Image- vs. histogram-based considerations in semantic segmentation of pulmonary hyperpolarized gas images

Nicholas J. Tustison[1], Talissa A. Altes[2], Kun Qing[3], Mu He[1], G. Wilson Miller[1], Brian B. Avants[1], Yun M. Shim[1], James C. Gee[4], John P. Mugler III[1], Jaime F. Mata[1]

[1]Department of Radiology and Medical Imaging, University of Virginia, Charlottesville, VA
[2]Department of Radiology, University of Missouri, Columbia, MO
[3]Department of Radiation Oncology, City of Hope, Los Angeles, CA
[4]Department of Radiology, University of Pennsylvania, Philadelphia, PA

Corresponding author:
Nicholas J. Tustison, DSc
Department of Radiology and Medical Imaging
University of Virginia
ntustison@virginia.edu

Abstract

Purpose: To characterize the differences between histogram-based and image-based algorithms for segmentation of hyperpolarized gas lung images.

Methods: Four previously published histogram-based segmentation algorithms (i.e., linear binning, hierarchical k-means, fuzzy spatial c-means, and a Gaussian Mixture Model with a Markov Random Field prior) and an image-based convolutional neural network were used to segment two simulated data sets derived from a public ($n = 29$) and a retrospective collection ($n = 51$) of hyperpolarized 129Xe gas lung images transformed by common MRI artefacts (noise and nonlinear intensity distortion). The resulting ventilation-based segmentations were used to assess algorithmic performance and characterize optimization domain differences in terms of measurement bias and precision.

Results: Although facilitating computational processing and providing discriminating clinically relevant measures of interest, histogram-based segmentation methods discard important contextual, spatial information and are consequently less robust, in terms of measurement precision, in the presence of common MRI artefacts relative to the image-based convolutional neural network.

Conclusions: Direct optimization within the image domain using convolutional neural networks leverages spatial information which mitigates problematic issues associated with histogram-based approaches and suggests a preferred future research direction. Further, the entire processing and evaluation framework, including the newly reported deep learning functionality, is available as open-source through the well-known Advanced Normalization Tools ecosystem.

Introduction

Historical overview of quantification

Early attempts at quantification of hyperpolarized gas images were limited to enumerating the number of ventilation defects or estimating the proportion of ventilated lung (1–3) which has evolved to more sophisticated techniques used currently. A brief outline of major contributions can be roughly sketched to include:

- binary thresholding based on relative intensities (4–6),
- linear intensity standardization via rescaling of the intensity histogram to a reference distribution based on healthy controls, i.e., "linear binning" (7,8),
- nonlinear intensity standardization using a customized hierarchical (9,10) or adaptive (11) k-means algorithm,
- nonlinear intensity standardization using fuzzy c-means (12) with spatial considerations based on local voxel neighborhoods (13), and
- Gaussian mixture modeling (GMM) of the intensity histogram with Markov random field (MRF) spatial prior modeling (14).

Given the functional nature of hyperpolarized gas images and the consequent sophistication of the segmentation task, these algorithmic approaches reduce the complex spatial image information to primarily intensity-only optimization considerations, contextualized in terms of the intensity histogram. Although facilitating computational processing, this simplifying transformation results in the loss of important spatial cues for identifying salient image features, such as ventilation defects (a well-studied correlate of lung pathophysiology), as spatial objects.

Each of these algorithms can be viewed as a type of MR intensity standardization (15) with varying degrees of flexibility and algorithmic sophistication. Due to hard threshold values, intensity-only approaches are unable to account for various MRI artefacts such as noise (16,17) and the intensity inhomogeneity field (18) which prevent such threshold values from distinguishing tissue types precisely consistent with that of human experts. These MR intensity nonlinearities have been well-studied (15,19–22) and are known to cause significant

intensity variation even in the same region of the same subject. As stated in (19):

> Intensities of MR images can vary, even in the same protocol and the same sample and using the same scanner. Indeed, they may depend on the acquisition conditions such as room temperature and hygrometry, calibration adjustment, slice location, B0 intensity, and the receiver gain value. The consequences of intensity variation are greater when different scanners are used.

Ignoring these nonlinearities is known to have significant consequences in the well-studied (and somewhat analogous) area of brain tissue segmentation in T1-weighted MRI (e.g., (23–25)) where the well-known relative intensities of major tissue types (i.e., cerebrospinal fluid (CSF), gray matter (GM), and white matter (WM)), which characteristically correspond to visible histogram peaks, as landmarks to determine the nonlinear intensity mapping (i.e., 1-D piecewise affine mapping) between structural features found within the histograms themselves (e.g., peaks and valleys) (15,21). However, in hyperpolarized gas imaging of the lung, no such characteristic structural features exist, generally, between histograms. Additionally, because of the functional nature of these images, the segmentation clusters that correspond to features of interest are not necessarily guaranteed to exist (e.g., ventilation defects in the case of healthy normal subjects with no lung pathology).

Linear binning is a simplified type of MR intensity standardization in which images from healthy controls are normalized to the range [0, 1] and then used to calculate the cluster intensity boundary values based on an aggregated estimate of the parameters of a single Gaussian fit. Subject images to be segmented are then rescaled to this reference histogram (i.e., a global affine 1-D transform). This mapping results in alignment of the cluster boundaries such that corresponding labels are assumed to have similar clinical interpretation. Variants of the well-known k-means algorithm constitute an algorithmic approach with additional flexibility over linear binning as it employs prior knowledge in the form of a generic clustering desideratum (i.e., minimizing within-cluster intensity variance) for optimizing a type of nonlinear MR intensity standardization. However, as with binary thresholding, both linear binning and k-means completely discard spatial context in optimizing voxelwise cluster membership.

Additional sophistication incorporating spatial considerations is found in the fuzzy spatial c-means (26) and Gaussian mixture-modeling (GMM) with a Markov random field (MRF) prior algorithms. The former, similar to k-means, optimizes over the within-class sample variance but includes a per-sample membership weighting (27) whereas the latter is optimized via the expectation-maximization (EM) algorithm (28). These algorithms have the advantage, in contrast to histogram-only algorithms, that the intensity thresholds between class labels are softened which demonstrates some relative robustness to certain imaging distortions, such as noise. However, all these algorithms are flawed in the inherent assumption that meaningful structure is found, and can be adequately characterized, within the associated image histogram in order to optimize a multi-class labeling.

Additionally, many of these segmentation algorithms use N4 bias correction (29), an extension of the nonuniform intensity normalization (N3) algorithm (18), to mitigate MR intensity inhomogeneity artefacts. Interestingly, N3/N4 iteratively optimizes towards a final solution using information from both the histogram and image domains. Based on the intuition that the bias field acts as a smoothing convolution operation on the original image intensity histogram, N3/N4 optimizes a nonlinear (i.e., deformable) intensity mapping based on histogram deconvolution. This nonlinear mapping is constrained such that its effects smoothly vary across the image. Additionally, due to the deconvolution operation, this mapping sharpens the histogram peaks which presumably correspond to distinct tissue types. While such assumptions are appropriate for the domain in which N3/N4 was developed (i.e., T1-weighted brain tissue segmentation) and while it is assumed that the enforcement of low-frequency modulation of the intensity mapping prevents new image features from being generated, it is not clear what effects N4 parameter choices have on the final segmentation solution, particularly for those algorithms that are limited to intensity-only considerations and less robust to the specified MR artefacts.

Motivation for current study

Investigating the assumptions outlined above, particularly those associated with the intensity mappings due to both the MR acquisition and inhomogeneity mitigation preprocessing, we

became concerned by the susceptibility of the histogram structure to such variations and the potential effects on current clinical measures of interest derived from these algorithms (e.g., ventilation defect percentage). Specifically, we noticed that histogram-based intensity perturbations can produce virtually little, if any, changes in the features of the image despite a relatively significant change in the histogram structure. Such effects imply that MR artefacts could profoundly impact histogram-based algorithmic performance. Figure 1 provides a sample visualization representing some of the structural changes that we observed when simulating these nonlinear mappings.

To briefly explore these effects further for the purposes of motivating additional experimentation, we provide a summary illustration from a set of image simulations in Figure 2 which are detailed later in this work and used for algorithmic comparison. Simulated MR artefacts were applied to each image which included both noise and nonlinear intensity mappings (and their combination) using two separate data sets: one in-house data set consisting of 51 129Xe gas lung images and the publicly available data described in (30) and made available at Harvard's Dataverse online repository (31) consisting of 29 hyperpolarized gas lung images and corresponding lung masks. These two data sets resulted in a total simulated cohort of 51 + 29 = 80 images ($\times$10 simulations per image $\times$3 types of artefact simulations). Prior to any algorithmic comparative analysis, we quantified the difference of each simulated image with the corresponding original image using the structural similarity index measurement (SSIM) (32). SSIM is a highly cited measure which quantifies structural differences between a reference and distorted (i.e., transformed) image based on known properties of the human visual system. SSIM has a range $[-1, 1]$ where 0 indicates no structural similarity and 1 indicates perfect structural similarity. We also generated the histograms corresponding to these images. Although several histogram similarity measures exist, we chose Pearson's correlation primarily as it resides in the same min/max range as SSIM with analogous significance. In addition to the fact that the image-to-histogram transformation discards important spatial information, from Figure 2 it should be apparent that this transformation also results in greater variance in the resulting information under common MR imaging artefacts, according to these measures. Thus, prior to any algorithmic considerations, these observations strongly suggest that optimizing in the domain of the histogram will be generally less informative and

less robust than optimizing directly in the image domain.

