Noname manuscript No.
(will be inserted by the editor)

Identifying Difficult Cases through Selective Iterative Classification and Label Propagation

Eura Shin, Samuel Berglin

the date of receipt and acceptance should be inserted later

Abstract Healthcare providers can leverage information on the diagnostic difficulty of a case to improve diagnostic confidence, improve the peer-review system, and train medical students. In this paper we use data from the Lung Image Database Consortium (LIDC) to determine the diagnostic difficulty of a lung nodules depicted in Computed Tomography (CT) scans. We discuss a formal approach to determining the most diagnostically difficult cases within the LIDC by building an ensemble of classifiers that reduce label uncertainty and consider the samples misclassified by these classifiers. We identify the most difficult cases and compare their characteristics to the remainder of the dataset. We find that hard cases are significantly different from non-hard cases with respect to ten separate low-level image features. Of these features, seven are related to the size of the nodule, two to the gabor texture features, and the other entropy. Radiologists also disagreed on how subtle these difficult nodules appear to be in the CT scans.

1 Introduction

Identifying the diagnostic difficulty of a case is essential to improving health care procedures. As medical imaging technology improves, radiologists are expected to manage an increasing volume of images while upholding the quality of their diagnoses. This is impossible without some sort of system to assess the diagnostic difficulty of an medical image, which can be used to guide the efficient allocation of resources with respect to: 1) assessing diagnostic confidence, 2) developing comprehensive peer-review systems, and 3) training medical students.

Early detection is crucial to the success of most cancer treatments. Radiologists will diagnose a malignant lung nodule by assessing a CT scan of the abnormality. These diagnoses may be confirmed by measures which are invasive and potentially dangerous to the patient, such as a biopsy or repeated exposure to radiation over time in the form of CT scans. In cases for which this follow up information is not available, multiple radiologists may be asked to assess the same image in order to reduce uncertainty. Unfortunately, in addition to be extremely expensive in terms of time and money, these diagnoses are highly variable [Rei18]. Considering this variability, it is essential that the diagnosis of a cancer is accompanied by a high diagnostic confidence. The ability to distinguish cases based on difficulty prior to evaluation will allow hospitals to efficiently allocate the use of radiologist's time. On a similar note, knowing the difficulty of a case will allow us to know how many experts are needed for a case, and therefore, better allocate the labels needed in building CADx systems.

Moreover, difficult cases could be used for radiologist evaluation and training. In radiology, physician performance is evaluated on diagnostic accuracy [MKY+09]. Conventionally, performance assessment is determined using only 10% of an expert's total diagnostic cases [Rei14]. A preferred

Address(es) of author(s) should be given

approach, proposed by Reiner, would be to outsource review cases to a neutral, external third party of experts using a standardized system [Rei14] where multiple radiologists independently assess the same image. To perform a comprehensive and fair evaluation of physician performance, these outsourced review cases must represent a diverse subset of medical cases in terms of difficulty. A randomized selection of images theoretically represents the full spectrum of the work required by a radiologist. However, [Rei17] speaks of a hybrid peer review selection system that combines random and targeted case selection, the latter of which "is intended to identify high-risk cases." Difficult cases can be used to assist in targeted case selection. These difficult cases could also be stored for a radiologist training scheme similar to hybrid peer review.

In this paper we introduce a method of identifying diagnostically difficult cases through two contributions. First, we present a selective iterative classification method that uses propagated labels to save labels and reduce noise within training sets for CADx systems. We then use the CADx systems trained on these noise-reduced labels to identify diagnostically difficult cases. We apply our approach on lung nodule Computed Tomography (CT) scans from the Lung Image Database Consortium (LIDC)[AIMB+11]. Notions of case difficulty are produced as a byproduct of training these noise-reduced classifiers. By drawing the connection between difficult classification tasks and difficult diagnostic cases, we are able to identify the most diagnostically difficult cases within the LIDC dataset. We then analyze the semantic characteristics of the most difficult cases indicated by our algorithms, with the expectation that the characteristics presented in this paper may be used to build classifiers that predict case difficulty.

2 Background

2.1 CADx Systems

Computer-Aided Diagnosis (CADx) systems act as “second readers” to assist medical professionals in making quicker and more accurate diagnoses. In order to classify a lung nodule, supervised classifiers use features extracted from CT scans to assign a label, or malignancy rating to a potential cancer. These classifiers are trained using labeled examples of the malignancy levels of tumors derived from real-life medical settings.

Because classifiers learn to identify malignant tumors from expert annotations, they are learning to classify tumors by predicting the behavior of these domain experts for new cases. Cases that are correctly labeled by a classifier are cases for which the feature information strongly relates to a training example with a corresponding label. In our application, this corresponds to a diagnosis from a radiologist. Similarly, cases that are misclassified are indicative of a lack of information or discrepancy between the features and label of a given instance. We expect that the cases that are difficult to classify in a machine learning setting are related to cases deemed difficult by medical professionals. More specifically, instances that are consistently misclassified across different machine learning algorithms will show high diagnostic variability amongst medical experts. This study discusses the use of CADx systems to distinguish between semantically hard and easy diagnostic cases in a medical setting.

2.2 Multiple Label Classification

The most challenging attribute of diagnosing medical images is lack of *ground truth*. This can only be ascertained through a patient biopsy, which is costly in time and resources. Instead each image may be diagnosed by multiple radiologists in an attempt to reduce label uncertainty. This presents what is known as the *noisy label problem* in machine learning. Specifically, each instance has a set of reference truth labels instead of a single ground truth label. This class of problems manifests in scenarios where the label itself relies on the interpretation of the annotator.

There has been extensive research to reduce label uncertainty in multiple label classification problems. [SPI08] acknowledge that repeated labeling does not directly improve the quality of classification models and instead place emphasis on the importance of selective labeling. Jin and Gharahmani have shown that a selective EM model for finding correct labels produces better classifiers than a naive, non-selective approach [JG03]. This is especially relevant in the field of medicine, where diagnoses can vary even amongst expert opinions. Mavandadi et al. improved the diagnosis of red blood cells potentially infected with malaria by training a model that incorporates an error probability associated with each expert labeler [MFY+12]. While these contributions address the issue of building classifiers on disagreeing expert opinions, they assume access to all annotator identities and do not attempt to minimize label cost. Huang et al. reduce label uncertainty with anonymous annotators but still do not attempt to reduce cost [HFSL18].

