0.884)
Unspiculated Malignant	0.766 (0.740, 0.791)	0.759 (0.732, 0.786)

Comparison between CRIS and CRRIS models. Table 4.24 and Table 4.25 present a classification comparison between CRIS and two CRRIS models. The first CRRIS model was trained with an ERM model on typical instances and a gDRO model on atypical instances (Table 4.24), while the second CRRIS model was trained with an ERM model on atypical instances and a gDRO model on typical instances (Table 4.25).

In Table 4.24, CRIS achieved a higher worst group accuracy than CRRIS, while having a similar overall accuracy with CRRIS that was trained with ERM on typical instances. From Table 4.25, we can observe that CRIS outperforms CRRIS, measured by both the worst group accuracy and overall accuracy. These results suggest that CRIS performs better than CRRIS when subclasses are spiculation-malignancy-based.

Table 4.24 Lung nodule malignancy classification results comparison between CRIS and CRRIS model with an ERM trained on typical instances on testing data using spiculation-malignancy-based subclasses.

Loss Function / Accuracy	CRIS	CRRIS (ERM trained on typical instances)
Overall	0.834 (0.823, 0.844)	0.842 (0.834, 0.850)
Unspiculated Benign	0.854 (0.838, 0.870)	**0.898 (0.884, 0.912)**
Spiculated Benign	0.724 (0.687, 0.760)	**0.780 (0.747, 0.813)**
Spiculated Malignant	**0.869 (0.854, 0.884)**	0.830 (0.812, 0.848)
Unspiculated Malignant	**0.759 (0.732, 0.786)**	0.696 (0.667, 0.724)

Table 4.25 Lung nodule malignancy classification results comparison between CRIS and CRRIS model with an ERM trained on typical instances on testing data using spiculation-malignancy-based subclasses

Loss Function / Accuracy	CRIS	CRRIS (ERM trained on atypical instances)
Overall Accuracy	**0.834 (0.823, 0.844)**	0.824 (0.814, 0.834)
Unspiculated Benign	**0.854 (0.838, 0.870)**	0.834 (0.819, 0.849)
Spiculated Benign	**0.724 (0.687, 0.760)**	0.699 (0.658, 0.741)
Spiculated Malignant	0.869 (0.854, 0.884)	**0.887 (0.874, 0.900)**
Unspiculated Malignant	**0.759 (0.732, 0.786)**	0.748 (0.726, 0.771)

Classification results summary using spiculation-malignancy-based subclass labels. When the model input are images or designed features, gDRO model outperforms ERM model measured by the worst group accuracy (Table 4.18 and Table 4.19). When the model input are images, implementing gDRO decreases the overall accuracy (Table 4.18). When the model input are a combination of

images and designed features or a combination of images and semantic features, there are no significant difference between ERM and gDRO (Table 4.20 and Table 4.21). CRIS model performs equally with gDRO model and CRIS model outperforms CRRIS model. Since CRIS requires less subclass labels during the training, *CRIS model is preferred when the subclass labels are spiculation-malignancy based.*

4.3.4 Classification Results with Malignancy-Likelihood-Based Subclasses

Comparison between ERM and gDRO models with different model input. Table 4.26 to Table 4.29 present classification results using different model inputs: images (Table 4.26), designed features (Table 4.27), a combination of images and designed features (Table 4.28), and a combination of images and semantic features (Table 4.29). The numbers in each cell represent the mean accuracy values, and the numbers in parentheses represent the 95% confidence interval across 30 trials. The numbers in bold indicate a significant difference between ERM and gDRO. The Moderately Suspicious group, defined as the group with the lowest accuracy using ERM loss, showed the worst performance. When images (Table 4.26), designed features (Table 4.27), or a combination of images and semantic features (Table 4.29) were used as input features, gDRO significantly improved the worst group accuracy compared to ERM. However, when the input features were a combination of images and designed features (Table 4.28), there

was no significant difference in the worst group accuracy between ERM and gDRO. There was no significant difference in overall accuracy between ERM and gDRO across all model inputs.

Table 4.26 Lung nodule malignancy classification results comparison on testing data between ERM and gDRO using malignancy-likelihood-based subclasses with images as model input

Loss Function / **Accuracy**	**ERM**	**gDRO**
Overall	0.850 (0.843, 0.856)	0.845 (0.837, 0.853)
Highly Unlikely	0.931 (0.918, 0.945)	0.915 (0.901, 0.928)
Moderately Unlikely	**0.861 (0.847, 0.875)**	0.821 (0.800, 0.842)
Moderately Suspicious	0.688 (0.663, 0.713)	**0.745 (0.719, 0.771)**
Highly Suspicious	0.948 (0.938, 0.958)	**0.952 (0.941, 0.963)**

Table 4.27 Classification results comparison on testing data between ERM and gDRO using malignancy-likelihood-based subclasses with designed features as model input

Loss Function / **Accuracy**	**ERM**	**gDRO**
Overall	0.858 (0.850, 0.865)	0.854 (0.842, 0.865)
Highly Unlikely	0.932 (0.920, 0.943)	0.916 (0.897, 0.935)
Moderately Unlikely	**0.897 (0.885, 0.909)**	0.843 (0.826, 0.860)
Moderately Suspicious	0.645 (0.628, 0.663)	**0.723 (0.706, 0.740)**
Highly Suspicious	0.978 (0.971, 0.986)	**0.988 (0.981, 0.994)**

Table 4.28 Lung nodule malignancy classification results comparison on testing data between ERM and gDRO using malignancy-likelihood-based subclasses with a combination of images and designed features as model input

Loss Function / Accuracy	ERM	gDRO
Overall	0.853 (0.844, 0.862)	0.848 (0.842, 0.854)
Highly Unlikely	0.936 (0.926, 0.946)	0.930 (0.918, 0.943)
Moderately Unlikely	0.853 (0.838, 0.867)	**0.826 (0.81 1, 0.841)**
Moderately Suspicious	0.703 (0.681,0.725)	0.726 (0.708, 0.745)
Highly Suspicious	0.963 (0.953,0.973)	0.965 (0.953, 0.976)

Table 4.29 Lung nodule classification results comparison on testing data between ERM and gDRO using malignancy-likelihood-based subclasses with a combination of images and semantic features as model input

Loss Function / Accuracy	ERM	gDRO
Overall	0.874 (0.867, 0.880)	0.872 (0.863, 0.881)
Highly Unlikely	0.948 (0.938, 0.958)	0.937 (0.924, 0.950)
Moderately Unlikely	0.864 (0.852, 0.876)	0.850 (0.832, 0.867)
Moderately Suspicious	0.750 (0.731, 0.770)	**0.777 (0.758, 0.795)**
Highly Suspicious	**0.978 (0.968, 0.988)**	0.976 (0.968, 0.983)

Comparison between ERM, gDRO and CRIS model. Table 4.30 and Table 4.31 compare the classification results using malignancy-based subclasses between ERM and CRIS and between gDRO and CRIS, respectively. It can be observed that implementing CRIS significantly increased the accuracy of the worst

group (moderately suspicious) at the expense of the overall accuracy when compared to the ERM and gDRO models.

Table 4.30 Lung nodule malignancy classification results comparison on testing data between ERM and CRIS model using malignancy-likelihood-based subclasses.

Loss Function / Accuracy	ERM	CRIS
Overall	**0.850 (0.843, 0.856)**	0.839 (0.832, 0.846)
Highly Unlikely	**0.931 (0.918, 0.945)**	0.904 (0.891, 0.917)
Moderately Unlikely	**0.861 (0.847, 0.875)**	0.804 (0.786, 0.822)
Moderately Suspicious	0.688 (0.663, 0.713)	**0.752 (0.732, 0.772)**
Highly Suspicious	**0.948 (0.938, 0.958)**	0.959 (0.947, 0.971)

Table 4.31 Classification results comparison on testing data between gDRO and CRIS model using malignancy-likelihood-based subclasses.

Loss Function / Accuracy	gDRO	CRIS
Overall	**0.845 (0.837, 0.853)**	0.839 (0.832, 0.846)
Highly Unlikely	**0.915 (0.901, 0.928)**	0.904 (0.891, 0.917)
Moderately Unlikely	**0.821 (0.800, 0.842)**	0.804 (0.786, 0.822)
Moderately Suspicious	0.745 (0.719, 0.771)	**0.752 (0.732, 0.772)**
Highly Suspicious	**0.952 (0.941, 0.963)**	0.959 (0.947, 0.971)

Comparison between CRIS and CRRIS models. Table 4.32 and Table 4.33 compare the classification results between CRIS and two CRRIS models. One CRRIS model uses an ERM model trained on typical instances and a gDRO model trained on atypical instances (Table 4.32), while the other CRRIS model uses an

ERM model trained on atypical instances and a gDRO model trained on typical instances (Table 4.33). Table 4.32 shows that CRIS outperforms the CRRIS model with ERM trained on typical instances in terms of the worst group accuracy. Table 4.33 indicates that while CRIS and CRRIS with ERM trained on atypical instances perform equally on the worst group accuracy, CRIS achieves a higher overall accuracy than CRRIS.