Ultimately, we are not claiming that these algorithms are erroneous, per se. Much of the relevant research has been limited to quantifying differences with respect to ventilation versus non-ventilation in various clinical categories and these algorithms have demonstrated the capacity for advancing such research through the use of clinically useful measures such as ventilation defect percentage. Furthermore, as the sample segmentations in Figure 3 illustrate, when considered qualitatively, each segmentation algorithm appears to produce reasonable segmentations even though the voxelwise differences are significant as are the corresponding histograms. However, the artefact issues influence quantitation in terms of core scientific measurement principles such as precision (e.g., reproducibility and repeatability (11,33)) and bias which are obscured in isolated considerations but become increasingly significant with multi-site (34) and large-scale studies. In addition, refinements in measuring capabilities correlate with scientific advancement so as acquisition and analysis methodologies improve, so should the level of sophistication and performance of the underlying measurement tools.

In assessing these segmentation algorithms for hyperpolarized gas imaging, it is important to note that human expertise leverages more than relative intensity values to identify salient, clinically relevant features in images—something more akin to the complex structure of deep-layered neural networks (37), particularly convolutional neural networks (CNN). Such models have demonstrated outstanding performance in certain computational tasks, including classification and semantic segmentation in medical imaging (38). Their potential for leveraging spatial information from images surpasses the perceptual capabilities of previous approaches and even rivals that of human raters (39). Importantly, CNN optimization occurs directly in the image space to learn complex spatial features, in contrast to the previously discussed methods where optimization (primarily) concerns image intensity-only information. We introduced a deep learning approach in (40) and further expand on that work for comparison with existing approaches below. Although we find its performance to be quite promising, more fundamental to this work than the network itself is simply pointing to the general potential associated with deep learning for analyzing hyperpolarized gas images *as spatial samplings of real-world objects*, as opposed to lossy representations of such objects.

In the spirit of open science, we have made the entire evaluation framework, including our novel contributions, available within the Advanced Normalization Tools software ecosystem (ANTsX) (41).

Methods

Hyperpolarized gas imaging acquisition

University of Virginia cohort

A retrospective dataset was collected consisting of young healthy ($n = 10$), older healthy ($n = 7$), cystic fibrosis (CF) ($n = 14$), interstitial lung disease (ILD) ($n = 10$), and chronic obstructive pulmonary disease ($n = 10$). MR imaging with hyperpolarized 129Xe gas was performed under an Institutional Review Board (IRB) approved protocol with written informed consent obtained from each subject. In addition, all imaging was performed under a Food and Drug Administration (FDA) approved physician's Investigational New Drug application. MRI data were acquired on a 1.5 T whole-body MRI scanner (Siemens Avanto, Siemens Medical Solutions, Malvern, PA) with broadband capabilities and a flexible 129Xe chest radiofrequency coil (RF; IGC Medical Advances, Milwaukee, WI; or Clinical MR Solutions, Brookfield, WI). During a ≤ 10 second breath-hold following the inhalation of ≈ 1000 mL of hyperpolarized 129Xe mixed with nitrogen up to a volume equal to 1/3 forced vital capacity (FVC) of the respective subject, a set of 15-17 contiguous coronal lung slices were collected to cover the entire lungs. Parameters of the gradient echo (GRE) sequence with a spiral k-space sampling with 12 interleaves for 129Xe MRI were as follows: repetition time msec / echo time msec, 7/1; flip angle, 20°; matrix, 128 × 128: in-plane voxel size, 4 × 4 mm; section slice thickness, 15 mm; and intersection gap, none. The data were deidentified prior to analysis. These data are available upon request and through a data sharing agreement.

Harvard Dataverse cohort

In addition to the data acquired at the University of Virginia, we also processed a publicly available lung dataset (31) available at the Harvard Dataverse and detailed in (30). These

data comprised the original 129Xe acquisitions from 29 subjects (10 healthy controls and 19 mild intermittent asthmatic individuals) with corresponding lung masks. In addition, seven artificially SNR-degraded images per acquisition were also part of this data set but not used for the analyses reported below. The image headers were corrected for proper canonical anatomical orientation according to Nifti standards and uploaded to the GitHub repository associated with this work.

Data simulations

Both datasets were transformed by adding Gaussian noise, nonlinear histogram-based intensity warping, and their combination. The peak signal-to-noise ratio (PSNR) is defined as

$$PSNR = 20 \cdot \log_{10}(\max(I_{original})) - 10 \cdot \log_{10}(\text{mse}(I_{original}, I_{simulated})), \quad (1)$$

where mse denotes the mean-squared error between the simulated image and the corresponding original image. The median PSNR values for the simulated UVa dataset are noise: 20.7dB, nonlinearities: 29.9dB, and noise and nonlinearities: 19.6dB. Analogous values for the Dataverse dataset are noise: 19.8dB, nonlinearities: 26.6dB, and noise and nonlinearities: 19.4dB.

Algorithmic implementations

In support of the discussion in the Introduction, we performed various experiments to compare the algorithms mentioned previously, viz. linear binning (7), hierarchical k-means (9), fuzzy spatial c-means (13), GMM-MRF (specifically, ANTs-based *Atropos* tailored for functional lung imaging) (14), and a trained CNN with roots in our earlier work (40), which we have dubbed "El Bicho." Note that we consider the binary thresholding variants to be simplified versions of linear binning and, therefore, omit them from explicit consideration in this work. A fair and accurate comparison between algorithms necessitates several considerations which have been outlined previously (42). In designing the evaluation study:

- All algorithms and evaluation scripts have been implemented using open-source tools

by the first author and have been made available as part of the GitHub repository corresponding to this work (https://github.com/ntustison/Histograms). Lung masks for the UVa data were created using segmentation functionality described in (40) and inspected/edited by one of the co-authors (M. H.). The lung masks for the Harvard Dataverse 129Xe data are publicly available with the online image repository (31).

- An important algorithmic hyperparameter is the number of ventilation clusters. In order to minimize differences in our set of evaluations, we merged the number of resulting clusters, post-optimization, to only three clusters: "ventilation defect," "hypo-ventilation," and "other ventilation" where the first two clusters for each output are the same as the original implementations and the remaining clusters are merged into the third category.

- Another significant issue was whether to apply N4 bias correction as a preprocessing step. We ultimately decided to include it for two reasons. First, it is explicitly used in multiple algorithms (e.g., (4,7,11,14,43)) despite the issues raised previously since it qualitatively improves image appearance. Another practical consideration for N4 preprocessing was due to the parameters of the reference distribution required by the linear binning algorithm (discussed in greater detail below). However, for completeness, we did run the same experiments detailed below using the uncorrected UVa images and the previously reported parameters for linear binning, and the results were similar. These results can also be found in the GitHub repository associated with this work.

- We extended the deep learning functionality first described in (40) to improve performance and provide a more clinically granular labeling (i.e., four clusters here instead of two in the previous work). This network is a 2-D U-net (44) with enhancements including additional training data with augmentation, attention gating (45), and recommended hyperparameters (46). These include four encoding/decoding layers with 32 filters at the base layer (and doubled at each subsequent layer). Training incorporated an 80/20 data split using categorical cross entropy and a multi-label Dice function (47)

$$Dice = 2\frac{\sum_r |S_r \cap T_r|}{\sum_r |S_r| + |T_r|} \tag{2}$$

where S_r and T_r refer to the source and target regions, respectively, as loss functions.

Results

We performed several comparative evaluations to probe the previously mentioned issues broadly categorized in terms of measurement bias and precision, with most of the focus being on the latter. Given the lack of ground-truth in the form of segmentation images, addressing issues of measurement bias is difficult. In addition to the fact that the number of ventilation clusters is not consistent across algorithms, it is not clear that the ventilation categories across algorithms have identical clinical definition. This prevents application of various frameworks accommodating the lack of ground-truth for segmentation performance analysis (e.g., (48)) to these data.

As mentioned in the Introduction, the cited algorithms have all demonstrated research utility and potential clinical utility. This is supported by our first evaluation which is based on diagnostic prediction of given clinical categories assigned to the imaging cohort using derived random forest models (49). This approach also provides an additional check on the validity of the algorithmic implementations. However, it is important to recognize that this evaluation is extremely limited as the underlying data are gross measures which do not provide accuracy estimates on the level of the algorithmic output (i.e., voxelwise segmentation).

Having established the general validity of the gross algorithmic output, we then switch to our primary focus which is the comparison of measurement precision between algorithms. We first analyzed the unique requirement of a reference distribution for the linear binning algorithm. Specifically, we quantify the effects of the choice of reference cohort on the clustering parameters for the linear binning algorithm. We then incorporate the trained El Bicho model in exploring additional aspects of measurement variance based on simulating both MR noise and intensity nonlinearities.