In radiology, the cost associated with annotating instances presents a need for medical imaging classifiers to prioritize labels for certain cases over others as a mechanism for cost reduction. Son and Kang consider an active learning scheme for regression that consists exploration (labeling unlabeled instances) and refinement (relabeling instances) [SK18]. Yan et al. employ a probabilistic model to automatically choose the most appropriate annotator to query in addition to choosing the most useful data points in an active learning scheme [YRFD11]. This method relies on the known identity of a labeler. In cases where the identity of the labelers are hidden and the quality of every label is considered equal, methods which measure annotator performance are not applicable.

More specific to the problem of efficiently reducing label uncertainty with anonymous annotators, Riely et al. use an active selective-iterative classification (SIC) approach in building a classification model to predict the malignancy of lung nodules [RSX+15]. This model is improved at each iteration; samples that are incorrectly labeled are given an additional label and the model is re-trained using this new information. This process leverages known labels to determine which cases' labels are uncertain as a cost reduction strategy. This method was directly inspired by the LIDC dataset. Our proposed method builds on SIC since it is cost saving and handles multiple labels.

2.3 Case Difficulty Analysis

[PSK+11] examine the relationship between the number of radiologists annotations per nodule and the number of clinically distinct nodules detected in a nodule detection dataset. They used real data and simulated data that varied the samples' difficulties to be uniformly distributed. They considered the difficulty of a sample to weight the likelihood of a "correct" response from the simulated annotators. [PSK+11] Paquerault et al. considered the need for different numbers of annotators for cases of varying difficulty but provided no metric for assessing the difficulty of a case prior to receiving labels.

Alberdi et al. investigate how human decision making is affected by CAD output, specifically for false-negatives [APSA04]. The authors ask *why* humans tend to make incorrect decisions following misleading output from a CAD system. They conclude that the poor sensitivity of human readers is caused by the difficulty of a case and misguiding output of a CAD system. They used human experts to evaluate the difficulties of these mammograms by requesting rankings for potential indicators of difficulty (eg, the technical quality of the films and the tissue density of the breasts) and the difficulty itself.

More similar to our own problem context, Lin et al. described a means of difficulty-based case selection for devising personalized radiologist training programs. [LYW14] The authors combine a collaborative filtering and content-based filtering technique to predict how difficult an unseen case would be for a given trainee. These recommender systems make use of a confidence rating given by the radiologist for each case. We aim to make similar difficulty predictions with no information on radiologist identity and associated confidence rating, but instead, using image features and the associated variability in radiologist ratings.

Using the same dataset, LIDC as this paper, Zamacona et al. examined case difficulty by adding labels in a non-selective fashion to iteratively build classifiers for CAD systems [ZRFR13]. After

having error variance of four iterations, they set thresholds of it to separate data into a difficult group and an easy group. In [ZNR+15], Zamacona et al. continued their work and provided a direct analysis of case difficulty without the need for human input. They used decision tree to predict the difficulty labels that were generated in the previous step and utilized the property of decision tree to find the important features that could differentiate difficult cases and easy cases. Instead of using error variance, Seidel et al. [SRFR14] adapted a similarity measure to weight the difference between two case errors and implemented hierarchical clustering to separate the data in terms of difficulty. We expand on the SIC approach given by Reily et al. [RSX+15], a more recent selective iterative approach that achieved better accuracies in CAD systems through noise reduction. To the best of our knowledge, the studies conducted by Zamacona et al., Seidel et al. and Reily et al. are the most directly applicable efforts towards using output from CAD systems as a metric for case difficulty. We acknowledge that the goal of Reily et al. was not directly to determine case difficulties using CAD classifiers. However, this method inherently executes difficulty detection by assessing accuracy at each iteration. It identifies cases for which the labels are not supported by the remainder of the data, indicating a difficult case.

3 Methods

3.1 Data

The LIDC dataset (publicly available at ncia.nci.nih.gov) is a collection of CT scans with annotated lesions and accompanying labels from up to four radiologists [AIMB+11]. Each of the radiologists provided ratings on nine semantic features: subtlety, internal structure, calcification, sphericity, margin, lobulation, spiculation, and malignancy.

The most notable characteristic of this dataset is uncertain labels. Each nodule image has up to four labels, where each label corresponds to a different radiologist rating. We train the classifiers to predict malignancy and reserve the remaining semantic ratings for the discussion of the results.

There is variability in these semantic labels because there was no forced consensus among radiologists when annotating the samples. In addition, only a fraction of these images have follow up biopsy data on the true malignancy of a nodule. Because the absolute truth for these labels are not given, no set of labels is necessarily "correct"; they may only be used to formulate a reference truth. A reference truth is derived from the image's currently known labels. If an image has four labels it will be formed from from the mode consensus of all four labels. Likewise, if only two labels are known it will be formed from those two.

Because our methodology is based on an iterative approach, we illustrate it using a subset of the LIDC dataset consisting of the 810 samples for which all four radiologists provided a malignancy rating.

3.1.1 Image Features

We used 64 features extracted from each image to stay consistent with prior work by Riely et al. Extracted features can be divided as follows: shape, size, intensity, and texture. The shape features include compactness, eccentricity, circularity, roughness, solidity, extent, elongation, and standard deviation of radial distance. The size features are area, convex area, perimeter, convex perimeter, equivalent diameter, major axis length, and minor axis length. The intensity features are the minimum, maximum, mean, and standard deviation of the gray level intensity of every pixel in both the segmented nodule and its background. We also used the absolute value of the difference between the means of the nodule and background intensities, denoted as intensity difference. Finally, the texture features consisted of five Markov features and twenty-four Gabor features.

