

UCLA

UCLA Electronic Theses and Dissertations

Title

Image quality metrics as a basis for characterizing and predicting CT protocol-induced variation in quantitative imaging

Permalink

<https://escholarship.org/uc/item/0f21z1r3>

ISBN

9798190920239

Author

Daly, Morgan

Publication Date

2026-09-08

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Image quality metrics as a basis for characterizing and predicting CT protocol-induced variation
in quantitative imaging

A dissertation submitted in partial satisfaction of the
requirements for the degree of Doctor of Philosophy
in Physics and Biology in Medicine

by

Morgan Allison Daly

2026

© Copyright by

Morgan Allison Daly

2026

ABSTRACT OF THE DISSERTATION

Image quality metrics as a basis for characterizing and predicting CT protocol-induced variation
in quantitative imaging

by

Morgan Allison Daly

Doctor of Philosophy in Physics and Biology in Medicine

University of California, Los Angeles, 2026

Professor Michael McNitt-Gray, Chair

Quantitative imaging in CT promises objective descriptors of disease from medical images, but clinical translation is limited by sensitivity to CT acquisition and reconstruction protocol. The same object imaged under different scanner models, reconstruction kernels, or dose levels can yield substantially different measurements, and for the texture features that carry much of radiomics' discriminative value, this protocol-induced variation can rival or exceed the biological differences of interest. The same dependence undermines deep learning models, whose performance often falls on data acquired outside the training distribution. Existing responses largely seek to reduce a biomarker's sensitivity to protocol, through harmonization, standardization, or stability-based feature selection. This dissertation takes a complementary approach: rather than reduce protocol sensitivity, it characterizes and predicts it in terms of physical, quantitative image quality metrics.

To enable this, an open-source CT texture phantom with 3D-printed modular inserts was designed and manufactured, producing distinct and reproducible radiomic signatures. The phantom was scanned together with an image quality phantom across 170 protocols spanning four scanner models from two vendors, pairing measurements of three-dimensional image quality (MTF, NPS, NEQ, and CNR) with radiomic texture features extracted under identical conditions. Analysis of this dataset established that texture features respond systematically to fundamental image quality characteristics, with relationships largely consistent across different physical textures. Regression models were then developed to predict radiomic feature values from image quality metrics. Finally, the framework was extended to a convolutional neural network classifying lung nodules, trained and evaluated across reconstruction kernels. The change in classifier performance was predictable from a single phantom-derived resolution metric, and a model fit on one scanner retained much of its accuracy when transferred to an unseen scanner, providing initial evidence that protocol-induced performance degradation can be anticipated from phantom measurements alone.

By describing protocol sensitivity in vendor-neutral physical terms, this work moves toward defining the range of CT conditions over which a quantitative imaging biomarker can be reliably applied, an imaging inclusion criterion expressed in measurable image quality rather than manufacturer labels, and toward predicting prospectively where such models would fail on out-of-distribution clinical data.

The dissertation of Morgan Allison Daly is approved.

Matthew S. Brown

John M. Hoffman

Grace Kim

Dan Ruan

Frédéric Noo

Michael McNitt-Gray, Comittee Chair

University of California, Los Angeles

2026

To my parents,
who did everything to make this possible.

Contents

Contents	vi
List of Figures	x
List of Tables	xv
List of Acronyms	xvii
1 Introduction	1
1.1 Motivation	1
1.2 Objective measurement of image quality in CT	3
1.2.1 Spatial resolution	4
1.2.2 Image noise	5
1.2.3 Limitations of image quality measurement	6
1.3 Approaches to protocol-induced variation	7
1.3.1 Improving generalizability	8
1.3.2 Standardization	8
1.3.3 Stability-based feature selection	9
1.3.4 Correction and harmonization	9
1.4 Characterizing and predicting protocol-induced variation	11
1.5 Specific aims	12
1.6 Overview	14
2 An open-source CT texture phantom for radiomic imaging studies: design, manufacture, and initial evaluation	16
2.1 Introduction	16
2.2 Materials and methods	19
2.2.1 Phantom design and physical realization	20
2.2.1.1 Phantom body specification and manufacturing	20
2.2.1.2 Modular insert design and 3D printing	22
2.2.2 Evaluation of manufacturing and metrology	25
2.2.2.1 Phantom body spatial uniformity	26
2.2.2.2 Texture insert reproducibility	29
2.2.2.3 Evaluation of texture insert set for non-redundancy	31
2.2.2.4 Sensitivity to imaging protocol variation	31
2.3 Results	32
2.3.1 Phantom body spatial uniformity	32
2.3.2 Texture insert reproducibility	34

2.3.3	Evaluation of texture insert set for non-redundancy	37
2.3.4	Sensitivity to imaging protocol variation	38
2.4	Discussion	41
2.5	Conclusions	47
3	Developing a dataset for investigations into the impact of CT acquisition and reconstruction conditions on quantitative imaging	48
3.1	Introduction	48
3.2	Methods	50
3.2.1	Phantoms	50
3.2.1.1	Image quality phantom	50
3.2.1.2	Texture phantom	52
3.2.2	Image acquisition	53
3.2.3	Image quality measurements	57
3.2.4	Image quality metric calculation	58
3.2.5	Radiomic feature measurements	61
3.3	Results	63
3.3.1	Image quality measurements and derived metrics	63
3.3.2	Texture phantom image data	67
3.3.3	Radiomics data	69
3.4	Public release of the dataset	71
3.5	Discussion	73
3.6	Conclusions	77
4	Investigating the effect of CT protocol-induced variation on radiomic texture features using physical image quality metrics	79
4.1	Introduction	79
4.2	Methods	81
4.2.1	Dataset	81
4.2.2	Radiomic feature selection	81
4.2.3	Image quality metric selection	82
4.2.4	Effect of image quality on texture appearance	83
4.2.5	Radiomic feature dependence on image quality metrics	84
4.3	Results	84
4.3.1	Radiomic feature selection	84
4.3.2	Image quality metric selection	86
4.3.3	Effect of image quality on texture appearance	87
4.3.4	Radiomic feature dependence on image quality metrics	91
4.4	Discussion	93
4.5	Conclusions	95

5	Predictive modeling of radiomic feature variation due to CT protocol using physical image quality metrics	96
5.1	Introduction	96
5.2	Methods	98
5.2.1	Dataset	98
5.2.2	Modeling approach	98
5.2.3	Model evaluation	99
5.2.4	Generalization assessments	99
5.2.5	Image quality across vendors	101
5.3	Results	102
5.3.1	Overall model performance	102
5.3.2	Generalization across scanners and vendors	105
5.3.3	Range in image quality across vendors	106
5.4	Discussion	109
5.5	Conclusions	111
6	Predicting CNN generalization across CT imaging conditions using physical image quality metrics	112
6.1	Introduction	112
6.2	Methods	115
6.2.1	Representative task	117
6.2.2	Imaging data and patient cohort	118
6.2.3	Dataset construction	119
6.2.3.1	Multi-reconstruction image dataset	119
6.2.3.2	Image quality measurements	120
6.2.3.3	Patch extraction	120
6.2.3.4	Dataset splits	122
6.2.4	Measuring CNN performance across CT imaging conditions	123
6.2.4.1	CNN training and evaluation	123
6.2.4.2	Change in CNN performance across CT imaging conditions	125
6.2.5	Modeling CNN performance as a function of image quality	126
6.2.5.1	Candidate model forms	126
6.2.5.2	Model fitting and evaluation	128
6.2.6	External validation on Force scanner	129
6.3	Results	130
6.3.1	CNN performance on OOD imaging conditions	130
6.3.2	Modeling CNN performance as a function of image quality	133
6.3.3	External validation on Force scanner	136
6.4	Discussion and conclusions	138
7	Conclusions	142

7.1	Summary of work	142
7.2	Future directions	143
A	Additional phantom manufacturing details for Chapter 2	146
A.1	Materials and manufacturing	146
A.2	Texture definitions and print settings	148
A.3	Estimated production costs	149
B	Additional results for Chapter 2	152
B.1	Additional results for texture insert reproducibility	152
B.2	Additional results for sensitivity to imaging protocol variation	155
C	Table of radiomic features included in Chapter 3	156
D	Distributions of radiomic features included in Chapter 3	160
E	Additional results for Chapter 4	165
F	Additional results for Chapter 6	167
	References	169

List of Figures

1.1	Comparison between low resolution (left) and high resolution (right) CT images and their corresponding MTFs (below). Reproduced from [27].	5
1.2	Illustration of the relationship between image noise appearance (a) and the corresponding NPS curves (b). The curves in (b) describe the noise power spectra of the image regions in (a). Moving up the rows, increasing noise magnitude in image space corresponds to increasing NPS amplitude in frequency space; moving across the columns, finer noise texture in image space corresponds to a shift of noise power toward higher spatial frequencies. Reproduced from [27].	7
1.3	Overview of the relationship between the Aims.	14
2.1	Renderings (a) and schematics (b) of the 3D printed phantom body shell design. The shell is printed with a depth of 44.0 mm to allow for shrinkage of the epoxy resin to a target depth of 40 mm. Both show the alignment features, including the keyed design for the insert cavities, grooves for laser leveling and alignment, and labels for positioning.	20
2.2	Manufactured phantom body. Shows the alignment features, including the keyed design for the insert cavities, grooves for laser leveling and alignment, and labels for positioning.	21
2.3	Schematics of the 3D printed modular inserts. The measurements of the insert exterior are shown in (a), with the corresponding wireframe showing the interior of an example texture design in (b). Both show the keyed lid design that ensures reproducible positioning of the inserts within the phantom body.	23
2.4	Images of the eight texture patterns that were designed and manufactured. The full CT image is shown left, with the inserts labeled. An enlarged view of the inserts is shown at right. Images are shown with window width 1300 HU, window level -450 HU.	24
2.5	Placement of uniformity ROIs. Shown with window width 1300 HU, window level -450 HU.	27
2.6	Uniformity measurements (mean values of ROIs in HU) for the four tested phantom bodies. Red, solid line illustrates the overall mean value (104.9 HU) and the gray, dashed lines illustrate ± 5 HU from the mean.	34
2.7	The three sets of 3D printed modular inserts, imaged at 360 mA and reconstructed with the Br62f kernel. Visual agreement on CT is excellent, including the alignment and orientation of the texture inserts relative to the phantom body across the three experimental setups and acquisitions. Images shown with window width 1300 HU and level -450 HU.	34
2.8	Principal component analysis of radiomic feature values measured on the eight texture inserts. Images of the textures themselves with labels overlaid are used as markers.	37

2.9	Images of the eight texture objects (labeled above) under each of the six imaging conditions (labeled left). Rows are ordered with the highest noise condition at the bottom to the lowest noise condition at the top. Images shown with window width 1300 HU and level -450 HU.	39
2.10	Radiomic features values for three illustrative features (first order mean, first order variance, NGTDM strength) for each texture insert across the six imaging conditions.	40
2.11	Radiomic feature values of all features (grouped by feature class on the y -axis), shown for each texture object and imaging condition (x -axis), using the same imaging condition order as in Figure 2.10. Feature values were min-max normalized to $[0, 1]$. Color variation between texture groupings reflects differences between the texture objects, while within-group variation reflects sensitivity to imaging protocol.	41
3.1	Example axial and coronal slices of the Corgi [®] phantom, used to measure image quality. Shown with window width 650 HU, window level 175 HU.	51
3.2	The Corgi phantom, developed for the rigorous assessment of CT image quality. Reproduced from [88].	52
3.3	The 3D-printed texture phantom with the texture insert objects labeled 1-8, with the PLA body ring. Shown with window width 1300 HU, window level -450 HU.	53
3.4	Illustrations of a selection of extracted image quality metrics.	58
3.5	Texture phantom. ROIs and labels placed on the four texture inserts used in experiments. (A) circular rods, (B) random circles, (C) checkerboard, and (D) square grid.	61
3.6	Measured image quality data from the highest CTDI _{vol} protocols of the dataset (≈ 17 mGy). Each line corresponds to a single imaging condition. From top to bottom: axial MTF, oblique MTF, axial NPS. These curves illustrate the range in CT protocol represented in the dataset.	65
3.7	Distributions of a selection of the extracted IQMs across the 170 protocols.	66
3.8	Close-up view of four of the texture objects (labeled A, B, C, D) on CT from a random selection of 12 protocols from the highest dose level.	67
3.9	Close-up view of four texture objects (labeled A, B, C, D) on CT from the range of radiation dose levels acquired. The smooth kernel (a) shows less variation than the sharp kernel (b).	68
3.10	Distributions (KDEs) of a selection of radiomic features across all 170 protocols. Each row contains three features for a given feature family. For each subplot, x -axis is feature value and y -axis is kernel density estimate.	70
4.1	Dendrogram of the clustering of the radiomic features. The hierarchical clustering is performed using Ward's linkage. The representative features are bolded.	85
4.2	Dendrogram of the image quality metrics showing the clusters. The 10 selected IQMs are shown bolded.	86

4.3	Close-up view of four of the texture objects (labeled A, B, C, D) on CT from 12 randomly selected protocols (selected from the 170 total).	88
4.4	Close-up view of four of the texture objects (labeled A, B, C, D) on CT from a range of 12 protocols (selected from the 170 total). Images are shown sorted by $MTF_{45}f_{75}$ (top), and noise magnitude (bottom).	90
4.5	Scatter plots of the four selected radiomic features (subplot rows) vs example image quality metrics (subplot columns) to illustrate some of the relationships captured. Each point represents an individual CT protocol. Color represents the physical texture objects, with labeling corresponding to the ROI labeling in Figure 4.4.	92
4.6	Spearman rank correlations between each of the four representative radiomic features and the selected IQMs, computed separately for the four texture objects (rows A–D within each feature block).	93
5.1	The mean test set R^2 value for each radiomic feature, averaged over the four texture objects. Error bars correspond to standard deviation.	104
5.2	Principal component analysis of the 10 image quality metrics across all protocols. The first two components captured 73 % of the total variance (PC1: 51 %, PC2: 22 %). Marker color corresponds to the scanner model and marker shape to the reconstruction kernel.	106
5.3	Distributions (kernel density estimation) of each of the IQMs used for modeling for each manufacturer.	107
6.1	Overview of the experimental approach. The CNN performance prediction model is developed on $(\Delta IQM, \Delta \mathcal{M})$ measurements made from the AS64 dataset. The model is then tested on an external validation dataset from the Force scanner to assess its ability to generalize to an unseen scanner model (here, measured using mean square error (MSE)).	116
6.2	Illustration of the classification task using data from the LUNA16 challenge. (a) Shows patches containing lung nodules, (b) shows patches without true nodules. Modified from Setio et al. [110].	117
6.3	Dataset splits. Model development was done with the AS64 data (blue). The Force data (red) was used as an external test set.	123
6.4	Schematic for training and evaluating a CNN within the same domain, source condition p_s	125
6.5	Schematic for evaluating the performance of a CNN trained on the source condition p_s when predicting on the target domain p_t	126
6.6	A sample patch containing a lung nodule (centered) reconstructed with each of the eight kernels from the Definition AS 64 dataset. As kernel sharpness increases, structured noise common in CT imaging can be seen emerging as streaks through the image.	130

6.7	The CNN classification performance ($AUPRC_t$) for each train-test condition pairing for the Definition AS64 test set. The diagonal of corresponds to CNN performance on its source domain ($p_t = p_s$), while all non-diagonal elements are OOD target domains ($p_s \neq p_t$).	131
6.8	The change in CNN performance relative to the condition it was trained on ($\Delta AUPRC_{s,t}$) for each train-test condition pairing for the Definition AS64 test set. The top-left to bottom-right diagonal is 0, as it corresponds to $p_t = p_s$	132
6.9	The best fit curves for each of the three model forms using $MTF_{0f_{75}}$ as the predictor IQM. Scatter points correspond to measured (ΔIQM , $\Delta AUPRC$) for all Definition AS64 train-test protocol with 95% confidence intervals. The legend indicates the R_{adj}^2 for each best fit curve, as well as the RMSE of the test data.	134
6.10	Evaluation of the CNN model pooling approach. The best fit curve (quadratic, $MTF_{0f_{75}}$) with actual datapoints labeled by the protocol the CNN was trained on (p_s).	135
6.11	The best fit curves for each of the three model forms using $MTF_{0f_{75}}$ as the predictor IQM. Scatter points correspond to measured (ΔIQM , $\Delta AUPRC$) for all Definition AS64 train-test protocol with 95% confidence intervals. The legend indicates the R_{adj}^2 for each best fit curve, as well as the RMSE of the test (Force) data.	137
B.1	Distributions of the inter-print CV values measured for each texture insert. Bars correspond to the number of radiomic features at a given CV value. The two thresholds used in the evaluation are indicated with vertical lines.	152
B.2	Distributions of the inter-slice CV values measured for each texture insert. Bars correspond to the number of radiomic features at a given CV value. The two thresholds used in the evaluation are indicated with vertical lines.	153
B.3	Distributions of CCC values measured between radiomic feature sets for each CT protocol comparison (relative to the Br46f, 360 mA reference protocol). Bars correspond to the number of radiomic features at a given CCC value. The 0.9 threshold used in the evaluation is indicated with a vertical line.	155
D.1	Distributions for all first order features.	160
D.2	Distributions for all GLCM features.	161
D.3	Distributions for all GLRLM features.	162
D.4	Distributions for all GLSZM features.	163
D.5	Distributions for all NGTDM features.	163
D.6	Distributions for all GLDM features.	164
E.1	Heatmap of the radiomic feature value Spearman correlations. The four representative radiomic features are shown bolded.	165
E.2	Heatmap of the image quality metric Spearman correlations. The 10 selected IQMs are shown bolded.	166

F.1	The CNN classification performance (AUPRC) across all eight target conditions p_t for each of the eight trained CNNs, evaluated on the Definition AS64 dataset. Each subplot corresponds to a CNN trained on a different p_s ; the red dashed line indicates the performance of the CNN on the $p_t = p_s$ test set (corresponding to the diagonal of the heatmap in Figure 6.7). Error bars represent bootstrapped 95% confidence intervals.	167
-----	---	-----

List of Tables

2.1	The imaging protocols (tube current and reconstruction kernel) used for each experiment.	25
2.2	Mean value $\pm$ standard deviation of the HU number measurements for each phantom body.	33
2.3	Texture insert reproducibility results. Inter-print CVs were averaged across all slice depths, and inter-slice CVs were averaged across all insert set replicates. Results are reported as % of features meeting CV thresholds (count/total).	35
3.1	Summary of the scan protocols' acquisition and reconstruction conditions (all FBP).	55
3.2	The CT scanner details and image acquisition parameters used for each of the four systems used. These parameters were used consistently for all acquisitions for all phantoms.	56
3.3	Image quality metrics calculated from measurements made with the Corgi phantom.	60
3.4	Radiomics feature categories, descriptions, and number of features extracted using Pyradiomics [76].	62
4.1	Representative image quality metrics selected after hierarchical clustering of the full 32-metric IQM set. One representative was selected from each cluster as the metric with the highest variance across all 170 protocols.	87
5.1	Train-test splits for the generalization experiments, with the number of protocols n in each subset.	101
5.2	Full-dataset prediction performance by feature class. Models were fit per feature and texture object and evaluated on a held-out test set. Values were aggregated over all four texture objects and all features in the class. Mean test R^2 ($\pm$ standard deviation) is reported, along with median R^2 and the percent of models exceeding $R^2 = 0.6$ and $R^2 = 0.8$	102
5.3	Generalization performance across train-test configurations. Median test R^2 over all reproducible features and four texture objects, with the fraction of models exceeding $R^2 = 0.6$ and $R^2 = 0.8$	105
5.4	Coverage of each test manufacturer's protocols by the training manufacturer's measured image quality metric ranges. "Out of range (≥ 1 metric)" is the fraction of test protocols outside the training range on at least one metric; "mean metrics out of range" is the average number of metrics (of 10 total) on which a test protocol falls outside the training range.	108
6.1	Patient cohort and nodule distribution by scanner model.	118
6.2	Image acquisition and reconstruction conditions for each scanner model.	120
6.3	Patch counts and prevalence across dataset splits.	122

6.4	Predictive model performance for $\Delta\text{MTF}_{0f_{75}}$ as the input feature, evaluated on the development dataset (Definition AS scanner) and held-out external validation dataset (Force scanner). Results are shown for each candidate model form with the best performing model form for each dataset bolded.	136
A.1	Detailed geometric definitions and printing specifications for the eight modular texture inserts.	149
A.2	Itemized cost breakdown of the capital equipment and consumable materials required to manufacture the four phantom bodies and three complete insert sets evaluated in this study. Costs reflect prices at the time of purchase and exclude local taxes, shipping, and tariffs.	151
B.1	Radiomic features with either inter-print or inter-slice CV above the stricter 0.1 threshold for Textures 1-7. The CV is provided where > 0.1 . Texture 8 was excluded, as it was solid insert rather than a heterogeneous texture pattern, making the reproducibility of radiomic texture features a less meaningful evaluation of manufacturing consistency.	154
C.1	Radiomics features and categories extracted from Pyradiomics [76].	156
F.1	Complete report of model fit results by IQM and fit shape, sorted by the quadratic model's train R^2_{adj} . Train refers to the AS64 data used to fit the model, Force refers to the fit model's performance on the external test set from the Force scanner. . . .	168

List of Acronyms

2D	two-dimensional
3D	three-dimensional
AAPM	American Association of Physicists in Medicine
ACR	American College of Radiology
AI	artificial intelligence
AUC	area under the curve
CAD	computer-aided detection
CCC	concordance correlation coefficient
CI	confidence interval
CNN	convolutional neural network
CNR	contrast-to-noise ratio
CT	computed tomography
CTDI	computed tomography dose index
DL	deep learning

DQE	detective quantum efficiency
ESF	edge spread function
FBP	filtered back projection
FOV	field of view
GLCM	gray level co-occurrence matrix
GLDM	gray level dependence matrix
GLRLM	gray level run length matrix
GLSZM	gray level size zone matrix
HU	Hounsfield unit
ICC	intraclass correlation coefficient
IID	independent and identically distributed
IR	iterative reconstruction
LSF	line spread function
MAE	mean absolute error
ML	machine learning
MSE	mean squared error

MTF	modulation transfer function
NEQ	noise equivalent quanta
NGTDM	neighborhood gray tone difference matrix
NPS	noise power spectrum
OOD	out-of-distribution
PCA	principal component analysis
PSF	point spread function
RMSE	root mean square error
ROC	receiver operating characteristic
ROI	region of interest
SNR	signal-to-noise ratio
TTF	task transfer function

Acknowledgments

First and foremost, I would first like to thank my advisor and Committee Chair, Dr. Michael McNitt-Gray, for his continued mentorship and support throughout the course of my doctoral studies. Mike is extraordinarily generous with his time and expertise, and he has met each and every question, concern, and hesitation that I have brought him with genuine thought and honest engagement. Those many hours of conversation have taught me so much, and have strengthened my confidence in my own intuition, with regard to my science, my writing, and my communication. His patience and willingness to invest in others make him an exceptional mentor, and he is someone I will continue to look to as a model for the kind of scientist and colleague I hope to become. I am deeply grateful to have worked with him over the years.