To summarize, we performed the following evaluations/experiments:

- Global algorithmic bias (in the absence of ground truth)

  - Diagnostic prediction

- Voxelwise algorithmic precision
  - Input/output variance based on reference distribution (linear binning only)
  - Effects of simulated MR artefacts on multi-site data

Diagnostic prediction

Due to the absence of ground-truth, we adopted the strategy from previous work (41,50) where we used cross-validation to build and compare prediction models from data derived from the set of segmentation algorithms. Specifically, we use pathology diagnosis (i.e., "CF," "COPD," and "ILD") as an established research-based correlate of ventilation levels from hyperpolarized gas imaging (e.g., (35,36,43)) and quantified the predictive capabilities of corresponding binary random forest classifiers (49) of the form:

$$Pathology\ vs.\ Healthy \sim \sum_{i=1}^{3} \frac{Volume_i}{Total\ volume} \tag{3}$$

where $Volume_i$ is the volume of the i^{th} cluster and $Total\ volume$ is total lung volume which is recognized as a multiple-cluster summation extension of the ventilation defect percentage. We used a training/testing split of 80/20. Due to the small number of subjects, we combined the young and old healthy data into a single category. 100 permutations were used where training/testing data were randomly assigned and the corresponding random forest model was constructed at each permutation.

The resulting receiver operating characteristic (ROC) curves for each algorithm and each diagnostic scenario are provided in Figure 4. All four algorithms perform significantly better than a random classifier. In the absence of ground truth, this type of evaluation does provide evidence that all these algorithms produce measurements which are clinically relevant although, it should be noted, that this is a very coarse assessment strategy given the global measures used (i.e., cluster volume percentage) and the general clinical categories employed. In fact, even spirometry measures can be used to achieve highly accurate diagnostic predictions

with machine learning techniques (51).

Effects of reference image set selection

One of the additional input requirements for linear binning over the other algorithms is the generation of a reference distribution. Therefore, we additionally investigated the influence of reference data set on the outcome of linear binning classification, since this is an integral aspect unique to this method. In addition to the output measurement variation caused by choice of the reference image cohort, this played a role in determining whether or not to use N4 preprocessing. As mentioned, a significant portion of N4 processing involves the deconvolution of the image histogram to sharpen the histogram peaks which decreases the standard deviation of the intensity distribution and can also result in a histogram shift. Using the original set of 10 young healthy data with no N4 preprocessing, we created a reference distribution according to (7), which resulted in an approximate distribution of $\mathcal{N}(0.45, 0.24)$. This produced 0 voxels being classified as belonging to Cluster 1 (Figure 5(a)) because two standard deviations from the mean is less than 0 and Cluster 1 resides in the region below -2 standard deviations. However, using N4-preprocessed images produced something closer, $\mathcal{N}(0.56, 0.22)$, to the published values, $\mathcal{N}(0.52, 0.18)$, reported in (7), resulting in a non-empty set for that cluster. This is consistent, though, with linear binning which does use N4 bias correction for preprocessing. We also mention that the Harvard Dataverse images used were preprocessed using N4 (30) which provides a third reason for its use on the University of Virginia image dataset (to maximize cross cohort consistency). In the case of the former image set, we did use the previously reported linear binning mean and standard deviation algorithm parameter values (i.e., $\mathcal{N}(0.52, 0.18)$). This was the only parameter difference between analyzing the two image sets.

The previous implications of the chosen image reference set also caused us to look at this choice as a potential source of both input and output variance in the measurements utilized and produced by linear binning. Regarding the former, we took all possible combinations of our young healthy control subject images and looked at the resulting mean and standard deviation values. As expected, there is significant variation for both mean and standard

deviation values (see top portion of Figure 6) which are used to derive the cluster threshold values. This directly impacts output measurements such as ventilation defect percentage. For the reference sets comprising eight or nine images, we compute the corresponding linear binning segmentation and estimate the volumetric percentage for each cluster. Then, for each subject, we computed the min/max range for these values and plotted those results cluster-wise on the bottom of Figure 6. This demonstrates that the additional requirement of a reference distribution is a source of potentially significant measurement variation for the linear binning algorithm.

Effects of MR-based simulated image distortions

As we mentioned in the Introduction, noise and nonlinear intensity artefacts common to MRI can have a significant distortion effect on the image with even greater effects seen with respect to change in the structure of the corresponding histogram. This final evaluation explores the effects of these artefacts on the algorithmic output on a voxelwise scale using the Dice metric (Equation (2)) which has a range of [0,1] where 1 signifies perfect agreement between the segmentations and 0 is no agreement.

Ten simulated images for each of the subjects of both the University of Virginia and Harvard Dataverse cohort were generated for each of the three categories of randomly generated artefacts: noise, nonlinearities, and combined noise and intensity nonlinearities. The original image as well as the simulated images were segmented using each of the five algorithms. Following our earlier protocol, we maintained the original Clusters 1 and 2 per algorithm and combined the remaining clusters into a single third cluster. This allowed us to compare between algorithms and maintain separate those clusters which are the most studied and reported in the literature. The Dice metric was used to quantify the amount of deviation, per cluster, between the segmentation produced by the original image and the corresponding simulated distorted image segmentation which is summarized in Figures 7 and 9. The algorithms were then compared, on a per-cluster and per-artefact basis, using one-way ANOVA followed by Tukey's Honest Significant Difference (HSD) test in Figures 8 and 10. The results of these tests are further visualized via simplified alluvial diagrams with the superior performing

algorithms, in terms of Dice overlap, listed on the left connecting to their worse performing counterparts on the right where the width of the connection is proportional to the overlap difference and colored by artefact type. The algorithms which exploit image-based spatial information, most notably El Bicho, demonstrate generally superior performance as compared with their histogram-only counterparts in both data sets. For example, in Cluster 1, for both datasets, the sole histogram-only algorithm that demonstrates any elevated pairwise performance is k-means but, proportionally, this significance is dwarfed by the performance of the algorithms which leverage spatial information. Additionally, it is apparent from these tests that El Bicho consistently provides the best performance across the specified clusters in the presence of MR-based image distortions.

Discussion

Over the past decade, multiple algorithms have been proposed for the segmentation of hyperpolarized gas images into clinically based functional categories. These algorithms are optimized using the histogram information primarily (with many using it exclusively) much to the relative detriment of algorithmic robustness and segmentation quality. This is due to the simple fact that these approaches discard, or do not optimally leverage, a vital piece of information essential for accurate quantitative image interpretation—the spatial relationships between voxel intensities. While simplifying the underlying complexity of the segmentation problem, these algorithms are deficient in leveraging the general modelling principle of incorporating as much available prior information to any solution method. In fact, this is a fundamental implication of the "No Free Lunch Theorem" (54)—algorithmic performance hinges on available prior information.

As illustrated in Figure 2, measures based on the human visual system seem to quantify what is understood intuitively; that image-based information is much more robust than its corresponding histogram-based information in the presence of image transformations, such as common MR artefacts. This observation is not intended to imply that the histogram-based approaches are useless in performing research. In fact, ventilation defect percentage is perhaps the most widely used clinical measurement reported in the literature and it is

easily quantified from the image histogram. Thus, even relatively simple histogram-only segmentation algorithms will provide some utility which was observed in the measurement bias experiments employing a variant of ventilation defect percentage to predict diagnostic accuracy. However, similar to the lossy relationship between the image and its corresponding histogram, such volumetric-based measures are lossy distillations of the segmentation information and might obscure important algorithmic characteristics and relative differences as well as discard potentially useful spatial information which is why additional experimentation explored measurement precision in the presence of MR artefacts.

Common MR artefacts of noise and intensity nonlinearities can produce quantifiable differences in the segmentation results and the degree of deviation (i.e., lack of measurement precision) largely corresponds to the algorithmic choice of optimization domain, i.e., image-based vs. histogram-based, with those algorithms leveraging the former providing improved segmentation repeatability. Notably, El Bicho generally yields the best segmentation overlap measures over the specified clusters and MR artefacts most likely due to optimization of the governing network weights over hierarchical image features found in the training set as opposed to strictly relative intensities and/or more simplistic neighborhood intensity information. In addition, this network demonstrates site acquisition generalizability as these performance gains are also seen in the Harvard Dataverse dataset.

In addition to motivating a renewed assessment of current algorithmic approaches to pulmonary hyperpolarized gas segmentation, there other avenues for further research. El Bicho was developed in parallel with the writing of this manuscript merely to showcase the incredible potential that deep learning can have in the field of hyperpolarized gas imaging (as well as to update our earlier work (40)). We certainly recognize and expect that alternative deep learning strategies (e.g., hyperparameter choice, training data selection, data augmentation, etc.) would provide comparable and even superior performance to what was presented with El Bicho. However, that is precisely our motivation for presenting this work—deep learning, generally, presents a much better alternative than histogram approaches as network training directly takes place in the image (i.e., spatial) domain and not in a transformed space where key information has been discarded. Just as important, deep learning provides other avenues

for research exploration and development. For example, given the relatively lower resolution of the acquisition image, exploration of the effects of deep learning-based super-resolution might prove worthy of application-specific investigation (55). Also, with the same network software libraries, high-performing classification networks can be constructed and trained which might yield novel insights regarding image-based characterization of disease. One additional modification that we did not explore in this work, but is extremely important, is the confound caused by multi-site data which has yet to be explored in-depth. With neural networks, such confounds can be handled as part of the training process or as an explicit network modification. Either would be important to consider for future work.