Table 4.32 Lung nodule malignancy classification results comparison between CRIS and CRRIS model with an ERM trained on typical instances on testing data using malignancy-based subclasses.

Loss Function / **Accuracy**	**CRIS**	**CRRIS (ERM trained on typical instances)**
Overall	0.839 (0.832, 0.846)	0.844 (0.835, 0.853)
Highly Unlikely	0.904 (0.891, 0.917)	0.919 (0.905, 0.933)
Moderately Unlikely	0.804 (0.786, 0.822)	**0.841 (0.824, 0.858)**
Moderately Suspicious	**0.752 (0.732, 0.772)**	0.713 (0.689, 0.736)
Highly Suspicious	0.959 (0.947, 0.971)	0.943 (0.931, 0.955)

Table 4.33 Lung nodule malignancy classification results comparison between CRIS and CRRIS model with an ERM trained on atypical instances on testing data using malignancy-based subclasses.

Loss Function / Accuracy	CRIS	CRRIS (ERM trained on atypical instances)
Overall	**0.839 (0.832, 0.846)**	0.817 (0.808, 0.827)
Highly Unlikely	**0.904 (0.891, 0.917)**	0.884 (0.866, 0.902)
Moderately Unlikely	**0.804 (0.786, 0.822)**	0.748 (0.721, 0.775)
Moderately Suspicious	0.752 (0.732, 0.772)	0.756 (0.732, 0.781)
Highly Suspicious	0.959 (0.947, 0.971)	**0.978 (0.969, 0.986)**

Classification results summary using malignancy-likelihood-based subclass labels. When the input features are images (Table 4.26) or designed features (Table 4.27) or a combination of images and semantic ratings (Table 4.29), gDRO significantly improves the worst group accuracy compared to ERM. However, when the input features are a combination of images and designed features (Table 4.28), there is no significant difference in the worst group accuracy between ERM and gDRO. Across all model inputs, there is no significant difference in the overall accuracy between ERM and gDRO. CRIS outperforms ERM and gDRO in terms of the worst group accuracy (Table 4.30 and Table 4.31) and achieves higher worst group accuracy and overall accuracy than two CRRIS models (Table 4.32 and Table 4.33), respectively. When subclass labels are malignancy-likelihood-based, CRIS is the preferred model for generalization.

4.4 Analysis of Results

In this section, I will first analyze my hidden stratification discovery and subgroup learning results, then, I will discuss my findings on three additional research questions. The first question is related to the performance of the trained ERM and gDRO models on nodules that have been confirmed as pathological (Section 4.4.2). The second question explores whether incorporating semantic features to select representative or atypical instances can improve the performance of the CRRIS model (Section 4.4.3). Lastly, the third question examines whether utilizing a pretrained model on a medical dataset can enhance the transfer learning process and lead to improved model performance (Section 4.4.4).

4.4.1 Hidden Stratification Discovery and Subgroup Learning Results Analysis

First, my results show that hidden stratification exists, as indicated by the different ERM model performances across stratification groups (e.g., see Table 4.8). A high degree of overlap between the clustered stratification and malignancy likelihood stratification (Tables 4.2 and 4.3) implies that the subclasses generated from clustering are semantically meaningful.

Second, incorporating gDRO loss into lung nodule malignancy classification models can significantly increase the worst group performance when subclass labels are clustering-based. When subclasses are spiculation-malignancy-based or

malignancy-likelihood-based, there are some exceptions when we use a combination of images and designed features or a combination of images and semantic features as model input (Tables 4.20 and 4.21). Both input features and stratification methods have a direct influence on classification results.

Third, using a combination of images and semantic features as the model input improves the overall accuracy and the worst group accuracy compared to other input features (Tables 4.8 to 4.11). Additionally, when the model input is a combination of images and semantic features, the ERM model generates higher accuracies on two benign subclasses, and the gDRO model produces higher accuracies on two malignant subclasses (Table 4.11).

Fourth, across all experimental designs, only when the subclasses are clustering-based, CRRIS with an ERM model trained on atypical instances achieved higher worst group accuracy with a tradeoff of overall accuracy compared to CRIS (Table 4.15). There is no evidence supporting the hypothesis that the CRRIS model can further boost the model generalization compared to the CRIS model. When we compare CRIS with gDRO, we can observe a significant increase of the worst group accuracy by implementing CRIS only when subclasses are malignancy-based (Table 4.31).

4.4.2 Lung Nodule Malignancy Classification Results with Pathological-Proven Examination Labels

The LIDC dataset also provides patient-level pathological-proven examination labels. However, since this study utilizes nodule level data, only patients with a single nodule were selected for analysis. A total of 34 patients met this criterion and their patient-level pathological-proven examination labels were used as the malignancy ground truth. Among these 34 nodules, 15 were excluded due to an indeterminate malignancy rating of 3, leaving 19 nodules for evaluating our model's ability to predict the pathological-proven labels. These 19 nodules were excluded from the training dataset and only used during the prediction phase.

Table 4.34 displays the correlation between radiologists' assessments and pathological-proven examination labels. The table shows that the false positive rate for malignancy semantic ratings is 33.33% (2 out of 6), while the false negative rate is 53.85% (7 out of 13).

Table 4.34 Relationship between pathological-proven lung nodule malignancy labels and semantic rating

Mode Rating Labels / **Pathological-proven Labels**	**Benign**	**Malignant**
Benign	4	2
Malignant	7	6

Table 4.35 illustrates the connection between pathological-proven labels and ERM prediction labels. After running the classification model 30 times, I obtained the same confusion matrix in every trial. The table indicates that the false positive

rate is 33.33% and the false negative rate is 38.46%. Comparing these results to those in Table 4.34, we can conclude that the ERM model significantly reduces the false negative rate from 53.85% to 38.46%, while maintaining the false positive rate when compared to radiologists' semantic ratings.

Table 4.35 Relationship between pathological-proven malignancy labels and ERM prediction labels

ERM Predicted Labels / **Pathological-proven Labels**	**Benign**	**Malignant**
Benign	4	2
Malignant	5	8

Table 4.36 displays the correlation between pathological-proven labels and gDRO prediction labels. Similar to the previous models, I ran the classification model 30 times and obtained the same confusion matrix in each trial. From Table 4.36, we can observe that the false positive rate is 33.33% and the false negative rate is 15.38%. Comparing these results to those in Tables 4.34 and 4.35, it is evident that the gDRO model provides a further reduction in the false negative rate.

Table 4.36 Relationship between pathological-proven lung nodule malignancy labels and gDRO prediction labels

Mode Rating Labels / **Pathological-proven Labels**	**Benign**	**Malignant**
Benign	4	2
Malignant	2	11

The results presented in Tables 4.34 to 4.36 indicate that in the LIDC dataset, radiologists annotated a significant number of false negative labels (Table 4.34). Using a CNN classifier with ERM or gDRO loss can help reduce false negative rate (Table 4.35 and Table 4.36).

In terms of running the CRIS model across 30 trials and comparing the pathological-proven labels and prediction labels, I obtained different confusion matrices. Table 4.37 and Table 4.38 present the confusion matrix with the highest overall accuracy and the confusion matrix with the lowest overall accuracy across the 30 trials. In the highest overall accuracy scenario, the false positive rate and false negative rate are both 33.33%. However, in the lowest overall accuracy scenario, the false positive rate and false negative rate are 50% and 53.85%, respectively.

When comparing these results to the ones obtained using human annotated semantic ratings (Table 4.34), it can be observed that the CRIS model does not necessarily decrease the false negative rate as indicated by the ERM and gDRO model results. In the highest overall accuracy scenario, the CRIS model decreases false negative rates. However, in the scenario with the lowest overall accuracy, it actually increases the false positive rate and maintains the false negative rate the same when compared with the human-annotated results.

Table 4.37 Relationship between pathological-proven lung nodule malignancy labels and CRIS prediction labels (the highest overall accuracy situation)

Mode Rating Labels / **Pathological-proven Labels**	**Benign**	**Malignant**
Benign	4	2
Malignant	3	10

Table 4.38 Relationship between pathological-proven lung nodule malignancy labels and CRIS prediction labels (the lowest overall accuracy situation)

Mode Rating Labels / **Pathological-proven Labels**	**Benign**	**Malignant**
Benign	3	3
Malignant	7	6

Table 4.39 and Table 4.40 present the results of lung nodule malignancy classification on pathologically proven lung nodules using the CRRIS model with an ERM model trained on atypical instances. The tables showcase the outcomes under the highest overall accuracy and the lowest overall accuracy scenarios across 30 trials. A similar pattern is observed with the CRIS model, where under the highest overall accuracy scenario, the CRRIS model with an ERM model trained on atypical instances significantly reduces the false negative rate from 53.85% (Table 4.34) to 23.08% when compared to human annotated ratings. However, in the lowest overall accuracy scenario, the CRRIS model with an ERM model trained on atypical instances increases the false positive rate from 33.33% to 66.67% while maintain the false negative the same.