Thank you to the other members of my dissertation committee for their time and valuable input: Dr. Matthew Brown, for introducing me to computer vision and for the opportunity to experience its full path to clinical translation, and for always being ready with a thoughtful question that brings us back the bigger picture; Dr. John Hoffman, for bringing the HD phantom into reality, and for greatly improving the quality of my computer programming; Dr. Grace Kim for excellent guidance on experimental design and statistical analysis, and for always meeting my updates with positivity and enthusiasm; Dr. Dan Ruan for bringing a rigorous quantitative perspective and insightful questions that pushed my work to be better; and Dr. Frédéric Noo for his remarkable generosity, repeatedly traveling from Salt Lake City to share his deep knowledge of CT physics and reconstruction, as well as attend my thesis proposal in person, and forwarding

me each fellowship and career opportunity that he finds. In addition to their feedback on the work itself, each of them went out of their way to support and encourage me over the years, for which I am deeply appreciative.

I would like to thank Dr. Andrew Hernandez and Dr. Ali Uneri for their support in the collection of the dataset that formed the basis of this dissertation. They have been wonderful collaborators over the past few years, and I am grateful for their insight and for the thoughtful feedback they gave on every draft I sent them.

I am truly grateful to all of my friends and colleagues I have met during my time in the Physics and Biology in Medicine Graduate Program and at the Center for Computer Vision and Imaging Biomarkers. I feel so fortunate to be a part of a community of people that is so welcoming and kind in addition to talented and dedicated. I would like to acknowledge all of my labmates for their support over the years, especially Wasil Wahi-Anwar and Nastaran Emaninejad for introducing me to everything CVIB in my first years, and to Youngwon Choi for imparting on me her deep learning expertise, which provided me with a far stronger foundation than I would have otherwise had, and for bringing with her so much joy and friendship. I also thank Rachel Petragallo, Andie Gong, Claudia Miller, Mike Lauria, Raymi Ramirez, Peter Boyle, John Hoffman, and Joshua Genender for their camaraderie and support over the years; I am grateful to have friends who understand both the demands of research and graduate student life, and how to step away and have fun.

I am also grateful to the faculty from the University of New Hampshire who provided the scaffolding that brought me to this point: Dr. Mark McConnell, who first introduced me to scientific research; Dr. Adam Boucher, whose teaching gave me the mathematical foundation to

pursue my goals; Dr. Jiadong Zang, who welcomed me into his lab and included me in the full process of research, from idea to publication, including my first coauthored papers; and the rest of the Department of Physics, which was the center of my undergraduate career.

Finally, thank you to my family and friends, who have loved and supported me endlessly through this journey. Graduate school has been a long road that has kept me three thousand miles away from home, but knowing they were there helped carry me through the difficulty and uncertainty. Most of all, I thank my parents Robert and Karen Daly, for providing everything that made it possible for me to be here today and for encouraging the independence to forge my own path. I love you.

I acknowledge that versions of the following manuscripts, either published, submitted, or in preparation, formed the basis of each chapter of this dissertation:

Chapter 2 is a version of M. A. Daly, M. F. McNitt-Gray, and J. M. Hoffman, “An open-source CT texture phantom for radiomic imaging studies: Design, manufacture, and initial evaluation,” *Medical Physics*, vol. 53, no. 8, e70613, 2026. DOI: [10.1002/mp.70613](https://doi.org/10.1002/mp.70613). MD was the primary author and contributed to the phantom design, data acquisition, and data analysis; MMG contributed to the data analysis and manuscript preparation; JH led design and manufacturing of the phantom and contributed to data acquisition, analysis, and manuscript preparation.

Chapter 3 is a version of M. A. Daly, J. M. Hoffman, A. M. Hernandez, A. Uneri, B. A. Varghese, J. Levy, and M. F. McNitt-Gray, “CT texture phantom dataset with paired image quality assessments for quantitative imaging,” *In press at Medical Physics*, 2026. DOI: [10.1002/mp.70633](https://doi.org/10.1002/mp.70633). MD was the primary author, contributed to the dataset acquisition and preparation,

and contributed to the phantom design; JH contributed to the dataset acquisition and preparation, and contributed to phantom design and manufacturing; AH contributed to dataset acquisition and preparation; AU performed image quality assessments; BV contributed to dataset acquisition and phantom design; JL contributed to phantom design and manufacturing; MMG contributed to dataset acquisition and provided guidance on CT protocols.

Chapter 5 is a version of M. A. Daly, J. M. Hoffman, A. M. Hernandez, A. Uneri, J. Levy, B. Varghese, and M. F. McNitt-Gray, "Predicting Radiomic Feature Variation Due to CT Protocol Using Measured Image Quality Metrics," in *2026 Joint AAPM/ COMP Annual Meeting & Exhibition*, Vancouver, BC: AAPM, 2026. MD was the primary author, performed modeling and analysis, contributed to the dataset acquisition and preparation, and contributed to the phantom design; JH contributed to the dataset acquisition and preparation, and contributed to phantom design and manufacturing; AH contributed to dataset acquisition and preparation; AU performed image quality assessments; BV contributed to dataset acquisition and phantom design; JL contributed to phantom design and manufacturing; MMG was PI.

Chapter 6 is a version of M. A. Daly, J. M. Hoffman, and M. F. McNitt-Gray, "Predicting CNN generalization across CT imaging conditions using physical image quality metrics," *In preparation*, 2026. MD was the primary author, performed modeling and analysis, and contributed to the dataset acquisition and preparation; JH contributed to the dataset acquisition and preparation, and contributed to code development; MMG was PI.

This work was supported in part by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) of the National Institute of Health (NIH) under award number T32EB002101-46,

as well as by the Edward J. Hoffman Fellowship.

Vita

Education

University of California, Los Angeles , Los Angeles, CA <i>M.S. Physics and Biology in Medicine</i>	2024
University of New Hampshire , Durham, NH <i>B.S. Physics, Minor in Applied Mathematics</i>	2018

Peer-reviewed publications

1. **M. A. Daly**, J. M. Hoffman, A. M. Hernandez, A. Uneri, B. A. Varghese, J. Levy, and M. F. McNitt-Gray, "CT texture phantom dataset with paired image quality assessments for quantitative imaging," *In press at Medical Physics*, 2026. DOI: [10.1002/mp.70633](https://doi.org/10.1002/mp.70633)
2. **M. A. Daly**, M. F. McNitt-Gray, and J. M. Hoffman, "An open-source CT texture phantom for radiomic imaging studies: Design, manufacture, and initial evaluation," *Medical Physics*, vol. 53, no. 8, e70613, 2026. DOI: [10.1002/mp.70613](https://doi.org/10.1002/mp.70613)
3. K.-P. Wong, S. Y. Homer, S. H. Wei, N. Yaghmai, O. A. E. Paz, T. J. Young, R. G. Buhr, I. Barjaktarevic, L. Shrestha, **M. Daly**, J. Goldin, D. R. Enzmann, and M. S. Brown, "Integration and evaluation of chest X-ray artificial intelligence in clinical practice," en, *Journal of Medical Imaging*, vol. 10, no. 5, p. 051805, Apr. 2023. DOI: [10.1117/1.JMI.10.5.051805](https://doi.org/10.1117/1.JMI.10.5.051805)
4. M. S. Brown, K.-P. Wong, L. Shrestha, M. Wahi-Anwar, **M. Daly**, G. Foster, F. Abtin, K. L. Ruchalski, J. G. Goldin, and D. Enzmann, "Automated Endotracheal Tube Placement Check Using Semantically Embedded Deep Neural Networks," *Academic Radiology*, vol. 30, no. 3, pp. 412–420, Mar. 2023. DOI: [10.1016/j.acra.2022.04.022](https://doi.org/10.1016/j.acra.2022.04.022)
5. M. McNitt-Gray, S. Napel, A. Jaggi, S. Mattonen, L. Hadjiiski, M. Muzi, D. Goldgof, Y. Balagurunathan, L. Pierce, P. Kinahan, E. Jones, A. Nguyen, A. Virkud, H. Chan, N. Emaminejad, M. Wahi-Anwar, **M. Daly**, M. Abdalah, H. Yang, L. Lu, W. Lv, A. Rahmim, A. Gastounioti, S. Pati, S. Bakas, D. Kontos, B. Zhao, J. Kalpathy-Cramer, and K. Farahani, "Standardization in Quantitative Imaging: A Multicenter Comparison of Radiomic Features from Different Software Packages on Digital Reference Objects and Patient Data Sets," *Tomography*, vol. 6, no. 2, pp. 118–128, Jun. 2020. DOI: [10.18383/j.tom.2019.00031](https://doi.org/10.18383/j.tom.2019.00031)

6. J.-X. Yu, **M. Daly**, and J. Zang, “Thermally driven topology in frustrated systems,” *Physical Review B*, vol. 99, no. 10, p. 104 431, Mar. 2019. DOI: [10.1103/PhysRevB.99.104431](https://doi.org/10.1103/PhysRevB.99.104431)
7. W.-T. Hou, J.-X. Yu, **M. Daly**, and J. Zang, “Thermally driven topology in chiral magnets,” *Physical Review B*, vol. 96, no. 14, p. 140 403, Oct. 2017. DOI: [10.1103/PhysRevB.96.140403](https://doi.org/10.1103/PhysRevB.96.140403)

Selected conference presentations

1. **M. A. Daly**, “Using Phantom Measurements of Image Quality to Predict Radiomic Feature Variability across Heterogeneous CT Protocols,” in *2026 Joint AAPM/ COMP Annual Meeting & Exhibition*, Vancouver, BC: AAPM (Invited Program), 2026
2. **M. A. Daly**, J. M. Hoffman, A. M. Hernandez, A. Uneri, J. Levy, B. Varghese, and M. F. McNitt-Gray, “Predicting Radiomic Feature Variation Due to CT Protocol Using Measured Image Quality Metrics,” in *2026 Joint AAPM/ COMP Annual Meeting & Exhibition*, Vancouver, BC: AAPM, 2026
3. **M. A. Daly**, J. M. Hoffman, A. M. Hernandez, A. Uneri, B. Varghese, J. Levy, D. J. Goodenough, and M. F. McNitt-Gray, “Developing a Dataset for Investigations into the Impact of CT Acquisition and Reconstruction Conditions on Quantitative Imaging Using Paired Image Quality and Radiomics Phantom Data: WE-1105-206-3,” *Medical Physics*, vol. 52, no. 10, e70059572–e70059573, 2025
4. J. M. Hoffman and **M. A. Daly**, “Feasibility of an open-standard, accessible, reproducible modular insert phantom for quantitative imaging,” in *CT Meeting 2024 Proceedings*, Bamberg, Germany, Aug. 2024, pp. 483–486
5. **M. Daly**, J. M. Hoffman, A. M. Hernandez, A. Uneri, J. H. Siewerdsen, P. E. Kinahan, N. B. Bevins, K. M. Kanal, D. A. Zamora, B. W. Maloney, J. A. Seibert, M. P. Supanich, M. Mahesh, J. M. Boone, and M. F. McNitt-Gray, “Harmonizing quantitative imaging feature values in CT using image quality metrics as a basis,” in *Medical Imaging 2024: Computer-Aided Diagnosis*, S. M. Astley and W. Chen, Eds., San Diego, United States: SPIE, Apr. 2024, p. 44. DOI: [10.1117/12.3006885](https://doi.org/10.1117/12.3006885)
6. **M. Daly**, M. W. Wahi-Anwar, G. Melendez-Corres, G. H. Kim, S. S. Raman, and M. F. McNitt-Gray, “Reducing dose in renal perfusion CT scans: The effects on quantitative imaging features for patients of different sizes,” in *Medical Imaging 2023: Physics of Medical Imaging*, R. Fahrig, J. M. Sabol, and L. Yu, Eds., San Diego, United States: SPIE, Apr. 2023, p. 78. DOI: [10.1117/12.2651479](https://doi.org/10.1117/12.2651479)

1 Introduction

1.1 Motivation

There is substantial interest in the development of quantitative imaging biomarkers, which have the potential to improve diagnostic precision, consistency, and standardization [1]. Radiomic features are quantitative descriptors of intensity, texture, and shape extracted from regions of interest within an image, and have shown potential as objective, reproducible measures of disease status and progression [2], [3], [4]. More recently, deep neural networks have emerged as a promising approach for image-derived quantitative imaging, with the potential to automate diagnostic tasks and capture image properties beyond what is accessible to either human observers or traditional computer vision approaches [5]. Of particular interest is computed tomography (CT) imaging, which reports X-ray attenuation on a calibrated scale relative to water (Hounsfield units), making it an inherently quantitative imaging modality [6].

Despite this promise, quantitative imaging biomarkers have translated slowly into clinical practice. Radiomic feature values have been shown to vary substantially across CT scanner models, reconstruction kernels, and dose levels [7], [8], [9], [10], [11], [12]. For the texture features that carry much of radiomics' discriminative promise, this protocol-induced variation can be comparable to or larger than the variation across a tumor population, such that an apparent difference between two patients may reflect the imaging protocol rather than the underlying

biology [13].

The same protocol sensitivity affects deep-learning-based quantitative imaging models. Although deep learning models often achieve strong performance on internal test data, this performance is rarely replicated on external datasets acquired with different protocols or at different institutions. Zech et al. reported that a pneumonia detection model dropped from an AUC of 0.93 on internal data to 0.82 on external data [14]; Wang et al. found that three published mammogram classification CNNs lost between 0.23 and 0.51 AUC across three external test sets [15]; and a systematic review by Yu et al. found that 24% of radiologic deep learning models with external validation reported a performance drop of at least 0.10 [16]. This phenomenon, known as *acquisition shift*, parallels the protocol sensitivity of handcrafted features: in both cases, the measurement depends on how the image was made, and that dependence undermines the biomarker's clinical use.

Acquisition shift is a particular form of dataset shift in which the distribution of image intensities changes because of differences in scanner hardware, acquisition settings, reconstruction algorithms, or patient positioning rather than differences in disease prevalence [17]. Unlike many sources of dataset shift, acquisition shift is often measurable and arises from known physical changes to the imaging system.

Standardizing CT protocols to address this problem is not fully achievable. While some efforts have been made to develop carefully standardized CT imaging protocols for clinical trials, such as the National Lung Screening Trial (NLST) [18], and the Subpopulations and Intermediate Outcome Measures in COPD Study (SPIROMICS) [19], protocols are generally the product of

vendor suggestion, local radiologists' familiarity with the clinical literature or personal preferences on image appearance [20]. This large degree of variation observed in clinical practice was reported as part of an analysis of the data collected by the Medical Imaging and Data Resource Center (MIDRC) [21], [22]. The Quantitative Imaging Biomarkers Alliance publishes protocol recommendations for this purpose, but have thus far been limited to volumetry applications [23]. Compounding this, scanner platforms differ fundamentally in detector design, beam filtration, and especially reconstruction kernels and algorithms, the definitions of which are largely proprietary, so there does not exist a straightforward way to select a protocol to match the image characteristics of another.

Existing work has largely sought to *reduce* a biomarker's sensitivity to protocol variations by using methods such as retrospective harmonization, image-domain correction, or stability-based feature selection. In this work, rather than reduce protocol sensitivity, we seek to *characterize and predict* CT protocol-induced variation in quantitative imaging using physical, scanner-independent measurements of image quality as a basis

1.2 Objective measurement of image quality in CT

Protocol parameters such as tube current, reconstruction kernel, and slice thickness influence image appearance through their effects on a relatively small number of fundamental image quality characteristics. Rather than describing protocols by vendor-specific acquisition and reconstruction settings, these physical characteristics provide a common, scanner-independent language for

describing the resulting images.

Objective image quality measurements provide a way to quantify these characteristics. In CT, image quality is primarily characterized in terms of spatial resolution, noise, and contrast, which together determine the visibility of anatomical structures and lesions.

1.2.1 Spatial resolution

Spatial resolution describes the ability of a CT system to distinguish small or closely spaced objects, and is determined by factors including the detector characteristics, focal spot size, image reconstruction, and pixel size [24]. Higher spatial resolution preserves finer anatomical detail, whereas lower spatial resolution results in increased blurring and loss of small structures.

The modulation transfer function (MTF) provides a quantitative description of spatial resolution [25], [26]. Defined as the magnitude of the Fourier transform of the point spread function, the MTF describes how object contrast is transferred as a function of spatial frequency.

Reconstruction kernels directly influence the MTF: smoother kernels attenuate high spatial frequencies, producing blurrier images with lower spatial resolution, whereas sharper kernels preserve higher spatial frequencies, improving the visibility of fine detail at the expense of increased image noise. The relationship between image appearance and the MTF is illustrated in Figure 1.1.

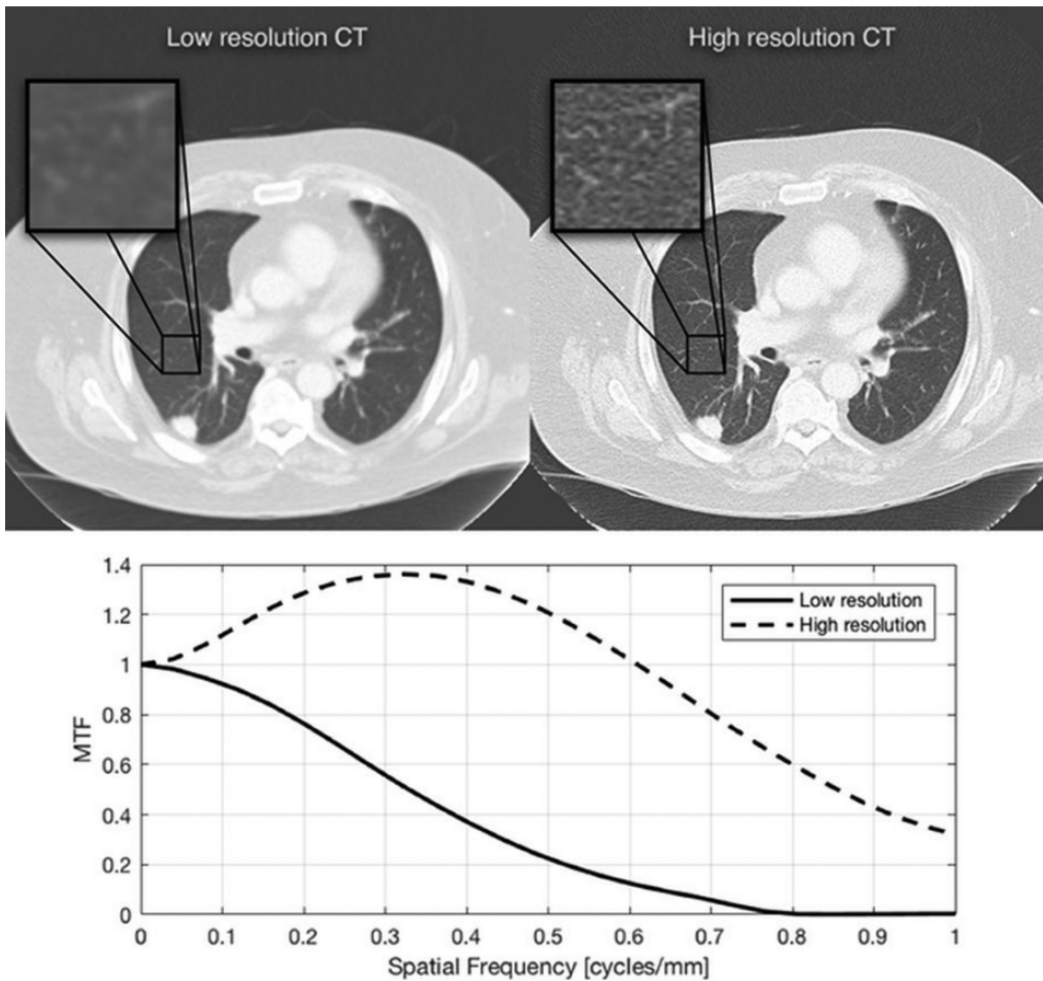


Figure 1.1: Comparison between low resolution (left) and high resolution (right) CT images and their corresponding MTFs (below). Reproduced from [27].

1.2.2 Image noise

Image noise in CT arises primarily from the finite number of detected X-ray photons, making its magnitude strongly dependent on the radiation dose. The image reconstruction process introduces spatial correlations in the noise, so neighboring pixels exhibit similar fluctuations rather than independent random variations, producing the characteristic textured appearance of CT noise [24].

Noise is commonly summarized as the standard deviation of pixel values within a uniform region, but this scalar describes only the magnitude of the fluctuations and not their spatial structure. Two images with the same pixel standard deviation can differ markedly in appearance depending on how the noise power is distributed across spatial frequencies. The noise power spectrum (NPS) captures this distribution, characterizing both the magnitude of the noise, through its integral, and its texture, through its shape.

The NPS is defined as the Fourier transform of the spatial autocovariance of image noise, and describes how noise variance is distributed across spatial frequencies [28]. Reconstruction kernels act directly on the NPS: smoother kernels suppress high-frequency noise and shift noise power toward lower spatial frequencies, producing coarser noise, while sharper kernels retain high-frequency power and produce finer-grained noise. The relationship between noise appearance and the NPS is illustrated in Figure 1.2.

1.2.3 Limitations of image quality measurement

The image quality measurements described above are well-posed for linear, shift-invariant systems. They are well-established and applicable for conventional filtered backprojection reconstruction, which is analytical. In this case, the MTF is independent of the object being imaged, so a single measured MTF characterizes the response to any input.

The introduction of iterative reconstruction, and more recently deep-learning reconstruction, complicates their use, because these methods are strongly nonlinear. In this nonlinear case, a

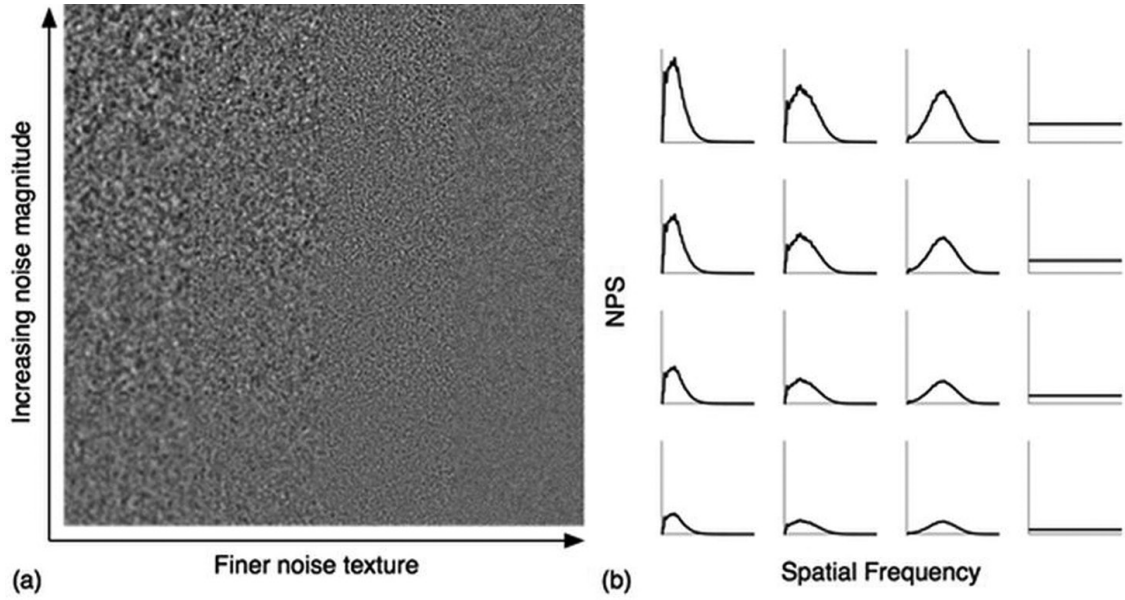


Figure 1.2: Illustration of the relationship between image noise appearance (a) and the corresponding NPS curves (b). The curves in (b) describe the noise power spectra of the image regions in (a). Moving up the rows, increasing noise magnitude in image space corresponds to increasing NPS amplitude in frequency space; moving across the columns, finer noise texture in image space corresponds to a shift of noise power toward higher spatial frequencies. Reproduced from [27].

measured MTF no longer represents the response to an arbitrary input object, and the system resolution becomes dependent on object contrast and background noise level. The NPS likewise becomes dose- and contrast-dependent [29]. Task-based metrics have been proposed that partially address this, but no single measurement fully characterizes a nonlinear system [30].