Admittedly, this work was limited in its exploration of MR artefacts. Noise variation was limited to a zero-mean Gaussian distribution and nonlinear intensity variation was explored strictly through smoothly varying histogram deformation. Inclusion of other noise models (e.g., shot, salt-and-pepper) might further characterize algorithmic differences and provide additional realistic data augmentation strategies. Specific to nonlinear intensity variation, a recent addition to the ANTsX ecosystem allows for the possible simulation of bias fields which would also expand data augmentation and, significantly, in the spirit of algorithmic parsimony, could potentially remove the dependency of N4 bias correction as an unnecessary preprocessing step.

Finally, although ventilation defect percentage has proven to be a compelling quantity for clinical studies, the results from the diagnostic prediction evaluation and the previous discussion implies that this popular measure does not fully leverage the spatial information of the segmentation information from any of these algorithms. Perhaps the results of this work, in addition to pointing to the need for rethinking algorithm innovation direction, also point to possibly investigating differentiating spatial patterns within the images as evidence of disease and/or growth and correlations with non-imaging data using sophisticated voxel-scale statistical techniques which intrinsically leverage spatial information (e.g., similarity-driven multivariate linear reconstruction (52,53)).

Acknowledgments

Support for the research reported in this work includes funding from the National Heart, Lung, and Blood Institute of the National Institutes of Health (R01HL133889).

References

1. Altes TA, Powers PL, Knight-Scott J, et al. Hyperpolarized 3He MR lung ventilation imaging in asthmatics: Preliminary findings. J Magn Reson Imaging 2001;13:378–84.

2. Lange EE de, Mugler JP 3rd, Brookeman JR, et al. Lung air spaces: MR imaging evaluation with hyperpolarized 3He gas. Radiology 1999;210:851–7 doi: 10.1148/radiology.210.3.r99fe08851.

3. Samee S, Altes T, Powers P, et al. Imaging the lungs in asthmatic patients by using hyperpolarized helium-3 magnetic resonance: Assessment of response to methacholine and exercise challenge. J Allergy Clin Immunol 2003;111:1205–11.

4. Shammi UA, D'Alessandro MF, Altes T, et al. Comparison of hyperpolarized 3He and 129Xe MR imaging in cystic fibrosis patients. Acad Radiol 2021 doi: 10.1016/j.acra.2021.01.007.

5. Thomen RP, Sheshadri A, Quirk JD, et al. Regional ventilation changes in severe asthma after bronchial thermoplasty with (3)He MR imaging and CT. Radiology 2015;274:250–9 doi: 10.1148/radiol.14140080.

6. Woodhouse N, Wild JM, Paley MNJ, et al. Combined helium-3/proton magnetic resonance imaging measurement of ventilated lung volumes in smokers compared to never-smokers. J Magn Reson Imaging 2005;21:365–9 doi: 10.1002/jmri.20290.

7. He M, Driehuys B, Que LG, Huang Y-CT. Using hyperpolarized 129Xe MRI to quantify the pulmonary ventilation distribution. Acad Radiol 2016;23:1521–1531 doi: 10.1016/j.acra.2016.07.014.

8. He M, Wang Z, Rankine L, et al. Generalized linear binning to compare hyperpolarized 129Xe ventilation maps derived from 3D radial gas exchange versus dedicated multislice gradient echo MRI. Acad Radiol 2020;27:e193–e203 doi: 10.1016/j.acra.2019.10.016.

9. Kirby M, Heydarian M, Svenningsen S, et al. Hyperpolarized 3He magnetic resonance functional imaging semiautomated segmentation. Acad Radiol 2012;19:141–52 doi: 10.1016/j.acra.2011.10.007.

10. Kirby M, Svenningsen S, Owrangi A, et al. Hyperpolarized 3He and 129Xe MR imaging in healthy volunteers and patients with chronic obstructive pulmonary disease. Radiology 2012;265:600–10 doi: 10.1148/radiol.12120485.

11. Zha W, Niles DJ, Kruger SJ, et al. Semiautomated ventilation defect quantification in exercise-induced bronchoconstriction using hyperpolarized helium-3 magnetic resonance imaging: A repeatability study. Acad Radiol 2016;23:1104–14 doi: 10.1016/j.acra.2016.04.005.

12. Ray N, Acton ST, Altes T, Lange EE de, Brookeman JR. Merging parametric active contours within homogeneous image regions for MRI-based lung segmentation. IEEE Trans Med Imaging 2003;22:189–99 doi: 10.1109/TMI.2002.808354.

13. Hughes PJC, Horn FC, Collier GJ, Biancardi A, Marshall H, Wild JM. Spatial fuzzy c-means thresholding for semiautomated calculation of percentage lung ventilated volume from hyperpolarized gas and 1 h MRI. J Magn Reson Imaging 2018;47:640–646 doi: 10.1002/jmri.25804.

14. Tustison NJ, Avants BB, Flors L, et al. Ventilation-based segmentation of the lungs using hyperpolarized (3)he MRI. J Magn Reson Imaging 2011;34:831–41 doi: 10.1002/jmri.22738.

15. Nyúl LG, Udupa JK. On standardizing the MR image intensity scale. Magn Reson Med 1999;42:1072–81 doi: 10.1002/(sici)1522-2594(199912)42:6<1072::aid-mrm11>3.0.co;2-m.

16. Andersen AH. On the Rician distribution of noisy MRI data. Magn Reson Med 1996;36:331–3 doi: 10.1002/mrm.1910360222.

17. Gudbjartsson H, Patz S. The Rician distribution of noisy MRI data. Magn Reson Med 1995;34:910–4 doi: 10.1002/mrm.1910340618.

18. Sled JG, Zijdenbos AP, Evans AC. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans Med Imaging 1998;17:87–97 doi: 10.1109/42.668698.

19. Collewet G, Strzelecki M, Mariette F. Influence of MRI acquisition protocols and image intensity normalization methods on texture classification. Magn Reson Imaging 2004;22:81–91 doi: 10.1016/j.mri.2003.09.001.

20. De Nunzio G, Cataldo R, Carlà A. Robust intensity standardization in brain magnetic resonance images. J Digit Imaging 2015;28:727–37 doi: 10.1007/s10278-015-9782-8.

21. Nyúl LG, Udupa JK, Zhang X. New variants of a method of MRI scale standardization. IEEE Trans Med Imaging 2000;19:143–50 doi: 10.1109/42.836373.

22. Wendt RE 3rd. Automatic adjustment of contrast and brightness of magnetic resonance images. J Digit Imaging 1994;7:95–7 doi: 10.1007/BF03168430.

23. Ashburner J, Friston KJ. Unified segmentation. Neuroimage 2005;26:839–51 doi: 10.1016/j.neuroimage.2005.02.018.

24. Avants BB, Tustison NJ, Wu J, Cook PA, Gee JC. An open source multivariate framework for n-tissue segmentation with evaluation on public data. Neuroinformatics 2011;9:381–400 doi: 10.1007/s12021-011-9109-y.

25. Zhang Y, Brady M, Smith S. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans Med Imaging 2001;20:45–57 doi: 10.1109/42.906424.

26. Chuang K-S, Tzeng H-L, Chen S, Wu J, Chen T-J. Fuzzy c-means clustering with spatial information for image segmentation. Comput Med Imaging Graph 2006;30:9–15 doi: 10.1016/j.compmedimag.2005.10.001.

27. Bezdek JC. Pattern recognition with fuzzy objective function algorithms. New York: Plenum Press; 1981.

28. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological) 1977;39:1–38.

29. Tustison NJ, Avants BB, Cook PA, et al. N4ITK: Improved N3 bias correction. IEEE Trans Med Imaging 2010;29:1310–20 doi: 10.1109/TMI.2010.2046908.

30. He M, Zha W, Tan F, Rankine L, Fain S, Driehuys B. A comparison of two hyperpolarized 129Xe MRI ventilation quantification pipelines: The effect of signal to noise ratio. Acad Radiol 2019;26:949–959 doi: 10.1016/j.acra.2018.08.015.

31. He M, Zha W, Tan F, Rankine L, Fain S, Driehuys B. SNR-degraded 129Xe ventilation MRI for the comparison of quantification methods. 2018 doi: 10.7910/DVN/FCQWP1.

32. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: From error visibility to structural similarity. IEEE Trans Image Process 2004;13:600–12 doi: 10.1109/tip.2003.819861.

33. Svenningsen S, McIntosh M, Ouriadov A, et al. Reproducibility of hyperpolarized 129Xe MRI ventilation defect percent in severe asthma to evaluate clinical trial feasibility. Acad Radiol 2020 doi: 10.1016/j.acra.2020.04.025.

34. Couch MJ, Thomen R, Kanhere N, et al. A two-center analysis of hyperpolarized 129Xe lung MRI in stable pediatric cystic fibrosis: Potential as a biomarker for multi-site trials. J Cyst Fibros 2019;18:728–733 doi: 10.1016/j.jcf.2019.03.005.

