No Nodule Left Behind: Evaluating Lung Nodule Malignancy Classification with Different Stratification Schemes

Thomas Zeng[a], Elias Furst[b], Yiyang Wang[c], Roselyne Tchoua[c], Jacob Furst[c], and Daniela Raicu[c]

[a]Carleton College
[b]Milwaukee School of Engineering
[c]DePaul University

ABSTRACT

Machine learning models have been widely used in lung cancer computer-aided diagnosis (CAD) studies. However, the heterogeneity in the visual appearance of lung nodules as well as lack of consideration of hidden subgroups in the data are significant obstacles to generating accurate CAD outcomes across all nodule instances. Previous lung cancer CAD models aim to achieve Empirical Risk Minimization (ERM), which leads to a high overall accuracy but often fails at predicting certain subgroups caused by the lung cancer heterogeneity. In this study, we aim to discover hidden lung nodule subgroups and enhance the malignancy classification performance of the worst-performance subgroup when compared to traditional ERM methods. We experiment with three different stratification methods for lung nodule subgroup discovery: clustering-based method, malignancy likelihood annotated by radiologists and spiculation ratings that were provided by radiologists. A high overlap between subgroup labels generated from the clustering-based approach and labels obtained from radiologists' semantic annotations indicates our discovered subgroups are semantically meaningful. We successfully improved the worst malignancy classification performance lung nodule subgroup through utilizing Group Distributionally Robust Optimization (gDRO) when compared to ERM as a baseline. Our study creates a framework for augmenting lung nodule malignancy classification under domain shift situations caused by the disease heterogeneity and underscores the necessity of addressing hidden stratification for future CAD schemes.

Keywords: Visual Appearance heterogeneity, Computer-Aided Diagnosis (CAD), Domain Shift Generalization, Group Distributionally Robust Optimization (gDRO)

1. INTRODUCTION

Lung cancer is a leading cause of death and the second most common form of cancer.[1] It has been shown that early screening can catch lung cancer in its early stages and thus drastically decrease mortality rates.[2] Hence, the development of Computer Aided Diagnosis (CAD) algorithms to help radiologists in this early screening is of vital importance.

In this regard, there has been a multitude of papers investigating the binary task of classifying lung nodules as malignant or benign.[3] In general, these classifiers purportedly perform very well, with recorded overall accuracy ranging between 85% and 95%. While at first glance this seems like excellent performance, it is not enough to only look at overall performance of the classifier. Specifically, as outlined by Oakden-Rayner et al.,[4] performance across subsets of the data is equally as important. Specific to lung nodules, there is great variation in their appearance[5,6] (Figure 1). Thus besides being separated by malignant or benign, there exist ways to further stratify nodules.

Hence, in this paper we seek to address this heterogeneity in the visual appearance of the lung nodule by comparing different methods of stratification methods on the Lung Image Database Consortium image collection (LIDC-IDRI) dataset – either using domain driven methods or unsupervised clustering methods. We furthermore use the group Distributionally Robust Optimization (gDRO) loss as proposed by Sagawa et al.[7] to see if we can improve performance across the worst performing subclass when compared to more canonical training methods such as Empirical Risk Minimization (ERM).

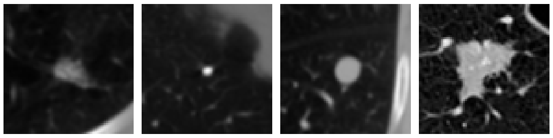


Figure 1. A demonstration of the heterogeneity of nodule visual appearances found in the dataset. The left two nodules are benign and the right two are malignant.

2. RELATED WORKS

Our work builds on three categories of previous work. Specifically, we build upon work in data stratification, classification model architectures, and loss function schemes.

2.1 Data Stratification

Oakden-Rayner et al.[4] identify the necessity of finding hidden stratification in the medical domain. They also describe three different methods:

1. **Schema completion:** have domain experts discuss and identify the stratification and hand label the samples
2. **Error auditing:** find patterns in the errors that an algorithm makes
3. **Algorithmic measurement:** use unsupervised learning methods to discover subclasses

As the focus of their paper is mainly to underscore stratification as a problem in the medical domain and find the methods to discover them, we seek to build upon this paper by looking at methods to increase performance on a stratified dataset.

In a similar vein to above, Sohoni et al.[8] provide an in-detail methodology, GEORGE, on how dimensionality reduction and clustering can be used to algorithmically discover subclass labels. To evaluate the efficacy of the method, the paper tests the methodology on canonical and synthetic datasets where the subclass labels are already well defined. Here we seek to apply this methodology to the LIDC dataset where data is less abundant and there exists no ground-truth way to subclass the labels. This is in contrast to commonly-used datasets such as Waterbirds,[7] which is constructed with four well-defined subclasses and contains almost five times more data than the LIDC dataset.

2.2 Classifier Architectures

With regard to lung nodule malignancy classifiers, there exists an eclectic mix of architectures. Monkam et al.[9] give an overview of some of the datasets and model architectures used. Some well performing classifiers trained specifically on the LIDC dataset include DeepLung[10] – which uses 3D Convolutional Neural Networks, NASLung[11] – which uses Neural Architecture Search to discover the architecture, PRoCAN[12]– which uses a Non-Local network with curriculum learning, Gated-Dilated networks that attend to nodules of varying size,[13] classifiers which use the lung-Nodules at different scales as input,[14] and lastly multi-task learning on the semantic features to improve performance of nodule classification.[15] More specific to our paper, Da et al.,[16] Zhao et al.,[17] Yu et al.,[18] and Nibali et al.[19] have also investigated the usage of transfer learning on the LIDC with different hyperparameters and backbone architectures such as ResNet, VGG or MobileNet. However, despite this vast literature on lung nodule malignancy classification on the LIDC, as previously mentioned, these models only optimize on the average performance of malignancy classification, whereas we will use a relatively simple model but explore the performance over hidden subtypes.

2.3 Loss Schemes

There is much existing literature investigating the use of various loss schemes for domain shift generalization. Previous work analyzed adversarial risk minimization,[20] but found that it was unable to produce a robust model, and instead tended to only overfit on the training distribution. Of great importance to our work is group Distributionally Robust Optimization[7] – a method to optimize over the worst class accuracy, given information about the structure of the data. Koh et al.[21] provide a benchmark – WILDS – measuring the performance of gDRO and other domain generalization schemes against the traditional ERM methods over various datasets that may be found in the "wild" i.e. the real word. However the main focus of this paper is on domain generalization – where the test set distribution is out of the domain of the training set. This is as opposed to our purposes, of maintaining robustness against domain shifts where the distribution of subclasses of data may differ across the train and test distribution. Lastly, these works have also mainly looked at datasets where the stratification is easily identifiable. This is not the case with LIDC and thus it is of interest to see how gDRO compares to ERM in this dataset where the stratification is not as apparent.

3. METHODS

Our methodology consists of stratification of the data, model selection and loss function for training as seen in Figure 2.

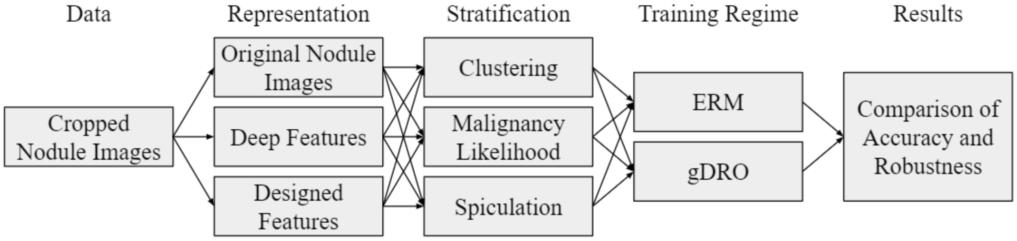


Figure 2. Methodology overview. Terms will be defined in order from left to right in the following subsections.