Table 4.39 Relationship between pathological-proven lung nodule malignancy labels and CRRIS with an ERM trained on atypical instances prediction labels (the highest overall accuracy situation)

Mode Rating Labels / **Pathological-proven Labels**	**Benign**	**Malignant**
Benign	4	2
Malignant	3	10

Table 4.40 Relationship between pathological-proven lung nodule malignancy labels and CRRIS with an ERM trained on atypical instances prediction labels (the lowest overall accuracy situation)

Mode Rating Labels / **Pathological-proven Labels**	**Benign**	**Malignant**
Benign	2	4
Malignant	7	6

Table 4.41 and Table 4.42 show the results of lung nodule malignancy classification on pathologically proven lung nodules using the CRRIS model with an ERM model trained on typical instances. Table 4.41 represents a confusion matrix with the highest overall accuracy while Table 4.42 represents a confusion matrix with the lowest overall accuracy. We can observe that under the highest overall accuracy scenario, the CRRIS model trained on typical instances significantly reduces the false negative rate from 53.85% (Table 4.34) to 15.38% when compared to human annotated ratings. Under the lowest overall accuracy scenario, the CRRIS model trained on typical instances decreases the false negative

rate from 53.85% to 38.46%, but increase the false positive rate from 33.33% to 50%.

Table 4.41 Relationship between pathological-proven lung nodule malignancy labels and CRRIS with an ERM trained on typical instances prediction labels (the highest overall accuracy situation)

Mode Rating Labels / **Pathological-proven Labels**	**Benign**	**Malignant**
Benign	4	2
Malignant	2	11

Table 4.42 Relationship between pathological-proven lung nodule malignancy labels and CRRIS with an ERM trained on typical instances prediction labels (the lowest overall accuracy situation)

Mode Rating Labels / **Pathological-proven Labels**	**Benign**	**Malignant**
Benign	3	3
Malignant	5	8

4.5.2 CRRIS Model using Semantic Features to Choose Representative Cases

Section 3.3.2 explains how I selected representative instances based on the Euclidean distance between the deep image feature vector of each instance and the deep image feature vector of the cluster center. As the LIDC dataset provides semantic features, I hypothesized that using semantic features to select representative cases (*Semantic-CRRIS*) would lead to a further increase in model performance, as measured by both the worst group accuracy and overall accuracy. More specifically, I compared the Jaccard distance of semantic features of each

lung nodule and of the cluster center. An instance is a representative case if the Jaccard distance between its semantic features and the semantic features of the cluster center is within the smallest p %, where p is a hyperparameter. I chose p as 50 since it gave me the best performance on the validation set. For all the tables in this section, numbers in each cell represent mean accuracy values and numbers in parentheses represent 95% confidence interval. Numbers in bold represent there is a significant difference between compared models.

Table 4.43 displays a comparison of lung nodule malignancy classification between CRIS and Semantic-CRRIS using clustering-based subclass labels. The results indicate that while overall accuracy remains unchanged, semantic-CRRIS significantly enhances the performance of the worst group.

Table 4.43 Lung nodule malignancy classification results comparison between CRIS and Semantic-CRRIS model with an ERM trained on atypical instances on testing data using clustering-based subclasses.

Loss Function / **Accuracy**	**CRIS**	**Semantic-CRRIS (ERM trained on atypical instances)**
Overall	0.835 (0.829, 0.841)	0.831 (0.822, 0.840)
Predominantly Most Likely Benign	0.980 (0.973, 0.987)	0.961 (0.947, 0.976)
Predominantly Moderately Likely Benign	0.778 (0.764, 0.792)	0.768 (0.750, 0.786)
Predominantly Moderately Likely Malignant	0.740 (0.721, 0.760)	**0.764 (0.743, 0.785)**
Predominantly Most Likely Malignant	0.985 (0.975, 0.994)	0.974 (0.959, 0.989)

Table 4.44 presents a comparison of lung nodule malignancy classification between CRRIS and Semantic-CRRIS. Unlike CRRIS, which selects representative instances using image feature Euclidean distance, Semantic-CRRIS improves the accuracy of both the worst and overall groups, albeit without statistical significance.

Table 4.44 Lung nodule malignancy classification results comparison between CRRIS and Semantic CRRIS model with an ERM trained on atypical instances on testing data using clustering-based subclasses.

Loss Function / **Accuracy**	**CRRIS (ERM trained on atypical instances)**	**Semantic-CRRIS (ERM trained on atypical instances)**
Overall	0.824 (0.816, 0.832)	0.831 (0.822, 0.840)
Predominantly Most Likely Benign	0.952 (0.938, 0.965)	0.961 (0.947, 0.976)
Predominantly Moderately Likely Benign	0.753 (0.734, 0.772)	0.768 (0.750, 0.786)
Predominantly Moderately Likely Malignant	0.754 (0.729, 0.779)	0.764 (0.743, 0.785)
Predominantly Most Likely Malignant	0.993 (0.985, 1.000)	0.974 (0.959, 0.989)

4.5.3 Transfer Learning with a Pretrained Model on a Medical Related Dataset

In Section 3.2.1, a transfer learning approach was introduced, utilizing a pretrained model on ImageNet, a large image database that does not include medical-related images such as CT, MRI, and X-ray. Since this study focuses on applications in the medical domain, an investigation was conducted to determine if employing a pretrained model on a medical-related database could further enhance

the model's performance. Mei et al.'s RadImageNet [135] dataset, comprising 1.35 million annotated medical CT scans and MRIs from 131,872 patients, was utilized for this purpose. Previous research [135] has indicated that pretrained models using RadImageNet outperform models trained on ImageNet for medical-related classification tasks. Hence, an evaluation of lung nodule classification performance was conducted using RadImageNet, and the results were compared with those obtained using ImageNet. It is noteworthy that all the pretrained models on RadImageNet employed a ResNet50 architecture, unlike the ResNet18 architecture used in this study. To ensure a fair comparison, the classification was re-run using a pretrained ResNet50 model on ImageNet.

Table 4.45 presents the results of lung nodule classification using ERM loss. It can be observed that for the LIDC dataset, there is no significant difference between utilizing a pretrained ResNet50 model on ImageNet and using a pretrained ResNet50 model on RadImageNet. The same conclusion can be drawn when employing the gDRO loss, as shown in Table 4.46.

Table 4.45 Comparison of lung nodule malignancy classification results using ERM loss between a pretrained ResNet50 model on ImageNet and a pretrained ResNet50 model on RadImageNet

Pretrained Dataset / **Accuracy**	**ImageNet**	**RadImageNet**
Overall	0.849 (0.842, 0.855)	0.856 (0.849, 0.863)
Predominantly Most Likely Benign	0.986 (0.979, 0.993)	0.993 (0.989, 0.997)
Predominantly Moderately Likely Benign	0.834 (0.814, 0.854)	0.850 (0.835, 0.865)
Predominantly Moderately Likely Malignant	0.695 (0.671, 0.719)	0.692 (0.669, 0.716)
Predominantly Most Likely Malignant	0.983 (0.975, 0.991)	0.982 (0.973, 0.991)

Table 4.46 Comparison of lung nodule malignancy classification results using gDRO loss between a pretrained ResNet50 model on ImageNet and a pretrained ResNet50 model on RadImageNet

Pretrained Dataset / **Accuracy**	**ImageNet**	**RadImageNet**
Overall	0.840 (0.829, 0.850)	0.840 (0.831, 0.849)
Predominantly Most Likely Benign	0.969 (0.958, 0.981)	0.973 (0.963, 0.983)
Predominantly Moderately Likely Benign	0.784 (0.756, 0.811)	0.795 (0.775, 0.816)
Predominantly Moderately Likely Malignant	0.758 (0.735, 0.781)	0.738 (0.713, 0.762)
Predominantly Most Likely Malignant	0.985 (0.973, 0.996)	0.979 (0.971, 0.987)

CHAPTER 5. Applications to Breast Cancer

5.1 Breast Cancer Histopathological Database (BreakHis)

Breast Cancer Histopathological Database (BreakHis) [136] comprises microscopic images of breast tumor tissue obtained from 82 patients using varying magnification levels (40X, 100X, 200X, and 400X). In this study, I evaluated my methodologies on 40X images. All BreakHis images can be divided into two superclass categories: Benign and Malignant. Within the benign category, images were further labeled as one of the four subclass categories: Adenosis (A), Fibroadenoma (F), Tubular Adenoma (TA), and Phyllodes Tumor (PT). Within the malignant category, images were further labeled as either Ductal Carcinoma (DC), Lobular Carcinoma (LC), Mucinous Carcinoma (MC), or Papillary Carcinoma (PC). Table 5.1 summarizes the distribution of 40X microscopic images. I resized all images to 460*700 and implemented standardized normalization for each image.