35. Mammarappallil JG, Rankine L, Wild JM, Driehuys B. New developments in imaging idiopathic pulmonary fibrosis with hyperpolarized xenon magnetic resonance imaging. J Thorac Imaging 2019;34:136–150 doi: 10.1097/RTI.0000000000000392.

36. Myc L, Qing K, He M, et al. Characterisation of gas exchange in COPD with dissolved-phase hyperpolarised xenon-129 MRI. Thorax 2020 doi: 10.1136/thoraxjnl-2020-214924.

37. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44 doi: 10.1038/nature14539.

38. Shen D, Wu G, Suk H-I. Deep learning in medical image analysis. Annu Rev Biomed Eng 2017;19:221–248 doi: 10.1146/annurev-bioeng-071516-044442.

39. Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The unreasonable effectiveness of deep features as a perceptual metric. In: 2018 IEEE/CVF conference on computer vision and pattern recognition.; 2018. pp. 586–595. doi: 10.1109/CVPR.2018.00068.

40. Tustison NJ, Avants BB, Lin Z, et al. Convolutional neural networks with template-based data augmentation for functional lung image quantification. Acad Radiol 2019;26:412–423 doi: 10.1016/j.acra.2018.08.003.

41. Tustison NJ, Cook PA, Holbrook AJ, et al. The ANTsX ecosystem for quantitative

biological and medical imaging. Sci Rep 2021;11:9068 doi: 10.1038/s41598-021-87564-6.

42. Tustison NJ, Johnson HJ, Rohlfing T, et al. Instrumentation bias in the use and evaluation of scientific software: Recommendations for reproducible practices in the computational sciences. Front Neurosci 2013;7:162 doi: 10.3389/fnins.2013.00162.

43. Santyr G, Kanhere N, Morgado F, Rayment JH, Ratjen F, Couch MJ. Hyperpolarized gas magnetic resonance imaging of pediatric cystic fibrosis lung disease. Acad Radiol 2019;26:344–354 doi: 10.1016/j.acra.2018.04.024.

44. Falk T, Mai D, Bensch R, et al. U-net: Deep learning for cell counting, detection, and morphometry. Nat Methods 2019;16:67–70 doi: 10.1038/s41592-018-0261-2.

45. Schlemper J, Oktay O, Schaap M, et al. Attention gated networks: Learning to leverage salient regions in medical images. Med Image Anal 2019;53:197–207 doi: 10.1016/j.media.2019.01.012.

46. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 2020 doi: 10.1038/s41592-020-01008-z.

47. Crum WR, Camara O, Hill DLG. Generalized overlap measures for evaluation and validation in medical image analysis. IEEE Trans Med Imaging 2006;25:1451–61 doi: 10.1109/TMI.2006.880587.

48. Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation. IEEE Trans Med Imaging 2004;23:903–21 doi: 10.1109/TMI.2004.828354.

49. Breiman L. Random forests. Machine Learning 2001;45:5–32.

50. Tustison NJ, Cook PA, Klein A, et al. Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements. Neuroimage 2014;99:166–79 doi: 10.1016/j.neuroimage.2014.05.044.

51. Badnjevic A, Gurbeta L, Custovic E. An expert diagnostic system to automatically identify asthma and chronic obstructive pulmonary disease in clinical settings. Sci Rep

2018;8:11645 doi: 10.1038/s41598-018-30116-2.

52. Avants BB, Tustison NJ, Stone JR. Similarity-driven multi-view embeddings from high-dimensional biomedical data. Nature Computational Science 2021 doi: 10.1038/s43588-021-00049-4.

53. Stone JR, Avants BB, Tustison NJ, et al. Functional and structural neuroimaging correlates of repetitive low-level blast exposure in career breachers. J Neurotrauma 2020;37:2468–2481 doi: 10.1089/neu.2020.7141.

54. Wolpert DH, Macready WG. No free lunch theorems for optimization. Trans. Evol. Comp 1997;1:67–82 doi: 10.1109/4235.585893.

55. Li Y, Sixou B, Peyrin F. A review of the deep learning methods for medical images super resolution problems. IRBM 2020 doi: https://doi.org/10.1016/j.irbm.2020.08.004.

List of Figures

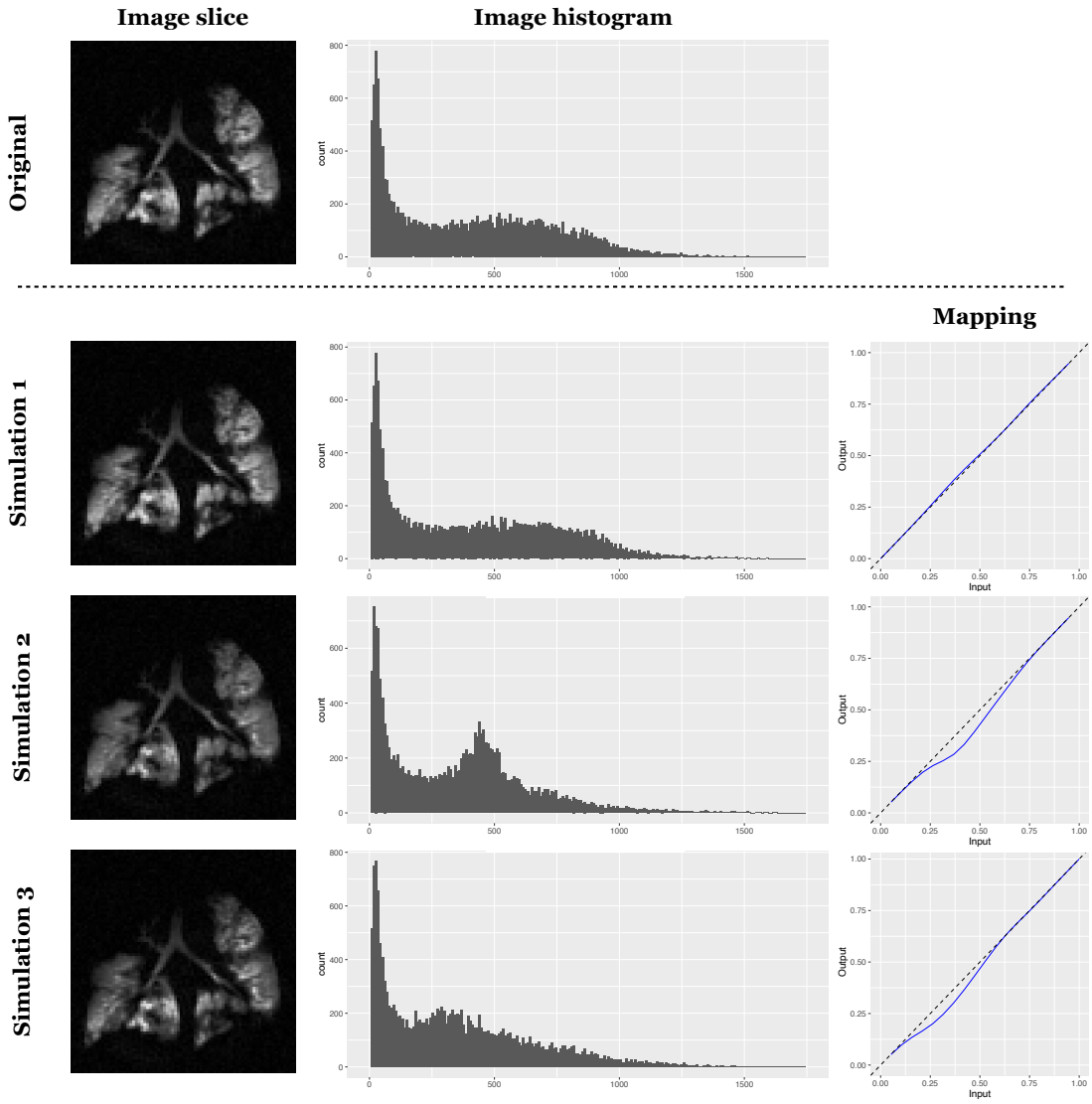


Figure 1: Illustration of the effect of MR nonlinear intensity warping on the histogram structure using a representative sampling of the simulations used in the experiments in this work. By simulating these types of nonlinear intensity changes, we can visualize both the image and the corresponding intensity histogram and investigate the effects on salient outcome measures. These simulated intensity mappings, although relatively small and difficult to distinguish in the image domain, can have an algorithmically consequential effect on the histogram structure.