1.3 Approaches to protocol-induced variation

There has been substantial work in response to the protocol-induced variation described in Section 1.1, which we termed acquisition shift. Most of these approaches aim to reduce a biomarker's sensitivity to imaging conditions: improving model generalizability across domains, standardizing

protocols at their source, selecting features that are stable across protocols, and correcting feature values after acquisition.

1.3.1 Improving generalizability

This problem has motivated substantial research on out-of-distribution (OOD) generalization, or domain generalization [31], [32]. Given access only to data from the available domains during training, the goal of OOD generalization is to maintain model performance across a broader set of domains, including unseen ones [33], most commonly by encouraging models to learn feature representations that are invariant across domains [34], [35].

These methods can extend the range of conditions a model tolerates, but they do not guarantee performance on an arbitrary unseen protocol. For the goal of developing clinically useful machine learning applications, generalizability has been criticized as overemphasized compared to improved understanding of whether a system will be useful in a given, specific case [36].

1.3.2 Standardization

As discussed in Section 1.1, standardization is not widely used in clinical practice outside of specific clinical trials, and consensus recommendations. As protocols are not directly comparable across vendors, developing these recommendations is not straightforward, and even a “within-standard” protocol on one scanner does not reproduce the image characteristics of another exactly. While standardization is an important approach to mitigate acquisition shift at the source, it has thus

far proven challenging to implement widely.

1.3.3 Stability-based feature selection

A common response to protocol-induced variation in radiomics studies is to retain only those radiomic features that are reproducible under the conditions of interest, discarding the rest before any modeling. The premise is that a biomarker built from features insensitive to protocol will itself be robust to protocol.

However, that a feature is sensitive to protocol does not imply that the feature is not informative or discriminative. Filtering on stability therefore may discard information relevant to the prediction task. Additionally, no feature is robust without bound. Protocol variation can always be taken to a regime in which any feature degrades; at the limit of vanishing dose, for instance, the image is dominated by noise and no feature can remain stable, as there is no longer a signal to measure. Robustness is thus not an intrinsic property of a feature but a statement about the range of conditions over which it holds.

This motivates characterizing how features respond to image quality, rather than filtering on whether they respond at all.

1.3.4 Correction and harmonization

Alternatively, normalization and harmonization techniques can be applied retrospectively. ComBat and its derivatives have provided an approach to harmonize feature values across a cohort of

patients scanned with different protocols[37], [38], [39]. However, it is not directly applicable to standardizing feature values for individual patients scanned across different settings.

Deep learning techniques have been proposed to convert images between different reconstruction kernels or to synthetically generate images with different technical parameters [40], [41]. While these approaches have shown promise in reducing feature variability, they suffer from a lack of interpretability and artifacts that may not be appropriate for all clinical applications (e.g., clinical trials). Furthermore, the generalization of these models to unseen protocols and scanner types remains a challenge [14], [15].

Physical metrics such as MTF and NPS have been proposed as a basis for harmonization through prospective protocol selection as a way of reducing variation in radiomic features across scanner models [42], [43]. These metrics have also been proposed as a basis for analytically predicting and correcting protocol-induced variation in radiomic features. In this framework, MTF and NPS serve as quantitative inputs that characterize the resolution and noise state of each imaging condition, enabling feature-level corrections that are grounded in the underlying physics rather than dependent on categorical protocol labels [44], [45].

1.4 Characterizing and predicting protocol-induced variation

The approaches of Section 1.3 all aim to *reduce* protocol sensitivity. A prior question, however, is whether that sensitivity can be *characterized* at all, in terms general enough to predict how a given biomarker will behave under a protocol it has not yet encountered.

A small number of studies have begun to relate radiomic feature behavior to physical image quality metrics rather than categorical protocol labels. Zarei et al. builds regression models relating radiomic feature stability to MTF- and NPS-derived metrics across simulated imaging conditions, and used these models prospectively to recommend protocol settings that minimize feature variation in lung nodule volumetry [46]. This direction is encouraging: it establishes that radiomic feature behavior is structured enough in image-quality space to support quantitative modeling, and that physical image quality metrics provide a more general basis for that modeling than categorical protocol descriptors.

However, this approach depends on a ground truth reference. While this framing can be applied to morphological features, it does not extend to texture features. A simulated nodule has a true volume, surface area, and sphericity computable from its geometry, independent of imaging; texture features are a property of the realized image and thus do not have a well-posed ground truth value. Additionally, as morphological features are computed from a segmentation mask, their protocol sensitivity acts through the segmentation boundary. Thus, any impact of protocol on morphological features is indirect, as it is mediated by the segmentation approach

(whether manual or automated).

1.5 Specific aims

To address these limitations, we propose an approach for predicting the variability of QIF values across different CT protocols based on measurable image quality metrics. By quantifying the relationship between protocol differences and QIF robustness in terms of fundamental image quality characteristics, rather than with categorical protocol labels, we aim to develop generalizable models that can predict the expected performance of imaging biomarkers across a wide range of CT acquisition and reconstruction conditions. This physics-based approach leverages the inherent link between image quality metrics and the perceptual and quantitative characteristics of the resulting images.

By quantifying the relationship between protocol differences and imaging biomarker robustness in terms of fundamental image quality characteristics, such as spatial resolution, noise, and contrast, we aim to develop generalizable models that can predict the expected performance of these tools across a wide range of CT acquisition and reconstruction conditions.

We propose three specific aims:

Specific Aim 1: Characterize the relationships between radiomic texture feature values and measured image quality metrics across CT protocols.

Specific Aim 2: Predict the variation in radiomic texture feature values between CT protocols from measured image quality metrics.

Specific Aim 3: Predict the variation in CNN classification performance between CT protocols from measured image quality metrics.

The significance of this work lies in the potential to facilitate the clinical adoption of quantitative imaging by providing a means to ensure reliable performance in the face of real-world heterogeneity in CT protocols. The successful development of such models would enable the definition of “imaging inclusion criteria” to identify the range of CT protocols over which a given imaging biomarker can be reliably applied, thereby facilitating its translation into clinical practice. Moreover, this range of CT protocols would be agnostic to scanner model or manufacturer, as it would be defined in terms of image quality features measurable with any CT scanner. This would be a significant step towards addressing the challenge of silent failure of quantitative imaging models on out-of-distribution clinical data.

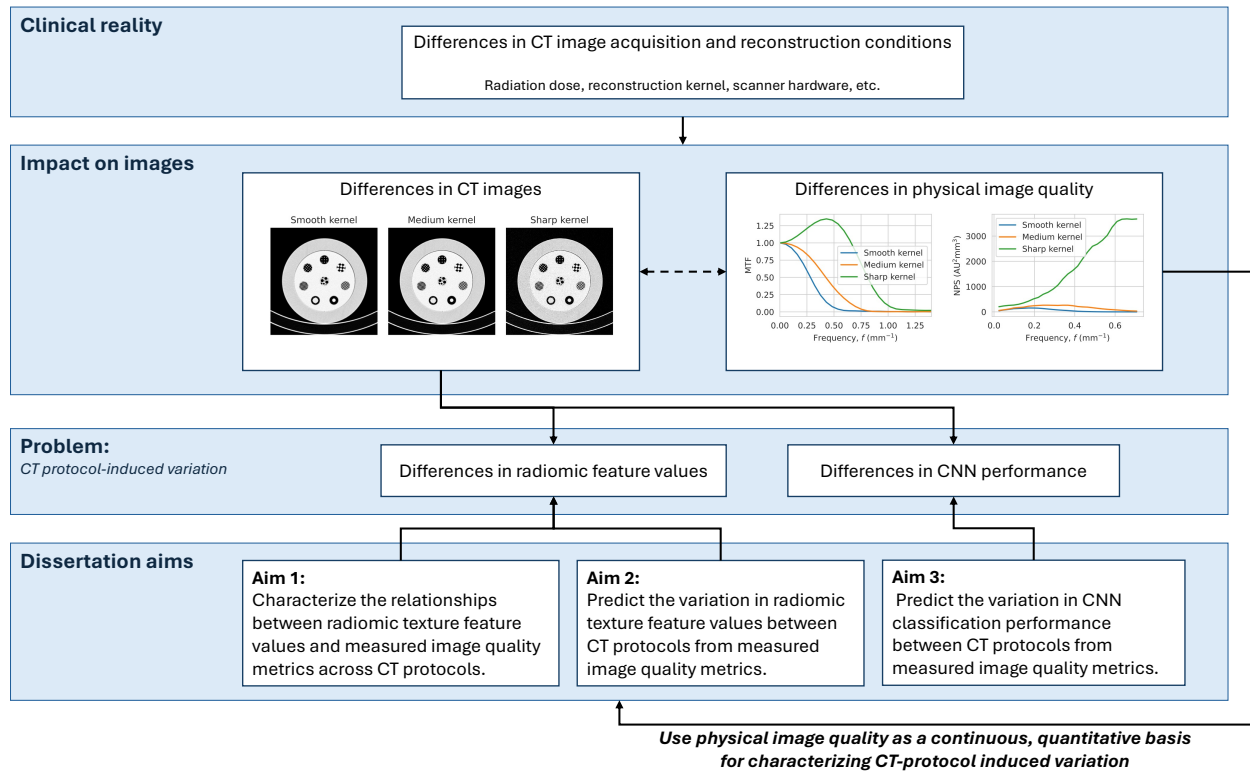


Figure 1.3: Overview of the relationship between the Aims.

1.6 Overview

Specific Aim 1 is addressed in Chapters 2 to 4. Chapter 2 describes the design, manufacture, and initial evaluation of an open-source CT texture phantom developed for radiomics research, with 3D-printed modular inserts producing distinct radiomic signatures. Chapter 3 describes a unique dataset of a CT texture phantom with paired image quality assessments for quantitative imaging. This dataset was generated by scanning the texture phantom described in Chapter 2 together with an image quality phantom across a diverse protocol space, measuring 3D image quality metrics (MTF, NPS) and extracting texture features from four texture objects under identical conditions.

This paired texture phantom and image quality (PTP-IQ) dataset enables direct investigation of the quantitative relationship between fundamental image quality metrics and image textures. The systematic variation of acquisition parameters across multiple scanner platforms creates a controlled experimental framework for addressing key challenges in reproducibility and robustness for quantitative imaging. Chapter 4 uses the data provided in Chapter 3 to perform the desired investigations into characterizing the resulting relationships, establishing that texture features respond systematically to fundamental image quality characteristics.

Specific Aim 2 is addressed by Chapter 5, which develops and evaluates regression models to predict texture feature values from physical image quality metrics. Using polynomial regression with ElasticNet regularization, we assess prediction accuracy and evaluate whether the models generalize across vendors and to unseen scanner platforms.

Specific Aim 3 is addressed in Chapter 6, which extends the framework from radiomic features to a deep learning model, investigating whether physical image quality metrics can predict out-of-distribution performance degradation of a convolutional neural network for lung nodule classification. Using clinical lung cancer screening CT reconstructed with multiple kernels, we model CNN performance change across imaging conditions as a function of measured image quality.

2 An open-source CT texture phantom for radiomic imaging studies: design, manufacture, and initial evaluation

This Chapter is based on the following manuscript: M. A. Daly, M. F. McNitt-Gray, and J. M. Hoffman, “An open-source CT texture phantom for radiomic imaging studies: Design, manufacture, and initial evaluation,” *Medical Physics*, vol. 53, no. 8, e70613, 2026. DOI: [10.1002/mp.70613](https://doi.org/10.1002/mp.70613).

2.1 Introduction

Quantitative imaging in computed tomography (CT) has produced promising results in medical imaging research, with applications ranging from tumor characterization to treatment response assessment. The field of radiomics, which involves the high-throughput extraction of quantitative features from regions of interest in medical images, has demonstrated potential to improve diagnostic accuracy and clinical decision making [2], [3], [4]. Texture-based radiomic features in particular, which capture spatial patterns of intensity variation within a region of interest, have shown utility across a range of clinical applications including the CT image-based diagnoses and prognoses in lung nodules [47], [48], [49], interstitial lung disease [50], [51], and bone [52],

[53], [54]. By converting medical images into mineable data, radiomics may enable the discovery of imaging biomarkers that reflect underlying pathophysiology not apparent to visual inspection alone.

However, the clinical translation of radiomic biomarkers has been impeded by challenges in reproducibility and standardization. Multiple studies have documented substantial variability in radiomic features due to differences in CT acquisition parameters, reconstruction algorithms, and scanner characteristics [7], [11], [55], [56]. This variability undermines the generalizability of radiomic models and limits their utility in multi-institutional studies and general clinical practice. The need for improved standardization approaches in efforts to develop imaging biomarkers has been highlighted by the Quantitative Imaging Biomarkers Alliance (QIBA) and the Image Biomarker Standardisation Initiative (IBSI) [57], [58], [59].

One barrier to the advancement of quantitative imaging is the lack of accessible standardized phantoms for validating radiomic measurements. Conventional CT phantoms, such as the American College of Radiology (ACR) CT accreditation phantom, serve important roles in routine quality assurance; however, they typically consist of uniform materials that are not suitable for evaluating texture-based radiomic features [60], [61], [62]. This limitation has led researchers to develop custom “handcrafted” phantoms consisting of cartridges filled with different materials to produce textures, such as the Credence Cartridge Radiomics (CCR) Phantom [13], [63], [64]. Phantoms that use filled cartridges as texture objects are inherently one-of-a-kind, as they cannot be manufactured reproducibly. This makes it challenging to directly compare results from different CT scanners, technical parameter settings, or institutions because differences could be due to

the physical texture objects, rather than the scanner or parameters, etc. Furthermore, extending existing work becomes essentially impossible without acquiring the original custom-built phantom used.

Recent advances in three-dimensional (3D) printing technology have made it a promising solution for producing custom phantoms for CT physics experiments. This approach enables the physical realization of computer-designed textures and shapes, providing a pathway to a substantially more reproducible manufacturing approach than traditional methods, such as filled cartridges. A 3D printed infill has been included as a subset of cartridges in some CT radiomics cartridge phantoms [13], [65], [66]. Other approaches have focused on directly printing entire phantom bodies or employed sophisticated approaches to print anatomical phantoms [67], [68], [69], [70]. However, these approaches often require professional-grade equipment, materials, and expertise that may not be readily available to many research groups. More importantly, these approaches were not developed with widespread reproducibility as a goal, and the designs have not been made publicly available.

The lack of a shared, reproducible phantom represents a substantial gap in the radiomics workflow. Without common reference objects, it becomes challenging to determine whether observed variations in radiomic measurements arise from biological differences, technical factors, or phantom inconsistencies.

The lack of a shared, reproducible phantom represents a substantial gap in the radiomics workflow. Without common reference objects, it becomes challenging to determine whether observed variations in radiomic measurements arise from biological differences, technical factors,

or phantom inconsistencies. To address this gap, we first present the design and manufacturing approach for an open-source, modular CT phantom body and set of eight texture inserts. We then provide an initial evaluation of the reproducibility of four manufactured phantom bodies based on measurements of uniformity, as well as the reproducibility across three replicate sets of inserts based on measurements of radiomic features. Finally, we characterize the phantom's utility as a platform for radiomic imaging studies by evaluating that the set of inserts provide non-redundant texture stimuli based on measured radiomic feature values, and evaluating the sensitivity of the inserts to CT imaging protocol variation.

2.2 Materials and methods

Section 2.2.1 presents the design and fabrication of the phantom system, comprising an epoxy-filled cylindrical body and a set of eight modular 3D-printed texture inserts. Section 2.2.2 then describes the evaluations performed to assess the reproducibility of the phantom manufacturing approach, as well as the phantom's utility as a platform for radiomic imaging studies. Additional manufacturing instructions and methods are provided in Appendices A.1 and A.2. The digital phantom design components (STL files) have been made publicly available at <https://doi.org/10.5281/zenodo.19340627>.

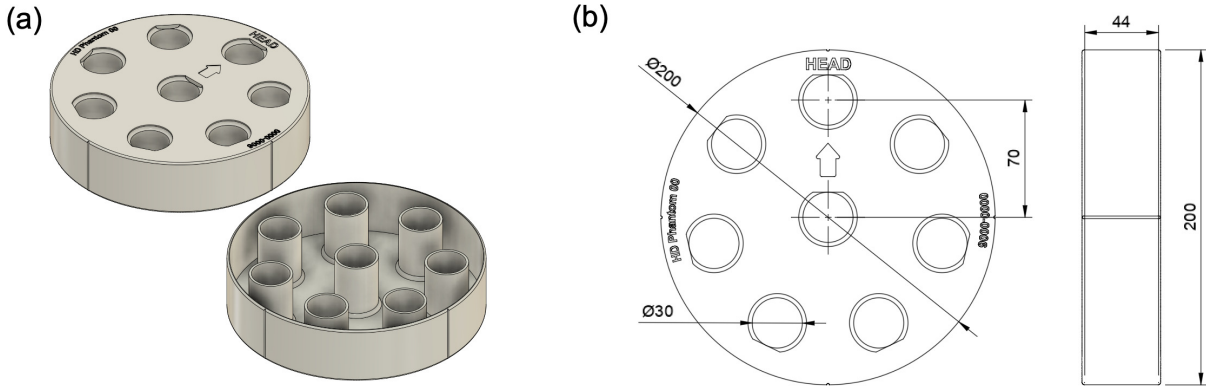


Figure 2.1: Renderings (a) and schematics (b) of the 3D printed phantom body shell design. The shell is printed with a depth of 44.0 mm to allow for shrinkage of the epoxy resin to a target depth of 40 mm. Both show the alignment features, including the keyed design for the insert cavities, grooves for laser leveling and alignment, and labels for positioning.

2.2.1 Phantom design and physical realization

2.2.1.1 Phantom body specification and manufacturing

The phantom body consists of a hollow, cylindrical 3D-printed shell filled with a liquid casting resin (200 mm diameter $\times$ 40 mm depth, shown in Figure 2.1). It features eight cylindrical cavities (30.0 mm diameter) designed to accept the modular texture inserts: seven arranged peripherally on a 70.0 mm radius circle, and one located centrally. The geometric configuration was optimized to minimize beam hardening and photon starvation artifacts by ensuring that no more than two (potentially high-density) inserts align along any single ray path when the phantom is positioned at the scanner isocenter. The phantom body dimensions were scaled to match standard commercial modules (e.g., Gammex 464, Sun Nuclear, Melbourne, FL), ensuring cross-compatibility with existing clinical positioning equipment and scatter-inducing body rings.

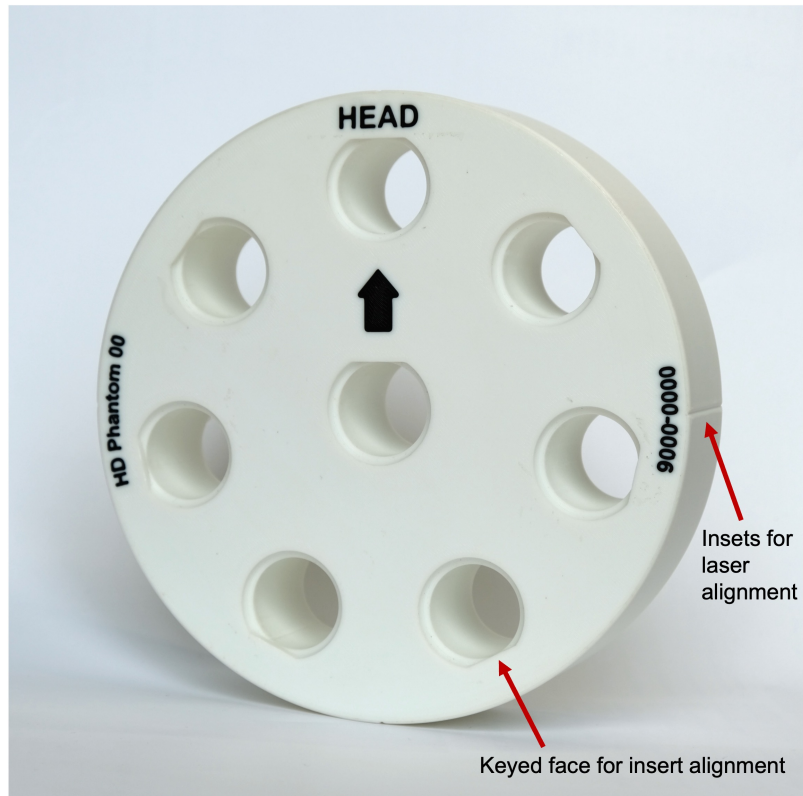


Figure 2.2: Manufactured phantom body. Shows the alignment features, including the keyed design for the insert cavities, grooves for laser leveling and alignment, and labels for positioning.

A photo of a completed phantom body is shown in Figure 2.2.

The specific properties of phantoms produced using this method will depend on the materials used for construction. PLA has been characterized as having CT attenuation values broadly similar to soft tissue, with reported HU values ranging from 115 to 200 HU when printed with 100% infill, varying linearly with density [70], [71], [72]. Prior work has found that the attenuation properties of PLA are stable over time and do not require special storage conditions, suggesting the inserts are suitable for longitudinal use [71]. A characterization of clear epoxy resin for phantom applications reported measurements of density as 1.17 g/cm^3 and an average CT number of

116±6.7 HU, although this was a 3D printed resin rather than a two-part liquid resin [73]. These materials are not expected to be water equivalent.

This hollow-shell-and-cast approach reduces total print time, mitigates thermal warping risk associated with large-volume fused deposition modeling (FDM) prints, and yields a more uniform attenuating background than 3D-printed infill; full material specifications and casting procedures are provided in Appendix A.1.