Table 5.1 BreakHis Image malignancy subclass Distribution

Superclass Labels	Subclass Labels	Number of Images
Benign	Adenosis (A)	114
	Fibroadenoma (F)	253
	Tubular Adenona (TA)	109
	Phyllodes Tumor (PT)	149
Malignant	Ductal Carcinoma (DC)	864
	Lobular Carcinoma (LC)	156
	Mucinous Carcinoma (MC)	205
	Papillary Carcinoma (PC)	145

5.2 Clustering-Based Hidden Stratification Discovery on BreakHis Dataset

The BreakHis dataset lacks the semantic features present in the LIDC dataset. Therefore, the spiculation-malignancy-based and malignancy-likelihood-based subclass labels introduced in Sections 4.2.2 and 4.2.3 are not available. However, the BreakHis dataset does provide subgroup labels within the malignant and benign categories (Table 5.1). These subgroup labels can be used to evaluate clustering-based hidden stratification discovery results.

Figure 5.1 (A) shows different silhouette coefficients for various numbers of clusters. The highest silhouette coefficients were obtained when the number of clusters was 2. Figure 5.1 (B) visualizes the two output clusters from the Gaussian Mixture Clustering model in UMAP space. Similar to the clustering result on all data points in the LIDC dataset (Figure 4.1), *we can see a clear separation between the malignant and benign superclass labels*.

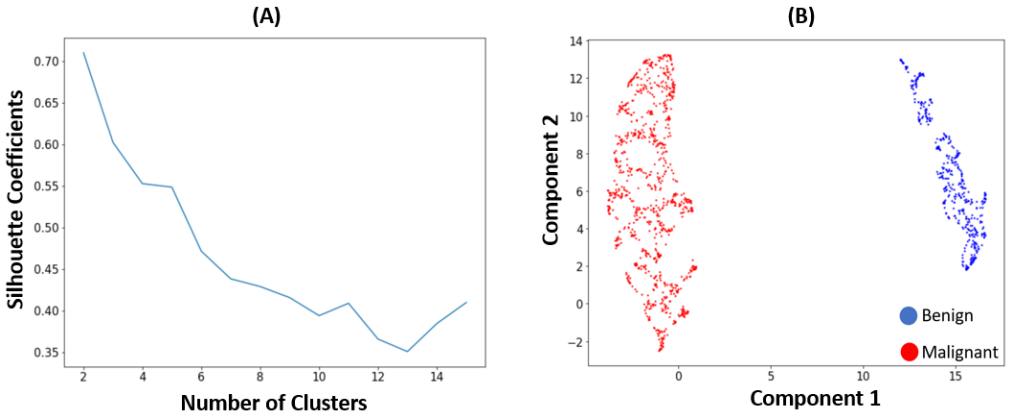

Figure 5.1 Clustering results for all data points in BreakHis. (A) Silhouette Coefficients for various numbers of clusters; (B) Visualization of two clusters in UMAP Space. Blue dots represent benign microscopic images and red dots represent malignant microscopic images. Components 1 and 2 represent two dimensions reduced from high-dimensional image feature space.

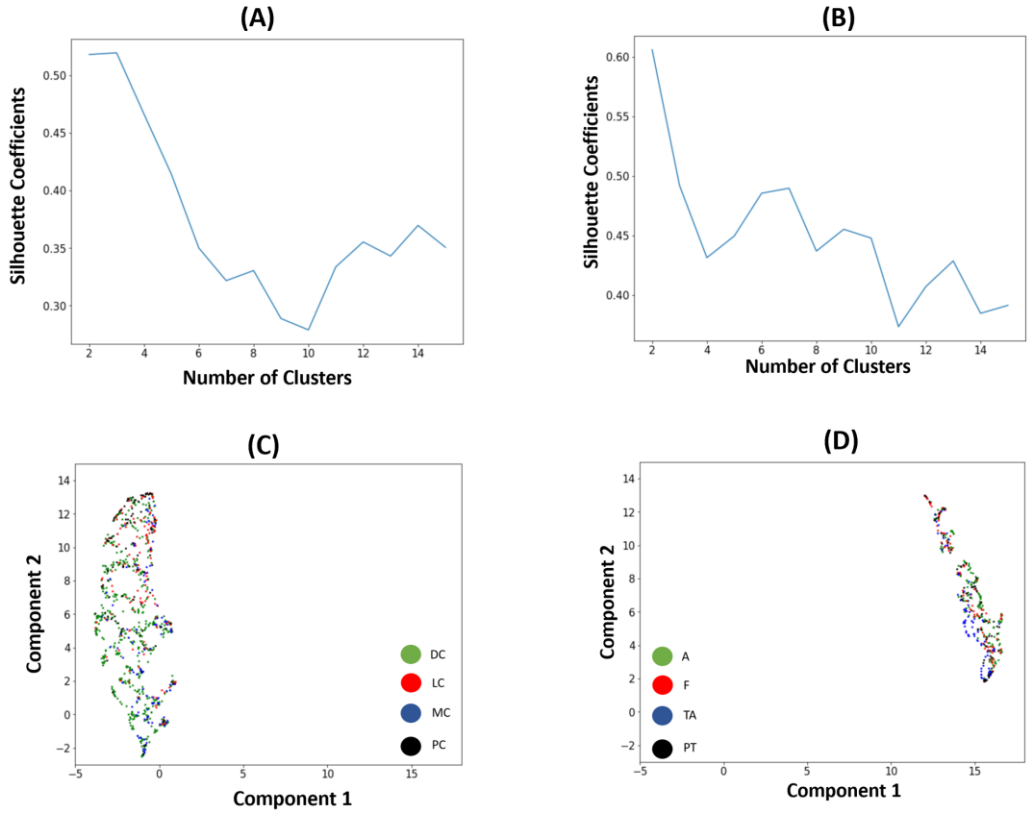

Figure 5.2 Clustering results and UMAP visualization on benign and malignant histopathological images. (A) and (B) show silhouette coefficients with various number of clusters using malignant and benign images respectively; (C) and (D) are visualizations of four subgroups labels by domain experts in UMAP space. Component 1 and Component 2 correspond to the two dimensions in which the high-dimensional image features were reduced using UMAP.

Figure 5.2 (A) shows that for malignant images, the optimal number of clusters is three and Figure 5.2 (B) indicates the optimal number of clusters is two for benign images. From Figure 5.2 (C) and Figure 5.2 (D), we see that there is no clear separation between ground truth subgroups in UMAP space. This result implies that the *subgroups generated from the clustering-based approach (three malignant subgroups and two benign subgroups) cannot be perceived by domain experts*.

Since there were no other ways to validate subgroups generated from the clustering-based approach, I chose to use the ground truth subgroup labels provided by domain experts directly for subgroup learning.

I further extracted a set of deep image features from a multi-class classification model trained with eight ground truth subclass labels listed in Table 5.1. Figure 5. 3 shows that in the UMAP space, we can clearly see a separation between eight ground truth subgroups (with a slight overlap between MC and PC).

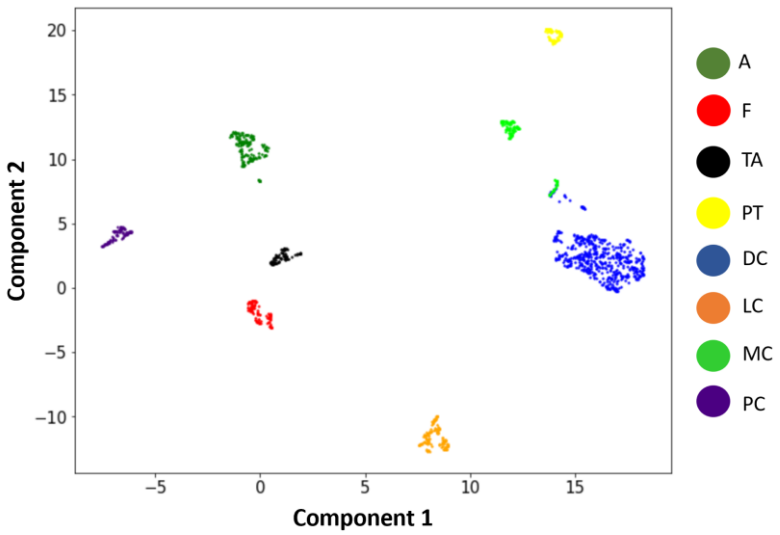

Figure 5.3 Two dimensional UMAP visualization for BreakHis dataset.The input of UMAP is deep image features extracted from a multiclass classifier. Components 1 and 2 correspond to the two dimensions in which the high-dimensional image features were reduced using UMAP.

5.3 Subtype Learning Results on BreakHis

This section first reports the binary microscopic image malignancy classification results (malignant vs. benign) on the testing dataset using superclass labels. After observing that there is no significant latent stratification phenomenon for the superclass classification problem in the BreakHis dataset, I further investigated the multi-class classification results of classifiers trained with subclass labels listed in Table 5.1. For all the tables in this section, numbers in each cell represent mean accuracy values and numbers in parentheses represent 95% confidence interval. Numbers in bold represent there is a significant difference between compared models.