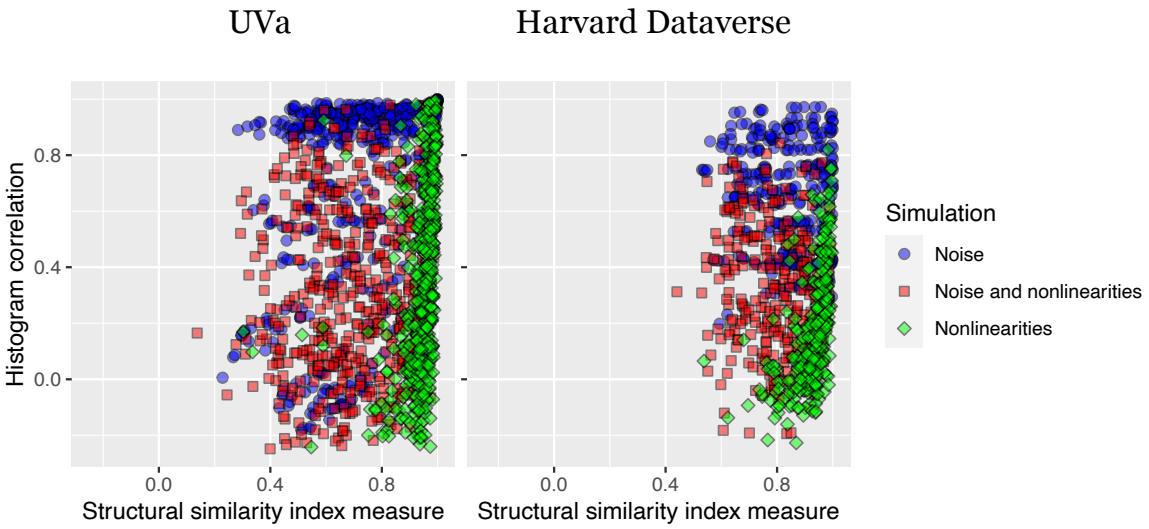


Figure 2: Multi-site: (left) University of Virginia (UVa) and (right) Harvard Dataverse 129Xe data. Image-based SSIM vs. histogram-based Pearson's correlation differences under distortions induced by the common MR artefacts of noise and intensity nonlinearities. For the nonlinearity-only simulations, the images maintain their structural integrity as the SSIM values remain close to 1. This is in contrast to the corresponding range in histogram similarity which is much larger. The effects with simulated Gaussian noise are similar where the range in histogram differences with simulated noise is much greater than the range in SSIM. Both sets of observations are evidence of the lack of robustness to distortions in the histogram domain in comparison with the original image domain.

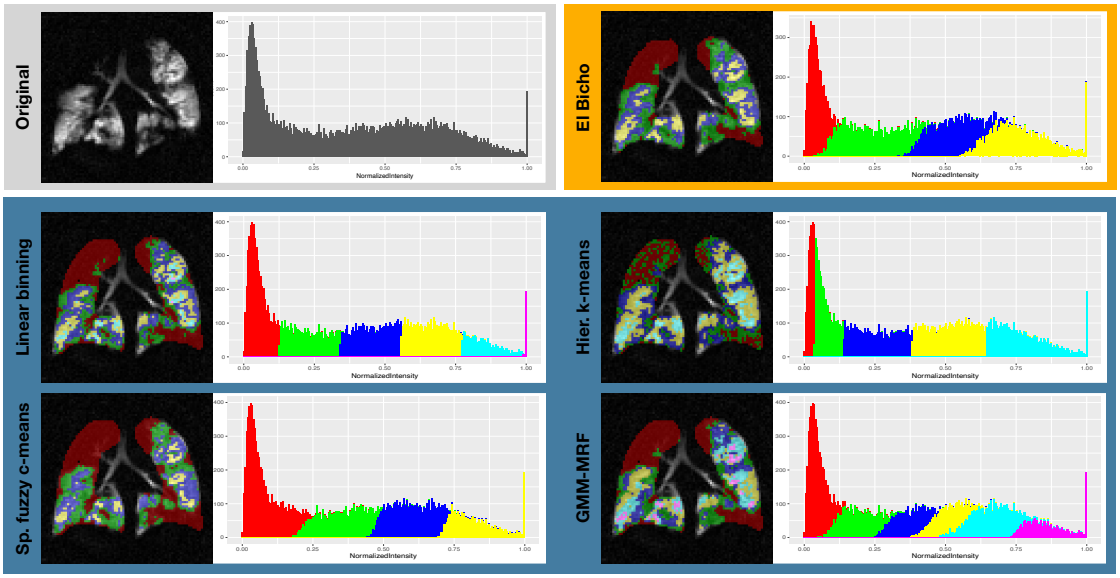


Figure 3: Illustration of sample segmentations produced by the four algorithms described above (i.e., linear binning, hierarchical k-means, spatial fuzzy c-means, and GMM-MRF) and the deep learning algorithm ("El Bicho") described below on a single cystic fibrosis subject. Also included are the corresponding segmentation histograms. Although quite disparate in the actual labeling of the lung and resulting histogram, each algorithm produces a reasonable parcellation.

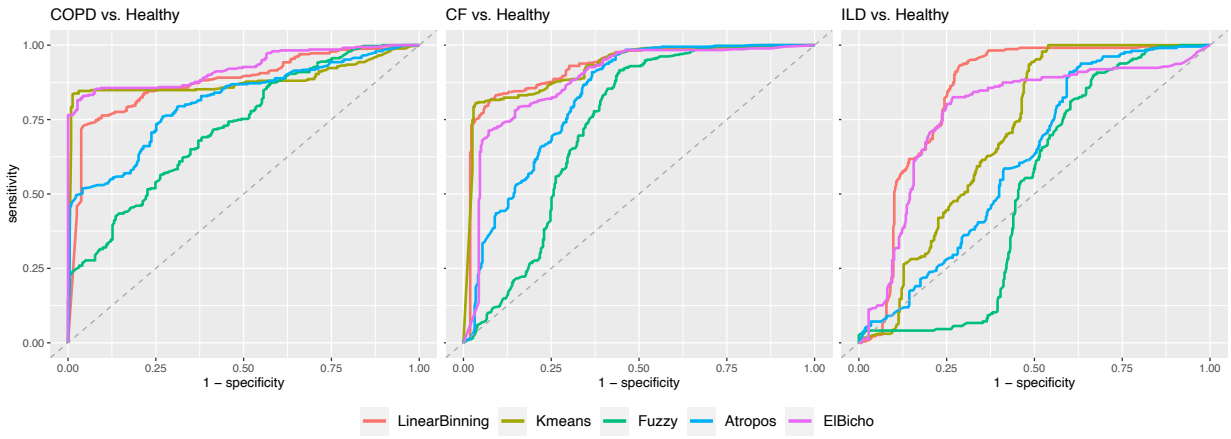
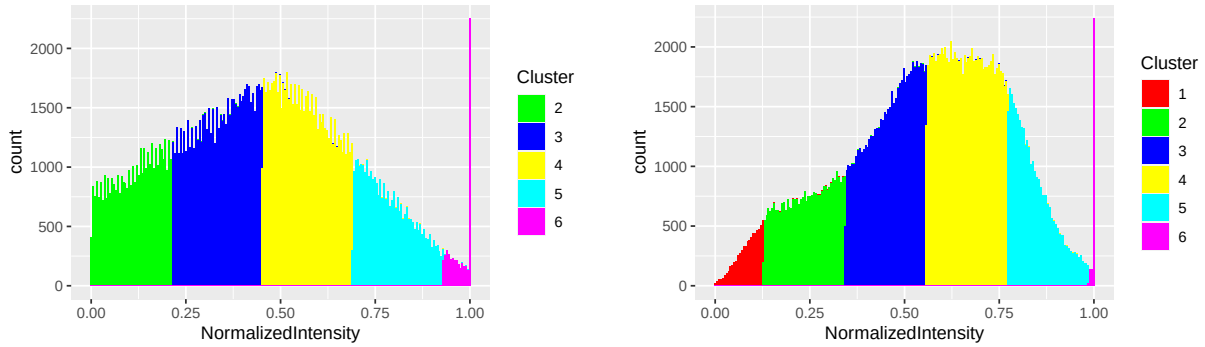


Figure 4: ROC curves resulting from the diagnostic prediction evaluation strategy involving randomly permuted training/testing data sets and predictive random forest models.



(a) Reference distribution (original images).

(b) Reference distribution (N4 images).

Figure 5: Ten young healthy subjects were combined to create two reference distributions, one based on the (a) original images and the other using (b) N4 preprocessing. Based on the generated mean and standard deviation of the aggregated samples, we label the resulting clusters in the respective histograms. Due to the lower mean and higher standard deviation of the original image set, Cluster 1 is not within the range of $[0, 1]$ for the resulting reference distribution which motivated the use of the N4 preprocessed image set.