Several phantom and insert setup constraints were integrated directly into the physical design to maximize the reproducibility of the phantom alignment and positioning across imaging sessions. External insets (or “grooves”) for horizontal and vertical alignment with the CT scanner’s external laser system were included to facilitate consistent phantom placement at isocenter. The phantom surface features a directional arrow and the text “HEAD” to enforce consistent superior-inferior orientation and scan direction. Importantly, the cavities for the modular inserts feature a flat, mechanically keyed face that permits only a single, predetermined orientation for each test object. This eliminates rotational variability during physical teardown and re-assembly, isolating sources of texture variance to desired experimental variables.

2.2.1.2 Modular insert design and 3D printing

The modular inserts’ exterior dimensions were designed to match the phantom body, including a “lid” (a 2 mm protrusion at the base of the insert, shown in Figure 2.3) to ensure consistent insertion depth, and a keyed face in the lid to ensure reproducible angular orientation across experimental setups. Test objects of interest (i.e., 3D-printable shapes and textures) were then

modeled within the cylindrical insert body. A schematic is shown in Figure 2.3.

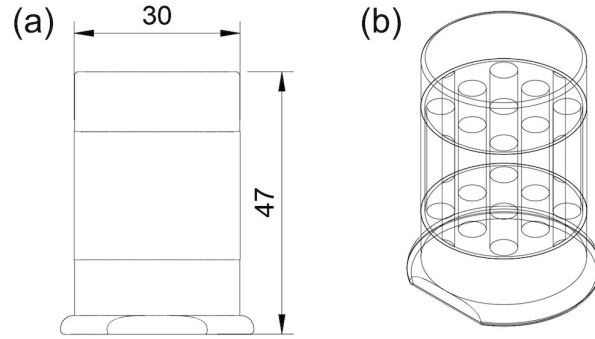


Figure 2.3: Schematics of the 3D printed modular inserts. The measurements of the insert exterior are shown in (a), with the corresponding wireframe showing the interior of an example texture design in (b). Both show the keyed lid design that ensures reproducible positioning of the inserts within the phantom body.

The eight texture inserts were designed as binary density distributions (utilizing 3D printed material and air), shown in Figure 2.4. All textures were extruded uniformly in the z -direction to eliminate partial-volume effects along the longitudinal length of the insert, ensuring that minor translational offsets in phantom setup or slice selection during ROI placement do not introduce slice-to-slice morphological variations during the reproducibility assessment.

Six inserts (Figure 2.4, Textures 2–7) were designed as mathematically defined geometric primitives: periodic rectangular and circular grid lattices of varying pitch (2.5–8.0 mm) and feature size (0.75–7.0 mm). Designs were intended to provide some range in texture appearance, such as scale, ratio of 3D printed material to air, and rounded or squared shapes. These regular structures were selected because their parameterization provides both a known digital ground truth and flexibility for future experiments. By defining the geometries through discrete variables

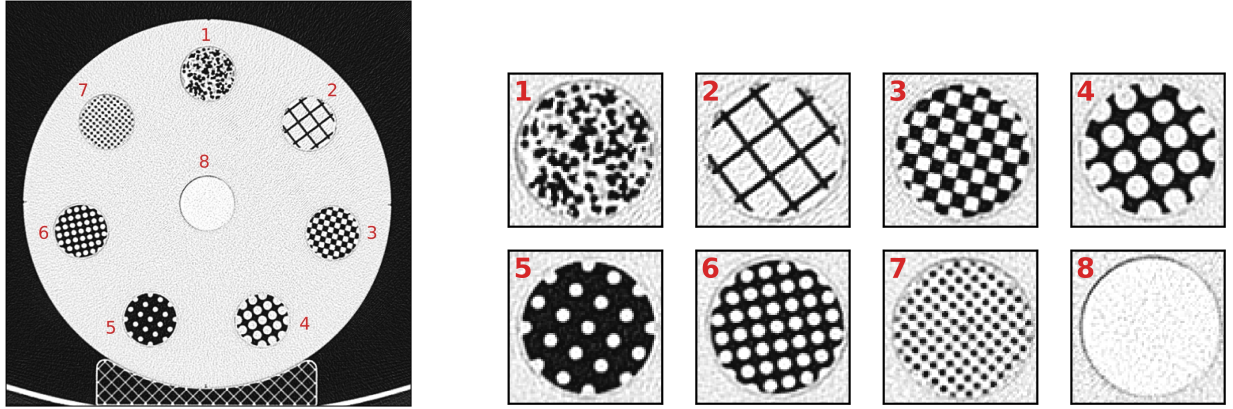


Figure 2.4: Images of the eight texture patterns that were designed and manufactured. The full CT image is shown left, with the inserts labeled. An enlarged view of the inserts is shown at right. Images are shown with window width 1300 HU, window level -450 HU.

(e.g., rod diameter, lattice spacing), the scale and spatial frequency of the textures can be easily adjusted as an additional experimental variable.

One insert was designed with a stochastic distribution of cylindrical elements with diameters ranging from 0.5 mm to 2.0 mm (Figure 2.4, Texture 1). Unlike handcrafted approaches to producing stochastic textures (e.g., granular fill materials), this insert retains a deterministic digital model, providing a reproducible and shareable ground truth despite its randomized appearance. This serves as a proof-of-concept for the platform’s extensibility toward more clinically realistic texture models. Finally, a solid 100% infill insert was included as a reference with no underlying structural heterogeneity (Figure 2.4, Texture 8).

All inserts were 3D printed using PLA, with air as the texture background. Definitions and print settings for all inserts are provided in Appendix A.2.

2.2.2 Evaluation of manufacturing and metrology

The following evaluations were performed to assess the reproducibility of the phantom manufacturing approach (Sections 2.2.2.1 and 2.2.2.2), as well as the phantom's utility as a platform for radiomic imaging studies (Sections 2.2.2.3 and 2.2.2.4). Four phantom bodies and three sets of modular inserts were manufactured as described in Section 2.2.1. Section 2.2.2.1 evaluates that the casting process produces a reproducible, homogeneous attenuating background by characterizing the spatial uniformity across the four epoxy phantom bodies. Section 2.2.2.2 evaluates the reproducibility of the 3D-printed texture inserts across three independent manufacturing cycles based on measurements of radiomic features. Section 2.2.2.3 assess that the eight inserts provide non-redundant stimuli spanning a range of radiomic signatures by characterizing their radiomic feature space coverage via principal component analysis (PCA). Section 2.2.2.4 investigates the ability of the platform to detect and express protocol-induced radiomic variations it is intended to measure by comparing radiomic feature values across different acquisition and reconstruction protocols.

Table 2.1: The imaging protocols (tube current and reconstruction kernel) used for each experiment.

Experiment	Tube current [mA]	Reconstruction kernel
Phantom body spatial uniformity (Section 2.2.2.1)	360	Br46f
Texture insert reproducibility (Section 2.2.2.2)	360	Br62f
Texture insert separability (Section 2.2.2.3)	360	Br62f
Sensitivity to imaging protocol (Section 2.2.2.4)	60, 360	Br32f, Br46f, Br62f

All images were acquired in helical mode with a Siemens Definition AS 64 (Siemens Health-

ineers, Forchheim, Germany) using a tube voltage of 120 kV, fixed tube current, and a pitch of 1.0. All images were reconstructed at a slice thickness of 1.0 mm, a reconstruction diameter of 219 mm, and a matrix size of 512×512 . The tube current and reconstruction kernel used are specified in Table 2.1. Doses were selected to be consistent with our institution's routine chest protocol ($\approx 10\text{--}15$ mGy) and a low-dose lung screening protocol (≈ 2 mGy).

2.2.2.1 Phantom body spatial uniformity

The spatial uniformity of the epoxy phantom bodies was measured to assess whether the casting process produced a homogeneous, reproducible attenuating background. The four phantom bodies were scanned using a tube current of 360 mA (CTDI_{vol} of 13.68 mGy) and images were reconstructed using a Siemens Br46f kernel. Insert holes were left unfilled; considering the modest phantom size (20 cm) and robust scatter rejection mechanisms on modern diagnostic CT scanners (anti-scatter grids, software correction), scatter-related effects on uniformity measurements were assumed to be negligible. Uniformity within each phantom body (intra-phantom uniformity) was evaluated through HU measured from different regions of interest (ROIs). Two sets of seven ROIs were placed at 40 mm and 80 mm radial distances to capture potential differences between the interior and periphery positions within the phantom. Three slices were measured within each phantom to capture potential axial heterogeneity in the z -direction. The measurements were then also used to assess the uniformity across four manufactured phantom bodies (inter-phantom uniformity).

The acceptance criteria was based on the ACR CT accreditation guidelines for uniformity,

which uses a maximum absolute deviation of ± 5 HU (and not exceeding ± 7 HU) between ROIs in a homogeneous phantom [74]. Though ACR uniformity only assesses differences between center and each periphery position in an assumed uniform object, we apply this ± 5 HU tolerance to the epoxy bodies to serve as a benchmark for phantom body homogeneity and agreement.

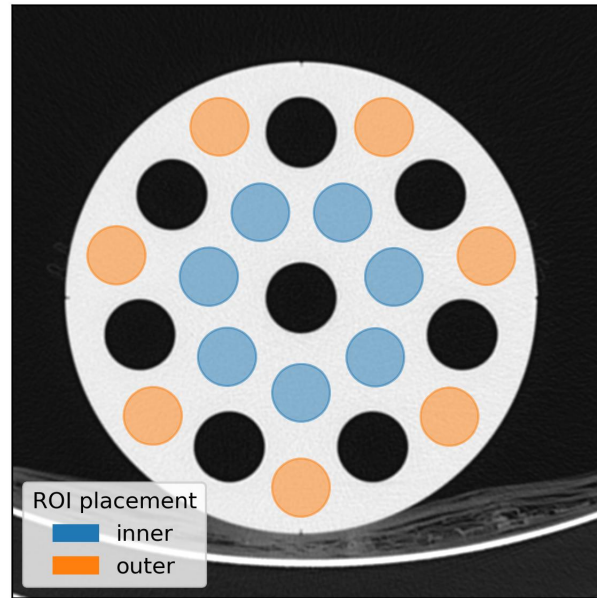


Figure 2.5: Placement of uniformity ROIs. Shown with window width 1300 HU, window level -450 HU.

Consistent ROI placement across all scans was ensured through the use of a 2D programmatic mask based on the digital ground-truth geometry, as shown in Figure 2.5. On each slice, fourteen circular ROIs (24 mm diameter) were extracted: seven positioned at a radial distance of 40 mm (“inner”) and seven at 80 mm (“outer”) from the phantom center. To capture potential axial heterogeneity, measurements were extracted from three image slices per phantom: the central slice, and slices offset by +5 mm and -5 mm. This sampling strategy was selected to provide

the minimum necessary data to calculate standard deviation and isolate z-axis variance while maintaining an efficient manual segmentation workflow.

The mean HU was calculated from each ROI. The overall mean HU for each phantom was reported ($\mu_{overall}$, averaged over all ROIs within a phantom), as well as the inner mean (μ_{inner} , averaged over the ROIs at the inner position) and outer mean (μ_{outer} , averaged over the ROIs at the outer position) to evaluate radial gradients. The intra-phantom uniformity (Δ_{max}) was defined as the maximum absolute difference in mean CT number between any two ROIs within a given phantom. To assess the reproducibility of the liquid-casting pipeline across independent builds, inter-phantom uniformity was evaluated by calculating the maximum absolute difference in $\mu_{overall}$ between any pair of the four manufactured bodies.

Additionally, to evaluate both inter-phantom and intra-phantom equivalence, a pairwise two one-sided tests (TOST) procedure was applied [75]. The TOST procedure was performed comparing the inner and outer ROIs within each individual phantom body to verify the absence of radial manufacturing gradients or severe beam hardening artifacts, as well as across all included phantom bodies to evaluate inter-phantom reproducibility. The TOST procedure tests whether the observed difference between groups is small enough to be considered clinically irrelevant—here, whether phantom-to-phantom and inner-to-outer differences fall within the ± 5.0 HU criteria. To account for potentially unequal variance between phantom bodies and between radial regions, the procedure used Welch's equations for standard error and degrees of freedom.

2.2.2.2 Texture insert reproducibility

To evaluate the reproducibility of the 3D-printing manufacturing approach, radiomic feature extraction was performed and compared across three independently manufactured sets of phantom inserts. While assessing radiomic stability serves as an indirect surrogate for evaluating the microscopic dimensional accuracy of the printed geometries, it represents a critical metric for validation; the primary utility of this phantom is to standardize quantitative imaging pipelines. Consequently, agreement of the extracted radiomic features across the replicate inserts was established as a benchmark for manufacturing reproducibility.

Three insert sets were independently printed and imaged with 360 mA (CTDI_{vol} of 13.68 mGy) and a Br62f reconstruction kernel. This high-frequency kernel was selected to maximize spatial resolution and preserve small details and edges within the textures, more accurately reflecting the true physical dimensionality of the printed objects. Between acquisitions the phantom was removed from the CT couch, all inserts were extracted and replaced with the next set, and the assembly was re-positioned using the scanner's external laser system. This approach ensures that measurements reflected the combined, realistic sources of variability: 3D printing, insert seating, and phantom positioning and alignment.

For image analysis, a rigid mask of 25.7 mm diameter circular ROIs (slightly smaller than the 30.0 mm diameter of the insert itself) derived from the digital ground truth geometry was manually placed over the eight texture inserts. Annotations were centered on the mid-plane of the insert volume and applied to slices at -3 mm and +3 mm from the center, yielding three

non-adjacent slices per insert.

From these ROIs, all available 2D radiomic texture features were extracted using *PyRadiomics* [76]. The extraction yielded a total of 91 features per ROI: 18 first-order features, and 73 second order features including gray level co-occurrence matrix (GLCM), gray level run length matrix (GLRLM), gray level size zone matrix (GLSZM), neighboring gray tone difference matrix (NGTDM), and gray level dependence matrix (GLDM) features.

Reproducibility across the three independent print-and-assembly cycles was quantified using the intraclass correlation coefficient (ICC), providing a measure of agreement where $ICC > 0.90$ is considered excellent [77], [78].

Additionally, the reproducibility was quantified using the coefficient of variation ($CV = \frac{\sigma}{\mu}$) for each texture object. To evaluate the physical fidelity of the 3D printing process, or the “inter-print” reproducibility, the CV for each radiomic feature was calculated across the three physical insert set replicates (e.g., Print 1, Print 2, and Print 3); the measurement was repeated at three slice locations, then averaged. To evaluate inter-slice (i.e., z -axis) uniformity, the CV was calculated across three distinct slice locations; the measurement was repeated for each of the three insert sets, then averaged. The features were subsequently stratified by radiomic feature class (first-order vs. second-order). The percentage of features meeting established CV reproducibility thresholds ($\leq 10\%$ and $\leq 15\%$) [23] was reported to provide an initial evaluation of platform manufacturing variability.

2.2.2.3 Evaluation of texture insert set for non-redundancy

As the texture inserts were designed with relatively simple two-material geometric patterns, it was necessary to confirm that geometric variations successfully translated into distinct radiomic signatures, rather than exhibiting high feature redundancy that would limit their utility.

To characterize the radiomic feature space coverage of the eight-insert set and confirm that the inserts provide non-redundant stimuli, a principal component analysis (PCA) was performed on the radiomic features extracted from each texture insert. Radiomic feature extraction was performed as described in Section 2.2.2.2 from images acquired with 360 mA (CTDI_{vol} of 13.68 mGy) and a Br62f reconstruction kernel. Features were standardized to zero mean and unit variance prior to decomposition, due to the sensitivity of PCA to variable scaling. The PCA was performed using the *scikit-learn* library [79]. The first two components were kept and used to visualize the relative spatial distribution of the eight textures in the rescaled PCA space. A well-dispersed, non-overlapping separation of the inserts across the principal axes of variance demonstrates that this simplified fabrication method successfully yields a diverse spectrum of radiomic stimuli.

2.2.2.4 Sensitivity to imaging protocol variation

Finally, we evaluated the sensitivity of the texture inserts to clinically relevant variations in acquisition and reconstruction protocol, assessing the platform’s ability to detect and express the protocol-induced radiomic variation it is intended to measure.

To evaluate protocol sensitivity, a single set of the 3D-printed inserts was imaged across six distinct clinical conditions, capturing common ranges of image noise and spatial resolution. Scans were acquired at 60 mA and 360 mA (yielding a $CTDI_{vol}$ of 2.28 mGy and 13.68 mGy, respectively) and reconstructed using three kernels spanning smooth to edge-enhancing frequencies (Br32f, Br46f, and Br62f). Radiomic feature extraction was performed identically to the reproducibility assessment described in Section 2.2.2.2 for each of these imaging conditions. Using Br46f and 360 mA as a reference condition (roughly equivalent to our institutional routine chest protocol), the radiomic feature reproducibility was quantified using concordance correlation coefficient (CCC) to assess pairwise agreement for each tube current and reconstruction kernel setting [80]. Consistent with established guidelines, a $CCC \geq 0.90$ was used as the threshold for reproducibility [81].

2.3 Results

2.3.1 Phantom body spatial uniformity

Across all 168 ROIs measured within the four manufactured phantom bodies, the overall dataset mean attenuation was 104.9 ± 1.5 HU. A summary of the measurements is provided in Table 2.2. All four phantom bodies met the target ± 5 HU criteria for intra-phantom uniformity, with Δ_{max} ranging from 2.7 to 4.4 HU. Additionally, all four phantom bodies met the inter-phantom reproducibility target of all ROIs falling within ± 5 HU of the overall mean, as shown in Figure 2.6. Furthermore, no visual beam hardening or streak artifacts were observed across the

Table 2.2: Mean value $\pm$ standard deviation of the HU number measurements for each phantom body.

Phantom body	μ_{inner} [HU]	μ_{outer} [HU]	$\mu_{overall}$ [HU]	Δ_{max} [HU]
1	103.5 ± 0.5	104.6 ± 0.6	104.1 ± 0.8	3.0
2	102.7 ± 0.6	103.7 ± 0.9	103.2 ± 0.9	4.4
3	105.3 ± 0.5	106.4 ± 0.7	105.8 ± 0.8	2.7
4	105.9 ± 0.7	106.8 ± 1.0	106.3 ± 1.0	3.7
Overall mean	104.4 ± 1.4	105.4 ± 1.5	104.9 ± 1.5	3.5 ± 0.7

epoxy volumes.

The pairwise TOST procedure confirmed both intra- and inter-phantom uniformity within the ± 5.0 HU criteria. For the worst-case inter-phantom comparison (Phantoms 2 and 4), the maximum mean divergence was -3.11 HU (90% CI: [-3.45, -2.76] HU), well within the equivalence criteria ($p < 0.001$). Intra-phantom spatial uniformity was similarly strong: for the worst-case phantom (Phantom 4), the inner-to-outer mean difference was -0.90 HU (90% CI: [-1.34, -0.45] HU), also falling entirely within the ± 5.0 HU criteria ($p < 0.001$). Thus, the four manufactured phantom bodies demonstrated excellent reproducibility and consistently high internal spatial uniformity.

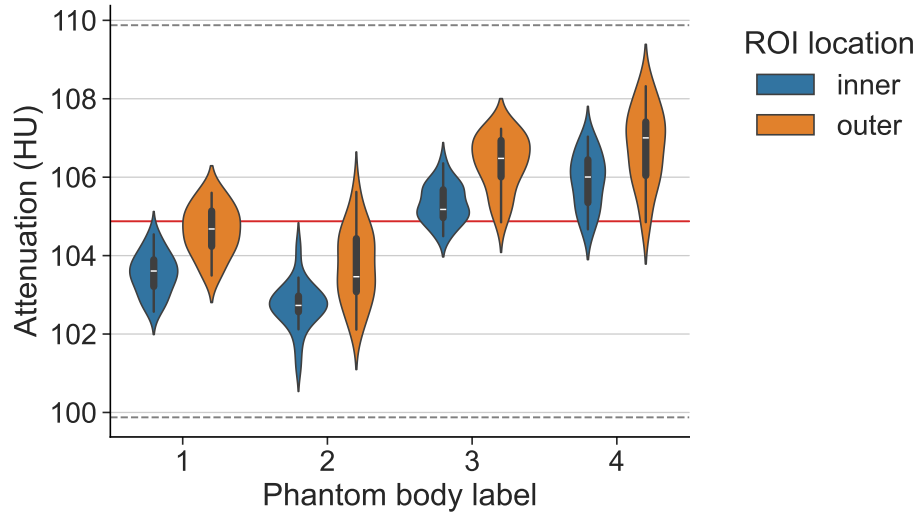


Figure 2.6: Uniformity measurements (mean values of ROIs in HU) for the four tested phantom bodies. Red, solid line illustrates the overall mean value (104.9 HU) and the gray, dashed lines illustrate ± 5 HU from the mean.

2.3.2 Texture insert reproducibility

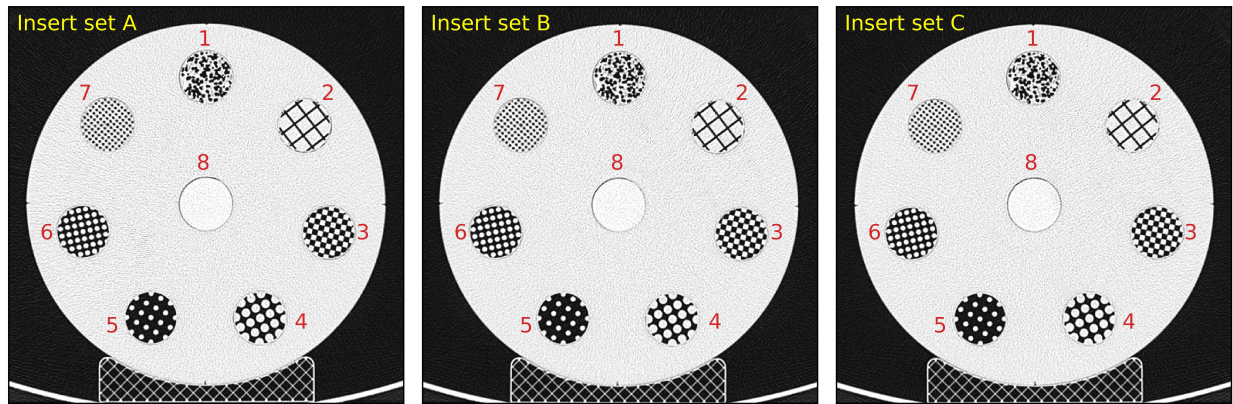


Figure 2.7: The three sets of 3D printed modular inserts, imaged at 360 mA and reconstructed with the Br62f kernel. Visual agreement on CT is excellent, including the alignment and orientation of the texture inserts relative to the phantom body across the three experimental setups and acquisitions. Images shown with window width 1300 HU and level -450 HU.

Visual inspection of CT images confirmed consistent insert alignment and orientation across

all three independent manufacturing sets (shown in Figure 2.7), with no macroscopic defects or differences between the sets observed.

Table 2.3: Texture insert reproducibility results. Inter-print CVs were averaged across all slice depths, and inter-slice CVs were averaged across all insert set replicates. Results are reported as % of features meeting CV thresholds (count/total).