3.1 Dataset

The LIDC dataset originally consisted of lung CT scans containing 2,669 nodules with annotations from four radiologists.[22] Each nodule in the dataset has a malignancy likelihood label on a scale from 1 (most likely benign) to 5 (most likely malignant) with 3 representing indeterminate malignancy. In this work we started with 2688 nodules but use only 1,488 of them as we drop any nodule with a mode malignancy of 3. In total we have 591 malignant nodules (malignancy likelihood of 4 or 5) and 897 benign nodules (malignancy likelihood of 1 or 2). The dataset also includes radiologist-generated features that rate the semantic features of each nodule such as degree of calcification, spiculation or lobulation. We obtain our reference-truth label for each rating by taking the mode of the four radiologist ratings, breaking ties by taking the ceiling of the mean of any tied modes.

We further divide the dataset into train, validation and test sets using a random 70%, 10% and 20% split respectively stratified by the malignancy likelihood ratings. This stratification is done under the assumption that a dataset overly biased towards malignancy labels of 1 or 5 may be easier to classify and thus overly optimistic of performance – and hence the stratification should prevent this. We report results on the test set.

3.2 Data Representations

In this paper, we run our experiments across three different representations of the data: designed features, deep features, and the images themselves. We do this to check the robustness of the dataset stratification and gDRO against the type of feature representation.

The designed features consists of 64 numerical features per nodule, calculated from the image pixels e.g. area, perimeter, circularity. This type of features belong to the field of "radiomics" where nodules are segmented and put into equations to produce various features.[23] This was the conventional way of creating features before the advent of deep learning. The specific equations we use were provided by Zinovev et al.,[24] which take the image data with nodule segmentations produced by the radiologists as input. This produces a standard set of features commonly used for the LIDC. We therefore include this data to observe whether the domain knowledge encoded in the choice of equations has an effect on the relative performance of ERM and gDRO.

The image feature consists of 71×71 dimensional crops of nodules along the transverse plane selecting the slice of maximum area and centered on the nodule centroid. As an additional prepossessing step, the images are also center cropped to dimensions of 51×51 and 31×31 and then upscaled to 71×71 using bilinear interpolation. These two cropped and upscaled representations are then combined with the original image crop to form a three-channel image. This step is inspired by work by Al-Shabi et al.[13] and Zhao et al.,[14] which have shown that the size and scale of the image can affect classifier performance. Furthermore our own ablation experiments show that this step performs a few percentage better than a three-channel image where each channel is the same image with no scaling.

Lastly, the deep features for the nodules are produced by using transfer learning on a ResNet18 model pretrained with ImageNet. The model is trained using a standard ERM loss scheme with the above described image features as input, as recommended by Rosenfeld et al.[25] Then by feeding in the images as input, we take the 512 dimensional activations of the last layer before the fully-connected classifier as features. This effectively serves to train and freeze the featurizing layers of the network, allowing us to evaluate the performance of ERM and gDRO solely on the classification layers of the model.

3.3 Models

For the designed features, we use a fully connected classifier. From our experiments, we find that the exact number of hidden layers and dimensions has no significant affect on the classification accuracy. Thus we only keep one hidden layer with dimension roughly half of the input layer. Specifically the fully connected classifier has an input layer of 64 dimensions, a hidden layer of 36 dimensions and a output layer of 2 dimensions for binary classification.

For the image features and deep features, we make use of a transfer learning model using ResNet18 as the backbone, as recommended by Nibali et al.[19] We use a single fully-connected layer to reduce ResNet18's 512 deep features to a binary classification of either malignant or benign.

For detailed list of the hyperparameters and specific training regimen used, see section A in the appendix.

3.4 Stratification

We run our experiments on three different stratification methods. Two of them use the semantic features determined by radiologists and one is a modified form of the GEORGE algorithm outlined by Sohoni et al.[8]

The first form of stratification we use is separating the nodules by spiculation. The choice of spiculation rather than another semantic feature is driven by the domain knowledge that spiculation is typically correlated with malignancy.[26,27] We also know that an ERM model learns spurious correlations,[8] in this case the positive correlation between spiculation and malignancy. We would then expect that the minority classes of nodules which go against this trend will be difficult for an ERM model to classify. The nodules are stratified by grouping nodules with a spiculation rating of 1 into an "unspiculated" category, while nodules with a spiculation rating greater than 1 are labelled as "spiculated". Examples of a nodule from each of the four subclasses can be seen in Figure 3.

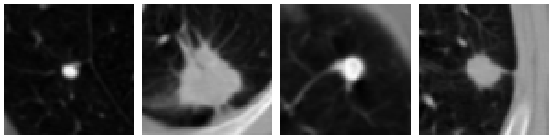


Figure 3. From left to right: Unspiculated Benign, Spiculated Malignant, Spiculated Benign, and Unspiculated Malignant nodules

The second method of stratification is derived by a semi-unsupervised clustering method adapted from the GEORGE method as described by Sohoni et al.[8] We first train a ResNet18 in the method as described in section

3.1 using ERM. We then extract the 512-dimensional activations from this model as the features to reduce using Uniform Manifold Approximation and Projection (UMAP) and cluster using Gaussian Mixture. The predicted clusters then become the subclass labels. For the GEORGE method, we use the silhouette coefficient as our metric to determine the optimal number of clusters.

Lastly we stratify nodules by radiologist rating of malignancy likelihood. This is motivated by the UMAP reduction we did for the labels derived by clustering during the GEORGE process. In the UMAP space, we notice that visualizing the points by malignancy seems to closely approximate the clusters we derived using the Gaussian Mixture modeling (see Section 4.2 for more details). Specifically, within the superclasses of malignant and benign, there are finer categories corresponding to the ratings given by the radiologists. After dropping nodules with a malignancy likelihood of 3, the remaining options in increasing order of malignancy correspond to ”Highly Unlikely”, ”Moderately Unlikely”, ”Moderately Suspicious”, and ”Highly Suspicious”, which are the labels used by the radiologists when rating the nodules.[28] For the remainder of this paper we will refer to these instead as ”Most Likely Benign”, ”Moderately Likely Benign”, ”Moderately Likely Malignant”, and ”Most Likely Malignant” for the sake of clarity.

3.5 ERM and gDRO

For model training, we utilize both ERM and gDRO as loss functions and compare them. The intuition between the two losses is a trade-off between overall performance of the model and worst class performance of the model. The general expectation is that ERM should result in greater overall accuracy than gDRO, but gDRO should result in a more “robust” model in the sense that the performance of the worst subclass should be greater.

The most typical per-dataset loss function is the Empirical Risk Minimization (ERM) scheme where each per-sample loss is weighted equally. More specifically, given that we have some training dataset

$$D = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$$

with n data points, where x_i is the observation and y_i is the label, a model F where $F(x_i) = \hat{y_i}$, and some per-sample loss function ℓ, ERM loss can be defined as the following:

$$\text{loss}_{\text{ERM}}(D) = \frac{1}{n}\sum_{i=1}^{n} \ell(F(x_i), y_i)$$

As seen in this scheme, we make the assumption that the per-sample loss for each $(x_i, y_i) \in D$ should be weighted equally. However the above assumption is not necessarily true, as a dataset might have so called “hidden stratification” where certain subsets of the training data might be different from others.

Hence, another per-dataset loss function – the group Distributionally Robust Optimization (gDRO) scheme has been proposed to correct this problem.[7] Specifically given that we have some training dataset D such that it can be divided into K disjoint sub-datasets $D_1, D_2, ...D_K$, then the gDRO loss can be defined as the following:

$$\text{loss}_{\text{gDRO}}(D) = \max_{k=1,...K} \text{loss}_{\text{ERM}}(D_k)$$

Specifically, here unlike ERM, the per-sample loss for each $(x_i, y_i) \in D$ is not weighted equally across the entire training dataset. Rather it is weighted equally among the subset D_k that (x_i, y_i) belongs to. And as the gDRO loss is the maximum loss among the sub-datasets, this theoretically should force the model to optimize along worst-class performance rather than the overall performance.

The specific implementation of gDRO we use is adapted from code provided by Gulrajani et al.[29] which implements the algorithm proposed by Sagawa et al.[7]

4. RESULTS

4.1 Spiculation Subclasses

Our first experiment is with stratifying the dataset by spiculation. In agreement with our hypothesis in section 3.4, the most common subclasses when stratifying by malignancy and spiculation are unspiculated benign and spiculated malignant. The sizes of each subclass can be seen in Table 1.