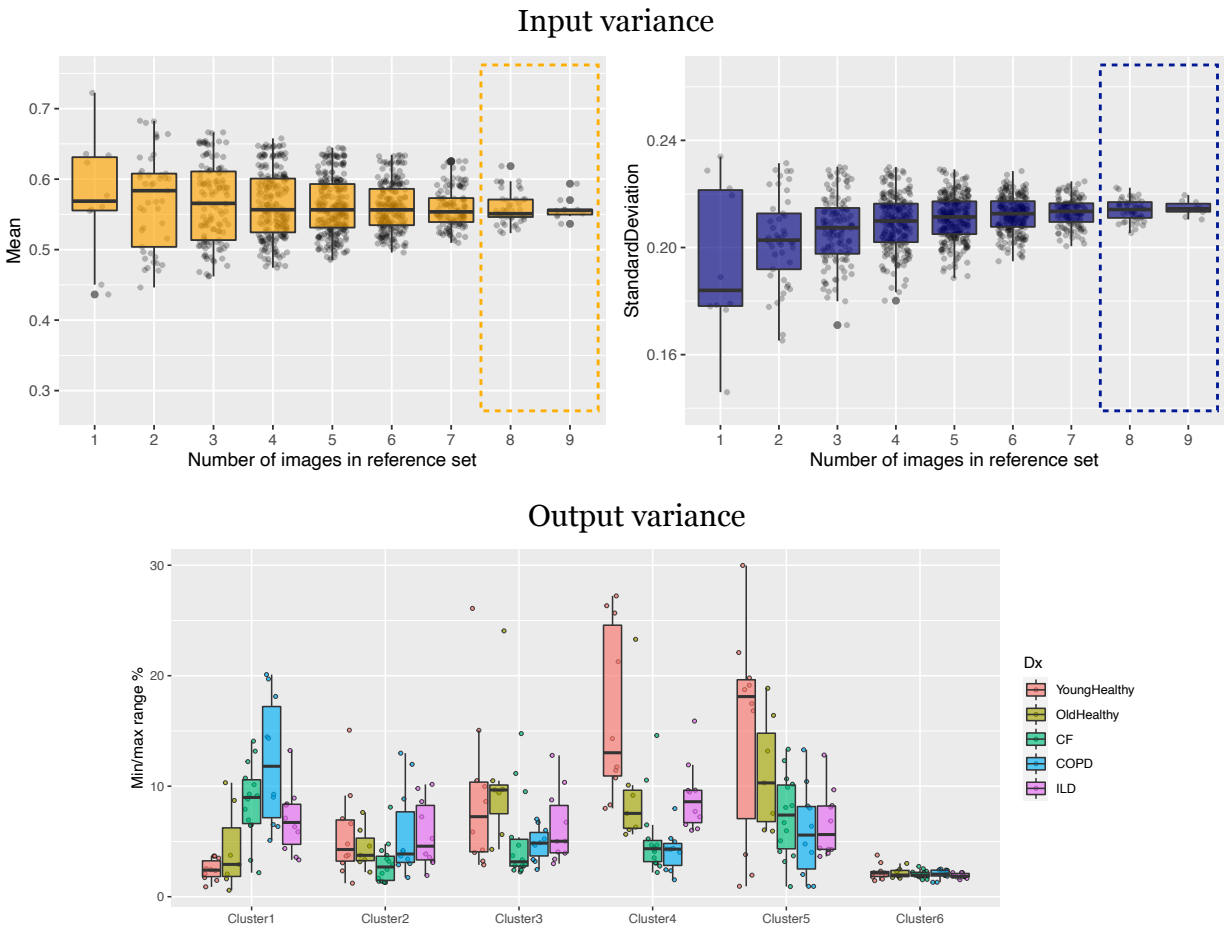


Figure 6: (Top) Variation of the mean (left) and standard deviation (right) over choice of reference set based on all different combinations of young healthy subjects per specified number of subjects. Although these parameters demonstrate convergence, there is still non-zero variation for any given set. (Bottom) This input variance is a source of output variance in the cluster volume plotted as the maximum range per subject as a percentage of total lung volume. We limit this exploration to reference sets with eight or nine images.

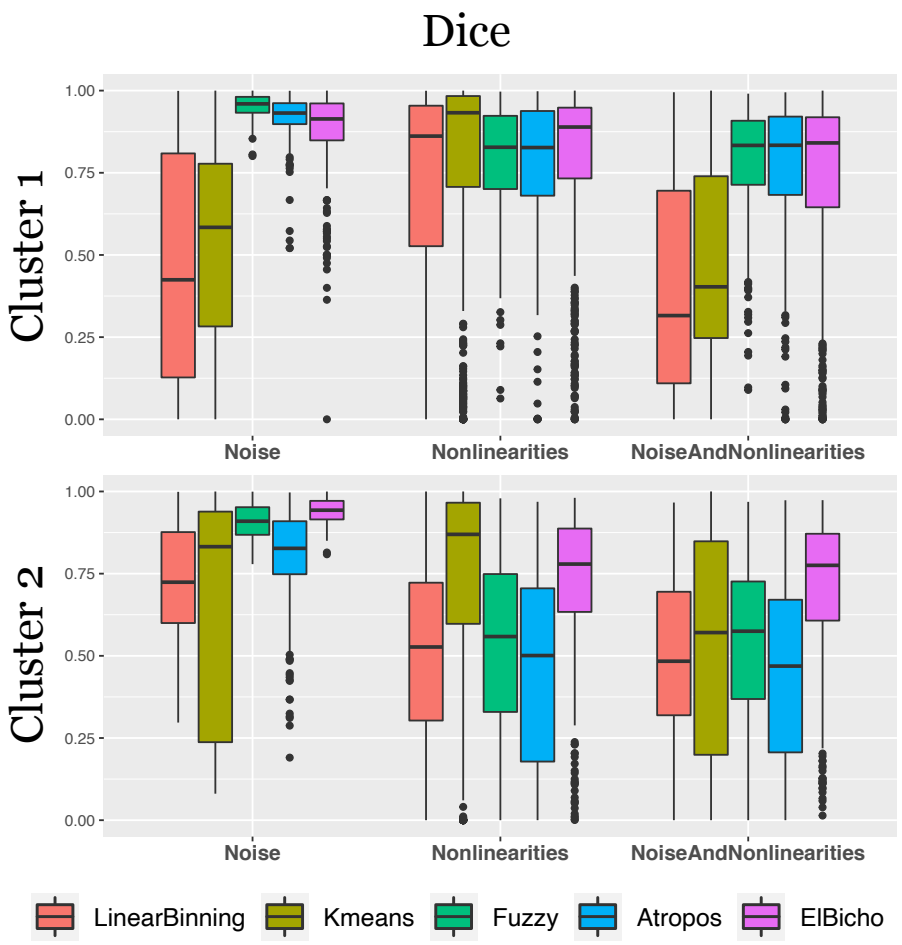


Figure 7: University of Virginia image cohort. Box plots illustrate the lack of segmentation overlap with reference segmentations caused by distortions produced by noise, histogram-based intensity nonlinearities, and their combination as measured by the Dice metric over all five algorithms. We provide the results of the two pathologically-relevant labels for comparison: "ventilation defect" (Cluster 1) and "hypo-ventilation" (Cluster 2).