Texture	Group	Inter-print CV		Inter-slice CV	
		< 15%	< 10%	< 15%	< 10%
1	1st Order	100.0 (18/18)	100.0 (18/18)	100.0 (18/18)	100.0 (18/18)
	2nd Order	100.0 (73/73)	95.9 (70/73)	100.0 (73/73)	98.6 (72/73)
2	1st Order	100.0 (18/18)	100.0 (18/18)	100.0 (18/18)	100.0 (18/18)
	2nd Order	100.0 (73/73)	97.3 (71/73)	100.0 (73/73)	98.6 (72/73)
3	1st Order	100.0 (18/18)	100.0 (18/18)	100.0 (18/18)	100.0 (18/18)
	2nd Order	100.0 (73/73)	100.0 (73/73)	100.0 (73/73)	100.0 (73/73)
4	1st Order	100.0 (18/18)	94.4 (17/18)	100.0 (18/18)	94.4 (17/18)
	2nd Order	100.0 (73/73)	95.9 (70/73)	98.6 (72/73)	95.9 (70/73)
5	1st Order	100.0 (18/18)	100.0 (18/18)	100.0 (18/18)	100.0 (18/18)
	2nd Order	100.0 (73/73)	95.9 (70/73)	95.9 (70/73)	94.5 (69/73)
6	1st Order	100.0 (18/18)	100.0 (18/18)	100.0 (18/18)	100.0 (18/18)
	2nd Order	100.0 (73/73)	100.0 (73/73)	98.6 (72/73)	93.2 (68/73)
7	1st Order	100.0 (18/18)	100.0 (18/18)	100.0 (18/18)	100.0 (18/18)
	2nd Order	94.5 (69/73)	89.0 (65/73)	100.0 (73/73)	94.5 (69/73)
8 ^a	1st Order	94.4 (17/18)	94.4 (17/18)	94.4 (17/18)	94.4 (17/18)
	2nd Order	71.2 (52/73)	63.0 (46/73)	67.1 (49/73)	60.3 (44/73)

^aNote: Solid insert, not a heterogeneous texture pattern

The extracted radiomic features demonstrated high agreement across the independent print-and-assembly cycles. The inter-print ICC across all eight textures exceeded 0.998 in all cases (mean ICC > 0.999), demonstrating very strong agreement well above the established threshold for excellent reproducibility (ICC > 0.90). Inter-slice ICC within individual inserts performed similarly, yielding a mean ICC > 0.999.

This stability is further reflected in the CV analysis, summarized in Table 2.3. For the first order features, all the heterogeneous texture inserts except for Texture 4 had 100.0% (18/18) of the first order features meet the strict $CV < 0.10$ reproducibility threshold across both inter-print and inter-slice evaluations. For Texture 4, 100.0% of the first order features met the $CV < 0.15$ threshold and 94.4% (17/18) of the first order features met the $CV < 0.10$ threshold in both inter-print and inter-slice evaluations. For Texture 8, the homogeneous solid insert, 94.4% of the first order features met the $CV < 0.10$ threshold. The radiomic feature that did not meet the reproducibility criteria in each of these cases was first order skewness, which relies on the third central moment of the intensity distribution and can be highly sensitive to minor fluctuations in image noise.

For the second order texture features, Textures 1–6 had 100.0% (73/73) of the second order features meet the inter-print $CV \leq 0.15$ criteria, and between 95.9% and 100.0% of the second order features meet the stricter $CV \leq 0.10$ criteria. Texture 7 exhibited slightly lower inter-print reproducibility, with 94.5% of the second order features meeting the $CV < 0.15$ threshold, and 89.0% of the second order features meeting the $CV < 0.10$ threshold. The inter-slice reproducibility for the second order features was also high, with 93.2% to 100.0% of the second order features meeting the $CV < 0.10$ threshold for all of the heterogeneous texture inserts. While Texture 8 displayed lower reproducibility for second order features, these results were not considered meaningful measures of the reproducibility of the physical objects, as there was not a heterogeneous texture pattern to measure and these variations likely reflect nuance in how radiomic textures are calculated in uniform objects.

Distributions of the inter-print and inter-slice CV, as well as the complete report of the features that did not meet the CV criteria, are provided in Appendix B.1. Overall, the inter-print reproducibility was comparable to inter-slice reproducibility, indicating that very low physical print-to-print manufacturing variance is similar to (or less than) the variation across printing layers.

2.3.3 Evaluation of texture insert set for non-redundancy

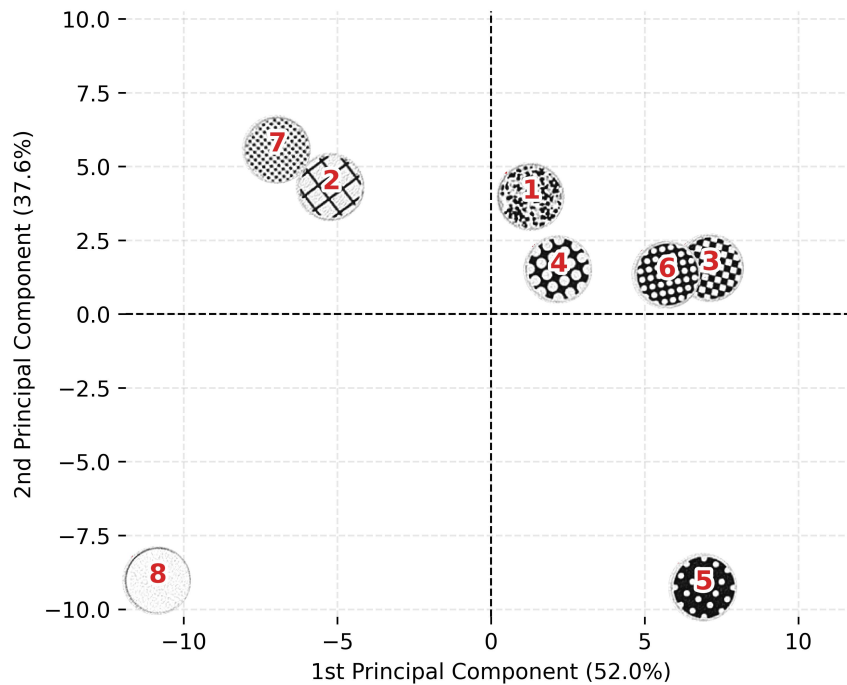


Figure 2.8: Principal component analysis of radiomic feature values measured on the eight texture inserts. Images of the textures themselves with labels overlaid are used as markers.

The PCA of the radiomic features showed that the first two principal components captured 89.6% of total radiomic variance (PC1: 52.0%; PC2: 37.6%), indicating the set of texture inserts selected is well-described by two dominant axes of variation. The distribution of inserts within

this reduced space is shown in Figure 2.8. The visualization shows that there is a spread in the distribution of the texture inserts within the feature space, confirming that the insert set provides a range of radiomic stimuli rather than redundant first-order variations. Most of the textures are well resolved along PC1, with PC2 separating Textures 5 and 8 from the other textures distinctly. These results show that Textures 5 and 8 are the most unique in the set, as they are each alone within their respective quadrants in the feature space with the greatest separation from any other texture insert. Conversely, certain inserts exhibit closer proximity in the feature space; for example, Textures 3 and 6 cluster together, reflecting their similar underlying physical geometry in texture pitch and element size. Rather than indicating a problematic redundancy across the entire insert set, this proximity is beneficial as it establishes a realistic spectrum of both highly distinct and distinct-but-similar radiomic stimuli. Overall, the visualizations confirm that the insert set provides diverse radiomic stimuli rather than redundant texture designs.

2.3.4 Sensitivity to imaging protocol variation

Visual assessment of the acquisitions (Figure 2.9) confirms that the selected protocols produced a clinically realistic spectrum of image noise and spatial resolution, with fine physical textures visibly blurring under smooth kernels and sharpening under edge-enhancing kernels.

The designed textures proved highly sensitive to variation in imaging protocol parameters, particularly reconstruction algorithms. Transitioning from the standard Br46f to the smooth Br32f kernel (with the reference tube current, 360 mA, held constant) yielded only 8.8% reproducible

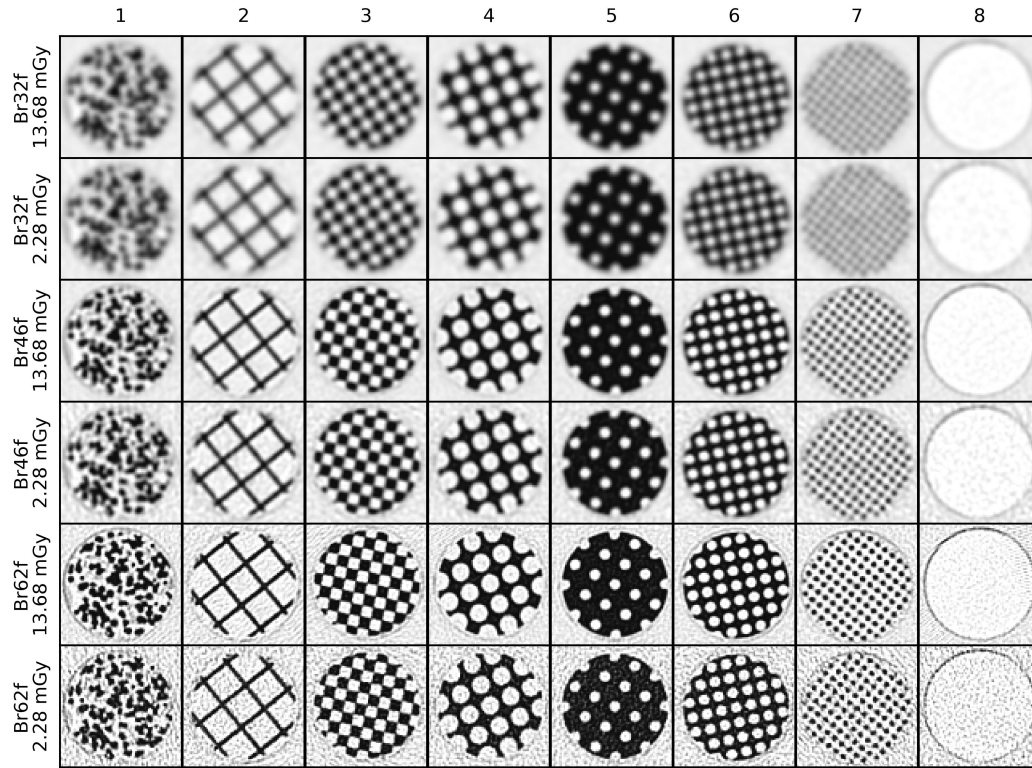


Figure 2.9: Images of the eight texture objects (labeled above) under each of the six imaging conditions (labeled left). Rows are ordered with the highest noise condition at the bottom to the lowest noise condition at the top. Images shown with window width 1300 HU and level -450 HU.

features, and transitioning to the sharp Br62f kernel yielded only 7.7%. When tube current was reduced from 360 mA to 60 mA (with the reference kernel, Br46f, held constant), 58.2% of the extracted radiomic features remained reproducible. Distributions of the CCC values are provided in Appendix B.2.

Together, Figure 2.10 and Figure 2.11 illustrate the measured radiomic feature values across all eight textures and imaging conditions. In Figure 2.10, raw feature values are shown for a representative subset of features, highlighting specific directional sensitivities to protocol variations. For example, first-order mean attenuation remains exceptionally stable across both dose

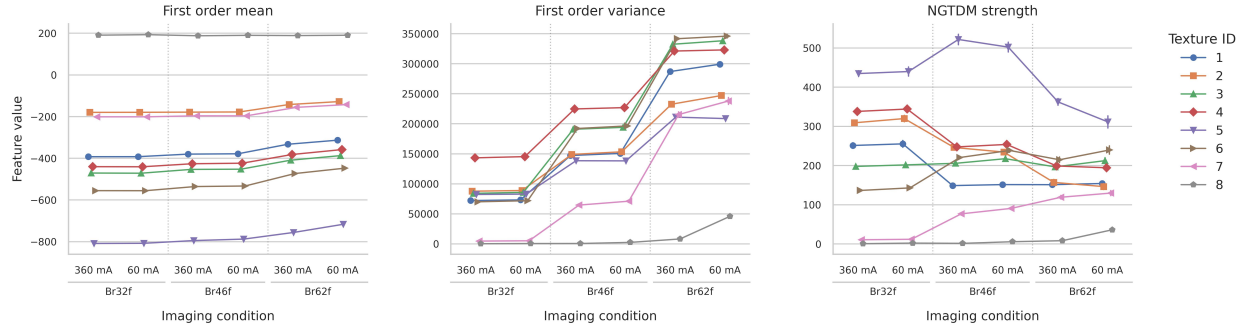


Figure 2.10: Radiomic features values for three illustrative features (first order mean, first order variance, NGTDM strength) for each texture insert across the six imaging conditions.

and kernel changes, aligning with the physical reality that mean HU is driven by the textures' material ratios (PLA vs. air) rather than reconstruction algorithms or dose. Conversely, first-order variance increases significantly under sharper kernels, consistent with the expected increase in image noise. Figure 2.11 extends this view to all extracted features, with feature values min-max normalized to $[0, 1]$ to enable visual comparison across features of differing scales. In this heatmap, macro-level color differences across texture groupings (columns, labeled above) reflect the degree to which features differ between texture objects, while color variation within a texture grouping reflects sensitivity to protocol. Features that appear uniform in color within a texture grouping are robust to protocol changes, whereas those exhibiting strong within-group variation are more sensitive to reconstruction kernel and dose. While investigations into the reproducibility of individual radiomic features was beyond the scope of this work, these figures illustrate the phantom's capacity to detect and measure this type of protocol-induced variation.

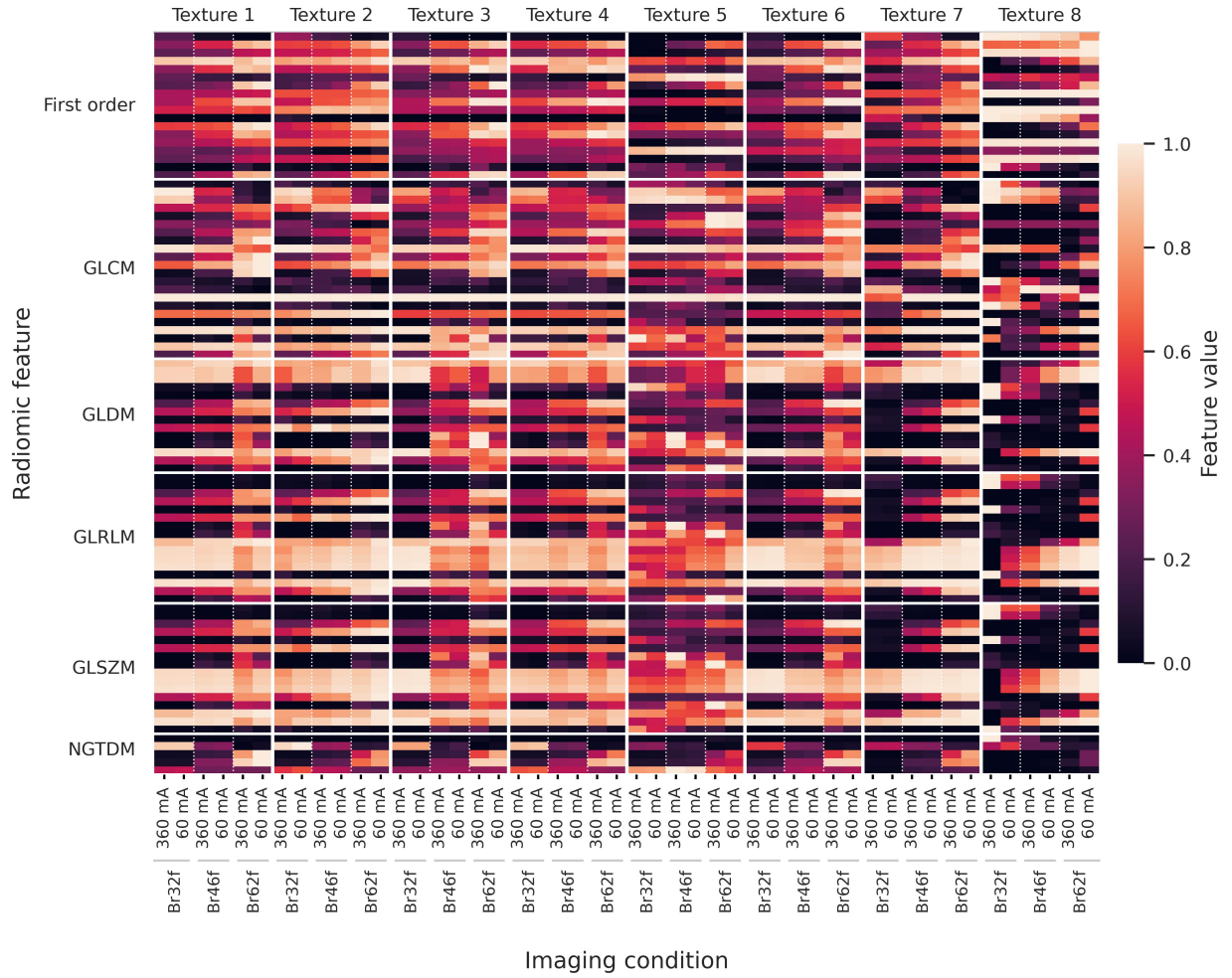


Figure 2.11: Radiomic feature values of all features (grouped by feature class on the y -axis), shown for each texture object and imaging condition (x -axis), using the same imaging condition order as in Figure 2.10. Feature values were min-max normalized to $[0, 1]$. Color variation between texture groupings reflects differences between the texture objects, while within-group variation reflects sensitivity to imaging protocol.

2.4 Discussion

This study demonstrates the feasibility of producing high-quality, reproducible phantoms for quantitative imaging using an accessible hybrid manufacturing approach. Results indicate that com-

binning 3D-printed molds with epoxy resin casting yields phantom bodies with high uniformity and reproducibility. The modular insert design enables rapid development and sharing of texture patterns, addressing a critical need in the radiomics community for standardized yet customizable validation tools. Importantly, the phantom system design and manufacturing provides a pathway to those tools with dramatically reduced costs and improved accessibility compared to commercial alternatives (approximately \$50 per phantom system; a cost breakdown is provided in Appendix A.3).

The epoxy casting approach produced four phantom bodies with spatial uniformity meeting our acceptance criteria and high inter-phantom consistency (all ROIs within ± 5 HU, and inter-phantom differences statistically indistinguishable within ± 5 HU with $p < 0.001$). A small but consistent radial gradient was observed, with outer ROIs approximately 1 HU higher than inner ROIs on average; this is likely attributable to minor beam hardening at the phantom periphery and falls well within acceptable limits. While this manufacturing run represents a best-case single-group scenario, it establishes that an individual research group can reliably produce a small batch of self-consistent phantom bodies using accessible equipment and standardized procedures.

The 3D printed texture inserts demonstrated excellent radiomic reproducibility across three independently manufactured and assembled sets, with a mean inter-print ICC > 0.999 across 91 radiomic features. This result is comparable to the ICC reported by Li et al. in test-retest analyses on a single texture phantom [65], suggesting that the print-to-print manufacturing variance achieved here is at the same level as the measurement repeatability floor for this class of phantom. In contrast, clinical test-retest studies in patients have reported much lower radiomic

feature reproducibility, underscoring the value of a controlled phantom reference for isolating protocol-induced variation from biological and positioning noise [12].

We note that Texture 7, which has the smallest and most densely packed grid elements of any structured insert, exhibited slightly lower inter-print reproducibility (94.5% at $CV < 0.15$). This suggests the 0.75 mm feature diameter of this insert approaches the practical resolution limits of the FDM process, where small dimensional deviations between print cycles have proportionally greater impact on the resulting radiomic signature. This is a useful boundary condition for future insert design: features substantially below 1.0 mm in any dimension should be treated with caution when using standard 0.4 mm nozzle FDM printing, and the 0.2 mm nozzle used for Texture 7 represents the lower practical limit with consumer-grade equipment. Future work should further characterize these limits for sub-millimeter texture designs.

The PCA of the radiomics features provides strong evidence that the insert set provides genuinely non-redundant radiomic stimuli, with each insert contributing unique information to the feature space. Textures 5 and 8 were particularly distinct, which could be attributed to each representing an extreme in terms of material compositions within the set, with Texture 5 having the highest ratio of air, and Texture 8 (the solid, uniform insert) having the highest ratio of PLA.

Finally, the phantom platform demonstrated the protocol sensitivity required of a quantitative imaging tool. While both tube current and reconstruction kernel had dramatic impacts on radiomic reproducibility, the effect of reconstruction kernel was greater: changing from the reference Br46f kernel to either the smooth Br32f or sharp Br62f kernel left only 8.8% or 7.7% of features reproducible, respectively, while reducing tube current from 360 mA to 60 mA left 58.2% of

features reproducible. This sensitivity to changes in imaging protocol, as well as this asymmetry in particular, is consistent with the broader radiomics literature, which identified reconstruction kernel as a dominant source of protocol-induced radiomic variation [7], [11], [55]. The fact that our synthetic textures exhibit this well-established pattern validates their suitability as reference stimuli: a phantom that did not respond to kernel changes in the expected direction would be of limited utility for protocol harmonization studies. The directional feature sensitivities observed for the individual radiomic features are also physically interpretable, such as the stability of first-order mean across all conditions versus the strong kernel-dependence of first-order variance, providing a useful sanity check on the measurement pipeline. The radiomic feature variation observed across the eight texture objects also supports the results from the texture insert set non-redundancy PCA evaluation.

A key distinguishing contribution of this platform relative to prior 3D-printed phantom work is its foundational commitment to open science. By dedicating the complete design files and manufacturing specification to the public domain under a CC0 license, we remove the licensing barriers that have historically restricted phantom-based harmonization efforts to well-resourced groups with access to proprietary hardware. Unlike prior approaches that employed 3D printing without public release of designs, this framework allows any research group with a consumer-grade FDM printer to replicate, validate, modify, and redistribute phantom designs without restriction. The modular architecture amplifies this benefit: new insert designs targeting specific clinical tasks or tissue types can be developed, validated, and shared independently of the phantom body, enabling community-driven expansion of the platform's insert library without requiring each group

to develop a complete phantom system from scratch.

The eight texture inserts evaluated in this work were selected to provide diverse and complementary radiomic signatures rather than to mimic specific pathologies. Consequently, the phantom is intended as a tool for studies involving radiomic feature reproducibility, protocol optimization methodologies, and scanner harmonization or quality assurance. The mean attenuation of the texture inserts evaluated ranged from -800 to 200 HU (shown in Figure 2.10), depending primarily on the texture's ratio of PLA to air. The materials used (PLA and clear epoxy resin) exhibited attenuation values within the range of 100 to 200 HU. We expect the inserts to be most relevant to quantitative imaging tasks involving heterogeneous tissue and, based on these attenuation characteristics, may be most applicable to contrast-enhanced soft tissues, lung, and abdominal organs. Applications in lung imaging may be of particular interest, as the lung is composed of heterogeneous patterns soft tissue and air that mirrors the PLA-and-air compositions of the texture inserts. However, the suitability of these textures for clinical applications should be evaluated in future work. Users are encouraged to design custom texture inserts tailored to specific clinical pathologies or modeling endpoints, which is supported by the phantom's modular nature and open-source design.