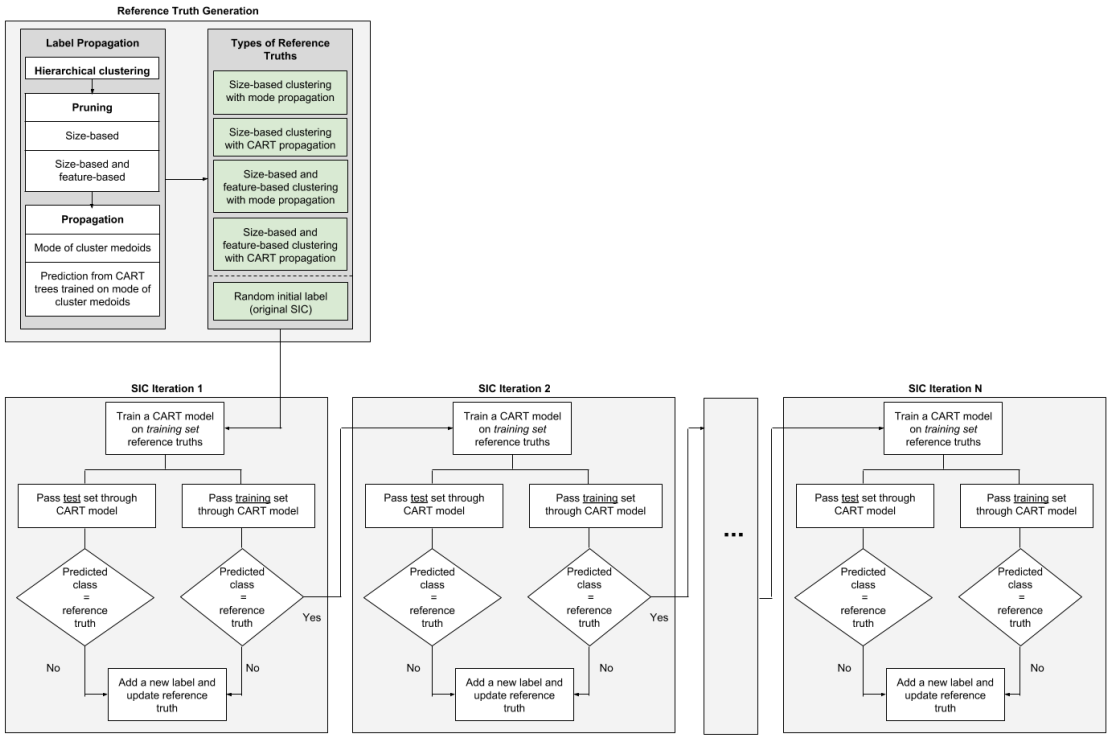

Fig. 1 Overview of the SIC and Propagated SIC methodology

3.1.2 Label Processing

The malignancy ratings range from 1 to 5 correspond to: 1. Highly Unlikely, 2. Moderately Unlikely, 3. Indeterminate, 4. Moderately Suspicious, 5. Highly Suspicious. These five classes are unbalanced within the LIDC dataset. The five possible malignancy labels of each nodule were mapped to three labels to balance the dataset as in Reily et al. We mapped labels $\{1, 2\} \rightarrow \{1\}$, $\{3\} \rightarrow \{2\}$, and $\{4, 5\} \rightarrow \{3\}$. Semantically, these new labels correspond to: 1. unlikely, 2. indeterminate, and 3. suspicious.

3.2 Selective Iterative Classification

Our analysis depends on the SIC approach developed by Reily et al.[RSX+15]. The SIC algorithm can be succinctly described through pseudo-code. An overview of our entire methodology, including both SIC and Propagated SIC, is given in Figure 1.

The algorithm shuffles the order of the labels to remove order bias. It then uses the first label as the initial "reference truth" to train a CART decision tree, the chosen supervised classifier. Although a decision tree was used, this approach works with any supervised classifier. On the next iteration, the misclassified training cases' labels are readjusted to be the mode of the first and second shuffled labels. This process is repeated until all labels have been used. We denote each repetition as an iteration. Since every point in this dataset has four possible labels, there are four possible iterations.

The purpose of this method is to save labels by only requesting more labels if the current label is not supported by the rest of the data and is misclassified. This way, the algorithm can leverage

Algorithm 1: Selective Iterative Classification

```
for each n in 1:N cases do
    Shuffle the P labels
end
Pick first label as initial reference truth
Train and store CART tree on initial reference truth
for p in 1:P possible labels do
    for n in 1:N cases do
        if Case n was misclassified at previous iteration p − 1 then
            Add new label to reference truth and take mode for new reference truth
        end
    end
    Train and store CART tree based on new labels
    Evaluate accuracy on test set
end
```

patterns found by the classifier to request labels only for "difficult" cases. Clearly, this saves labels when compared to asking for multiple opinions for every case. It also reduces label uncertainty, since asking for more annotations for cases that are already supported by prior data can introduce label uncertainty.

As mentioned above, any supervised classifier can be used in this process. However, for comparison with prior methods, we continue to use CART via the rpart package in R. We allowed trees to grow deep and then pruned according to cross validated error (within the training set). Trees were pruned to be the smallest tree within one standard error of the tree with the smallest cross validated error. In order to facilitate deep initial growth, we set the minimum number of observations in a parent node to 3 and the complexity parameter (cp) to 10^{-10}. These parameters cause overfit trees that can then be pruned.

We repeated this process 10 times as in Reily et al. Each time, we randomly split data into training (80%) and testing (20%), which were maintained when comparing methods.

3.3 Propagated Selective Iterative Classification

Using the standard SIC process, labels to be used in the first iteration are chosen at random. In this work, we present four separate methods for label assignment that stem from variations in clustering (size based and size-feature based) and label propagation (mode and CART). In total we consider five methods of label assignment:

1. Random
2. Size clustering with mode propagation
3. Size clustering with CART propagation
4. Size and feature clustering with mode propagation
5. Size and feature clustering with CART propagation

We use these five variations of cluster guided label propagation to improve the quality of the labels at the first iteration of the SIC process. Our process clusters the data and then uses the information within these clusters *propagate* labels. This process of using information from a representative sample allows us to forgo the need for information from other like cluster members.

3.3.1 Clustering Techniques

We introduce two methods of forming clusters that both begin with a hierarchical clustering of the feature data. These methods diverge in pruning this dendrogram; one method is size-based and the other size and feature-based. This hierarchical clustering followed by pruning approach is similar

Algorithm 2: Propagated selective iterative classification

1 **for** each n in 1:N cases **do**
2 | Shuffle the P labels
3 **end**
4 Cluster the data
5 **for** each cluster C **do**
6 | Propagate label l, constructed from the medoid of C, to all other points in C as reference truth
7 **end**
8 Enter SIC (Algorithm 1) line 5

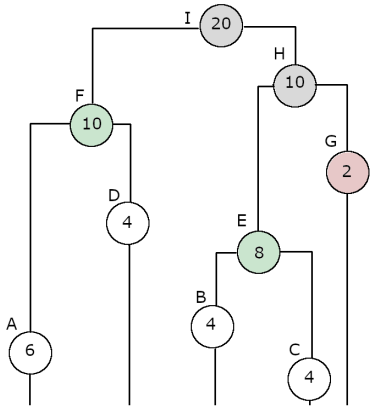

Fig. 2 Example dendrogram of hierarchical clustering with algorithm results; chosen clusters in green, ignored clusters in grey, orphan clusters in red.

to the pruning method described in Dasgupta and Hsu [DH08], although size and feature density is the control parameter, rather than purity. We refrain from using label purity as in Dasgupta et al. because we wish to conserve labels. We use these methods to ensure all clusters are large enough to effectively save labels but small enough to ensure points within the same cluster are similar.