Table 1. Subclass counts (Spiculation)

Counts	Unspiculated	Spiculated	Total
Benign	781	116	897
Malignant	204	387	591
Total	802	686	1488

Comparing performance of ERM and gDRO on the spiculated subclasses using the deep features, designed features and images (Table 2), we find that ERM does better than gDRO at a statistically significant level for all feature representations in terms of overall performance. However, while gDRO does allow the designed features model to improve on the worst group performance as compared to ERM, this is not the case for the deep feature model. This suggests that the feature representation of model inputs influences the effect gDRO has on worst class performance. The worst group in all cases is Unspiculated Malignant, which agrees with our hypothesis that one of the minority classes (Spiculated Benign or Unspiculated Malignant) would have the worst performance.

Table 2. Overall accuracy and subclass sensitivities for designed feature, deep features and images averaged over 100 trials. All reported values have a standard error less than 0.01. Bold numbers indicate better performance significant at $p < 0.05$.

Feature representation	**Designed Features**		**Deep Features**		**Images**	
Model	ERM	gDRO	ERM	gDRO	ERM	gDRO
Overall accuracy	**0.889**	0.863	**0.865**	0.864	**0.873**	0.870
Subclass sensitivities	–	–	–	–	–	–
Unspiculated Benign	**0.921**	0.851	**0.929**	0.928	**0.899**	0.890
Spiculated Benign	**0.847**	0.804	0.926	0.926	0.907	0.897
Spiculated Malignant	**0.900**	0.891	0.835	0.834	0.883	0.884
Unspiculated Malignant (Worst)	0.775	**0.896**	0.642	0.643	0.739	0.750

4.2 Clustered Subclasses

For our second experiment of clustered labels, when we directly apply the GEORGE method to the training dataset embeddings, we get that the maximal silhouette coefficient of 0.85 occurs when we only have two clusters. Specifically, the two clusters produced are exactly defined by the superlabels of benign or malignant as seen in Figure 4.

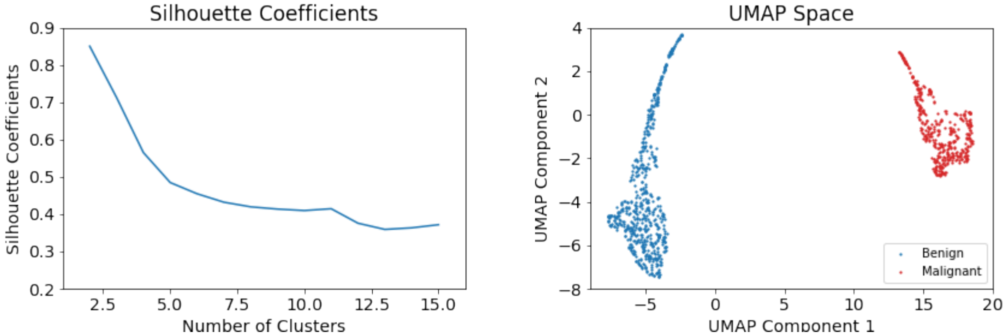


Figure 4. Left: Line plot of average silhouette scores across clusters of 2-15 clusters when done only on all nodules from the training dataset. Right: UMAP visualization of the training embeddings and the resulting clusters when fitting the Gaussian Mixture model for only 2 clusters

As our objective is to get subtype labels, the above derived clusters are insufficient. They only delineate the superclasses, which provides no subclass information. Thus we repeat the process of clustering respectively on the malignant nodules and the benign nodules since the high silhouette coefficient is indicative of them being separate types. Here, we get similar results for both types of nodules with the highest silhouette coefficient resulting when each type of nodule is further partitioned into two clusters. Using these clusters, we result with four subclass, two under the label of malignant and two under the label of benign as seen in Figure 5.

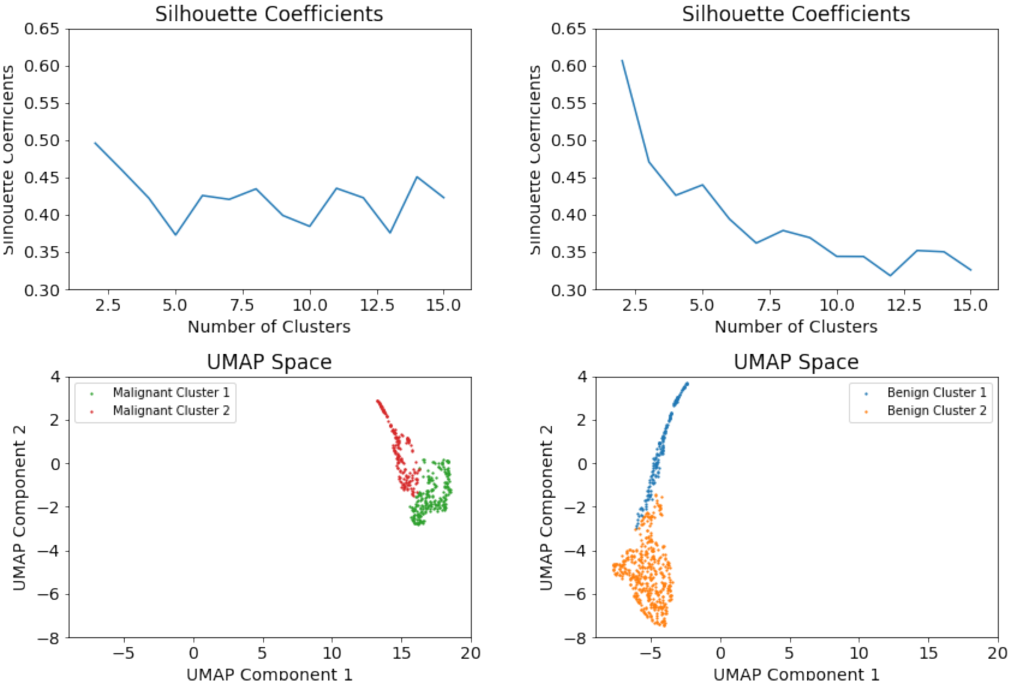


Figure 5. Top to Bottom: Line plot of average silhouette scores across clusters of 2-15 clusters when done only on the Malignant Nodules (left) and Benign Nodules (right) from the training dataset, UMAP visualization of the training embeddings and the resulting clusters when fitting the Gaussian Mixture model for only 2 clusters on Malignant (left) and Benign (right)

We notice that the clusters derived from the above method can be approximated by the "clusters" we would get if we label the UMAP feature reduction by their corresponding malignancy likelihood. Specifically for each malignancy likelihood of 1 (least likely),2,4 and 5(most likely), there exists a cluster derived from GEORGE that is predominantly of that malignancy likelihood. By running multiple trials, we can see that the randomization of the transfer model's initial weights causes the UMAP reduction and therefore the clustering to produce different results (Figure 6).

Hence utilizing this relationship, we attempt to decrease the variability of our generated clusters. Specifically we separately train 50 ResNet18 models to derive 50 sets of activation feature that we cluster. By looking at the counts of malignancy likelihood ratings in each cluster, we assign each cluster to either Predominantly Most Likely Benign, Predominantly Moderately Likely Benign, Predominantly Moderately Likely Malignant or Predominantly Most Likely Malignant. We then calculate the mode of the cluster labels of all 50 trials for each nodule to derive a "stable" subtyping of the nodules, as seen in Figure 7.

As the derivation of "stable" clusters relies on the prior that the unsupervised clusters is noisily approximated by malignancy likelihood, we test this prior with the hypothesis that the average silhouette score of our "stable" subtyping should be higher than the average silhouette score if we only used malignancy likelihood as labels. For this, we again run 50 experiments where each time we train a transfer learning model to extract features and calculate the silhouette scores in the UMAP reduced space of both the clusters by likelihood and clusters by mode-of-50-trials labels. We then compute a single-sided one sample T-Test and get a p-value of 1.23×10^{-42} which is significant at the 0.05 threshold level. This allows us to conclude our "stable" clustering does give higher silhouette scores – and as a result generates better clusters – compared to only stratifying by malignancy likelihood.