Comparison between ERM and gDRO models. Table 5.2 presents the results of binary microscopic image classification (malignant vs. benign) on the testing data. Unlike what we observed for the LIDC dataset, we did not find a significant hidden stratification phenomenon for the BreakHis dataset. The worst group (Adenosis) still has a mean overall accuracy 0.893. Implementing gDRO did not lead to a significant improvement in model performance, as measured by both the overall accuracy and the worst group accuracy.

Table 5.2 Binary histopathological image malignancy classification results on the testing data

Loss Function / Accuracy	ERM	gDRO
Overall	0.953 (0.945, 0.960)	0.947 (0.939, 0.955)
Fibroadenoma (Benign)	0.955 (0.932, 0.978)	0.979 (0.968, 0.990)
Tubular adenoma (Benign)	0.938 (0.909, 0.967)	0.947 (0.921, 0.972)
Adenosis (Benign)	0.893 (0.854, 0.932)	0.929 (0.908, 0.949)
Phyllodes tumor (Benign)	0.934 (0.904, 0.963)	0.947 (0.924, 0.970)
Ductal carcinoma (Malignant)	0.996 (0.993, 0.998)	0.971 (0.960, 0.982)
Mucinous carcinoma (Malignant)	0.951 (0.926, 0.977)	0.955 (0.935, 0.974)
Lobular carcinoma (Malignant)	0.978 (0.962, 0.995)	0.964 (0.945, 0.983)
Papillary carcinoma (Malignant)	0.976 (0.961, 0.992)	0.970 (0.951, 0.989)

Table 5.3 shows the multi-class microscopic image classification results, and the training subclass labels are listed in Table 5.1. We can observe that the Lobular carcinoma subgroup has a significantly worse classification accuracy when compared with other subgroups, indicating a hidden stratification phenomenon. Implementing gDRO significantly increases the accuracy of the Lobular carcinoma subgroup, which had the worst performance in the initial model.

Table 5.3 Subclass histopathological image classification results on the testing data

Loss Function / Accuracy	ERM	gDRO
Overall	0.911 (0.905, 0.917)	0.903 (0.894, 0.911)
Fibroadenoma (Benign)	0.970 (0.957, 0.983)	0.948 (0.927, 0.969)
Tubular adenoma (Benign)	0.915 (0.889, 0.941)	0.907 (0.879, 0.935)
Adenosis (Benign)	0.928 (0.897, 0.960)	0.913 (0.874, 0.952)
Phyllodes tumor (Benign)	0.943 (0.919, 0.968)	0.910 (0.878, 0.943)
Ductal carcinoma (Malignant)	0.942 (0.932, 0.952)	0.929 (0.917, 0.941)
Mucinous carcinoma (Malignant)	0.943 (0.919, 0.967)	0.918 (0.884, 0.952)
Lobular carcinoma (Malignant)	0.769 (0.730, 0.809)	**0.849 (0.807, 0.891)**
Papillary carcinoma (Malignant)	0.906 (0.876, 0.935)	0.902 (0.877, 0.926)

Comparison between ERM, gDRO and CRIS models. Table 5.4 shows the classification comparison between the ERM and CRIS models. We can observe that the CRIS model significantly improves the accuracy of the worst-performing group (Lobular Carcinoma) while maintaining overall accuracy.

Table 5.4 Subclass histopathological image classification comparison between ERM and CRIS models on the testing dataset

Loss Function / Accuracy	ERM	CRIS
Overall	0.911 (0.905, 0.917)	0.903 (0.895, 0.911)
Fibroadenoma (Benign)	0.970 (0.957, 0.983)	0.956 (0.941, 0.971)
Tubular adenoma (Benign)	0.915 (0.889, 0.941)	0.895 (0.860, 0.930)
Adenosis (Benign)	0.928 (0.897, 0.960)	0.890 (0.854, 0.926)
Phyllodes tumor (Benign)	**0.943 (0.919, 0.968)**	0.906 (0.868, 0.945)
Ductal carcinoma (Malignant)	**0.942 (0.932, 0.952)**	0.917 (0.903, 0.931)
Mucinous carcinoma (Malignant)	0.943 (0.919, 0.967)	0.911 (0.878, 0.944)
Lobular carcinoma (Malignant)	0.769 (0.730, 0.809)	**0.815 (0.766, 0.865)**
Papillary carcinoma (Malignant)	0.906 (0.876, 0.935)	0.881 (0.850, 0.912)

Table 5.5 presents the classification comparison between the gDRO and CRIS models. We can see that CRIS model significantly improves the worst group accuracy while maintaining overall accuracy.

Table 5.5 Subclass histopathological image classification comparison between ERM and CRIS Models on the testing dataset

Loss Function / Accuracy	gDRO	CRIS
Overall	0.911 (0.905, 0.917)	0.903 (0.895, 0.911)
Fibroadenoma (Benign)	0.970 (0.957, 0.983)	0.956 (0.941, 0.971)
Tubular adenoma (Benign)	0.915 (0.889, 0.941)	0.895 (0.860, 0.930)
Adenosis (Benign)	0.928 (0.897, 0.960)	0.890 (0.854, 0.926)
Phyllodes tumor (Benign)	**0.943 (0.919, 0.968)**	0.906 (0.868, 0.945)
Ductal carcinoma (Malignant)	**0.942 (0.932, 0.952)**	0.917 (0.903, 0.931)
Mucinous carcinoma (Malignant)	0.943 (0.919, 0.967)	0.911 (0.878, 0.944)
Lobular carcinoma (Malignant)	0.769 (0.730, 0.809)	**0.815 (0.766, 0.865)**
Papillary carcinoma (Malignant)	0.906 (0.876, 0.935)	0.881 (0.850, 0.912)

Table 5.6 compares the classification result between CRIS model and CRRIS model with an ERM model trained on atypical instances. We can observe that although the CRRIS model significantly improves the worst group's performance, it decreases the overall accuracy.

Table 5.6 Subclass histopathological image classification comparison between CRIS and CRIS model with an ERM trained on atypical instances

Loss Function / Accuracy	CRIS	CRRIS (ERM trained on atypical instances)
Overall	**0.903 (0.895, 0.911)**	0.886 (0.877, 0.895)
Fibroadenoma (Benign)	**0.956 (0.941, 0.971)**	0.932 (0.915, 0.949)
Tubular adenoma (Benign)	0.895 (0.860, 0.930)	0.894 (0.863, 0.924)
Adenosis (Benign)	**0.890 (0.854, 0.926)**	0.847 (0.797, 0.897)
Phyllodes tumor (Benign)	0.906 (0.868, 0.945)	0.920 (0.896, 0.943)
Ductal carcinoma (Malignant)	**0.917 (0.903, 0.931)**	0.888 (0.872, 0.905)
Mucinous carcinoma (Malignant)	**0.911 (0.878, 0.944)**	0.854 (0.813, 0.894)
Lobular carcinoma (Malignant)	0.815 (0.766, 0.865)	**0.890 (0.855, 0.926)**
Papillary carcinoma (Malignant)	0.906 (0.876, 0.935)	0.891 (0.860, 0.922)

5.4 Analysis of Results

First, for a binary malignancy classification task, I did not observe a significant hidden stratification phenomenon (Table 5.2) in the BreakHis dataset as I did in the LIDC dataset (Table 4.8). Implementing gDRO does not significantly improve the worst group accuracy nor does it significantly decrease the overall accuracy. When comparing the binary classification results between the two datasets, we can conclude that a classification model with ERM loss is adequate if the worst group performance is comparable to the performance of other subgroups.

Second, by utilizing deep image features extracted from a binary malignancy classification model, I did not achieve clustering-based subgroups that could be

verified by domain experts (Figure 5.2). For my future work, I plan to explore various image feature embeddings to enhance subgroup discovery.

Third, for a multiclass classification task using ground truth subgroup labels provided by domain experts, we can observe a hidden stratification phenomenon (Table 5.3) and implementing gDRO algorithm significantly improves the worst group performance.

Fourth, the CRIS model significantly improves the worst group performance compared to the base ERM and gDRO models (Tables 5.4-5). The CRRIS model with an ERM model trained on atypical instances can also significantly increase the worst group performance compared to the CRIS model (Table 5.6).

CHAPTER 6. Summary and Future Works

In this study, I proposed a hidden stratification discovery and subgroup learning scheme that enhances machine learning model generalization ability. I applied this methodology to two applications: lung cancer malignancy classification and breast cancer malignancy classification. In the lung cancer application, clustering-based semantically meaningful subgroups were discovered (Table 4.2-3) while in the breast cancer application, clustering-based subgroups could not be identified by domain experts (Figure 5.2).

In both applications, implementing gDRO significantly mitigated the hidden stratification phenomena, leading to improved worst group accuracy (Table 4.8 and Table 5.3). Moreover, training a model with ERM loss and gDRO loss sequentially further increased the worst group performance (Table 4.16, Table 4.17, and Table 5.4-6). In the breast cancer application, we observe that a classification model with ERM loss is adequate if the worst group performance is comparable to the performance of other subgroups (Table 5.2).