This initial evaluation was performed on a typical clinical CT scanner system, using the conditions described in Table 2.1. While the phantom body and modular insert architecture are expected to remain applicable across CT platforms, the specific texture inserts evaluated here represent only an initial validation set and can be refined or expanded as scanner resolution and quantitative imaging requirements evolve. Whether the measured reproducibility of the manu-

facturing approach presented in this work is impacted by substantially higher spatial resolutions available with systems such as photon-counting CT (PCCT) remains to be evaluated. However, because the phantom specification is fundamentally digital and does not mandate a particular fabrication method, if needed the platform can be adapted to match the ultra-high spatial resolutions achievable with next-generation PCCT systems through high-resolution additive manufacturing, micromachining, or other computer-aided manufacturing techniques.

Several limitations of this work should be noted. First, the reproducibility analysis intentionally represents a best-case manufacturing scenario: conducted by a single group using a single printer and single material supplier. This baseline was a necessary first step, but future multi-center studies will be needed to quantify the inter-institution variability introduced by different printers, filament batches, and casting environments. Some increase in variability is expected across sites, and it remains an open question whether stricter material standardization (e.g., specifying filament density tolerances or epoxy formulation) will be required for cross-site harmonization. Second, this was a single time point investigation, and thus did not include evaluation of the longitudinal stability of the phantom materials. While prior work has found that the attenuation properties of PLA are stable over time and do not require special storage conditions [71], a longitudinal stability study is warranted before this platform is used for studies spanning multiple years. Finally, the current insert set was intentionally designed as a simplified baseline: binary material distributions with no z -axis structural variation and no attempt to mimic specific biological tissues or clinical pathologies. This simplification was appropriate for the initial evaluation of the reproducibility of the manufacturing approach and validating the platform's basic metrological properties, but

it means the inserts do not yet provide the clinically relevant or application-specific stimuli that would be needed for targeted radiomic studies (e.g., inserts designed to emulate the textural properties of lung nodules, liver lesions, or other tissues of clinical interest). Developing such inserts, including designs that incorporate z -axis structural complexity or stochastic heterogeneity representative of specific pathologies, represents the most important next step for the platform. The modular open-source architecture is specifically designed to support these types of extensions.

2.5 Conclusions

The open-source, accessible phantom presented here addresses critical barriers to quantitative imaging research. By combining 3D printing technology with epoxy resin casting, we have created a solution that is simultaneously affordable, customizable, and reproducible. Serving as a rigorous middle ground between expensive commercial products and non-reproducible hand-made alternatives, this platform successfully demonstrated the key properties we require for our radiomic texture experiments: uniformity, radiomic reproducibility, and radiomic sensitivity to clinical protocol variations. By releasing the phantom and insert design into the public domain (<https://doi.org/10.5281/zenodo.19340627>), this tool remains unencumbered for the entire research community, fostering collaborative development and long-term standardization efforts. The immediate practical consequence is that any research group with access to a consumer-grade FDM printer can replicate, validate, and extend this platform without restriction.

3 Developing a dataset for investigations into the impact of CT acquisition and reconstruction conditions on quantitative imaging

This Chapter is based on the following manuscript: M. A. Daly, J. M. Hoffman, A. M. Hernandez, A. Uneri, B. A. Varghese, J. Levy, and M. F. McNitt-Gray, “CT texture phantom dataset with paired image quality assessments for quantitative imaging,” *In press at Medical Physics*, 2026. DOI: [10.1002/mp.70633](https://doi.org/10.1002/mp.70633). Submitted as a Medical Physics Dataset Article.

3.1 Introduction

Quantitative imaging features (QIFs) extracted from computed tomography (CT) images hold significant promise for precision medicine applications, including disease characterization, treatment response assessment, and outcome prediction [2], [3], [82]. However, the clinical translation of QIF-based biomarkers faces substantial challenges due to variability introduced by differences in image acquisition and reconstruction protocols across scanners and institutions [12], [13]. This protocol-dependent variability can obscure true biological signals and compromise the reproducibility of radiomics studies [83].

Texture-based QIFs, or radiomic texture features, have shown utility in predicting patient

outcomes and treatment response [4], [47], [84]. The relationship between CT image quality and image textures remains incompletely characterized, limiting efforts to develop effective harmonization strategies. While previous studies have demonstrated that acquisition parameters significantly influence texture features [7], [9], [66], [85], systematic datasets enabling comprehensive investigation of the relationships between image quality and radiomic texture features are lacking. Existing phantom studies typically focus on either image quality assessment or texture analysis in isolation, preventing direct correlation between image quality and QIF variability.

To address this gap, we have developed a dataset consisting of DICOM images of the 3D printed texture phantom presented in Chapter 2 with paired image quality measurements acquired under matched specified conditions across multiple CT scanner models and protocols. This included 3D measurements of the modulation transfer function (MTF) and the noise power spectrum (NPS), as well as contrast-to-noise ratio (CNR), as well as scalar summary metrics extracted from these measurements. Comprehensive quantification of 3D MTF and 3D NPS is essential for accurately characterizing the three-dimensional signal and noise transfer properties of modern CT systems [28], [86]. From the texture phantom, radiomic features were extracted from four texture objects for each condition.

This dataset enables researchers to directly correlate these detailed image quality measurements with QIF measurements, particularly texture features that are inherently sensitive to image quality variations. This dataset is restricted to filtered back projection (FBP) reconstruction kernels. While iterative reconstruction is increasingly common in clinical practice, traditional image quality measurements may not fully characterize system performance under non-linear processing

conditions, as measurements often become dependent on object contrast and noise levels [30].

The primary objectives of this dataset are to: (1) provide a standardized resource for investigating the relationship between image quality metrics and radiomic texture features (and between image quality and image textures more generally), (2) enable the development and validation of harmonization methods that account for underlying image quality differences, and (3) facilitate optimization of acquisition protocols for quantitative imaging applications. This dataset is used for the analyses of Chapters 4 to 6 of this dissertation.

3.2 Methods

An image quality assessment phantom and a CT texture phantom were scanned using matched conditions for 170 protocols across four scanners. The protocols capture a wide range of radiation dose levels and reconstruction kernels. Measurements of 3D spatial resolution, 3D noise characteristics, and CNR were made for each protocol.

3.2.1 Phantoms

3.2.1.1 Image quality phantom

Image quality characterization was performed using the Corgi[®] phantom (Phantom Laboratory, Salem, NY) that enables simultaneous measurement of CNR, 3D MTF, and 3D NPS using automated analysis software (described in Section 3.2.3) [87]. Corgi[®] contains 13.5 cm diameter, 4.0 cm thick cylindrical modules composed of polyurethane (PUR). For this study, a circular annulus

was used, making the outer phantom diameter 20.0 cm. Corgi[®] is comprised of blank modules which are uniform in composition, and test modules that are constructed with 20 mm diameter cylindrical inserts with hemispherical end-caps with known attenuation properties: polyvinyl chloride (PVC, 958 HU), polyoxymethylene (POM, 342 HU), polycarbonate (PC, 101 HU), nylon (PA, 31 HU), and polystyrene (PS, -27 HU) at 140 kV [88]. The blank modules are used for 3D NPS measurements, and the test modules are used for 3D MTF and CNR measurements. Example CT images of the are shown in Figure 3.1, and a schematic showing the test modules is provided in Figure 3.2.

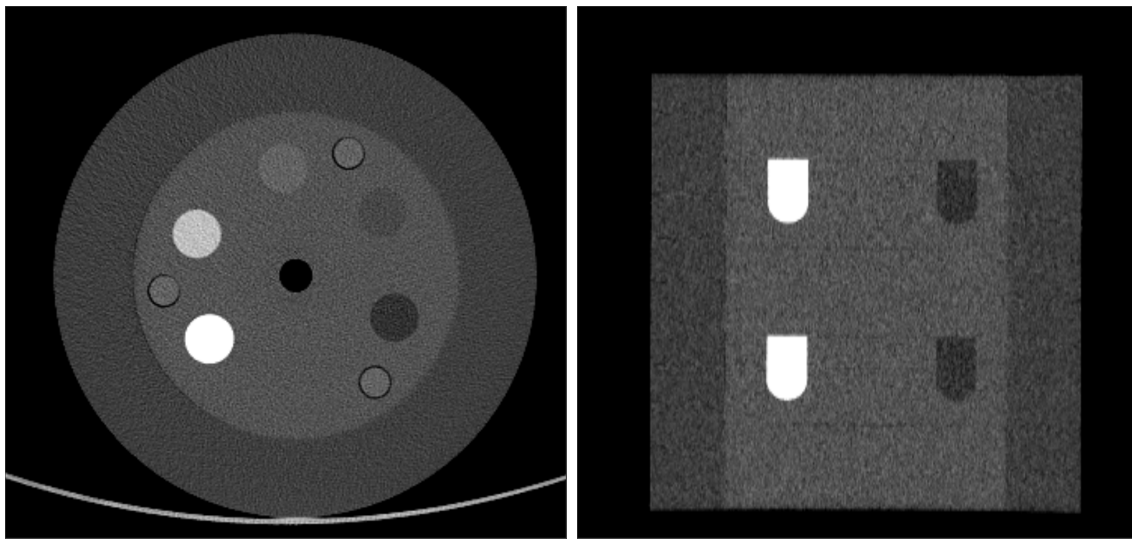


Figure 3.1: Example axial and coronal slices of the Corgi[®] phantom, used to measure image quality. Shown with window width 650 HU, window level 175 HU.

The 3D measurement capability is particularly important for CT systems, as it accounts for potential anisotropy in spatial resolution and noise characteristics that may result from the helical acquisition geometry, variations in the effective focal spot size, or application of apodization kernel

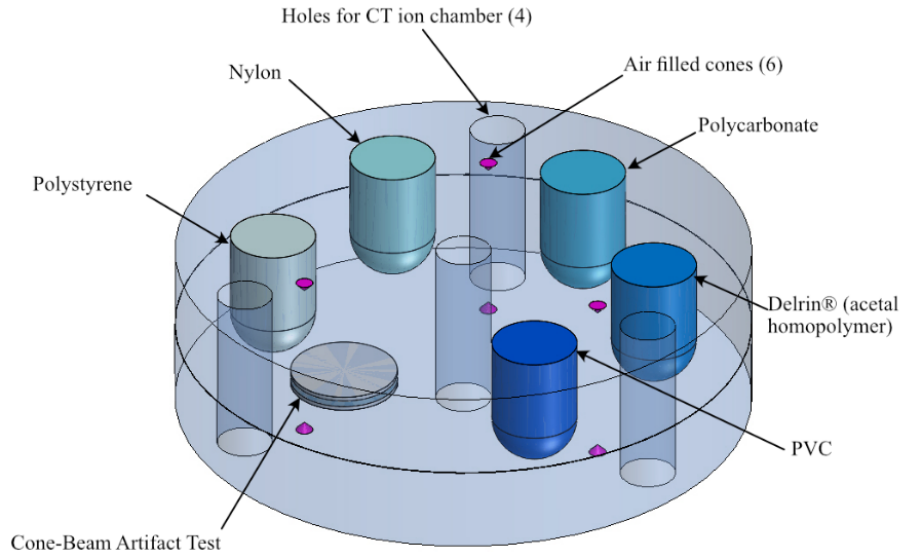


Figure 3.2: The Corgi phantom, developed for the rigorous assessment of CT image quality. Reproduced from [88].

during reconstruction [28], [86], [89]. This comprehensive approach provides a more complete characterization of image quality than traditional 2D measurements.

3.2.1.2 Texture phantom

To facilitate reproducibility and enable customization, we developed an open-source 3D-printed texture phantom, as described in Chapter 2 [90]. The specific modular inserts used differ from those described in Chapter 2 and are shown in Figure 3.3. The STL files for this set of inserts is provided at <https://doi.org/10.5281/zenodo.18718137> [91]. A PLA body ring was used to increase the total diameter of the phantom to 27.5 cm.

Inserts 1, 2, and 3 were uniformly spaced 3 mm circular rods with varied pitch (4, 6, 8 mm).

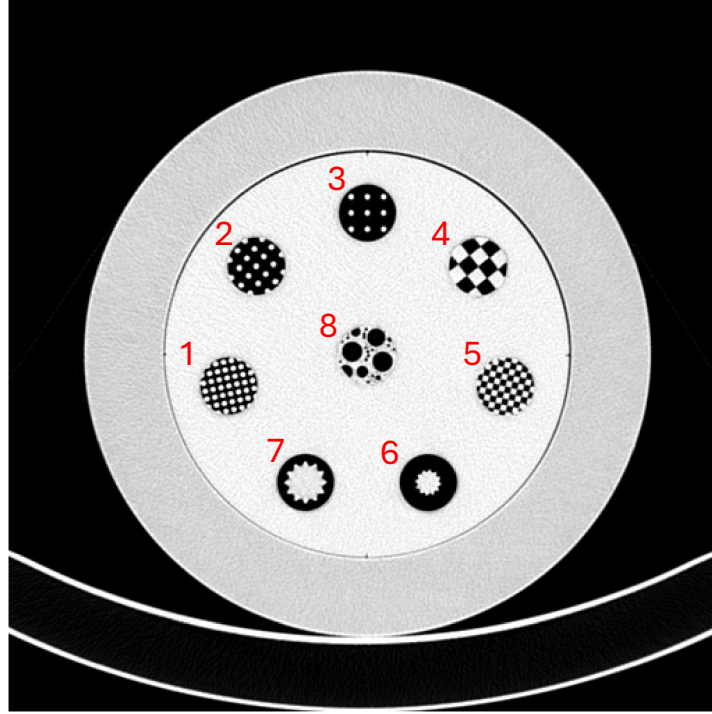


Figure 3.3: The 3D-printed texture phantom with the texture insert objects labeled 1-8, with the PLA body ring. Shown with window width 1300 HU, window level -450 HU.

Inserts 4 and 5 were checkerboard patterns with different scales: grid square side lengths of 3 mm and 7 mm. Inserts 6 and 7 were spiculated solid objects of different scales. Insert 8 was a texture composed of circles with random size and positioning. Inserts 1, 4, 5, and 8 were used for radiomic feature analysis (described in Section 3.2.5). The positioning and orientation of the inserts within the phantom was kept consistent across all acquisitions.

3.2.2 Image acquisition

Image acquisition was performed at two academic medical centers using four CT scanner models from two manufacturers: SOMATOM Force and SOMATOM Definition AS64 (Siemens Healthi-

neers, Forchheim, Germany) and Revolution EVO and Revolution Apex (GE Healthcare, Wauke-
sha, Wisconsin).

The image quality phantom and the two texture phantoms were scanned under matched conditions with careful alignment for a diverse set of CT protocols. Scans were acquired at five $CTDI_{vol}$ levels and reconstructed with eight reconstruction kernels. This was repeated across the four different CT scanner systems. All other parameters were kept constant across acquisitions from the same scanner, and matched as closely as possible between scanner models. Imaging conditions were selected with the goal of capturing varied image quality (i.e., noise and spatial resolution) within a range of clinical practice for chest and abdomen/pelvis CT examinations. We did not specifically consider head and neck CT examinations in this protocol selection.

The $CTDI_{vol}$ levels ranged from 2.1 mGy to 17.1 mGy. The $CTDI_{vol}$ values were selected to approximate that of the lowest dose CT examinations (lung cancer screening) through that of more typical diagnostic CT examinations (chest or abdomen/pelvis, depending on patient size) [92], [93]. This range will capture a range of clinically realizable noise conditions. All scans used 120 kV; the desired $CTDI_{vol}$ levels were obtained by varying only tube current.

Reconstruction kernels were chosen to capture a range of clinically relevant FBP reconstructions. For the GE scanners, this included all eight kernels available. For the Siemens scanners, eight kernels were selected that ranged from a smoothing kernel (e.g., Br32d) to a sharp over-enhancing kernel (e.g., Br64d). All reconstructions employed FBP algorithms. The kernels and dose levels selected for each scanner model can be found in Table 3.1.

Scans were acquired with a slice thickness of 1.0 mm on Siemens scanners and 1.25 mm

Table 3.1: Summary of the scan protocols' acquisition and reconstruction conditions (all FBP).

Vendor	Scanner	CTDI _{vol} [mGy]	Reconstruction Kernels	<i>n</i>
GE	EVO	2.2, 4.2, 6.2, 10.6, 13.7, 17.0	BONE, BONEPLUS, CHST, DETAIL, EDGE, LUNG, STANDARD	42
	Apex	2.1, 4.3, 6.5, 10.7, 13.9, 17.1	BONE, BONEPLUS, CHST, DETAIL, EDGE, LUNG, SOFT, STANDARD	48
Siemens	Force	2.1, 4.3, 6.4, 10.7, 17.1	Br32d, Br36d, Br40d, Br44d, Br49d, Br54d, Br59d, Br64d	40
	Definition	2.1, 4.1, 6.4, 10.5, 16.9	Br32f, Br36f, Br40f, Br43f, Br49f, Br54f, Br59f, Br62f	40
Total number of protocols:				170

on GE scanners, as these were the closest thin slice thicknesses available. Thin slice thickness was selected as a higher resolution is generally preferred for quantitative imaging to reduce partial volume effects. Additionally, the provision of thin-slice data offers maximum flexibility, as thick-slice datasets (e.g., 2.5 mm or 5.0 mm) can be readily synthesized from thin-slice images through simple averaging. This allows users to simulate various clinical viewing conditions from the provided data. Conversely, generating thin slices from thick-slice acquisitions is not possible without the loss of high-frequency information or the introduction of interpolation artifacts.

All protocols used 120 kV, 0.5 second rotation time, pitch of 1.0, and 350 mm field of view. Tube current modulation was disabled to ensure that acquisition was consistent across slices, and enable precise and reproducible pairing between texture phantom scans and image quality measurements. Detailed descriptions of the four CT scanner systems and the parameters that were held constant across acquisitions are provided in Table 3.2.

Table 3.2: The CT scanner details and image acquisition parameters used for each of the four systems used. These parameters were used consistently for all acquisitions for all phantoms.

Parameter	DICOM tag	Scanner 1	Scanner 2	Scanner 3	Scanner 4
Manufacturer	(0008, 0070)	SIEMENS	SIEMENS	GE MEDICAL SYSTEMS	GE MEDICAL SYSTEMS
Model	(0008, 1090)	SOMATOM Definition AS64	SOMATOM Force	Revolution EVO	Revolution Apex
Software Version	(0018, 1020)	syngo CT VB20A	syngo CT VB20A	cj_kl.89	ca- dence_ct_25.44
Slice thickness (mm)	(0018, 0050)	1.0	1.0	1.25	1.25
Tube voltage (kV)	(0018, 0060)	120	120	120	120
Reconstruction diameter (mm)	(0018, 1100)	350	350	350	350
Rows	(0028, 0010)	512	512	512	512
Columns	(0028, 0011)	512	512	512	512
Pixel Spacing	(0028,0030)	0.68	0.68	0.68	0.68

3.2.3 Image quality measurements

Image quality was measured using the Corgi[®] phantom. The 3D MTF, 3D NPS, and CNR were computed using automated software evaluation as described in Siewerden and Uneri et al [87].

Spatial resolution was measured both in-plane (also referred to as “axial” or “2D”) and out-of-plane (“oblique” or “3D”). The edge spread function (ESF) was measured from the PVC insert and used to calculate the MTF for the 2D and 3D measurements. The MTF measurements were made on the highest dose scans to minimize uncertainty introduced by image noise. For a given scanner system and kernel, the spatial resolution measurements made on the highest dose scan were then propagated to the lower dose scans. For linear FBP, it is reasonable to assume that MTF is independent of dose.

The 3D NPS was measured on the phantom’s uniform module, composed of homogeneous PUR. The NPS curves are obtained by radially sampling the 3D NPS in the axial and oblique directions. The NPS was measured for each of the 170 conditions.

The CNR measurements were made on each insert (PVC, POM, PC, PA, PS) to assess contrast detectability across a range of attenuations. The CNR was calculated as:

$$CNR = \frac{\mu_{ROI} - \mu_{BACKGROUND}}{\sigma_{BACKGROUND}} \quad (3.1)$$

where μ_{ROI} is the attenuation measured within the insert, $\mu_{BACKGROUND}$ is the attenuation measured in the background, and $\sigma_{BACKGROUND}$ is the noise (standard deviation) of the background. The measured attenuation, contrast, and CNR were reported for each insert. The CNR was measured for each of the 170 conditions.

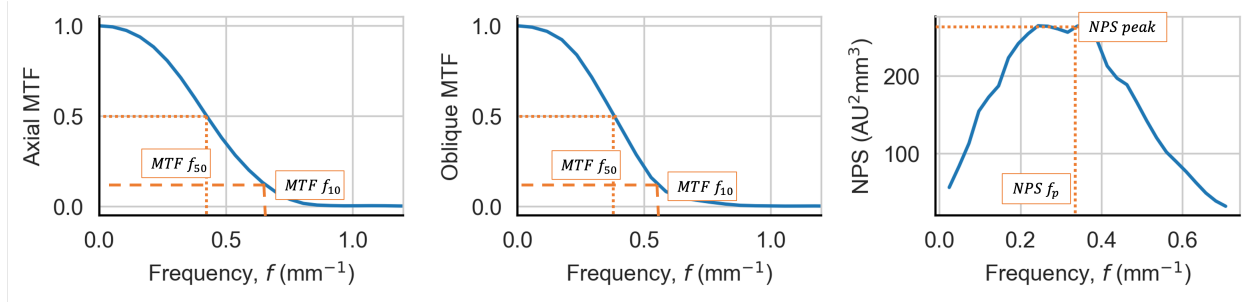


Figure 3.4: Illustrations of a selection of extracted image quality metrics.

3.2.4 Image quality metric calculation

From the extracted image quality measurements, we calculated a set of scalar parameters that sought to characterize the shape and behavior of the MTF and NPS curves. We refer to these derived scalar parameters as “image quality metrics” or IQMs. Figure 3.4 shows axial MTF, oblique MTF, and NPS curves for an example CT protocol, with a selection of IQMs annotated for illustrative purposes. The complete set of computed IQMs is listed in Table 3.3.

Resolution (MTF). From each axial and oblique MTF curve, the spatial frequencies at which the MTF falls to 10%, 25%, and 50% were extracted ($MTF_{f_{10}}$, $MTF_{f_{25}}$, $MTF_{f_{50}}$). To quantify the degree of edge enhancement, the peak MTF value (MTF_{peak}) and the frequency at which it occurs ($MTF_{f_{peak}}$) were also recorded. Overenhancing kernels produce an MTF that peaks above 1.0 ($MTF_{peak} > 1$), whereas conventional kernels do not ($MTF_{peak} \approx 1$); the peak value therefore provides a continuous descriptor that distinguishes overenhancing from conventional kernels.