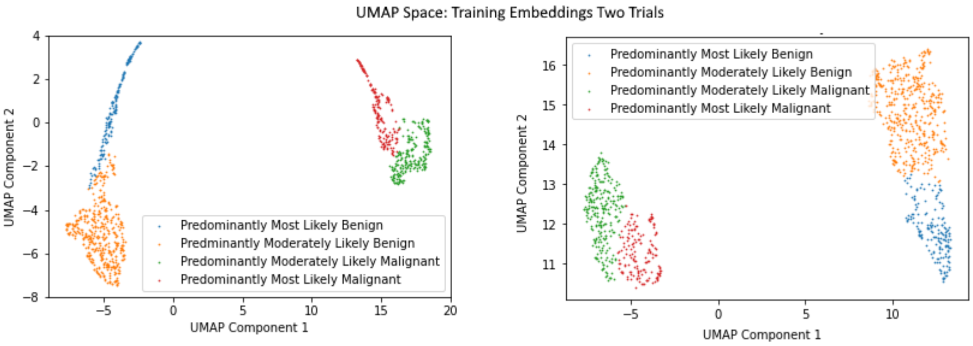


Figure 6. The results of two trials of the clustering process. Each trial produces a different UMAP embedding, which causes the clusters to differ. The axes for the two graphs are not equivalent as they refer to two different UMAP feature spaces.

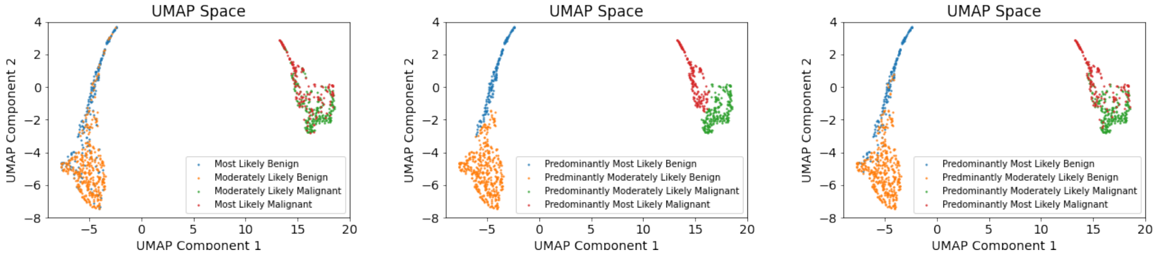


Figure 7. From left to right: a UMAP visualization of the train embeddings labeled by their malignancy likelihood values, a UMAP visualization of the derived clusters when clustering is done once (specifically on these exact embeddings that are visualized), a UMAP visualization of the clusters derived from the mode value of 50 trials

Thus using an adaptation of the GEORGE method, we are able to calculate stable sub-type labels with some interpretability as they are noisily approximated by malignancy likelihood ratings (Table 3).

Table 3. Subclass counts of clustered labels grouped by the malignancy likelihood ratings (LB = Likely Benign, LM = Likely Malignant)

Cluster Counts	Moderately LB (2)	Most LB (1)	Total
Predominantly Moderately LB	514	118	632
Predominantly Most LB	48	217	265
–	Moderately LM (4)	Most LM (5)	–
Predominantly Moderately LM	281	86	367
Predominantly Most LM	75	149	224

When comparing results for ERM and gDRO using the clustered labels (Table 4), we find that gDRO does increase worst class performance for all types of features. However the difference between ERM and gDRO is much more pronounced when using the designed features than it is when using the deep features or images. This again shows that the nature of the feature representation has am effect on relative ERM and gDRO performance. Also from the table, we see that both ERM and gDRO achieved perfect or near-perfect test accuracy on the extreme subclasses (Predominantly Most Likely Benign and Predominantly Most Likely Malignant), aligning with our hypothesis that these subclasses are easy to learn. This suggests that the clustering methodology is effective in separating out nodules that are easy for classifiers to learn and identify. Lastly, as expected by our intuition about the clusters, Predominantly Moderately Likely Malignant proved to be the worst-performing subclass. The "Moderately Likely" subclasses are intuitively more ambiguous in the feature space and are harder to classify. Since the malignant class is smaller than the benign class, we expect it to perform worse.

Table 4. Overall accuracy and subclass sensitivities for designed feature, deep features and images averaged over 100 trials. All reported values have a standard error less than 0.01. Bold numbers indicate better performance significant at $p < 0.05$. LB is Likely Benign, LM is Likely Malignant.

Feature representation	**Designed Features**		**Deep Features**		**Images**	
Model	ERM	gDRO	ERM	gDRO	ERM	gDRO
Overall accuracy	**0.888**	0.865	0.865	**0.865**	0.874	0.871
Subclass sensitivities	–	–	–	–	–	–
Predominantly Most LB	**0.977**	0.915	1.00	1.00	0.973	0.971
Predominantly Moderately LB	**0.882**	0.800	**0.901**	0.897	**0.871**	0.855
Predominantly Moderately LM (Worst)	0.795	**0.875**	0.667	**0.675**	0.749	**0.769**
Predominantly Most LM	1.00	1.00	1.00	1.00	0.981	0.982

4.3 Malignancy Likelihood Subclasses

For our last stratification type, we use the malignancy likelihood labels given by radiologists as a cross comparison against the clustered labels. As seen in Table 5, the majority of nodules are either moderately likely benign or moderately malignant. At the same time they are also our worse performing subclasses for both the deep features and designed features (Table 6).

Table 5. Subclass counts of Moderately Likely and Most Likely Nodules by malignancies

Counts	Moderately Likely	Most Likely	Total
Benign	562	335	897
Malignant	356	235	591
Total	918	570	1488

Table 6. Overall accuracy and subclass sensitivities for designed feature, deep features and images averaged over 100 trials. All reported values have a standard error less than 0.01. Bold numbers indicate better performance significant at $p < 0.05$.

Feature representation	**Designed Features**		**Deep Features**		**Images**	
Model	ERM	gDRO	ERM	gDRO	ERM	gDRO
Overall accuracy	**0.889**	0.863	**0.876**	0.875	0.870	0.871
Subclass sensitivities	–	–	–	–	–	–
Most Likely Benign	**0.888**	0.865	0.941	0.941	0.945	0.944
Moderately Likely Benign	**0.889**	0.787	**0.868**	0.863	0.869	0.865
Moderately Likely Malignant (Worst)	0.762	**0.838**	0.758	**0.761**	0.744	0.752
Most Likely Malignant	1.00	1.00	0.979	0.979	0.958	0.958

Specifically, if we again compare results from ERM and gDRO on the different types of features, the performance closely mirrors that of the clustered labels. The only difference being that the performance on the Most Likely classes are lower than the predominantly most likely classes and inversely the Moderately Likely classes are higher than the predominantly moderately likely classes. This may be because the clusters are better separated in the feature space than the malignancy labels are, as we determined using silhouette coefficients in Section 4.2.

These results again corroborates our view that the clustered labels roughly indicate the "hardness" of a nodule to be classified and the malignancy likelihoods noisily approximate these clusters. At the same time, the inability of ERM to perform well on these borderline malignant subclasses – despite the fact that the counts of borderline malignant nodules dominate the counts of more well-defined malignancy nodules – suggests that there may exist further stratification in the borderline malignant nodules that we have not found.

5. DISCUSSION

5.1 Overall Findings

We proposed a novel CAD scheme that stratifies lung nodules into different subgroups and enhances the malignancy classification performance on the worst-performing subgroup. In terms of stratifying the LIDC dataset, we find that all of the stratification methods we employ are valid subtypings as indicated by the differing performance of our ERM models on each subclass. Similarly the high degree of overlap between the clustered stratification and malignancy likelihood stratification (Table 3) indicate that the clustering method is semantically meaningful. Therefore we can conclude that all three stratification methods tested in this paper provide information about the underlying structure of the data.