In the lung cancer application, compared to the malignancy rating distributions obtained using UMAP (Table 4.2), we observe less overlap between the clustered stratification and the malignancy likelihood stratification (Table 4.4). This result is consistent with findings in the literature that UMAP has an advantage over PCA in preserving the cluster relationships of data points in high-dimensional space [122].

Through evaluating the model performance on pathologically proven examination-labeled nodules, we found that using a CNN classifier with ERM or gDRO loss can help reduce the false negative rate (Table 4.35 and Table 4.36) when compared with human experts.

We systematically evaluated the performance of the model under domain shift scenarios by considering various model inputs, deep learning architectures, and training strategies. We observed that when semantic features are available, utilizing semantic features further improves the overall accuracy and the worst-group accuracy (Table 4.11, Table 4.21, and Table 4.29).

As part of the future work, the investigation of different image feature extraction techniques will be prioritized to enhance the elucidation of hidden subgroups. In the context of lung cancer application, the utilization of image features derived from the final convolutional layer of a binary malignancy classification model exhibits the potential to unveil unlabeled subgroups within the UMAP space. However, when applied to the breast cancer application, employing the same feature set does not yield clustering-based subgroups as identified by domain experts. To address this limitation, an exploration into disentangled representation learning (DRL) approaches will be conducted. These approaches aim to decompose the input data into disentangled factors, wherein each factor corresponds to a relevant variable in the data generating process. The hypothesis

put forth is that extracting disentangled representations from histopathological images can assist in identifying unlabeled semantic meaningful subgroups.

In this study, the implementation of the gDRO algorithm primarily focuses on improving the worst group performance. However, it is important to consider scenarios where multiple subgroups exhibit similar levels of underperformance. To address this concern, a proposed modification involves designing an enhanced gDRO loss function capable of improving the performance of multiple desired subgroups simultaneously. By incorporating this modification, we aim to achieve a more comprehensive enhancement across multiple subgroups rather than solely focusing on the worst-performing group.

In addition to the aforementioned objectives, the future direction of this research aims to expand the study's scope by addressing critical aspects such as enhancing fairness and reducing bias in machine learning. In this study, subgroups refer to different disease subtypes due to disease heterogeneity. However, subgroups can also encompass different ethnicities and races. This research will focus not only on improving the overall performance but also on the performance of subsets generated by different ethnicities and races. One approach to enhance the worst group performance is gDRO, which has shown promising results. However, future work will also include exploring different techniques for worst group optimization.

References

[1] L. Oakden-Rayner, J. Dunnmon, G. Carneiro, and C. Ré, "Hidden stratification causes clinically meaningful failures in machine learning for medical imaging," in *Proceedings of the ACM conference on health, inference, and learning*, 2020, pp. 151-159.

[2] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang, "Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization," presented at the International Conference on Learning Representations (ICLR), Virtual Conference, 2020.

[3] S. G. Armato III *et al.*, "The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans," *Medical Physics,* vol. 38, no. 2, pp. 915-931, 2011.

[4] N. Sohoni, J. Dunnmon, G. Angus, A. Gu, and C. Ré, "No subclass left behind: Fine-grained robustness in coarse-grained classification problems," *Advances in Neural Information Processing Systems,* vol. 33, pp. 19339-19352, 2020.

[5] C.-Y. Hsieh, M. Xu, G. Niu, H.-T. Lin, and M. Sugiyama, "A pseudo-label method for coarse-to-fine multi-label learning with limited supervision," presented at the International Conference on Learning Representations (ICLR), New Orleans, Louisiana, United States, 2019.

[6] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," *IEEE Transactions on Biomedical Engineering,* vol. 63, no. 7, pp. 1455-1462, 2015.

[7] Z. Shen *et al.*, "Towards out-of-distribution generalization: A survey," *arXiv preprint arXiv:2108.13624,* 2021.

[8] W. Hu, G. Niu, I. Sato, and M. Sugiyama, "Does distributionally robust supervised learning give robust classifiers?," in *International Conference on Machine Learning*, 2018: PMLR, pp. 2029-2037.

[9] J. Wen *et al.*, "Subtyping brain diseases from imaging data," *arXiv preprint arXiv:2202.10945,* 2022.

[10] J. Wen *et al.*, "Multi-scale semi-supervised clustering of brain images: Deriving disease subtypes," *Medical Image Analysis,* vol. 75, p. 102304, 2022.

[11] A. Sotiras, S. M. Resnick, and C. Davatzikos, "Finding imaging patterns of structural covariance via non-negative matrix factorization," *Neuroimage,* vol. 108, pp. 1-16, 2015.

[12] A. Ezzati, A. R. Zammit, C. Habeck, C. B. Hall, and R. B. Lipton, "Detecting biological heterogeneity patterns in ADNI amnestic mild cognitive impairment based on volumetric MRI," *Brain Imaging and Behavior,* vol. 14, no. 5, pp. 1792-1804, 2020.

[13] J. Chen, L. Milot, H. Cheung, and A. L. Martel, "Unsupervised clustering of quantitative imaging phenotypes using autoencoder and gaussian mixture

model," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019: Springer, pp. 575-582.

[14] Z. Yang, J. Wen, and C. Davatzikos, "Surreal-GAN: Semi-Supervised Representation Learning via GAN for uncovering heterogeneous disease-related imaging patterns," presented at the International Conference on Learning Representations (ICLR), Virtual, 2022.

[15] J. Wu *et al.*, "Unsupervised Clustering of Quantitative Image Phenotypes Reveals Breast Cancer Subtypes with Distinct Prognoses and Molecular PathwaysImaging Subtypes of Breast Cancer," *Clinical Cancer Research,* vol. 23, no. 13, pp. 3334-3342, 2017.

[16] M. Fan *et al.*, "Tumour heterogeneity revealed by unsupervised decomposition of dynamic contrast-enhanced magnetic resonance imaging is associated with underlying gene expression patterns and poor survival in breast cancer patients," *Breast Cancer Research,* vol. 21, no. 1, pp. 1-16, 2019.

[17] M.-A. Schulz, M. Chapman-Rounds, M. Verma, D. Bzdok, and K. Georgatzis, "Inferring disease subtypes from clusters in explanation space," *Scientific Reports,* vol. 10, no. 1, pp. 1-6, 2020.

[18] J. N. Weinstein *et al.*, "The cancer genome atlas pan-cancer analysis project," *Nature Genetics,* vol. 45, no. 10, pp. 1113-1120, 2013.

[19] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems,* vol. 30, 2017.

[20] S. Arslanturk, S. Draghici, and T. Nguyen, "Integrated Cancer Subtyping using Heterogeneous Genome-Scale Molecular Datasets," in *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 2020, vol. 25, pp. 551-562.

[21] P. Vasudevan and T. Murugesan, "Cancer Subtype Discovery Using Prognosis-Enhanced Neural Network Classifier in Multigenomic Data," *Technology in Cancer Research & Treatment,* vol. 17, 2018.

[22] P. Anderson, R. Gadgil, W. A. Johnson, E. Schwab, and J. M. Davidson, "Reducing variability of breast cancer subtype predictors by grounding deep learning models in prior knowledge," *Computers in Biology and Medicine,* vol. 138, p. 104850, 2021.

[23] X. Zhang *et al.*, "Data-driven subtyping of Parkinson's disease using longitudinal clinical records: a cohort study," *Scientific Reports,* vol. 9, no. 1, pp. 1-12, 2019.

[24] Z. Xu *et al.*, "Identifying sub-phenotypes of acute kidney injury using structured and unstructured electronic health record data with memory networks," *Journal of Biomedical Informatics,* vol. 102, p. 103361, 2020.

[25] Z. Xu *et al.*, "Subphenotyping depression using machine learning and electronic health records," *Learning Health Systems,* vol. 4, no. 4, p. e10241, 2020.

[26] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, "Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data," *Machine Learning,* vol. 52, no. 1, pp. 91-118, 2003.