Noise (NPS). From the NPS curve, $NPS(f)$, was summarized by its magnitude and its frequency content. The noise magnitude σ characterizes the overall noise level, and NPS_{max} gives the peak noise magnitude. The frequency content was summarized by the peak frequency f_P and the average frequency f_A :

$$f_P = \arg \max_f [NPS(f)], \quad (3.2)$$

$$f_A = \frac{\int f NPS(f) df}{\int NPS(f) df}, \quad (3.3)$$

where $f = \sqrt{f_x^2 + f_y^2}$ is the radial spatial frequency and $NPS(f)$ is the radially rebinned 1D NPS.

Detectability (NEQ). The frequency-dependent noise-equivalent quanta (NEQ) combines resolution and noise into a single detectability descriptor:

$$NEQ(f) = a^2 \frac{MTF(f)^2}{NPS(f)}, \quad (3.4)$$

where a is the mean pixel value, computed from ROIs placed in the uniform module of the phantom [94].

Spectral shape descriptors. In addition to the single-frequency metrics above, the overall shape of the MTF, NPS, and NEQ curves was summarized by treating each curve as a distribution over radial frequency and computing its skewness and kurtosis, which capture the asymmetry and peakedness of the curve, respectively. This yielded the shape descriptors MTF_{skew} , $MTF_{kurtosis}$, NPS_{skew} , $NPS_{kurtosis}$, NEQ_{skew} , and $NEQ_{kurtosis}$.

Contrast detectability (CNR). As described in Section 3.2.3, CNR was measured on each of the five inserts (PVC, POM copolymer, nylon, polycarbonate, polystyrene) to characterize contrast detectability across a range of attenuations (CNR_{PVC} , CNR_{POM} , CNR_{PC} , CNR_{PA} , CNR_{PS}).

Each IQM was computed for all 170 imaging conditions.

Table 3.3: Image quality metrics calculated from measurements made with the Corgi phantom.

Image quality metric	Symbol	Unit
[Axi. Obl.] MTF's frequency at 10%	$\text{MTF}_{\{axi obl\}}f_{10}$	mm^{-1}
[Axi. Obl.] MTF's frequency at 25%	$\text{MTF}_{\{axi obl\}}f_{25}$	mm^{-1}
[Axi. Obl.] MTF's frequency at 50%	$\text{MTF}_{\{axi obl\}}f_{50}$	mm^{-1}
[Axi. Obl.] MTF's frequency at 75%	$\text{MTF}_{\{axi obl\}}f_{75}$	mm^{-1}
[Axi. Obl.] MTF's frequency at 90%	$\text{MTF}_{\{axi obl\}}f_{90}$	mm^{-1}
[Axi. Obl.] MTF's frequency at peak	$\text{MTF}_{\{axi obl\}}f_{peak}$	mm^{-1}
[Axi. Obl.] MTF peak	$\text{MTF}_{\{axi obl\}peak}$	1
[Axi. Obl.] MTF skew	$\text{MTF}_{\{axi obl\}skew}$	1
[Axi. Obl.] MTF kurtosis	$\text{MTF}_{\{axi obl\}kurtosis}$	1
Noise magnitude	σ	HU
NPS peak	NPS_{max}	HU^2mm^3
NPS's frequency at average	$\text{NPS}f_{avg}$	mm^{-1}
NPS's frequency at peak	$\text{NPS}f_{peak}$	mm^{-1}
NPS skew	NPS_{skew}	1
NPS kurtosis	$\text{NPS}_{kurtosis}$	1
NEQ maximum	NEQ_{max}	1
NEQ skew	NEQ_{skew}	1
NEQ kurtosis	$\text{NEQ}_{kurtosis}$	1
CNR of PVC	CNR_{PVC}	1
CNR of POM copolymer	CNR_{POM}	1
CNR of nylon	CNR_{PC}	1
CNR of polycarbonate	CNR_{PA}	1
CNR of polystyrene	CNR_{PS}	1

3.2.5 Radiomic feature measurements

The radiomic features were measured using the texture phantom described in the Section 3.2.1.2.

Circular regions of interest (ROIs) with standardized dimensions were placed within four texture inserts at three central, non-adjacent axial locations as shown in Figure 3.5 Sampling central, non-adjacent slices mitigated potential artifacts from slice-to-slice correlation or edge effects.

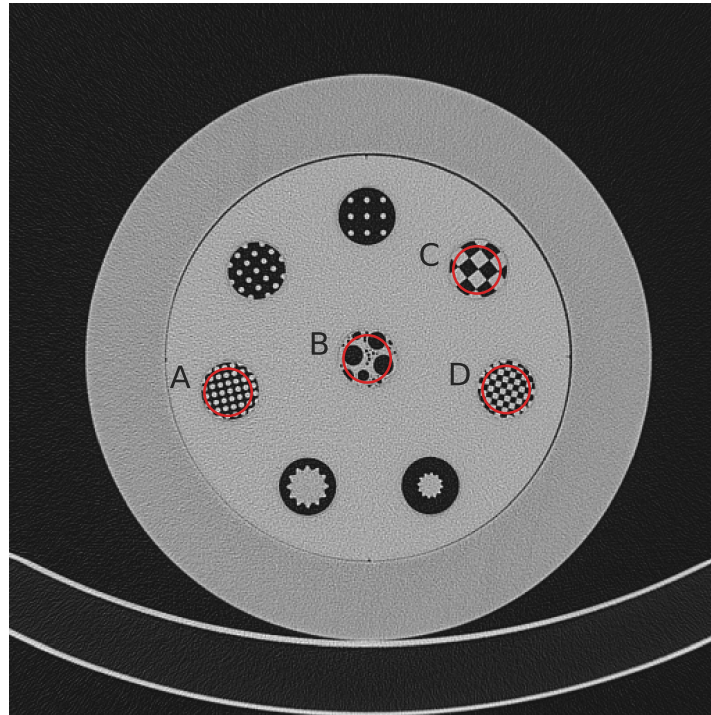


Figure 3.5: Texture phantom. ROIs and labels placed on the four texture inserts used in experiments. (A) circular rods, (B) random circles, (C) checkerboard, and (D) square grid.

The radiomic features were extracted from each ROI using PyRadiomics software [76]. First order and second order texture features were extracted, for a total of 94 QIFs. Shape features

were excluded, as all ROIs were identical. The categories consisted of First Order Statistics (First Order), Gray Level Co-occurrence Matrix (GLCM), Gray Level Run Length Matrix (GLRLM), Gray Level Size Zone Matrix (GLSZM), Gray Level Dependence Matrix (GLDM), and Neighboring Gray Tone Difference Matrix (NGTDM), as described in Table 3.4. A complete list of the radiomic features used can be found in Table C.1. Features comply with feature definitions as described by the Imaging Biomarker Standardization Initiative (IBSI) [59], to improve the reproducibility and generalizability of our results.

Table 3.4: Radiomics feature categories, descriptions, and number of features extracted using Pyradiomics [76].

Category	Description	Count
First Order Statistics (First Order)	Characterize the distribution of voxel intensities without regard to spatial relationships, including measures of central tendency, dispersion, and distribution shape.	19
Gray Level Co-occurrence Matrix (GLCM)	Quantify spatial relationships between neighboring voxels, capturing local texture patterns and directional dependencies.	24
Gray Level Run Length Matrix (GLRLM)	Characterize consecutive voxels with identical gray levels, sensitive to linear texture patterns and directionality.	16
Gray Level Size Zone Matrix (GLSZM)	Quantify regions of connected voxels with identical gray levels, capturing information about texture homogeneity and region sizes.	16
Gray Level Dependence Matrix (GLDM)	Characterize dependence of voxels at specified distances, capturing larger-scale texture patterns.	14
Neighboring Gray Tone Difference Matrix (NGTDM)	Assess local intensity differences, providing measures of texture coarseness and complexity.	5
Total features		94

3.3 Results

3.3.1 Image quality measurements and derived metrics

The MTF curves measure spatial resolution of the images as a function of spatial frequency. The NPS curves provide a measure of the images' noise as a function of spatial frequency, where the overall magnitude corresponds to the image's noise magnitude, and the shape of the curve corresponds to the noise texture.

The measured image quality curves from the highest dose protocols of the dataset ($\text{CTDI}_{vol} \approx 17$) are shown in Figure 3.6. Evaluating a single dose level is sufficient to capture the majority of the dataset's range in image quality, as varying dose does not impact MTF or noise texture (only noise magnitude). The measured axial and oblique MTF curves demonstrate a wide range of spatial resolution, consistent with the range of reconstruction kernels included in the dataset (ranging from smoothing to sharpening). Peak MTF values exceeding 1.0 indicate edge-enhancing reconstruction. The oblique MTF curves (middle) are compressed to lower spatial frequencies compared to the axial MTF curves (top), reflecting the reduced resolution in the slice-thickness direction relative to the in-plane direction. The NPS curves (bottom) illustrate the expected tradeoff between spatial resolution and noise: sharper kernels, which yield higher spatial resolution, are associated with substantially elevated noise magnitude and a shift of noise power toward higher spatial frequencies. Line style denotes scanner model (see legend), with multiple curves per scanner corresponding to the set of reconstruction kernels measured; the four scanners span a broadly comparable range of imaging conditions. Taken together, these curves confirm that

the imaging conditions sampled span a meaningful range of the resolution-noise tradeoff space, providing a diverse set of protocols for evaluating relationships between image quality and texture images.

Figure 3.7 shows the distributions of a representative selection of the extracted IQMs across the 170 imaging conditions. The conditions span a wide range of both spatial resolution and noise confirming that the dataset captures a meaningful diversity of image quality. Negative CNR corresponds to $\mu_{\text{ROI}} < \mu_{\text{BACKGROUND}}$. This range in image quality, and selection of metrics that capture that variation, enabling the characterization of IQM–radiomic relationships in the following chapters.

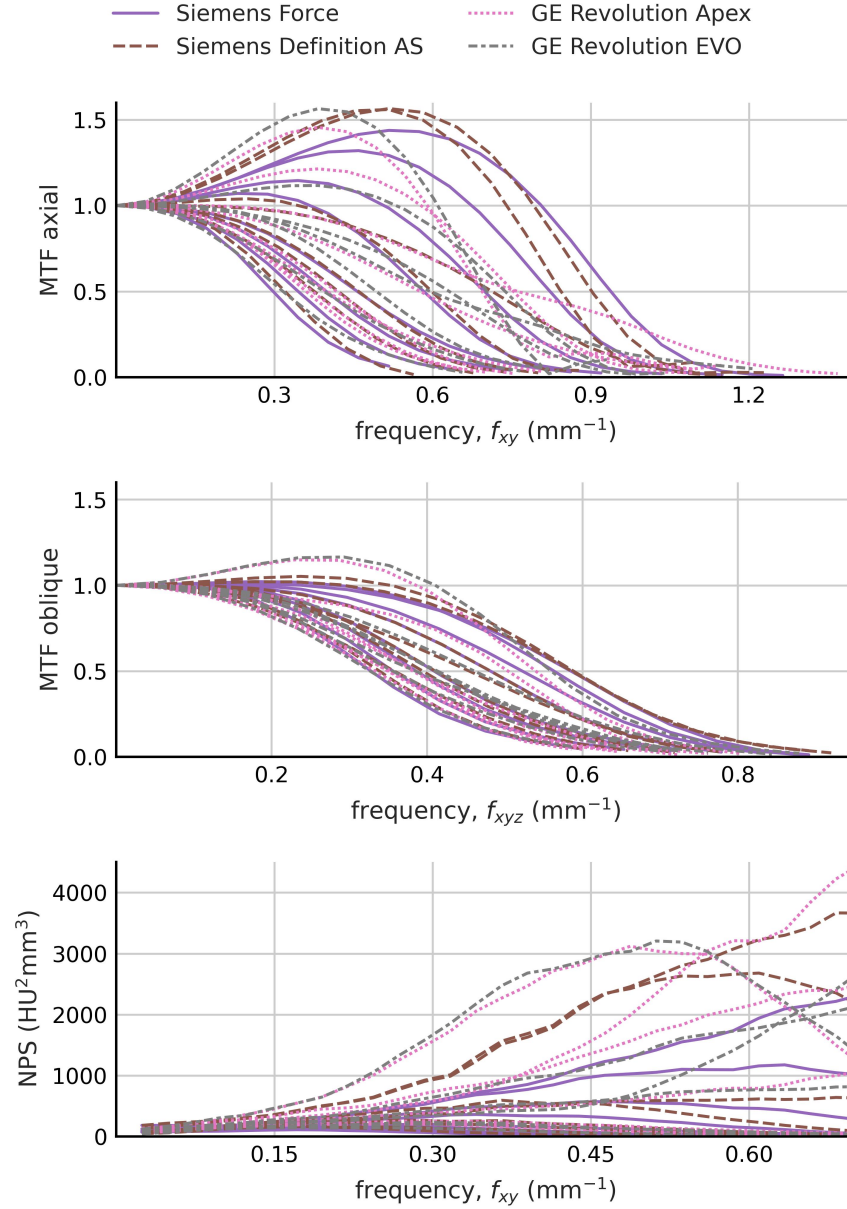
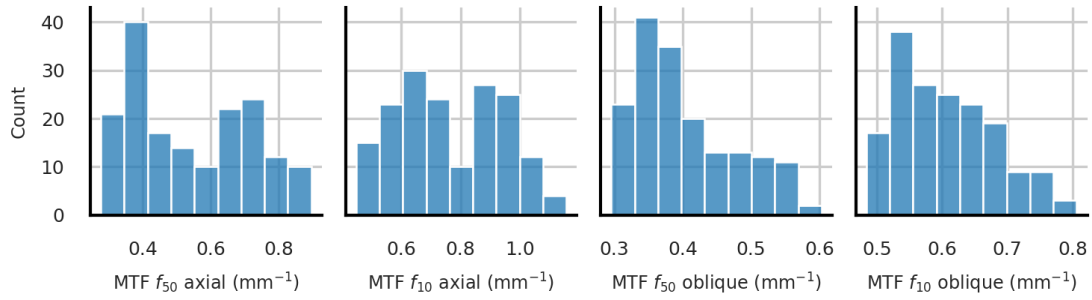
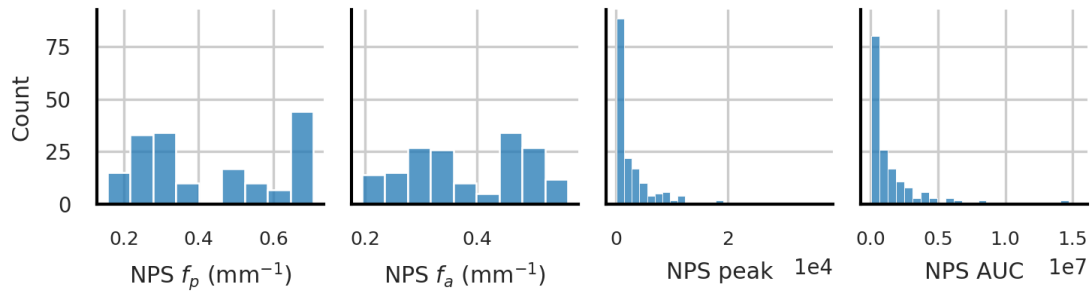


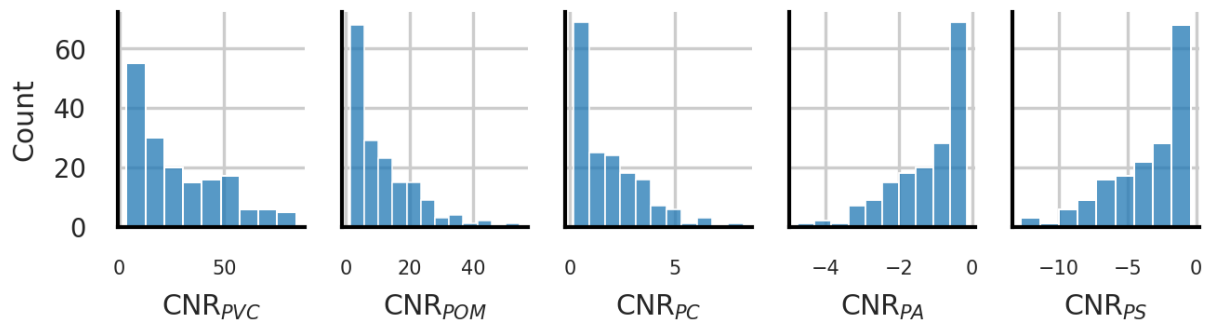
Figure 3.6: Measured image quality data from the highest CTDI_{vol} protocols of the dataset (≈ 17 mGy). Each line corresponds to a single imaging condition. From top to bottom: axial MTF, oblique MTF, axial NPS. These curves illustrate the range in CT protocol represented in the dataset.



(a) Distribution of image quality metrics derived from the modulation transfer function (MTF) measured from the dataset.



(b) Distribution of image quality metrics derived from the noise power spectrum (NPS) measured from the dataset.



(c) Distribution of contrast-to-noise (CNR) values measured from the dataset.

Figure 3.7: Distributions of a selection of the extracted IQMs across the 170 protocols.

3.3.2 Texture phantom image data

The image data from the texture phantom across the 170 protocols illustrates the impact of imaging protocol on the image appearance qualitatively. Figure 3.8 shows close-up images of the four texture objects for a selection of protocols from the highest dose level. For all four texture objects, labeled A through D, variations in the reconstruction kernel substantially alter the sharpness and noise. Smooth kernels, such as Apex SOFT and Force Br32d, yield images with reduced visible noise but present blurred structural boundaries across the checkerboard and circular patterns. Sharper kernels, such as the Apex EDGE or Force Br64d, enhance edge definition and preserve high-spatial-frequency details, but also increase the amount of noise in the image.

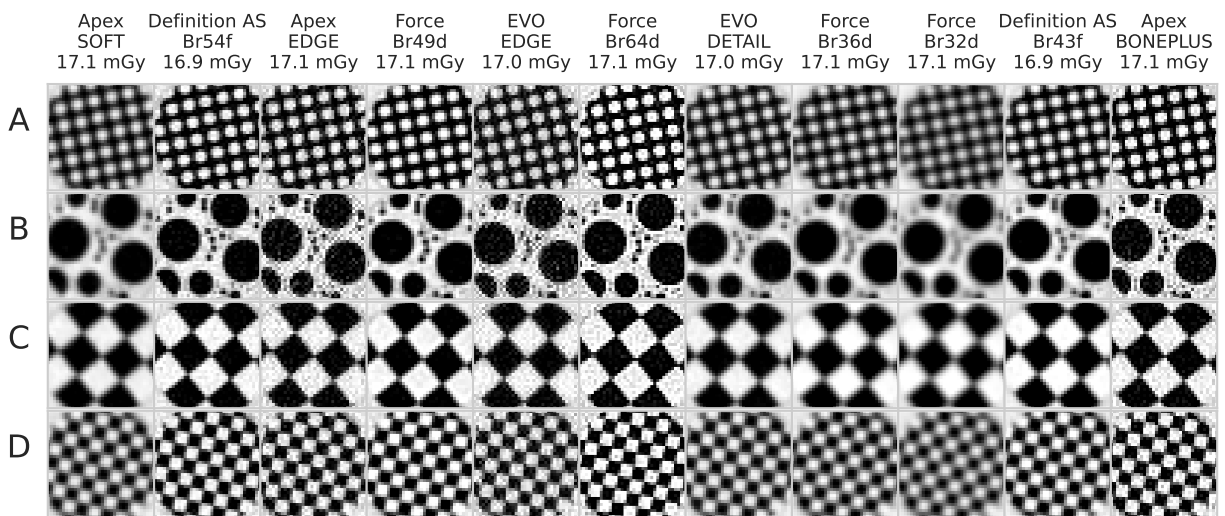
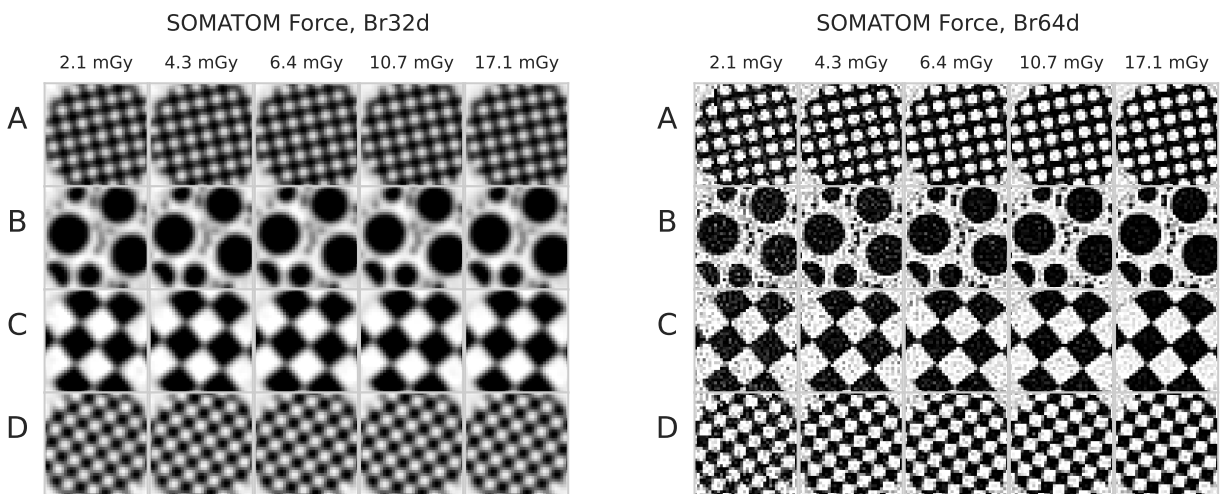


Figure 3.8: Close-up view of four of the texture objects (labeled A, B, C, D) on CT from a random selection of 12 protocols from the highest dose level.

Figure 3.9 shows close-up images of the four texture objects at each dose level for a smooth kernel (a) and a sharp kernel (b) to show the combined influence of radiation dose and kernel is

further elucidated by comparing acquisitions across varying dose levels.

Figure 3.9 shows the effect of dose reduction from 17.1 mGy down to 2.1 mGy for a smooth kernel (Br32d) and a sharp kernel (Br64d) from the Force scanner. For the smooth kernel (a), the qualitative appearance of the texture objects remains relatively uniform across the dose range, as the smoothing properties of the algorithm compensate for the increase in noise due to dose reduction. The sharp kernel (b) displays a stronger sensitivity to radiation dose. As the dose decreases, the noise increases substantially.



(a) Images across the five dose levels used reconstructed with a Br32d kernel.

(b) Images across the five dose levels used reconstructed with a Br64d kernel.

Figure 3.9: Close-up view of four texture objects (labeled A, B, C, D) on CT from the range of radiation dose levels acquired. The smooth kernel (a) shows less variation than the sharp kernel (b).