Our malignancy classification result (Table 2) using gDRO suggests that we can improve the performance on the worst-performance lung nodule group, which is often misclassified by traditional CAD models with ERM loss. Although, the degree of improvement can be fairly modest as seen for the Deep and Image features. Our classification result is consistent with the previous finding that gDRO sacrifices overall accuracy while improving the worst-performance group sensitivity.

In terms of comparison of ERM and gDRO, one of our major findings is that gDRO requires good feature representations to improve worst-group performance. gDRO has the greatest effect when using the designed features. We theorize that this is because the designed features are created using domain knowledge, and are therefore very effective and information-dense features. For instance, they directly incorporate information about the size and shape of the nodule, while a CNN must learn these features for itself. We further hypothesize that the ResNet18 was not able to find a “good” representation of the LIDC data because there were too few training samples to create effective features. Previous work has shown that without specific modifications, deep CNNs require a very large amount of data to learn an effective model.[30,31] Thus we may need to use a more intricate architecture to work around the limitation of the small amount of data we have.

Our experimental results agree with the idea put forth by Rosenfeld et al.[25] that an image classification model can be broken down into a featurizer and a classifier, and that a robust model depends strongly on the ability of the featurizer to learn good features. Since our model could not form good features from the limited data, gDRO has only a small effect on the worst-group performance.

5.2 Future Works

Although we tested three stratification methods with varying levels of domain-specific information, we believe that there are more methods to explore and that some of them will provide better performance from gDRO. Future work should focus on finding and testing new methods for stratification, as well as potentially combining previous methods to achieve a combination of domain- and data-driven subclasses. Similar reasoning applies to the representation of the data, as we have shown that it is at least as important for gDRO as the choice of subclasses. Possible avenues of future work include combining designed and deep features, using more intricate model architectures, testing different sets of designed features, and building models that can learn features on par with the designed features.

5.3 Reproducibility

We train and test our models using PyTorch. The code we use to run our experiments can be found at https://github.com/mtzig/LIDC_GDRO.

ACKNOWLEDGMENTS

This work was supported by NSF CISE REU Program Award Number 1950894.

REFERENCES

[1] Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E., and Forman, D., "Global cancer statistics," *CA: a cancer journal for clinicians* **61**(2), 69–90 (2011).

[2] Jemal, A., Center, M. M., DeSantis, C., and Ward, E. M., "Global patterns of cancer incidence and mortality rates and trendsglobal patterns of cancer," *Cancer epidemiology, biomarkers & prevention* **19**(8), 1893–1907 (2010).

[3] Kumar, D., Wong, A., and Clausi, D. A., "Lung nodule classification using deep features in ct images," in [*2015 12th conference on computer and robot vision*], 133–138, IEEE (2015).

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

[5] Hu, H.-D., Wan, M.-Y., Xu, C.-H., Zhan, P., Zou, J., Zhang, Q.-Q., and Zhang, Y.-Q., "Histological subtypes of solitary pulmonary nodules of adenocarcinoma and their clinical relevance," *Journal of Thoracic Disease* **5**(6), 841 (2013).

[6] Seidelman, J. L., Myers, J. L., and Quint, L. E., "Incidental, subsolid pulmonary nodules at ct: etiology and management," *Cancer Imaging* **13**(3), 365 (2013).

[7] Sagawa*, S., Koh*, P. W., Hashimoto, T. B., and Liang, P., "Distributionally robust neural networks," in [*International Conference on Learning Representations*], (2020).

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

[9] Monkam, P., Qi, S., Ma, H., Gao, W., Yao, Y., and Qian, W., "Detection and classification of pulmonary nodules using convolutional neural networks: a survey," *Ieee Access* **7**, 78075–78091 (2019).

[10] Zhu, W., Liu, C., Fan, W., and Xie, X., "Deeplung: Deep 3d dual path nets for automated pulmonary nodule detection and classification," in [*2018 IEEE winter conference on applications of computer vision (WACV)*], 673–681, IEEE (2018).

[11] Jiang, H., Shen, F., Gao, F., and Han, W., "Learning efficient, explainable and discriminative representations for pulmonary nodules classification," *Pattern Recognition* **113**, 107825 (2021).

[12] Al-Shabi, M., Shak, K., and Tan, M., "Procan: Progressive growing channel attentive non-local network for lung nodule classification," *Pattern Recognition* **122**, 108309 (2022).

[13] Al-Shabi, M., Lee, H. K., and Tan, M., "Gated-dilated networks for lung nodule classification in ct scans," *IEEE Access* **7**, 178827–178838 (2019).

[14] Zhao, J., Zhang, C., Li, D., and Niu, J., "Combining multi-scale feature fusion with multi-attribute grading, a cnn model for benign and malignant classification of pulmonary nodules," *Journal of digital imaging* **33**(4), 869–878 (2020).

[15] Li, X., Kao, Y., Shen, W., Li, X., and Xie, G., "Lung nodule malignancy prediction using multi-task convolutional neural network," in [*Medical Imaging 2017: Computer-Aided Diagnosis*], **10134**, 551–557, SPIE (2017).

[16] Da Nóbrega, R. V. M., Peixoto, S. A., da Silva, S. P. P., and Rebouças Filho, P. P., "Lung nodule classification via deep transfer learning in ct lung images," in [*2018 IEEE 31st international symposium on computer-based medical systems (CBMS)*], 244–249, IEEE (2018).

[17] Zhao, X., Qi, S., Zhang, B., Ma, H., Qian, W., Yao, Y., and Sun, J., "Deep cnn models for pulmonary nodule classification: model modification, model integration, and transfer learning," *Journal of X-ray Science and Technology* **27**(4), 615–629 (2019).

[18] Yu, J., Yang, B., Wang, J., Leader, J. K., Wilson, D. O., and Pu, J., "2d cnn versus 3d cnn for false-positive reduction in lung cancer screening," *Journal of Medical Imaging* **7**(5), 051202 (2020).

[19] Nibali, A., He, Z., and Wollersheim, D., "Pulmonary nodule classification with deep residual networks," *International journal of computer assisted radiology and surgery* **12**(10), 1799–1808 (2017).

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

[21] Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., et al., "Wilds: A benchmark of in-the-wild distribution shifts," in [*International Conference on Machine Learning*], 5637–5664, PMLR (2021).

[22] Armato III, S. G., McLennan, G., Bidaut, L., McNitt-Gray, M. F., Meyer, C. R., Reeves, A. P., Zhao, B., Aberle, D. R., Henschke, C. I., Hoffman, E. A., et al., "The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans," *Medical physics* **38**(2), 915–931 (2011).

[23] Kumar, V., Gu, Y., Basu, S., Berglund, A., Eschrich, S. A., Schabath, M. B., Forster, K., Aerts, H. J., Dekker, A., Fenstermacher, D., et al., "Radiomics: the process and the challenges," *Magnetic resonance imaging* **30**(9), 1234–1248 (2012).

[24] Zinovev, D., Raicu, D., Furst, J., and Armato III, S. G., "Predicting radiological panel opinions using a panel of machine learning classifiers," *Algorithms* **2**(4), 1473–1502 (2009).

[25] Rosenfeld, E., Ravikumar, P., and Risteski, A., "Domain-adjusted regression or: Erm may already learn features sufficient for out-of-distribution generalization," *arXiv preprint arXiv:2202.06856* (2022).

[26] Hancock, M. C. and Magnan, J. F., "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* **3**(4), 044504 (2016).

[27] Wahidi, M. M., Govert, J. A., Goudar, R. K., Gould, M. K., and McCrory, D. C., "Evidence for the treatment of patients with pulmonary nodules: when is it lung cancer?: Accp evidence-based clinical practice guidelines," *Chest* **132**(3), 94S–107S (2007).

[28] McNitt-Gray, M. F., Armato III, S. G., Meyer, C. R., Reeves, A. P., McLennan, G., Pais, R. C., Freymann, J., Brown, M. S., Engelmann, R. M., Bland, P. H., et al., "The lung image database consortium (lidc) data collection process for nodule detection and annotation," *Academic radiology* **14**(12), 1464–1474 (2007).

[29] Gulrajani, I. and Lopez-Paz, D., "In search of lost domain generalization," in [*International Conference on Learning Representations*], (2021).