Both methods begin with a hierarchical clustering of the data, using average linkage to build the dendrogram.

Our first clustering method is size-based pruning. It starts at the lowest level on the dendrogram and collects clusters of adequate size as it continues making horizontal cuts up the tree. Consider Figure 2. For a size parameter of eight, the algorithm starts by considering the clusters A, D, B, C, G. The clusters B, C and A, D do not independently contain at least eight points and are replaced by the next node in the tree, resulting in a new set of clusters F, E, G. Because the algorithm does not allow a single point to belong to multiple clusters, the algorithm stops considering clusters beyond a cluster that meets the size requirement. This may result in orphan clusters such as G. Samples within these orphan clusters are considered unclustered and are given an initial random label as in the first iteration of the SIC process. The algorithm for this pruning method is described in Algorithm 3.

Note that the level on the dendrogram L is only used to traverse the clusters of the dendrogram to find the smallest clusters that are larger than eight and represent unique data. We postpone discussion of the parameter S, as it relies on our method of label propagation, which is discussed later.

Our second clustering method, size and feature-based pruning, is an extension of size-based pruning. The feature addition introduces a more stringent criterion for accepting a cluster, which would take place at line four of Algorithm 3. We require f features to be *tightly packed* in C_i. The purpose of this is to ensure that each cluster has at least f features that are very similar to one another. We call a feature a tightly packed if its standard deviation within the cluster is less than a third of the standard deviation of the feature in the dataset. We set $f = 30$ out of the 64 image features in the LIDC dataset because we found that this parameter returned reasonable cluster sizes (10 to 20 cases) while amounting for nearly half of the total image features. We explored $f = 10, 20, 30,$ and 40. We wanted clusters to contain about 10 to 20 instances so labels were only

Algorithm 3: Size-based pruning for hierarchical clustering

input : Dataset of N points, size parameter S
output : List of clusters
initialize: $L = N$, where L is the level on the dendrogram of 1 (top) to N (bottom) levels
$U = u_1, ..., u_N$, where U is the list of points that have not been clustered

```
while |U| > 0 and L > 1 do
    cut tree at L, resulting in clusters C_1...C_L
    for C_i in C_1...C_L do
        if all points in C_i are in U and |C_i| ≥ S then
            Add C_i to the list of clusters
            U = U − C_i
        end
    end
    L − −
end
return List of clusters
```

propagated to nearby neighbors while still saving labels. Note that the value of f only affects the pruning procedure; all features are used when training all other CART classifiers.

3.3.2 Propagation Techniques

Assume there are K clusters, $C_1...C_K$, where each C_i corresponds to a subset of the data. We take advantage of the image similarities within a cluster by assigning the label of a representative sample to the remaining points within a cluster. We call this process *label propagation*. The medoid is used as the representative sample within each cluster. For each cluster C_i the respective medoid m_i is the closest point to the center of the cluster, which is the mean of all feature vectors of C_i. Our two propagation techniques utilize these medoids differently.

Our first propagation technique is called mode propagation. We take the mode of all labels for each medoid m_i and assign it to the remaining points in C_i. This relies solely on cluster information to determine how labels are propagated.

Our second propagation technique is called CART propagation. Instead of propagating the mode of each medoid within a cluster, we train a CART tree on the modes of the set of medoids. In training a CART model on these representative samples, this method expects that the subset formed by these cluster medoids form a well-distributed representation of the entire dataset. The remaining points in the training set are then passed through this model and the resulting classification is assigned as the propagated label for these points.

3.3.3 Size Parameter

Calculation of the size parameter S for our clustering algorithm is straightforward for datasets like LIDC with a uniform number of labels for every point. It is parametrized by $S = \left\lceil \frac{P}{Q} \right\rceil$, where P is the number of labels per point and Q is the maximum fraction of labels to be used within each cluster. In our dataset with $P = 4$, we use $Q = 1/2$ to save at least half the labels for each cluster. Thus, $S = 8$. As a result, at most half the labels are necessary to label the entire cluster when compared to taking single random labels for each observation. Adjusting Q, and thus S, adjusts how assertively labels are propagated to neighbors. If Q is smaller, S increases and the final clusters are larger. This saves labels, but decreases the representativeness of the propagated label for the points near the cluster's border.

All four cluster-based variations of the SIC algorithm (size-mode, size-CART, feature-mode, feature-CART) exploit the image feature similarities between points of the same cluster to reduce the total number of labels necessary, T. For a dataset D with N observations, SIC requires N labels in the first iteration (one label for every point). For a clustering resulting in K clusters $C_1 \dots C_K$, the

total number of initial labels can be described in 1 where l_x is the number of labels for a given point x and m_i is the medoid of cluster c_i. In our dataset all of our points have four labels, so $\forall x \in D : l_x = 4$.

$$T = N - |\cup_{i=1}^{K} C_i| + \sum_{i=1}^{K} l_{m_i} \tag{1}$$

3.4 Hard Cases

To study the relationship between cases deemed difficult by the classifier and cases on which expert annotators disagree, we present a formal definition for what constitutes a *hard* case. We define the hardest cases using results from aggregated the five SIC techniques. A case is **instance-hard** if it is misclassified by a *single* SIC method *on every iteration*. To measure how *consistently* these cases are considered difficult by the given SIC process, each method is run for $t = 10$ trials. A case may be instance-hard up to ten times. Because we have five versions of the SIC algorithm (random, size-mode, size-CART, feature-mode, feature-CART), we repeat this process five times for each case, every time with a different algorithm. The number of times a case is considered hard can now range from 0 - 50 (5 variations $\times$ 10 trials) times. We call the number of times a cases is instance-hard across all five methods its **difficulty level**.

The **hardest cases** are defined to be those above p-level of difficulty, where p is a certain percentile in the list of cases sorted by their level of difficulty. Changing p influences how many cases are considered the hardest. For example, $p = 80$ defines hard cases to be the top 20% most difficult in the difficulty level vector.

Note that the five methods of label assignment discussed in 3.3 will produce different classifiers throughout the SIC process. These distinct classifiers will have independent considerations of hard cases when passed the same training and testing data. We take advantage of this variation across classifiers to form a comprehensive study of hard cases within the LIDC dataset.