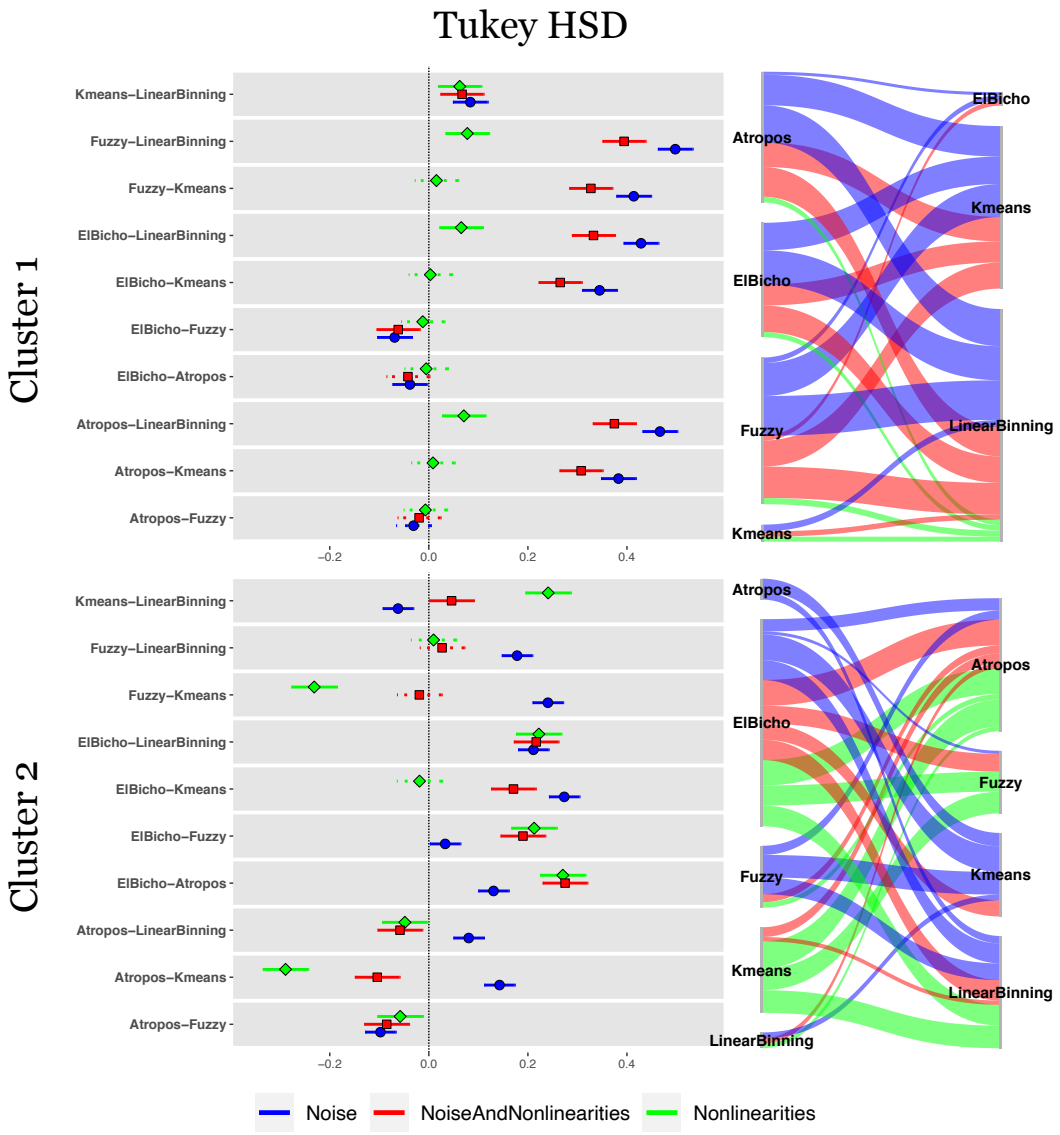


Figure 8: University of Virginia image cohort. (Left) Results from Tukey's test following one-way ANOVA to compare the resulting overlaps between algorithms (cf Figure 7). Higher positive values indicate increased robustness to simulated image distortions. A solid line indicates statistical significance at the 0.05 level whereas the dashed line indicates no statistically significant difference. (Right) To further visualize the Tukey results, a simplified alluvial diagram is used to provide connections illustrating relative performance between algorithms where the algorithms listed on the left have improved performance relative to their connected algorithms on the right with the width of the connection being proportional to difference in performance.

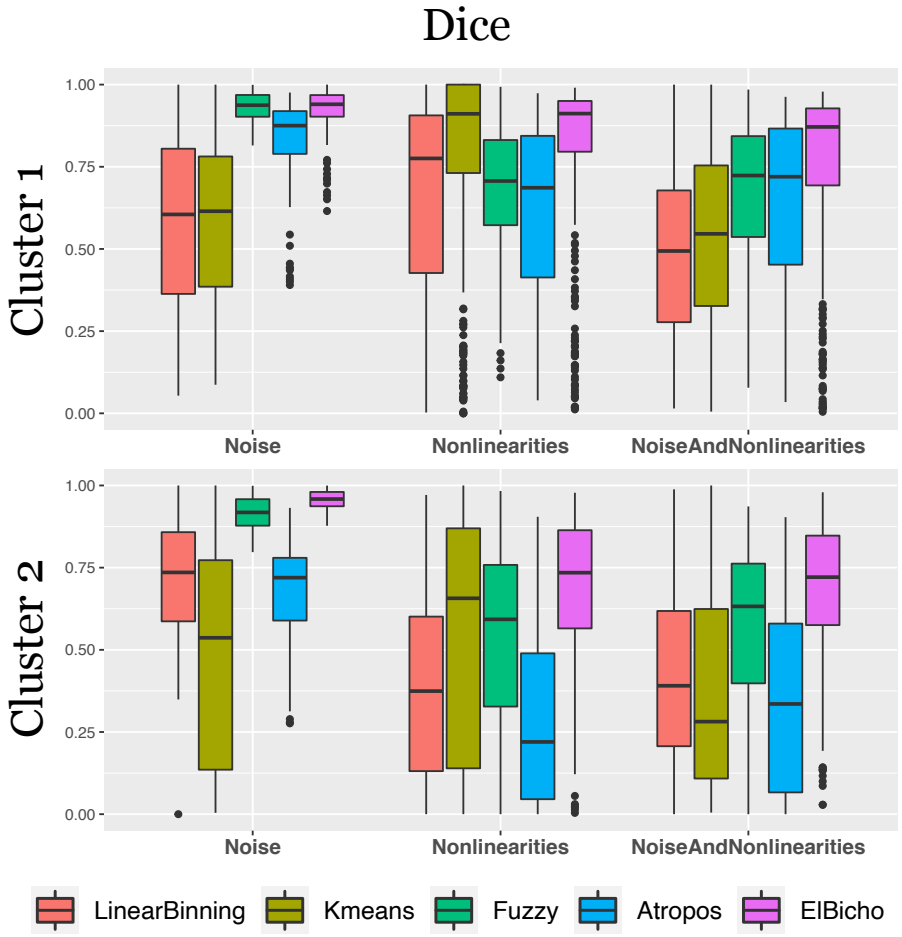


Figure 9: Harvard Dataverse image cohort. Box plots illustrate the lack of segmentation overlap with reference segmentations caused by distortions produced by noise, histogram-based intensity nonlinearities, and their combination as measured by the Dice metric over all five algorithms. We provide the results of the two pathologically-relevant labels for comparison: "ventilation defect" (Cluster 1) and "hypo-ventilation" (Cluster 2).

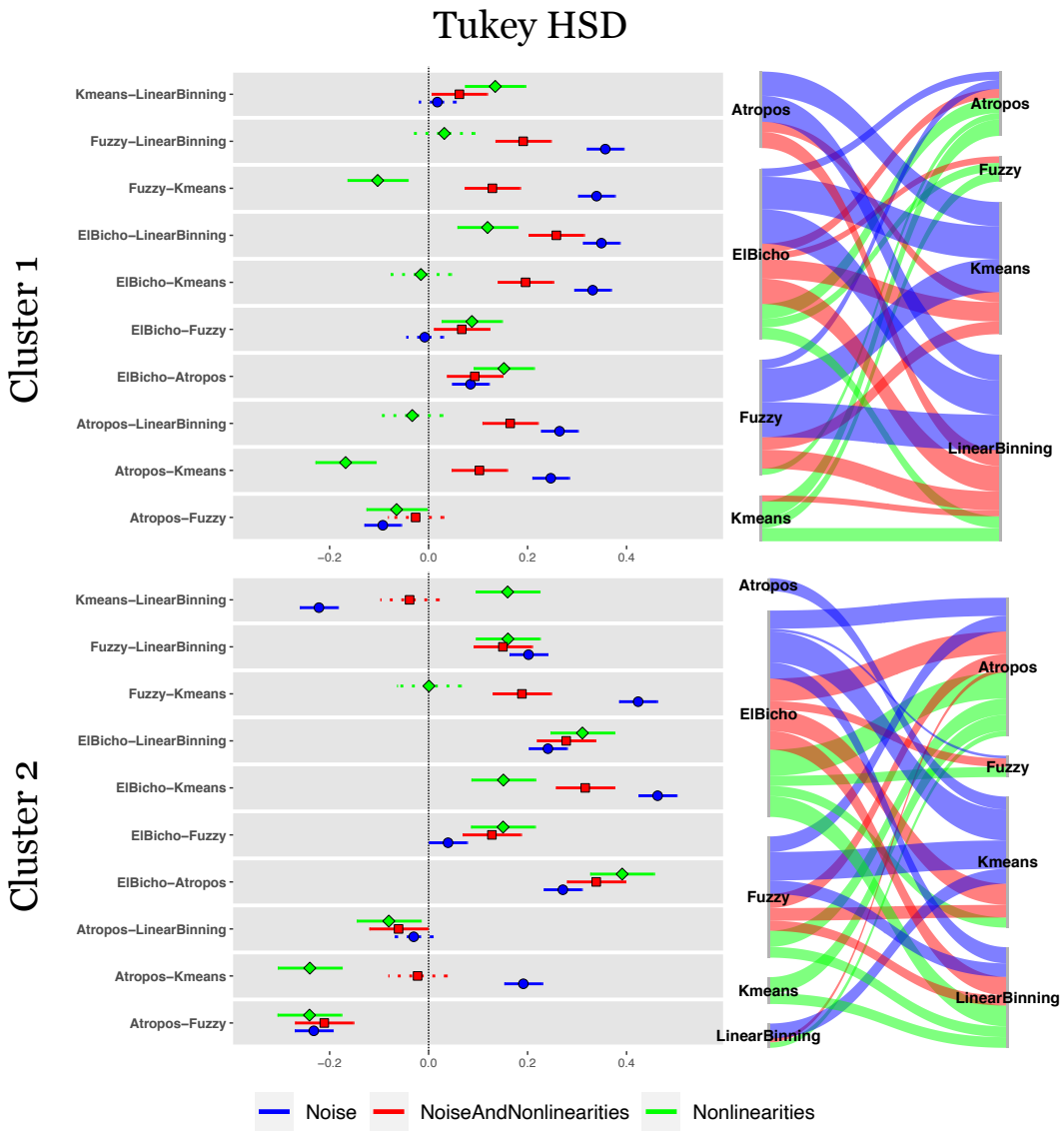


Figure 10: Harvard Dataverse image cohort. (Left) Results from Tukey's test following one-way ANOVA to compare the resulting overlaps between algorithms (cf Figure 9). Higher positive values indicate increased robustness to simulated image distortions. A solid line indicates statistical significance at the 0.05 level whereas the dashed line indicates no statistically significant difference. (Right) To further visualize the Tukey results, a simplified alluvial diagram is used to provide connections illustrating relative performance between algorithms where the algorithms listed on the left have improved performance relative to their connected algorithms on the right with the width of the connection being proportional to difference in performance.