[30] Keshari, R., Ghosh, S., Chhabra, S., Vatsa, M., and Singh, R., "Unravelling small sample size problems in the deep learning world," in [*2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM)*], 134–143, IEEE (2020).

[31] Liu, S. and Deng, W., "Very deep convolutional neural network based image classification using small training sample size," in [*2015 3rd IAPR Asian conference on pattern recognition (ACPR)*], 730–734, IEEE (2015).

APPENDIX A. EXPERIMENTAL CONSTANTS

Each data representation (designed features, deep features, and images) requires the use of a different model architecture:

- Designed features: Fully connected neural network with layer sizes 64, 36, 2
- Deep features: Fully connected neural network with layer sizes 512, 64, 36, 2
- Images: Pre-trained ResNet18 with learning rate scheduler, single fully connected layer of size 2

The Designed Features and Deep Features models are trained for 100 epochs with a batch size of 128, which we determined is enough time for them to achieve good validation accuracy in all cases without overfitting. The Image models, owing to their relative size and complexity, needed a different training regime than the smaller fully-connected models. They are trained for only 15 epochs, as they are able to learn and overfit the data much quicker. They also made use of a learning rate scheduler to mitigate overfitting. The learning rate scheduler used is PyTorch's ReduceLROnPlateau, with mode='max', factor=0.2 and patience=2. The training data for the Image models is also augmented with a flip over the X-axis and $90^\circ, 180^\circ, 270^\circ$ rotations.

All models use the same optimizer, namely, PyTorch's Adam optimizer with a learning rate of 0.0005 and a weight decay of 0.005. gDRO uses $\eta = 0.01$ as the hyperparameter governing the scaling of the subclass loss weights. These hyperparameters and optimizers are developed from those provided by Yu et al.[18] with further cross-validation tuning. For more details about the algorithm, refer to the paper by Sagawa et al.[7]

APPENDIX B. RESULTS AND STATISTICAL ANALYSIS

For each test, results are collected for ERM and gDRO. Each test consists of 100 trials of ERM and 100 of gDRO. In each trial, the model is trained for 100 epochs for the deep and designed features models or 15 epochs for the images model. The results presented in this paper all represent the model's test performance after the final training epoch. The model's test-set sensitivities on each subclass, as well as the overall test-set accuracy, are recorded for each epoch and each trial. The model is reinitialized at the beginning of each trial, and the initial model weights are randomized for each trial. In addition, the data batching is randomized across trials, although the train/cross-validation/test split remains constant over all experiments.

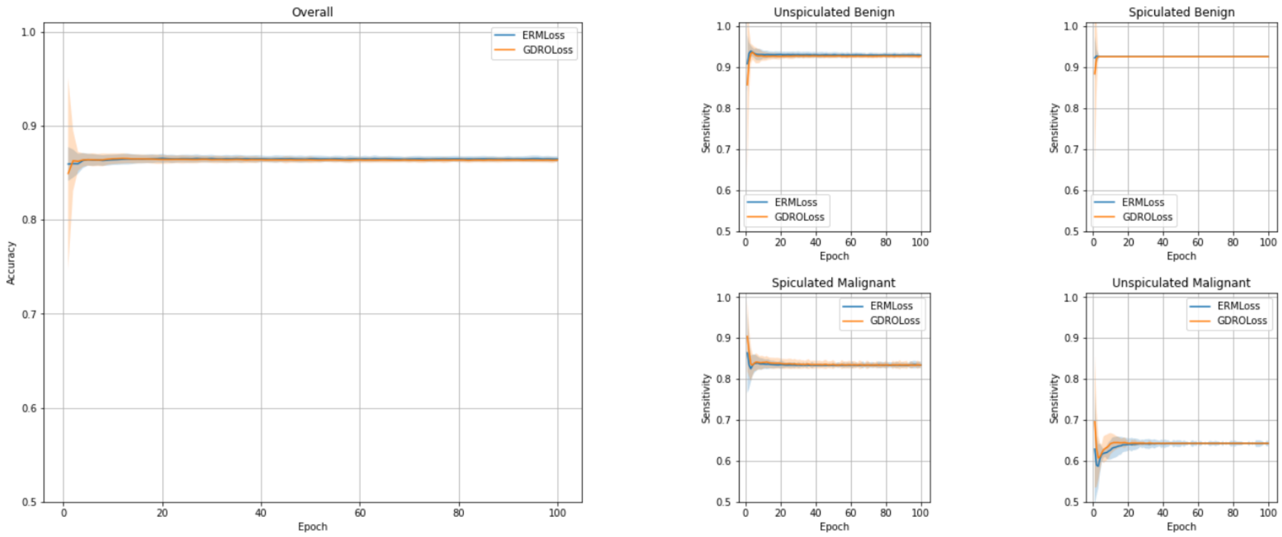
The results were analyzed using a two-sample two-tailed t-test on the difference in mean sensitivities between ERM and gDRO for each subclass, using a significance threshold of $p < 0.05$. As the t-test assumes normality, the data were tested for normality using a Q-Q plot and were found to be approximately normally distributed.

APPENDIX C. PERFORMANCE OF ERM AND GDRO ACROSS EPOCHS

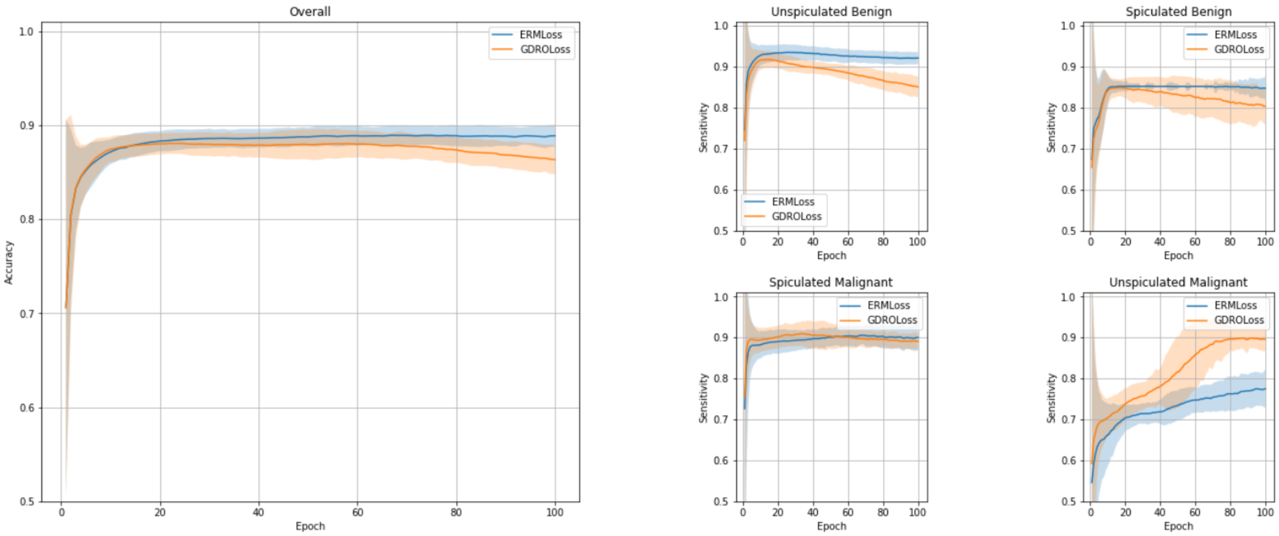
While the results presented in this paper only represent the model's test performance on the final epoch, the model's accuracy and subclass sensitivities were recorded for every epoch. Below are the graphs of the mean sensitivity/accuracy for each epoch, for every experiment whose results were previously featured in this paper. Each graph contains two lines, one representing ERM and the other representing gDRO. In addition, there is a shaded area representing the region within one standard deviation of the mean sensitivity/accuracy.