4 Results

In the following section, we first provide a performance overview of the five variations of the SIC algorithm. Next, we perform an visual analysis of the 2.5% identified hardest nodules in the data set by analyzing the CT images and their placement in an LDA plot. Finally, we perform statistical test to differentiate between the top 30% hard and non-hard nodules with respect to their image features and semantic characteristics.

4.1 Results of SIC and Propagated SIC algorithms

We begin by applying the five algorithms discussed in 3 to the testing set derived from the LIDC data. In addition to these five algorithms we implement a standard CART tree trained on the consensus of all radiologist labels for comparison of the SIC performance. These results are displayed in Table 1

4.2 Visual exploration of the 2.5% hardest nodules

A natural direction is examining the image features of the hardest cases to see if they are outliers. In Figure 3, we plot the top 2.5% (25 cases) hardest cases in red on the 2d space reduced from

Label Assignment	Iteration 1		Iteration 2		Iteration 3		Iteration 4	
	Acc.%	Labels Used	Acc.%	Labels Used	Acc.%	Labels Used	Acc.%	Labels Used
CART using mode-labeled consensus							73.49	3316
Random (SIC) [RSX+15]	56.20	829.0	73.73	1089.0	77.77	1249.3	79.88	1385.2
Size - Mode	50.06	448.4	69.40	714.6	76.87	869.0	80.60	972.6
Size - CART	47.89	448.4	61.14	548.4	67.11	625.7	70.54	689.5
Feature and Size - Mode	51.75	635.9	68.55	927.1	75.66	1118.2	79.94	1259.2
Feature and Size - CART	48.37	635.9	62.71	900.6	70.48	1094.7	74.40	1246.1

Table 1 Results for CART and SIC classifiers. Accuracy on test set (in %) is abbreviated to "Acc." to preserve space.

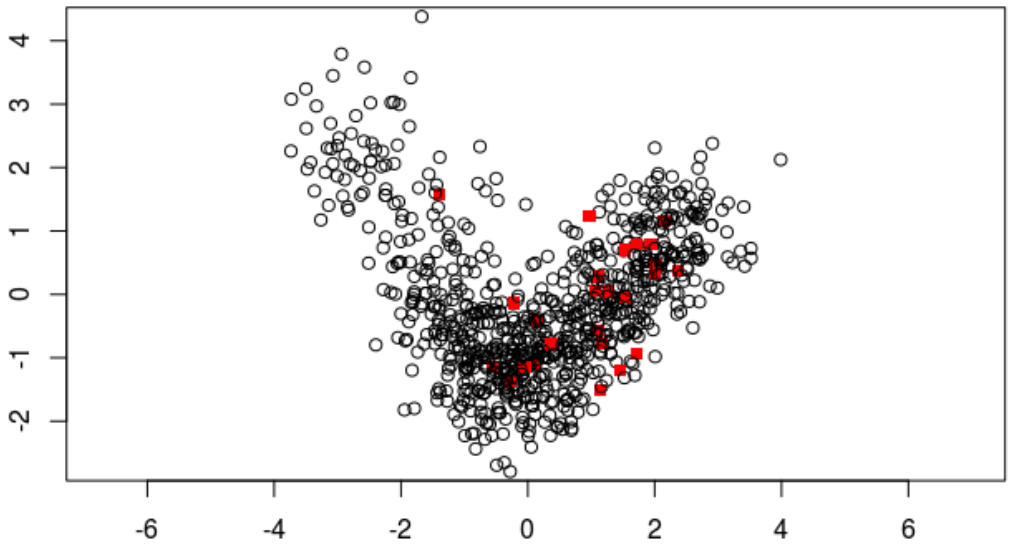

Fig. 3 Plot of feature data on 2d LDA space. The red points indicate the top 2.5% hardest observations. The hardest cases are not distinguishable in the linear space, leading to further analysis of the image features and semantic characteristics below."

the 64d feature space through linear discriminant analysis (LDA). In only Figure 3 and 4, we set $p = 97.5$, therefore plotting only the top 2.5% hardest cases, to prevent cluttering and to allow for visual distinction between the hardest and remaining cases. We use a 2D space since we have only three levels of malignancy (described in 3.1.2). We use LDA for dimensionality reduction while striving for class separation. To train the LDA model, we use the mode of label set for each point. In the linear space we cannot see separability between hard and non-hard cases. Future work will consider visualizations in nonlinear spaces through kernel discriminant analysis.

Figure 4 displays the CT scans of the twenty-five hardest nodules associated with $p = 97.5$. From the plot in Figure 3, we recognize that these twenty-five nodules are well-distributed throughout the feature space (in red), amongst the rest of the nodules in the dataset. For each nodule, we include the malignancy and subtlety ratings from all four radiologists. We choose to include these two semantic characteristics because of the results later discussed in Table 2, which indicate that subtlety ratings are more variable for the $p = 70$ hardest cases.