[27] A. V. Kapp and R. Tibshirani, "Are clusters found in one dataset present in another dataset?," *Biostatistics,* vol. 8, no. 1, pp. 9-31, 2007.
[28] X. Liu, P. Sanchez, S. Thermos, A. Q. O'Neil, and S. A. Tsaftaris, "Learning disentangled representations in the imaging domain," *Medical Image Analysis,* p. 102516, 2022.
[29] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," *Stat,* vol. 1050, p. 1, 2014.
[30] M. A. Kramer, "Nonlinear principal component analysis using autoassociative neural networks," *AIChE Journal,* vol. 37, no. 2, pp. 233-243, 1991.
[31] I. Higgins *et al.*, "beta-vae: Learning basic visual concepts with a constrained variational framework," presented at the International Conference on Learning Representations (ICLR), 2017.
[32] C. P. Burgess *et al.*, "Understanding disentangling in beta-VAE," *arXiv preprint arXiv:1804.03599,* 2018.
[33] H. Kim and A. Mnih, "Disentangling by factorising," in *International Conference on Machine Learning*, 2018: PMLR, pp. 2649-2658.
[34] M. Yang, F. Liu, Z. Chen, X. Shen, J. Hao, and J. Wang, "CausalVAE: Disentangled representation learning via neural structural causal models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9593-9602.
[35] X. Shen, F. Liu, H. Dong, Q. Lian, Z. Chen, and T. Zhang, "Disentangled generative causal representation learning," *arXiv preprint arXiv:2010.02637,* 2020.
[36] Y. Zhu, M. R. Min, A. Kadav, and H. P. Graf, "S3vae: Self-supervised sequential vae for representation disentanglement and data generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6538-6547.
[37] R. Couronné, P. Vernhet, and S. Durrleman, "Longitudinal self-supervision to disentangle inter-patient variability from disease progression," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2021: Springer, pp. 231-241.
[38] W. Grathwohl and A. Wilson, "Disentangling space and time in video with hierarchical variational auto-encoders," *arXiv preprint arXiv:1612.04440,* 2016.
[39] W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised learning of disentangled and interpretable representations from sequential data," *Advances in Neural Information Processing Systems,* vol. 30, 2017.
[40] Y. Li and S. Mandt, "Disentangled sequential autoencoder," presented at the International Conference on Machine Learning (ICML), Stockholmsmässan, Sweden, 2018.
[41] F. Yang, R. Meng, H. Cho, G. Wu, and W. H. Kim, "Disentangled Sequential Graph Autoencoder for Preclinical Alzheimer's Disease Characterizations from ADNI Study," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2021: Springer, pp. 362-372.

[42] P. K. Gyawali, Z. Li, S. Ghimire, and L. Wang, "Semi-supervised learning by disentangling and self-ensembling over stochastic latent space," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019: Springer, pp. 766-774.

[43] J. Irvin *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of the AAAI conference on artificial intelligence*, 2019, vol. 33, no. 01, pp. 590-597.

[44] I. Goodfellow *et al.*, "Generative adversarial nets," *Advances in Neural Information Processing Systems,* vol. 27, 2014.

[45] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," *Advances in neural information processing systems,* vol. 29, 2016.

[46] S. Mukherjee, H. Asnani, E. Lin, and S. Kannan, "Clustergan: Latent space clustering in generative adversarial networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, no. 01, pp. 4610-4617.

[47] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401-4410.

[48] W. Nie *et al.*, "Semi-supervised stylegan for disentanglement learning," in *International Conference on Machine Learning*, 2020: PMLR, pp. 7360-7369.

[49] Y. Tang, Y. Tang, Y. Zhu, J. Xiao, and R. M. Summers, "A disentangled generative model for disease decomposition in chest x-rays via normal image synthesis," *Medical Image Analysis,* vol. 67, p. 101839, 2021.

[50] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2097-2106.

[51] T. Xia, A. Chartsias, and S. A. Tsaftaris, "Pseudo-healthy synthesis with pathology disentanglement and adversarial learning," *Medical Image Analysis,* vol. 64, p. 101719, 2020.

[52] K. Kobayashi *et al.*, "Decomposing normal and abnormal features of medical images for content-based image retrieval of glioma imaging," *Medical Image Analysis,* vol. 74, p. 102227, 2021.

[53] A. Ben-Cohen, R. Mechrez, N. Yedidia, and H. Greenspan, "Improving CNN training using disentanglement for liver lesion classification in CT," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2019: IEEE, pp. 886-889.

[54] M. Gravina, S. Marrone, M. Sansone, and C. Sansone, "DAE-CNN: Exploiting and disentangling contrast agent effects for breast lesions classification in DCE-MRI," *Pattern Recognition Letters,* vol. 145, pp. 67-73, 2021.

[55] I. Kobyzev, S. J. Prince, and M. A. Brubaker, "Normalizing flows: An introduction and review of current methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 43, no. 11, pp. 3964-3979, 2020.
[56] Google. "Generative Adversarial GAN." https://developers.google.com/machine-learning/gan/problems (accessed.
[57] J. Lucas, G. Tucker, R. B. Grosse, and M. Norouzi, "Don't blame the Elbo! a linear Vae perspective on posterior collapse," *Advances in Neural Information Processing Systems,* vol. 32, 2019.
[58] P. Esser, R. Rombach, and B. Ommer, "A disentangling invertible interpretation network for explaining latent representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9223-9232.
[59] L. Deng, "The mnist database of handwritten digit images for machine learning research," *IEEE Signal Processing Magazine,* vol. 29, no. 6, pp. 141-142, 2012.
[60] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3730-3738.
[61] A. Sankar *et al.*, "GLOWin: A flow-based invertible generative framework for learning disentangled feature representations in medical images," *arXiv preprint arXiv:2103.10868,* 2021.
[62] R. Wang, P. Chaudhari, and C. Davatzikos, "Harmonization with flow-based causal inference," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2021: Springer, pp. 181-190.
[63] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2414-2423.
[64] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 172-189.
[65] K. Li, L. Yu, S. Wang, and P.-A. Heng, "Unsupervised retina image synthesis via disentangled representation learning," in *International Workshop on Simulation and Synthesis in Medical Imaging*, 2019: Springer, pp. 32-41.
[66] L. Zuo *et al.*, "Unsupervised MR harmonization by learning disentangled representations using information bottleneck theory," *NeuroImage,* vol. 243, p. 118569, 2021.
[67] A. Chartsias *et al.*, "Disentangled Representation Learning in Cardiac Image Analysis," *Medical Image Analysis,* vol. 58, p. 101535, 2019.
[68] C. Bass *et al.*, "Icam-reg: Interpretable classification and regression with feature attribution for mapping neurological phenotypes in individual scans," *arXiv preprint arXiv:2103.02561,* 2021.
[69] H. Li, T. Loehr, A. Sekuboyina, J. Zhang, B. Wiestler, and B. Menze, "Domain adaptive medical image segmentation via adversarial learning of disease-specific spatial patterns," *arXiv preprint arXiv:2001.09313,* 2020.

[70] J. Wang, C. Lan, C. Liu, Y. Ouyang, W. Zeng, and T. Qin, "Generalizing to unseen domains: A survey on domain generalization," in *International Joint Conference on Artificial Intelligence (IJCAI)*, Montreal, Canada, 2021.

[71] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International Conference on Machine Learning (ICML)*, 2015: PMLR, pp. 1180-1189.

[72] Y. Ganin *et al.*, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research,* vol. 17, no. 1, pp. 2096-2030, 2016.

[73] H. Li, S. J. Pan, S. Wang, and A. C. Kot, "Domain generalization with adversarial feature learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5400-5409.

[74] Y. Li *et al.*, "Deep domain generalization via conditional invariant adversarial networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 624-639.

[75] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2223-2232.

[76] R. Gong, W. Li, Y. Chen, and L. V. Gool, "Dlow: Domain flow for adaptation and generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2477-2486.

[77] C. Chen, Q. Dou, H. Chen, J. Qin, and P. A. Heng, "Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation," *IEEE Transactions on Medical Imaging,* vol. 39, no. 7, pp. 2494-2505, 2020.

[78] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto, "Unified deep supervised domain adaptation and generalization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5715-5725.

[79] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *European Conference on Computer Vision*, 2010: Springer, pp. 213-226.

[80] F. Zhou, Z. Jiang, C. Shui, B. Wang, and B. Chaib-draa, "Domain generalization with optimal transport and metric learning," *Neurocomputing,* vol. 456, pp. 469-480, 2021.

[81] J. Wang, Y. Chen, W. Feng, H. Yu, M. Huang, and Q. Yang, "Transfer learning with dynamic distribution adaptation," *ACM Transactions on Intelligent Systems and Technology (TIST),* vol. 11, no. 1, pp. 1-25, 2020.

[82] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *European Conference on Computer Vision*, 2016: Springer, pp. 443-450.

[83] C. Zhou *et al.*, "Improving the generalization of glaucoma detection on fundus images via feature alignment between augmented views," *Biomedical Optics Express,* vol. 13, no. 4, pp. 2018-2034, 2022.

[84] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Transactions on Neural Networks,* vol. 22, no. 2, pp. 199-210, 2010.

[85] T. Grubinger, A. Birlutiu, H. Schöner, T. Natschläger, and T. Heskes, "Domain generalization based on transfer component analysis," in *International Work-Conference on Artificial Neural Networks*, 2015: Springer, pp. 325-334.

[86] K. Muandet, D. Balduzzi, and B. Schölkopf, "Domain generalization via invariant feature representation," in *International Conference on Machine Learning*, 2013: PMLR, pp. 10-18.

[87] C. Gan, T. Yang, and B. Gong, "Learning attributes equals multi-source domain generalization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 87-97.

[88] A. Van Opbroek, H. C. Achterberg, M. W. Vernooij, and M. De Bruijne, "Transfer learning for image segmentation by combining image weighting and kernel learning," *IEEE Transactions on Medical Imaging,* vol. 38, no. 1, pp. 213-224, 2018.

[89] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International Conference on Machine Learning*, 2017: PMLR, pp. 1126-1135.