C.1 Spiculation Subclasses

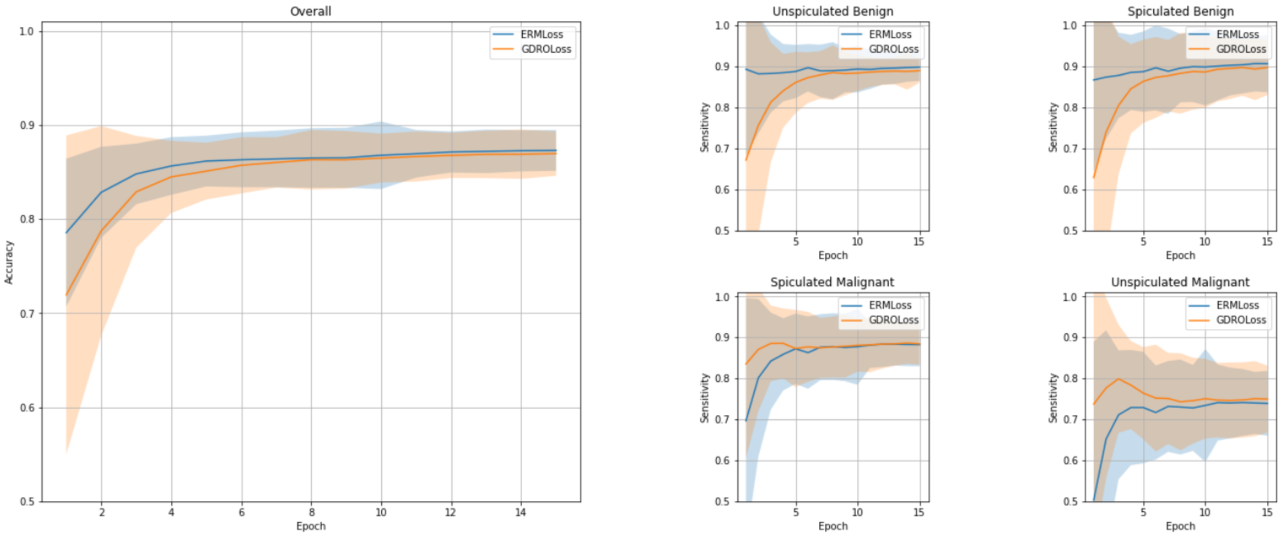
Deep Features



Designed Features

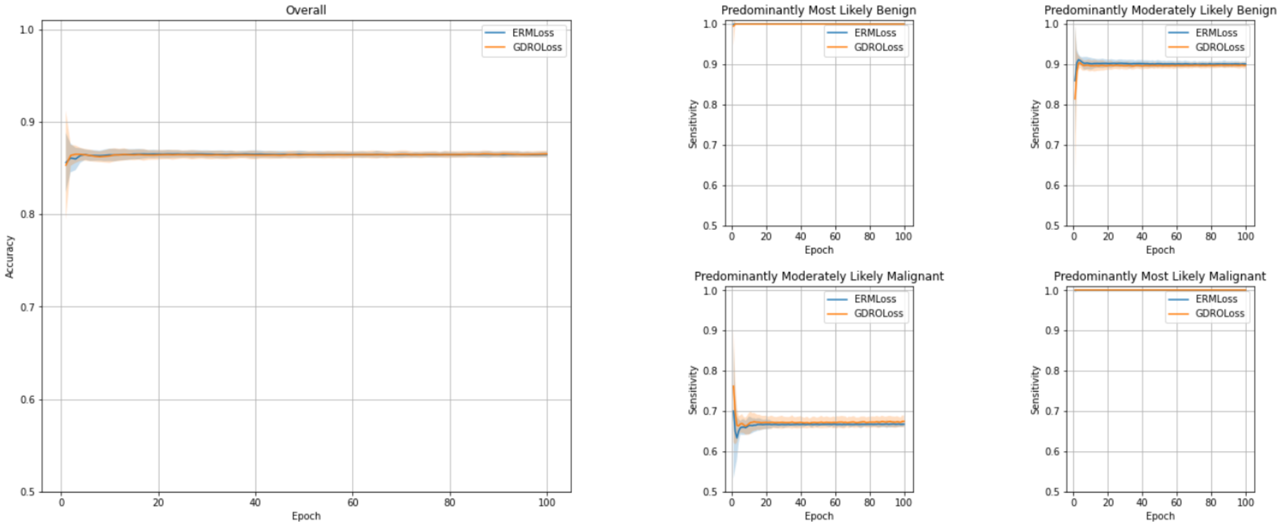

Overall
Accuracy
Epoch
1.0
0.9
0.8
0.7
0.6
0.5
0
20
40
60
80
100
ERMLoss
GDROLoss
Unspiculated Benign
Spiculated Benign
Spiculated Malignant
Unspiculated Malignant
Sensitivity