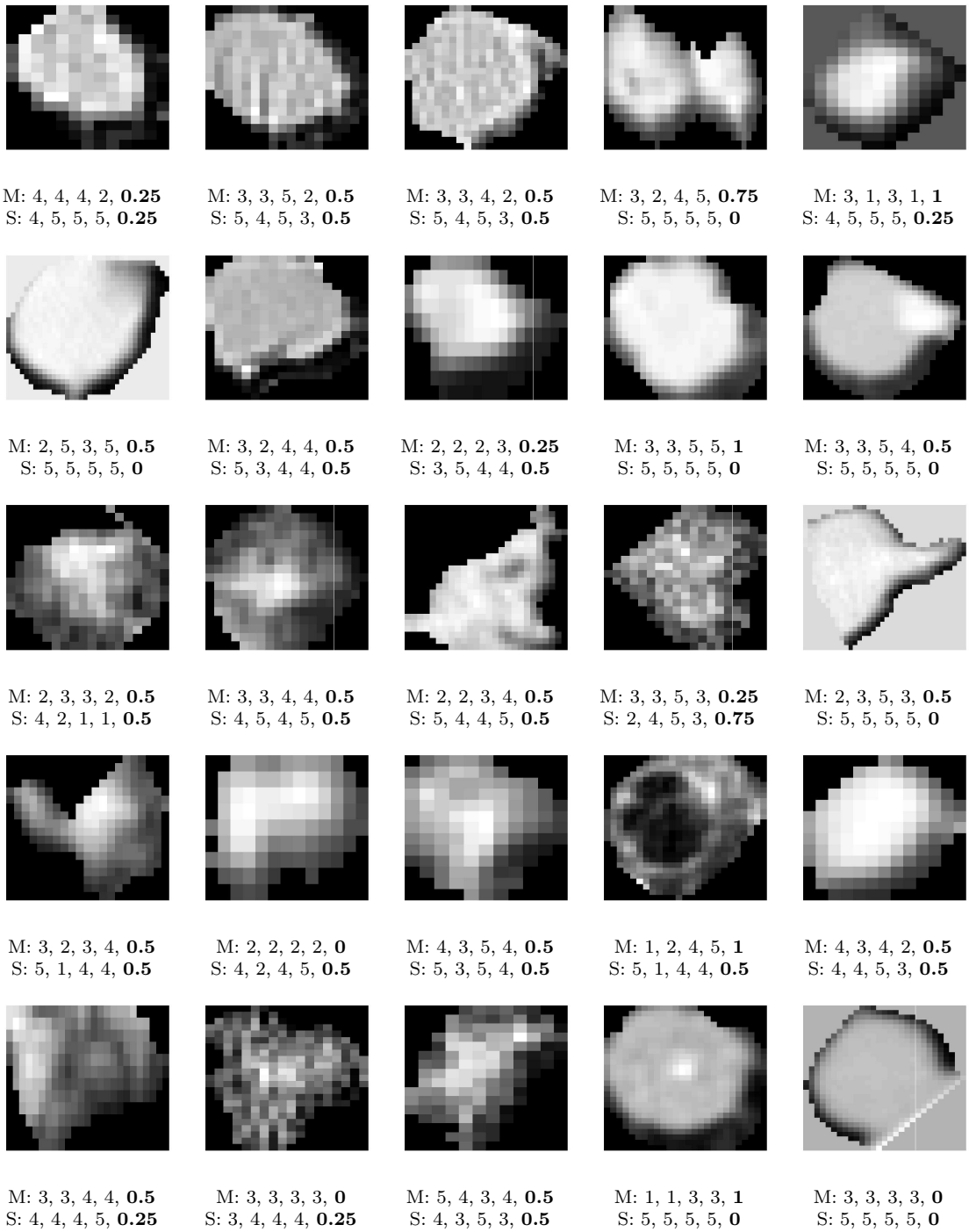

Fig. 4 Top 2.5% hardest hard nodules. The corresponding malignancy (M) and subtlety (S) ratings from each radiologist, along with the associated mean variance range in bold, is below each image. Malignancy ratings range from 1 to 5, with 1 as highly unlikely and 5 highly suspicious. Subtlety ratings also range from 1 to 5, with 1 as extremely subtle and 5 as obvious.

4.3 Statistical analysis of hard nodules

Analysis of the semantic ratings for the top 30% of hardest cases yields interesting results. For the remainder of our analysis, we set $p = 70$ and call the top 30% (273 instances) our samples the **hardest cases**.

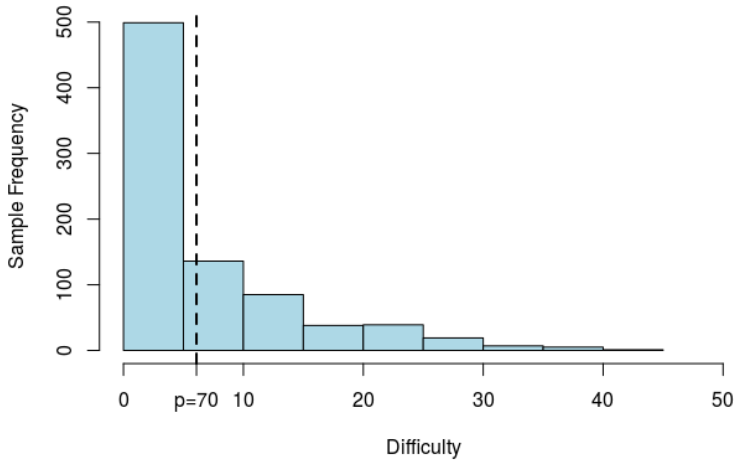

Fig. 5 Frequency of samples with respect to difficulty level.

Semantic Characteristic	Hard Case Average VR	Non-Hard Case Average VR	P-value
Subtlety	0.408	0.367	0.004
Internal Structure	0.007	0.006	0.416
Calcification	0.080	0.062	0.097
Sphericity	0.389	0.400	0.797
Margin	0.370	0.350	0.102
Lobulation	0.280	0.319	0.993
Spiculation	0.231	0.265	0.986
Texture	0.179	0.153	0.031
Malignancy	0.402	0.401	0.482

Table 2 Results for testing if the hardest cases (top 30%) have a significantly larger VR than non-hard cases for various semantic characteristics.

The maximum difficulty of a sample is fifty, indicating a sample was misclassified fifty times. Figure 5 displays the number of samples per sample difficulty level. This represents the number of misclassification across all iterations of an iterative algorithm, for all five SIC algorithm variations, and ten trials. The cutoff for our value of $p = 70$ is included in Figure 5. The samples to the right of this line correspond to our 273 hardest cases.

We expect the semantic characteristics to be more varied for the hardest cases (i.e. radiologists are more likely to disagree on the hardest nodules). Formally, we define variance in the semantic characteristics for hardest and non-hard cases using the mean variance range (VR). VR for an ordinal set S with mode m is defined through an indicator function $I(\cdot)$ as follows.

$$VR(S) = \frac{\sum_{s \in S} I(s \neq m)}{|S|} \tag{2}$$

In order to test for variability, we use a t-test for difference of means. Formally, if μ_{hard} is the VR of the hardest cases and μ_{rest} is the VR for the rest of the dataset, we are testing the alternative $H_a : \mu_{hard} > \mu_{rest}$. We also apply a Bonferroni correction since we are testing 9 independent hypotheses, rather than a single hypothesis. This results in a significance threshold of $\alpha = 0.05/9 \approx 0.005$. Table 2 supports this hypothesis on one semantic characteristic, subtlety.