[90] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Learning to generalize: Meta-learning for domain generalization," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[91] Y. Balaji, S. Sankaranarayanan, and R. Chellappa, "Metareg: Towards domain generalization using meta-regularization," *Advances in Neural Information Processing Systems,* vol. 31, 2018.

[92] Y. Du *et al.*, "Learning to learn with variational information bottleneck for domain generalization," in *European Conference on Computer Vision*, 2020: Springer, pp. 200-216.

[93] M. Mancini, S. R. Bulo, B. Caputo, and E. Ricci, "Best sources forward: domain generalization through source-specific nets," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018: IEEE, pp. 1353-1357.

[94] M. Segu, A. Tonioni, and F. Tombari, "Batch normalization embeddings for deep domain generalization," *arXiv preprint arXiv:2011.12672,* 2020.

[95] X. Zhang, L. Zhou, R. Xu, P. Cui, Z. Shen, and H. Liu, "Domain-Irrelevant Representation Learning for Unsupervised Domain Generalization," *arXiv preprint arXiv:2107.06219,* 2021.

[96] Y. Liao, R. Huang, J. Li, Z. Chen, and W. Li, "Deep semisupervised domain generalization network for rotary machinery fault diagnosis under variable speed," *IEEE Transactions on Instrumentation and Measurement,* vol. 69, no. 10, pp. 8064-8075, 2020.

[97] C. S. Perone, P. Ballester, R. C. Barros, and J. Cohen-Adad, "Unsupervised domain adaptation for medical imaging segmentation with self-ensembling," *NeuroImage,* vol. 194, pp. 1-11, 2019.
[98] R. Zhang, Q. Xu, C. Huang, Y. Zhang, and Y. Wang, "Semi-Supervised Domain Generalization for Medical Image Analysis," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, 2022: IEEE, pp. 1-5.
[99] W. van Amsterdam, J. Verhoeff, P. de Jong, T. Leiner, and M. Eijkemans, "Eliminating biasing signals in lung cancer images for prognosis predictions with deep learning," *NPJ Digital Medicine,* vol. 2, no. 1, pp. 1-6, 2019.
[100] J. Peters, P. Buhlmann, and N. Meinshausen, "Causal inference using invariant prediction: identification and confidence intervals," *Methodology,* 2015.
[101] N. Pfister, P. Bühlmann, and J. Peters, "Invariant causal prediction for sequential data," *Journal of the American Statistical Association,* vol. 114, no. 527, pp. 1264-1276, 2019.
[102] C. Heinze-Deml, J. Peters, and N. Meinshausen, "Invariant causal prediction for nonlinear models," *Journal of Causal Inference,* vol. 6, no. 2, 2018.
[103] M. McClellan, B. J. McNeil, and J. P. Newhouse, "Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality?: analysis using instrumental variables," *The Journal of the American Medical Association (JAMA),* vol. 272, no. 11, pp. 859-866, 1994.
[104] H. Theil, "Estimation and simultaneous correlation in complete equation systems," in *Henri Theil's Contributions to Economics and Econometrics*: Springer, 1992, pp. 65-107.
[105] E. E. Pracht, J. J. Tepas III, B. G. Celso, B. Langland-Orban, and L. Flint, "Survival advantage associated with treatment of injury at designated trauma centers: a bivariate probit model with instrumental variables," *Medical Care Research and Review,* vol. 64, no. 1, pp. 83-97, 2007.
[106] D. Rothenhäusler, N. Meinshausen, P. Bühlmann, and J. Peters, "Anchor regression: Heterogeneous data meet causality," *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* vol. 83, no. 2, pp. 215-246, 2021.
[107] M. Oberst, N. Thams, J. Peters, and D. Sontag, "Regularizing towards causal invariance: Linear models with proxies," in *International Conference on Machine Learning*, 2021: PMLR, pp. 8260-8270.
[108] D. C. Castro, I. Walker, and B. Glocker, "Causality matters in medical imaging," *Nature Communications,* vol. 11, no. 1, pp. 1-10, 2020.
[109] M. Basseville, "Divergence measures for statistical data processing—An annotated bibliography," *Signal Processing,* vol. 93, no. 4, pp. 621-633, 2013.
[110] L. Rüschendorf, "The Wasserstein distance and approximation theorems," *Probability Theory and Related Fields,* vol. 70, no. 1, pp. 117-129, 1985.
[111] H. Namkoong and J. C. Duchi, "Stochastic gradient methods for distributionally robust optimization with f-divergences," *Advances in Neural Information Processing Systems,* vol. 29, 2016.

[112] T. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang, "Fairness without demographics in repeated loss minimization," in *International Conference on Machine Learning*, 2018: PMLR, pp. 1929-1938.
[113] A. Sinha, H. Namkoong, and J. Duchi, "Certifying Some Distributional Robustness with Principled Adversarial Training," in *International Conference on Learning Representations*, 2018.
[114] J. Byrd and Z. Lipton, "What is the effect of importance weighting in deep learning?," in *International Conference on Machine Learning*, 2019: PMLR, pp. 872-881.
[115] J. Liu, Z. Hu, P. Cui, B. Li, and Z. Shen, "Heterogeneous risk minimization," in *International Conference on Machine Learning*, 2021: PMLR, pp. 6804-6814.
[116] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant Risk Minimization," *Stat,* vol. 1050, p. 27, 2020.
[117] S. Chang, Y. Zhang, M. Yu, and T. Jaakkola, "Invariant rationalization," in *International Conference on Machine Learning*, 2020: PMLR, pp. 1448-1458.
[118] M. Koyama and S. Yamaguchi, "Out-of-distribution generalization with maximal invariant predictor," *arXiv preprint arXiv: 2008.01883,* 2020.
[119] H. Ye, C. Xie, Y. Liu, and Z. Li, "Out-of-distribution generalization analysis via influence function," *arXiv preprint arXiv:2101.08521,* 2021.
[120] S. Targ, D. Almeida, and K. Lyman, "Resnet in resnet: Generalizing residual architectures," presented at the International Conference on Learning Representations (ICLR), 2016.
[121] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision,* vol. 115, no. 3, pp. 211-252, 2015.
[122] L. McInnes, J. Healy, N. Saul, and L. Großberger, "UMAP: Uniform Manifold Approximation and Projection," *Journal of Open Source Software,* vol. 3, no. 29, p. 861, 2018.
[123] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics,* vol. 2, no. 4, pp. 433-459, 2010.
[124] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics,* vol. 20, pp. 53-65, 1987.
[125] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural Computation,* vol. 29, no. 9, pp. 2352-2449, 2017.
[126] B. Qiu, D. Raicu, J. Furst, and R. Tchoua, "CompNet: A Designated Model to Handle Combinations of Images and Designed features," *arXiv preprint arXiv:2209.14454,* 2022.
[127] A. Efthymiou, S. Rudinac, M. Kackovic, M. Worring, and N. Wijnberg, "Graph Neural Networks for Knowledge Enhanced Visual Representation of Paintings," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3710-3719.

[128] S. Bianco, "Large age-gap face verification by feature injection in deep networks," *Pattern Recognition Letters,* vol. 90, pp. 36-42, 2017.
[129] S. Bianco, D. Mazzini, P. Napoletano, and R. Schettini, "Multitask painting categorization by deep multibranch neural network," *Expert Systems with Applications,* vol. 135, pp. 90-101, 2019.
[130] T. H. Nguyen, H. R. Zhang, and H. L. Nguyen, "Improved Worst-Group Robustness via Classifier Retraining on Independent Splits," *arXiv preprint arXiv:2204.09583,* 2022.
[131] M. F. McNitt-Gray *et al.*, "The Lung Image Database Consortium (LIDC) data collection process for nodule detection and annotation," *Academic Radiology,* vol. 14, no. 12, pp. 1464-1474, 2007.
[132] P. Opulencia, D. S. Channin, D. S. Raicu, and J. D. Furst, "Mapping LIDC, RadLex™, and lung nodule image features," *Journal of Digital Imaging,* vol. 24, no. 2, pp. 256-270, 2011.
[133] M. C. Hancock and J. F. Magnan, "Lung nodule malignancy classification using only radiologist-quantified image features as inputs to statistical learning algorithms: probing the Lung Image Database Consortium dataset with two statistical learning methods," *Journal of Medical Imaging,* vol. 3, no. 4, p. 044504, 2016.
[134] M. M. Wahidi, J. A. Govert, R. K. Goudar, M. K. Gould, and D. C. McCrory, "Evidence for the treatment of patients with pulmonary nodules: when is it lung cancer?: ACCP evidence-based clinical practice guidelines," *Chest,* vol. 132, no. 3, pp. 94S-107S, 2007.
[135] X. Mei *et al.*, "RadImageNet: An open radiologic deep learning research dataset for effective transfer learning," *Radiology: Artificial Intelligence,* vol. 4, no. 5, p. e210315, 2022.
[136] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "Breast cancer histopathological image classification using convolutional neural networks," in *2016 international joint conference on neural networks (IJCNN)*, 2016: IEEE, pp. 2560-2567.