Images

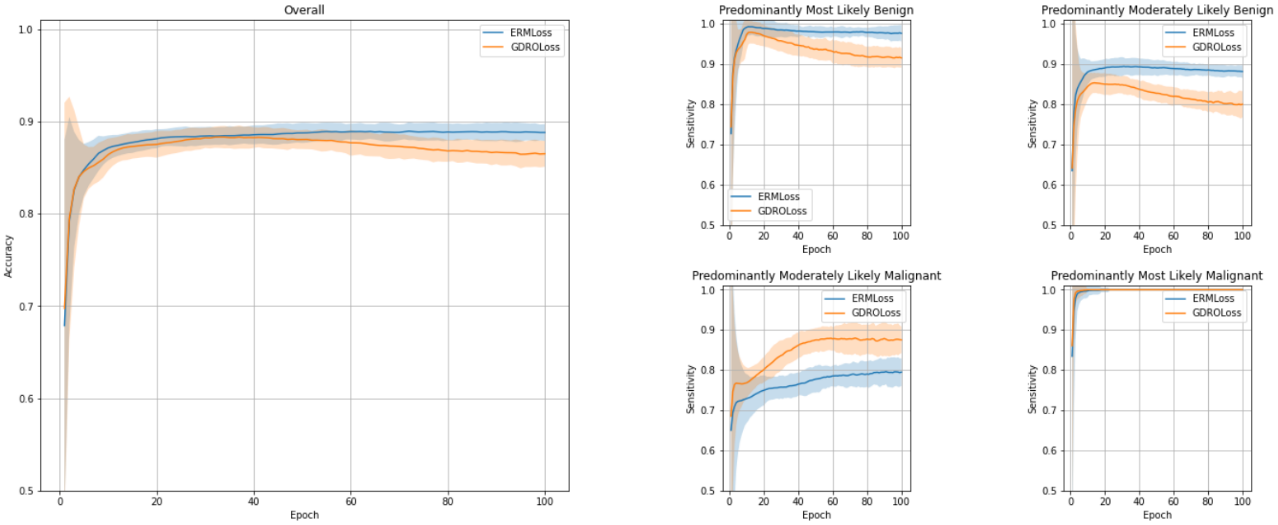


C.2 Cluster Subclasses

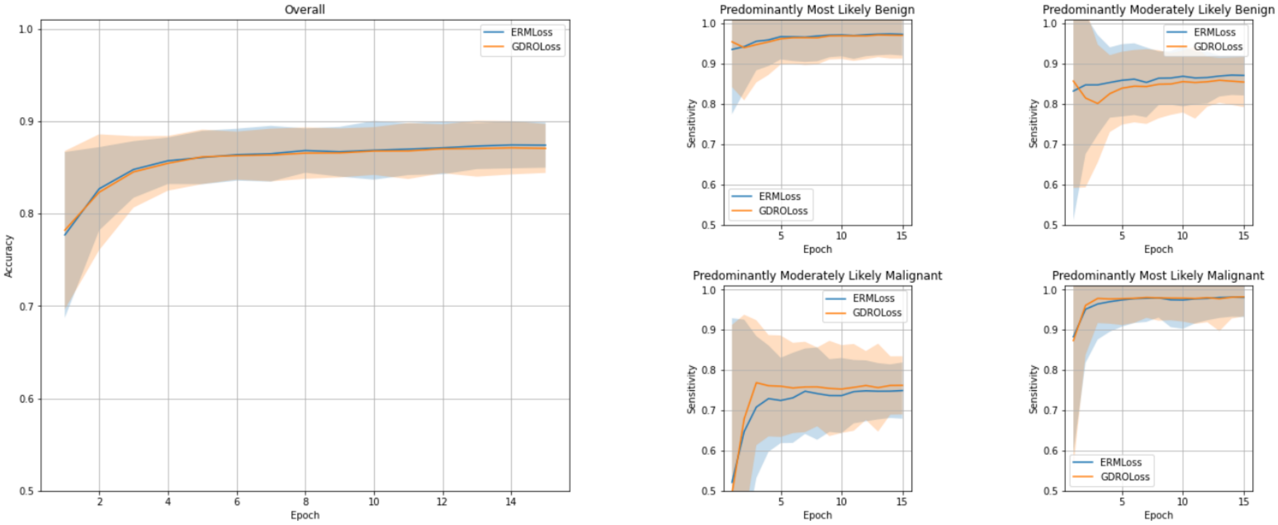
Deep Features



Designed Features

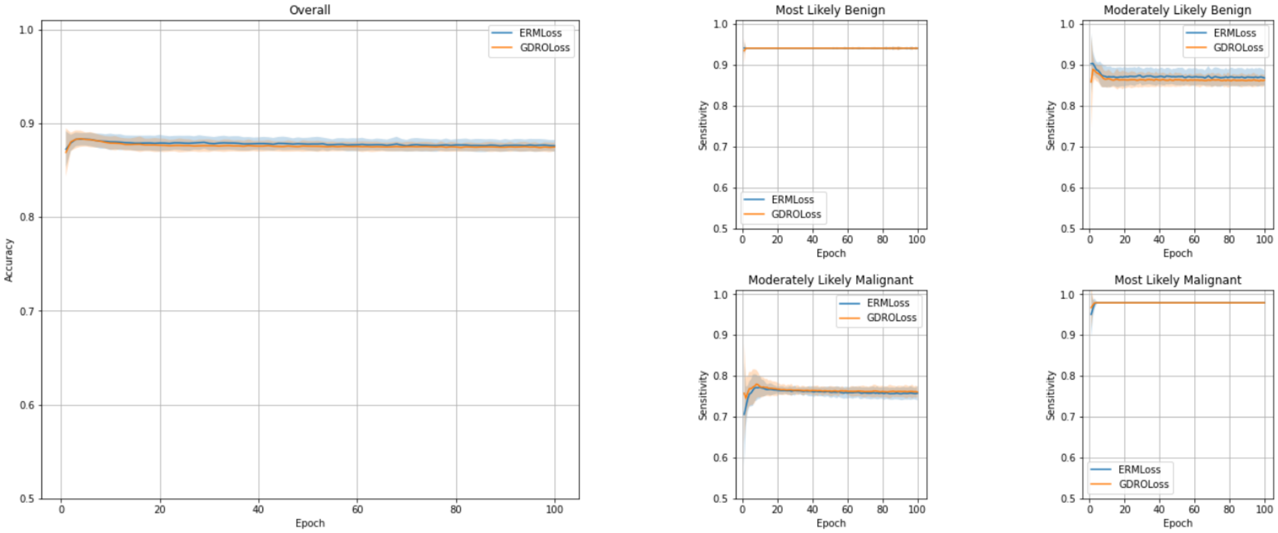


Images

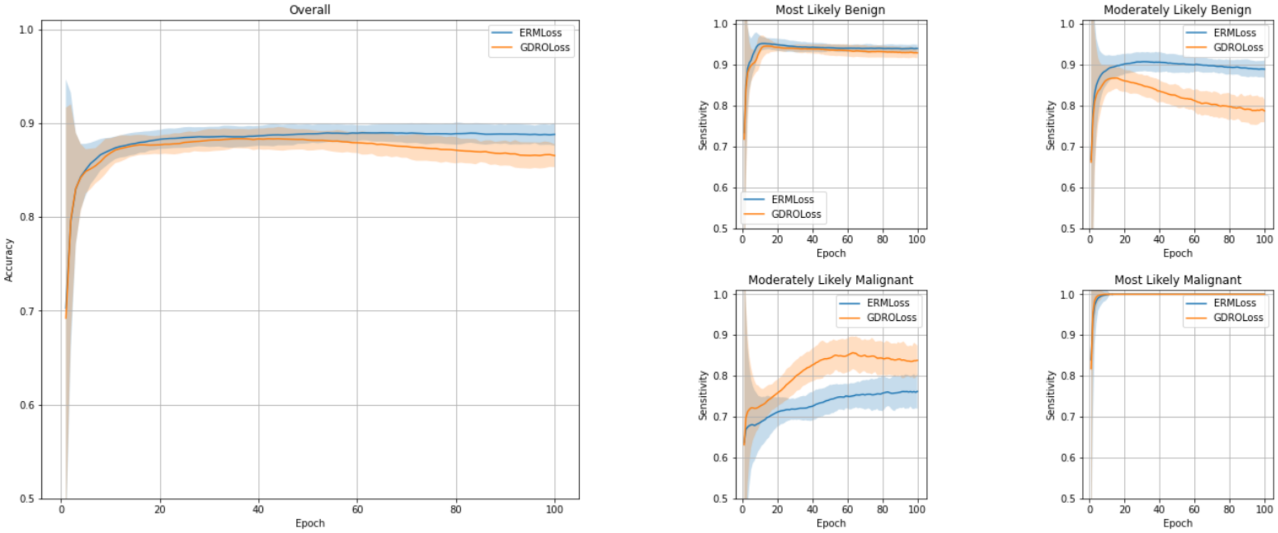


C.3 Malignancy Subclasses

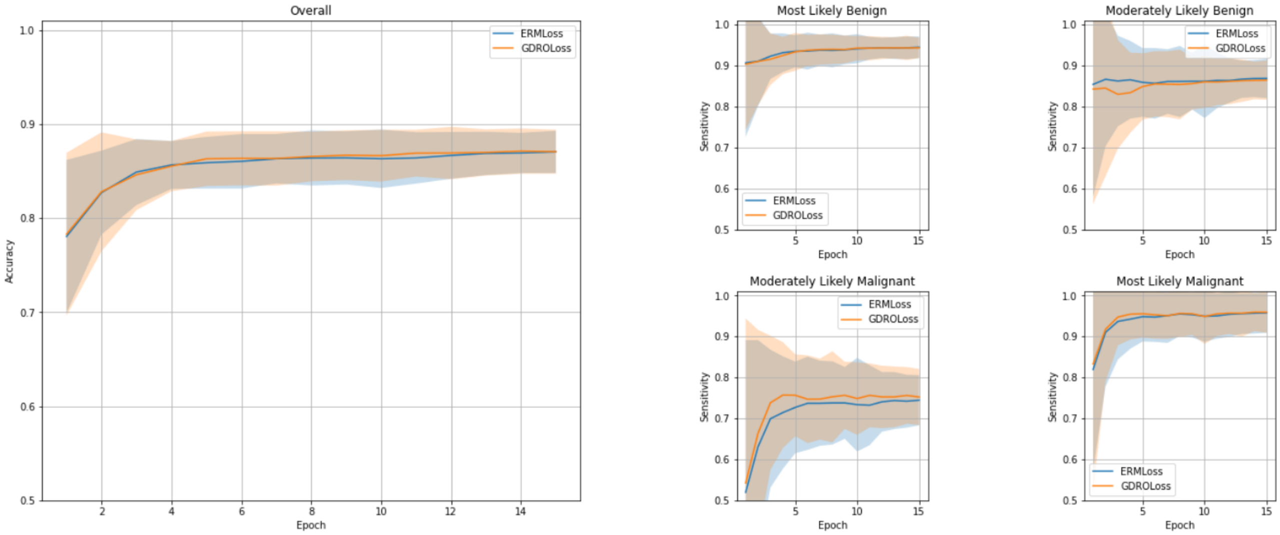
Deep Features



Designed Features



Images


Overall
Accuracy
Epoch
0.5
0.6
0.7
0.8
0.9
1.0
2
4
6
8
10
12
14
ERMLoss
GDROLoss
Most Likely Benign
Moderately Likely Benign
Moderately Likely Malignant
Most Likely Malignant
Sensitivity
Epoch
5
10
15