The image features as well as two imaging parameters of the hard cases are also assessed for differences. The imaging parameters were exposure time and slice thickness of the scan. We conducted a series of t-tests for difference in means for each. Results are shown in Table 3. In each test, we have $H_0 : \mu_{\text{hard}} = \mu_{\text{rest}}$ and $H_a : \mu_{hard} \neq \mu_{rest}$ and apply a Bonferroni correction to get $\alpha = 0.05/(64 \text{ image feature tests} + 2 \text{ image parameter tests}) = 0.05/66 \approx$ 7e-4. We suspect

Feature	P-value	Feature	P-value	Feature	P-value	Feature	P-value
Area	2.42e-08	Convex Area	1.94e-08	Perimeter	3.00e-07	Convex Perimeter	6.19e-07
Equiv Diameter	5.06e-07	Major Axis Length	4.25e-06	Minor Axis Length	1.59e-07	Elongation	5.79e-01
Compactness	1.53e-02	Eccentricity	3.07e-01	Solidity	5.63e-02	Extent	2.17e-01
Circularity	3.89e-01	Radial Distance SD	2.58e-02	Roughness	1.75e-03	Min Intensity	9.87e-01
Max Intensity	5.96e-01	Mean Intensity	1.72e-01	SD Intensity	4.48e-02	Min Intensity BG	8.35e-01
Max Intensity BG	5.96e-01	Mean Intensity BG	7.43e-01	SD Intensity BG	9.35e-01	Intensity Difference	2.40e-02
markov 1	5.81e-01	markov 2	8.30e-01	markov 3	2.52e-01	markov 4	2.47e-01
markov 5	3.50e-01	gabormean 0 0	8.82e-02	gabor SD 0 0	1.22e-02	gabor mean 0 1	1.28e-02
gabor SD 0 1	7.14e-04	gabor mean 0 2	7.20e-01	gabor SD 0 2	1.26e-01	gabormean 1 0	1.05e-02
gabor SD 1 0	2.44e-03	gabor mean 1 1	2.13e-02	gabor SD 1 1	4.31e-04	gabor mean 1 2	1.58e-02
gabor SD 1 2	9.56e-04	gabor mean 2 0	4.26e-02	gabor SD 2 0	1.30e-02	gabor mean 2 1	1.48e-03
gabor SD 2 1	7.92e-04	gabor mean 2 2	7.74e-01	gabor SD 2 2	7.96e-02	gabor mean 3 0	2.17e-02
gabor SD 3 0	2.51e-02	gabor mean 3 1	1.25e-01	gabor SD 3 1	3.32e-03	gabor mean 3 2	1.46e-02
gabor SD 3 2	2.51e-02	Contrast	1.91e-03	Correlation	2.54e-02	Energy	4.66e-02
Homogeneity	5.16e-03	Entropy	7.13e-05	3rd order moment	5.09e-02	Inverse Variance	4.04e-03
Sum Average	1.32e-01	Variance	3.01e-02	Cluster Tendency	5.03e-02	Max Probability	1.78e-01
Exposure Time	1.78e-01	Slice Thickness	2.53e-01				

Table 3 Test of significant for difference in image features between difficult and remaining cases

that difficult nodules would be smaller. A smaller size would make nodules more difficult since they would have to be viewed at lower resolution to be as large as bigger nodules. Thus, for features positively related to size we set $H_a : \mu_{hard} < \mu_{rest}$. We find significant differences between the hardest and remaining samples for area, convex area, perimeter, convex perimeter, equivalent diameter, major axis length, minor axis length, gabor SD 0_1, gabor SD 1_1, and entropy.

5 Discussion

In order to isolate the diagnostic difficulty of cases using classifiers, we need to improve the quality of the label set. This will ensure that a case is misclassified because the features are inconsistent with the expected malignancy level, rather than due to noise in labels. Table 1 shows that the five methods of propagation followed the SIC implementation of CART improves classification accuracy when compared to a standard CART classifier, despite using fewer labels. We attribute this improvement in accuracy to the reduction in label uncertainty executed by SIC; the algorithm is able to choose informative labels that do not introduce more noise. We acknowledge that there is still a possibility that label noise unrelated to case difficulty remains in the dataset. Because of our clear reduction in uncertainty, we are lead to believe that cases misclassified by the SIC variations are more likely to be difficult or inconsistent cases, rather than cases that are misclassified due to noise in training data for the classifier.

In Figure 3, we investigate the spacial relationship of the hard cases with respect to the rest of the samples. Intuitively, one would expect that hard cases will be abnormal cases, or outliers, and located far from the rest of the data in the feature space. Clearly hard cases are distributed throughout the relevant feature space. Although they are necessary for the classification of malignancy levels, we find that all sixty-four features may not be useful in distinguishing difficult cases.

Figure 5 shows the distribution of the LIDC dataset with respect to our definition of difficulty. In the histogram, we observe that a majority of labels have a difficulty level of less than five. Our analysis of the 30% of hardest cases encompasses the samples that have at least five misclassification across all trials.

Although hard cases are well distributed throughout the feature space, we explore the idea that this is caused by the high dimensionality of the space, namely that all sixty-four features may not contribute to distinguishing between hard and easy cases. In Figure 3, by analyzing the difference of individual image features between the hardest and remaining cases, we ask ourselves *why* our hardest samples are considered difficult by our classifiers. Note that this is different than the approach taken in [ZNR+15], since they use a predefined notion of difficulty to save annotations while we use an annotation saving approach with no predefined notion of difficulty. The primary goal of Zamacona et al. was to save labels, while we examine the differences between easy and hard cases since the label saving SIC algorithm outputs measures of difficulty as a result. We

apply a difference of means test to the sixty-four image features and two imaging parameters. We immediately notice that image features related to size all resulted in significant values; Area, Convex Area, Perimeter, Convex Perimeter, Equivalent Diameter, Major Axis Length, and Minor Axis Length. This suggests that our hard cases are significantly smaller than the other nodules. In addition to size, two texture features derived from Gabor filters, gaborSD_0_1 and gaborSD_1_1, as well as entropy are significantly different in our hardest cases. Our findings on the significance of entropy and size in determining the difficulty of an image feature are consistent with prior work in difficulty detection, namely Zamacoma et al. They found that "these features [entropy, size, and contrast] play a large role in distinguishing between easy and hard cases.[ZNR+15]"

Next, we explore whether this difference in low-level image features translates to a difference in high-level semantic characteristics of the lung nodules. We analyze the variance of the nine semantic characteristic ratings provided for each nodule by the radiologists in Table 2. Although all 64 image features are necessary for the classification of malignancy levels, we find that only subtlety displays significantly higher variation amongst radiologist ratings for the hardest cases. Subtlety is defined as the "difficulty in detection—refers to the contrast between the lung and its surroundings.[OCRF11]" Our identified image features related to size, texture, and entropy align well with the notion of variable subtlety. Images that are smaller in size, variable in texture, and highly random are more likely to be variable in terms of their shape distinction.

We find it interesting that malignancy is not also significantly variable. It is possible that this high variability in subtlety does not carry over to a high variability in malignancy because of a lack of correlation between the nine semantic ratings. It would be interesting to test this by observing whether cases with higher ratings for a characteristic, such as spiculation, directly relates to higher malignancy ratings.

6 Conclusion

Classifiers built to predict malignancy ratings fail to label samples which are small, variable in texture, and with high entropy. We demonstrated that the hardest cases (top 30%) manifest these low-level image features, and that these features correspond with a significant variation in the semantic characteristic of subtlety.

We reduced noise by selecting the labels used in by the classifiers. This ensured that our algorithm's hard cases are also seen as more difficult by radiologists. The SIC approaches offer a unique glimpse into a difficulty of a nodule, since they assess the relationship between each case's label and its consistency with the rest of the dataset multiple times.

We explored the image features of these diagnostically difficult cases to find that they are significantly smaller in size than the rest of the nodules in the dataset. In addition to entropy and two gabor texture features, all seven image features related to the size of the nodule were shown as statistically smaller than the remaining nodules. This verifies intuition that smaller nodules are more likely to be misinterpreted by radiologists. We note that the number of cases that are considered the hardest cases will depend on the threshold parameter p, and therefore, changing this parameter may result in slightly different significance in testing.

The sole usage of feature data is an interesting extension of this idea of difficulty. Perhaps difficulty can be effectively modeled from the image features we identified in this analysis: seven size-related features, entropy, and gabor texture features. In addition to these features, subtlety ratings for each sample may contribute to building strong case difficulty classifiers. Training and testing such classifiers would require reliable labels on the difficulty of a nodule. While "difficult" and "not difficult" could easily be modeled as a binary classification problem, in the future we expect to explore this implementation using probabilistic classifiers to produce the *level of difficulty* for each unseen case.

References

AIMB+11. Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, 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.

APSA04. Eugenio Alberdi, Andrey Povyakalo, Lorenzo Strigini, and Peter Ayton. Effects of incorrect computer-aided detection (cad) output on human decision-making in mammography. *Academic radiology*, 11(8):909–918, 2004.

DH08. Sanjoy Dasgupta and Daniel Hsu. Hierarchical sampling for active learning. In *Proceedings of the 25th international conference on Machine learning*, pages 208–215. ACM, 2008.

HFSL18. Wenying Huang, Songhe Feng, Lijuan Sun, and Congyan Lang. Partial label learning via low rank representation and label propagation. In *Proceedings of the 10th International Conference on Internet Multimedia Computing and Service*, page 32. ACM, 2018.

JG03. Rong Jin and Zoubin Ghahramani. Learning with multiple labels. In *Advances in neural information processing systems*, pages 921–928, 2003.

LYW14. Hongli Lin, Xuedong Yang, and Weisheng Wang. A content-boosted collaborative filtering algorithm for personalized training in interpretation of radiological imaging. *Journal of digital imaging*, 27(4):449–456, 2014.

MFY+12. Sam Mavandadi, Steve Feng, Frank Yu, Stoyan Dimitrov, Karin Nielsen-Saines, William R Prescott, and Aydogan Ozcan. A mathematical framework for combining decisions of multiple experts toward accurate and remote diagnosis of malaria using tele-microscopy. *PloS one*, 7(10):e46192, 2012.

MKY+09. Shmuel Mahgerefteh, Jonathan B Kruskal, Chun S Yam, Arye Blachar, and Jacob Sosna. Peer review in diagnostic radiology: current state and a vision for the future. *Radiographics*, 29(5):1221–1231, 2009.

OCRF11. Pia Opulencia, David S Channin, Daniela S Raicu, and Jacob D Furst. Mapping lidc, radlex^{TM}, and lung nodule image features. *Journal of digital imaging*, 24(2):256–270, 2011.

PSK+11. Sophie Paquerault, Berkman Sahiner, Anna Kettermann, Laura M Yarusso, Lubomir M Hadjiiski, and Heang-Ping Chan. Analysis of the number of distinct findings obtained by multiple readers in an mrmc study: When do findings obtained from the addition of new readers become redundant, or otherwise negligible? In *Medical Imaging 2011: Image Perception, Observer Performance, and Technology Assessment*, volume 7966, page 79661M. International Society for Optics and Photonics, 2011.

Rei14. Bruce I Reiner. A crisis in confidence: A combined challenge and opportunity for medical imaging providers. *Journal of the American College of Radiology*, 11(2):107–108, 2014.

Rei17. Bruce I Reiner. Redefining the practice of peer review through intelligent automation part 1: Creation of a standardized methodology and referenceable database. *Journal of digital imaging*, 30(5):530–533, 2017.

Rei18. Bruce I Reiner. Quantifying analysis of uncertainty in medical reporting: Creation of user and context-specific uncertainty profiles. *Journal of digital imaging*, pages 1–4, 2018.

RSX+15. Amelia Riely, Kyle Sablan, Thomas Xiaotao, Jacob Furst, and Daniela Raicu. Reducing annotation cost and uncertainty in computer-aided diagnosis through selective iterative classification. In *Medical Imaging 2015: Computer-Aided Diagnosis*, volume 9414, page 94141K. International Society for Optics and Photonics, 2015.

SK18. Youngdoo Son and Seokho Kang. Regression with re-labeling for noisy data. *Expert Systems with Applications*, 2018.

SPI08. Victor S Sheng, Foster Provost, and Panagiotis G Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 614–622. ACM, 2008.

SRFR14. Mike Seidel, Alexander Rasin, Jacob D Furst, and Daniela S Raicu. Towards achieving diagnostic consensus in medical image interpretation. In *2014 IEEE International Conference on Data Mining Workshop*, pages 771–780. IEEE, 2014.

YRFD11. Yan Yan, Romer Rosales, Glenn Fung, and Jennifer G Dy. Active learning from crowds. In *ICML*, volume 11, pages 1161–1168, 2011.

ZNR+15. Jose R Zamacona, Ronald Niehaus, Alexander Rasin, Jacob D Furst, and Daniela S Raicu. Assessing diagnostic complexity: an image feature-based strategy to reduce annotation costs. *Computers in biology and medicine*, 62:294–305, 2015.

ZRFR13. Jose R Zamacona, Alexander Rasin, Jacob D Furst, and Daniela S Raicu. Reducing classification cost through strategic annotation assignment. In *Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on*, pages 287–294. IEEE, 2013.