3.3.3 Radiomics data

Feature sensitivity to protocol variations showed significant heterogeneity. Figure 3.10 shows kernel density estimates (KDEs) of a selection of radiomic features, grouped by feature family, with separate distributions for each of the four texture inserts (A–D). The separation of the four distributions (textures) varied by feature. For example, first order mean and median, and several GLCM and GLRLM features show distinct, only partially overlapping distributions across textures, indicating the ability to discriminate between the texture objects. Other features, including GLCM MaximumProbability, GLSZM ZonePercentage, and GLDM GrayLevelNonUniformity, show substantial overlap across all four textures.

The width of each per-texture distribution reflects the variation in that feature's value across the 170 imaging protocols. For several features, such as GLDM GrayLevelVariance and GLRLM RunEntropy, this protocol-induced spread is large enough that the per-texture distributions are no longer clearly separable, indicating that CT protocol can introduce variation similar in scale to the variation that distinguishes one physical texture from another.

The measured distributions of radiomic feature values for all features are provided in Appendix D.

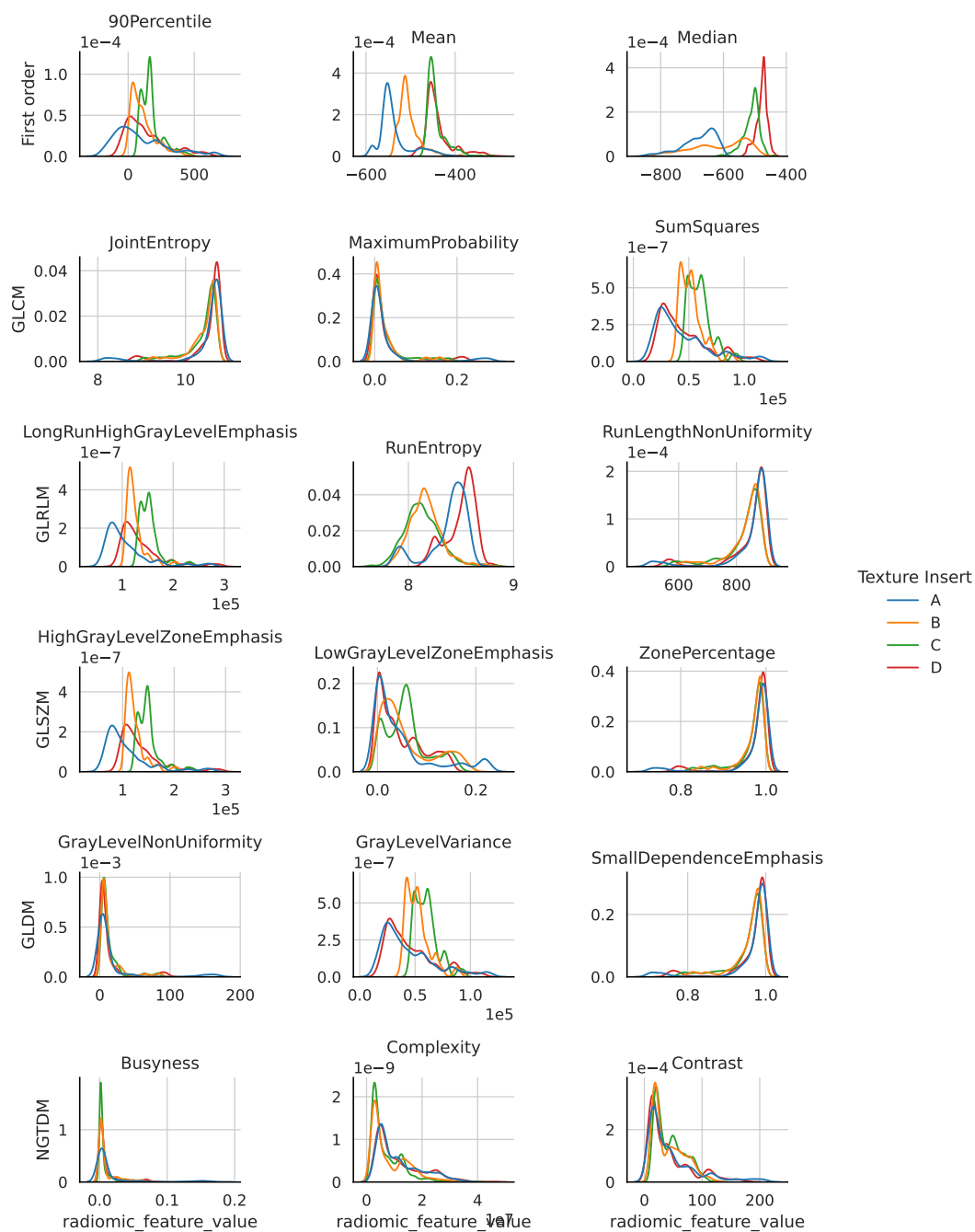


Figure 3.10: Distributions (KDEs) of a selection of radiomic features across all 170 protocols. Each row contains three features for a given feature family. For each subplot, x -axis is feature value and y -axis is kernel density estimate.

3.4 Public release of the dataset

As part of this effort, a version of this dataset was curated and publicly released [95]. The paired texture phantom and image quality (PTP-IQ) includes DICOM image data and HDF5 files containing image quality measurements, but does not include the derived image quality metrics or the extracted radiomic feature values discussed in this Chapter. The public dataset includes a second texture phantom (a radiomics liver phantom [64]) that was not used within this dissertation. All files have been made available via Zenodo (<https://doi.org/10.5281/zenodo.16820900>) [96].

While the radiomics liver phantom provides additional textures and modules that are distinct from the 3D-printed phantom included in this Chapter, it is a custom-made object that was not designed to be reproducible. Because reproducibility and open accessibility were prioritized throughout this dissertation, we restricted our analyses to the 3D-printed phantom (presented in Chapter 2), whose design can be reproduced by other groups from the publicly available specification.

The public dataset includes 154 of the 170 imaging conditions described in this Chapter. This is because (1) the public dataset was restricted to five dose levels per scanner, and (2) certain conditions that were not acquired for the radiomics liver phantom.

The total dataset size is 17.57 GB uncompressed. Prior to upload, each subdirectory was compressed using gzip (<https://www.gnu.org/software/gzip/>), after which the total size was 9.59 GB. The repository is organized as follows:

```
ptp-iq-dataset/  
  
  metadata.csv  
  
  radiomics-liver-phantom/  
  
    {protocol_id}/  
  
      {instance_number}.dcm  
  
  3d-printed-texture-phantom/  
  
    {protocol_id}/  
  
      {instance_number}.dcm  
  
  image-quality-data/  
  
    {protocol_id}.h5  
  
  examples/  
  
    example-data/  
  
      environment.yaml  
  
      data_reading_example.ipynb
```

The use of the DICOM data format, HDF5 file format, gzip compression, and Zenodo repository align with Findable, Accessible, Interoperable, Reusable (FAIR) data management principles [97].

3.5 Discussion

This dataset provides a unique resource for the quantitative imaging community by enabling direct investigation of the quantitative relationship between fundamental image quality metrics and image textures. The systematic variation of acquisition parameters across multiple scanner platforms creates a controlled experimental framework for addressing key challenges in reproducibility and robustness for quantitative imaging.

One important application of this dataset is in feature robustness analysis, where image features (e.g., radiomics or handcrafted features) can be extracted and evaluated for stability across a range of acquisition and reconstruction protocols. This can inform the selection of QIFs that are reproducible for some relevant range of protocols. Conversely, the dataset may also inform protocol optimization for specific quantitative imaging tasks: if particular features are known to be predictive at a given operating point, one could determine the protocol conditions under which those features remain stable, guiding protocol recommendations tailored to a given clinical application.

More specifically, this dataset enables the analysis of feature robustness and reproducibility in terms of measurable descriptors of image quality, rather than relying solely on protocol descriptions that depend on manufacturer-specific terminology. While image quality measurements themselves are the multi-factorial result of radiation dose, reconstruction kernel, and vendor-specific system design, they provide a common, quantitative basis for characterizing the imaging conditions.

Currently, the reliance on categorical labels (e.g., ‘Br40d’ vs. ‘Bone’ kernels) limits the gen-

eralizability of radiomic studies, as these labels lack a physical definition and vary significantly between manufacturers. We hypothesize that physical image quality measurements provide a standardized, manufacturer-independent coordinate system for comparing image acquisition and reconstruction conditions. By studying radiomic features within this physical space, we collapse the high-dimensional variability of proprietary hardware and software into a continuous set of comparable metrics that will enable a more direct translation and application of research results between studies. These approaches could be particularly important for achieving multi-center clinical deployment of quantitative imaging-based predictive models; an objective that has remained challenging despite promising results in research studies.

In practice, the image quality metrics provided in this dataset could be incorporated into harmonization pipelines in several ways. Physical metrics such as MTF and NPS have been proposed as a basis for harmonization through prospective protocol selection as a way of reducing variation in radiomic features across scanner models [42], [43]. These metrics have also been proposed as a basis for analytically predicting and correcting protocol-induced variation in radiomic features. In this framework, MTF and NPS serve as quantitative inputs that characterize the resolution and noise state of each imaging condition, enabling feature-level corrections that are grounded in the underlying physics rather than dependent on categorical protocol labels [44], [45]. The DICOM image subset alone also constitutes a valuable resource for validating or benchmarking image-based harmonization and domain adaptation techniques: because the physical object being scanned is held constant across all imaging conditions, protocol-induced variation is the sole source of differences between images, providing a degree of experimental control not achievable

with clinical data [98], [99], [100].

The remainder of this dissertation explores some of these applications directly. In Chapter 4, we use the dataset to characterize the relationships between physical image quality metrics and radiomic feature values across the full protocol space. In Chapter 5, we develop and evaluate models that predict radiomic feature variation from these metrics. Finally, in Chapter 6, we extend the framework to predict the out-of-distribution performance of a deep learning model across CT imaging conditions.

A number of potential extensions to this dataset are possible. Future contributions and collaborations are supported by hosting on Zenodo, a stable open-access repository maintained by CERN, which provides persistent archival and a permanent DOI to ensure the long-term accessibility of this dataset. For example, external research groups could contribute annotations or region-of-interest segmentations for the texture phantoms. This could be extended to the release of the associated pre-computed radiomic feature measurements to supplement the existing data.

Extending the dataset to additional texture objects is possible if a group has access to one of the scanners used in the dataset. With careful matching of the protocol, the dataset's image quality data could be directly applied to study different texture phantoms. As the phantoms used in this study consist of relatively simple texture patterns, extending to texture patterns representing specific pathologies, or anthropomorphic phantom configurations for anatomical realism could be of interest for targeted clinical investigations. For example, PixelPrint enables the 3D printing of patient-specific lung phantoms with accurate attenuation profiles and textures [70].

This dataset differs from previously published CT phantom datasets in several important respects. For example, Kalendralis et al. released a multicenter CT phantom dataset intended for radiomics reproducibility tests [62]. However, Kalendralis et al. dataset used the COPDGene Phantom II (Phantom Laboratory, Salem, NY) and the Abdominal Phantom (CIRS, Norfolk, VA) which are composed of uniform density materials and were not designed for radiomics studies. Additionally, their dataset does not include image quality measurements. The Credence Cartridge Radiomics (CCR) Phantom dataset includes a variety of textures and was explicitly crafted to study radiomics features [13]. However, it also lacks accompanying measurements of image quality metrics, limiting its utility for studies that seek to correlate texture-based features with objective image quality.

There are several limitations of the dataset that should be acknowledged. First, while the dataset captures protocol variability across scanner models, manufacturers, dose levels, and reconstruction kernels, other sources of variability were not systematically explored. These factors, such as slice thickness, beam energy, patient size, or positioning, can also significantly influence image appearance and feature values and should be considered in future studies.

The phantoms used in this study represent a limited set of texture patterns. While they provide a controlled and reproducible means of assessing texture sensitivity, they may not fully capture the complexity and heterogeneity found in biological tissues. There are currently no standard reference objects for assessing image texture in CT, and most commercially available phantoms are composed of uniform materials. The texture objects included in this dataset represent a preliminary step toward filling that gap, but additional efforts are needed to develop

and standardize texture phantoms for quantitative imaging.

This dataset does not include images reconstructed using iterative reconstruction algorithms. Although iterative methods are widely used in clinical practice, their non-linear behavior makes it more challenging to characterize in terms of image quality. Future extensions of the dataset may consider inclusion of iterative reconstructions, along with additional measures of image quality required to characterize them, such as contrast- and noise-dependent measures of spatial resolution, noise nonuniformity index, and noise inhomogeneity index; this may require defining a task-specific context [30].

3.6 Conclusions

We have provided a public dataset of CT images from two texture phantoms that have been acquired from multiple scanner models and a wide range of acquisition protocols, along with the associated image quality measurements from each of those scanner models and protocols. This unique resource enables systematic investigation of relationships between fundamental image quality metrics and quantitative imaging features, addressing a critical need in radiomics harmonization research. The dataset’s standardized acquisition design, paired image quality measurements, and open-source components facilitate reproducible research toward robust quantitative CT imaging. By making the PTP-IQ dataset freely available, we aim to accelerate development of characterization and harmonization methods essential for clinical translation of QIF-based biomarkers.

In the remainder of this dissertation, we use this dataset to characterize and model radiomic feature variation due to CT protocol using physical image quality (Chapters 4 and 5), and to predict the effect of CT protocol variation on the out-of-distribution performance of a deep learning model (Chapter 6).

4 Investigating the effect of CT protocol-induced variation on radiomic texture features using physical image quality metrics

4.1 Introduction

Radiomic features are quantitative descriptors of intensity, texture, and shape extracted from regions of interest within medical images, and have shown promise as imaging biomarkers of disease status and progression [2], [3]. Their clinical translation is limited, however, by sensitivity to CT acquisition and reconstruction protocol: the same physical object imaged under different scanner models, reconstruction kernels, or dose levels can yield substantially different feature values, and for texture features this protocol-induced variation can rival or exceed the biological differences the features are meant to capture [13].

In Chapter 1, we proposed approaching that variation not by reducing a biomarker’s sensitivity to it, as harmonization, standardization, and stability-based feature selection do, but by characterizing and predicting it in terms of physical, scanner-independent image quality metrics. In Chapter 3 we developed a dataset designed to enable this, but did not analyze it. It comprises paired CT scans of a texture phantom (presented in Chapter 2) and an image quality phantom

(Corgi, The Phantom Laboratory), acquired under 170 protocols spanning four scanner models from two vendors (GE and Siemens), a range of dose levels (CTDI_{vol} approximately 2 to 17 mGy), and seven or eight reconstruction kernels per scanner. For each protocol, 3D image quality measurements (MTF, NPS, NEQ, CNR) were made from the image quality phantom, and radiomic features extracted from the texture objects under identical conditions. This chapter is an initial characterization of that dataset: it reduces the radiomic feature and image quality metric spaces to representative subsets, builds intuition for how radiomic features respond to image quality, and establishes that systematic relationships between the two exist.

The aims of this chapter are to: (1) reduce the radiomic feature and image quality metric spaces to non-redundant representative subsets; and (2) demonstrate that radiomic features respond systematically to variation in measured image quality. We additionally examine whether these responses are consistent across the four texture objects, which would suggest the relationships reflect general properties of the feature definitions rather than texture-specific effects.

The analyses in this chapter are primarily descriptive: the aim is to establish that the relationships are present and to characterize their form, not yet to model them. Establishing that these relationships exist, and characterizing their form, builds the foundation for the predictive modeling pursued in Chapter 5.

4.2 Methods

4.2.1 Dataset

The dataset used was described in Chapter 3, and consisted of CT scans of a radiomics texture phantom with paired image quality data for 170 protocols spanning four scanner models from two vendors (GE and Siemens). $CTDI_{vol}$ ranged from approximately 2 to 17 mGy, sampled at five levels on the Siemens scanners and six on the GE scanners. Eight reconstruction kernels were acquired per scanner, except the GE Revolution EVO, which had seven. This gave between 40 and 48 protocols per scanner; the full set of dose levels and kernels for each is listed in Table 3.1.

4.2.2 Radiomic feature selection

As an initial step to reduce the large radiomic feature space, we eliminated features that were not reproducible between slices. This is a common first step in radiomics studies, as high intra-class variability is not recommended for imaging biomarkers [1].

We calculated the coefficient of variation (CV) across slices for each texture object. The CV is given as

$$CV_{slice} = \frac{\sigma_{slices}}{\mu_{slices}} \quad (4.1)$$

where σ_{slices} is the standard deviation and μ_{slices} is the mean calculated over three non-adjacent slices. We eliminated features with a $CV > 10\%$ [101].

In addition, four radiomic features were selected to serve as representative examples for the

initial investigation. Hierarchical clustering on the features' Spearman rank-order correlations was performed using Ward's linkage [102]. Spearman correlation was preferred over Pearson correlation for robustness to non-linear monotonic relationships between features. The distance between any two metrics was defined as $1 - |\rho|$, where ρ is the Spearman correlation coefficient, so that strongly correlated metrics in either direction are treated as close. Flat clusters were formed by cutting the dendrogram at a distance threshold of $t = 1.0$, corresponding to a minimum absolute Spearman correlation of $|\rho| > 1 - t$ for any two metrics to be assigned to the same cluster. All implementations were done with the Python scikit-learn package [79].

4.2.3 Image quality metric selection

A total of 32 image quality metrics (IQMs) were computed from the MTF, NPS, CNR, and NEQ measurements described in Chapter 3, summarized in Table 3.3. Many of these metrics are expected to be highly correlated: for example, MTF evaluated at successive spatial frequency points (f_{10} , f_{25} , f_{50} , f_{90}) along the same measurement direction will vary together as the reconstruction kernel changes. Retaining all 32 metrics would introduce substantial redundancy into the regression models developed in Chapter 5, inflating the feature space without adding independent information.

To reduce the feature space, we performed hierarchical clustering on the Spearman rank-order correlation matrix of the 32 IQMs, computed across all $N = 170$ protocols as described in Section 4.2.2. Hierarchical clustering was performed using average linkage, in which the distance

between two clusters is the mean pairwise distance between all members of each cluster, with $t = 0.10$.

One representative metric was selected from each cluster as the metric with the highest variance across all 170 protocols, computed on min-max scaled data. Variance was used as the selection criterion in preference to a target-dependent measure such as mutual information, so that the selected IQM set is fixed across all radiomic features and texture labels rather than optimized for a specific modeling target.

4.2.4 Effect of image quality on texture appearance

To visualize how image quality variation affects the appearance of CT texture images, we selected a subset of 12 protocols representing a range of measured MTF and NPS values in the dataset. For each protocol, we extracted representative image patches from the four texture objects (A, B, C, D) and displayed them side-by-side.

We created visualizations of the protocols displayed in random order to show the diversity of image appearance across the dataset, and compared those to visualizations of the protocols sorted by a single image quality metric to demonstrate systematic trends in texture appearance as a function of that metric. These visualizations provide qualitative evidence that changes in spatial resolution and noise texture produce visually perceptible and systematic changes in texture appearance that correspond to changes in measured image quality.

4.2.5 Radiomic feature dependence on image quality metrics

To quantitatively assess how radiomic features respond to image quality variation, we created scatter plots of radiomic feature values versus individual image quality metrics. For each of the four selected radiomic features, the feature values extracted from each texture object were plotted as a function of selected image quality metrics.

Each point in these scatter plots represents a single protocol, with color indicating which texture object (A, B, C, or D) the feature was extracted from. This visualization approach allows us to assess: (1) whether relationships between image quality and radiomic features exist, (2) the functional form of these relationships (linear, exponential, logarithmic, etc.), (3) whether relationships are consistent across different texture types (indicated by similar trends for all four colors), and (4) the relative strength of different image quality metrics as predictors of radiomic feature variation.

Finally, the Spearman correlation coefficient was calculated between each radiomic feature-texture object target and each selected IQM and visualized with a heatmap.

4.3 Results

4.3.1 Radiomic feature selection

A total of 52 features (out of the original 89) remained after applying the *CV* threshold of 10%. This eliminated features with greater inter-slice variability in order to better isolate the impact of

image quality as a source of variation.

From this subset of 52 radiomic features, four representative features were selected for illustrative analyses. This was done based on the results of the hierarchical clustering shown in Figure 4.1. A threshold of 1.0 was applied to form the clusters. The full Spearman correlation matrix heatmap is provided in Appendix E.

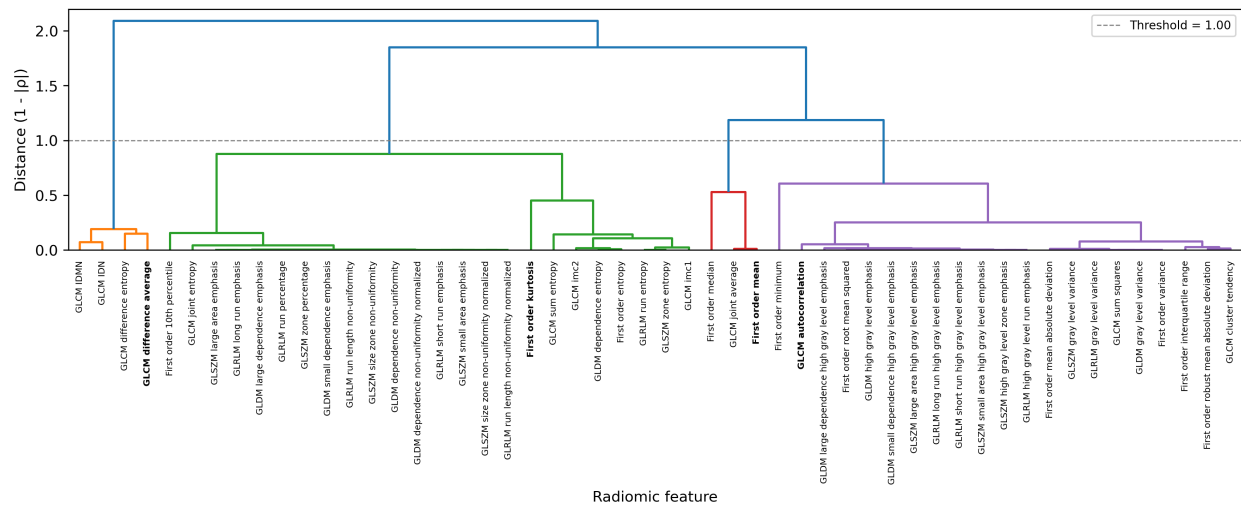


Figure 4.1: Dendrogram of the clustering of the radiomic features. The hierarchical clustering is performed using Ward's linkage. The representative features are bolded.

Based on these results, the representative radiomic features selected were:

1. First order mean
2. First order kurtosis
3. GLCM autocorrelation
4. GLCM inverse difference normalized (IDN)

4.3.2 Image quality metric selection

This procedure reduced the 32 IQMs to $N_{IQM} = 10$ representative metrics. The dendrogram with the clusters indicated by color is shown in Figure 4.2. The selected set are indicated with bolding, and listed in Table 4.1. The selected set spans all major measurement categories: axial and oblique MTF, NPS, NEQ and CNR, confirming that the clustering procedure did not preferentially discard any physical measurement type. The full Spearman correlation matrix heatmap is provided in Appendix E.

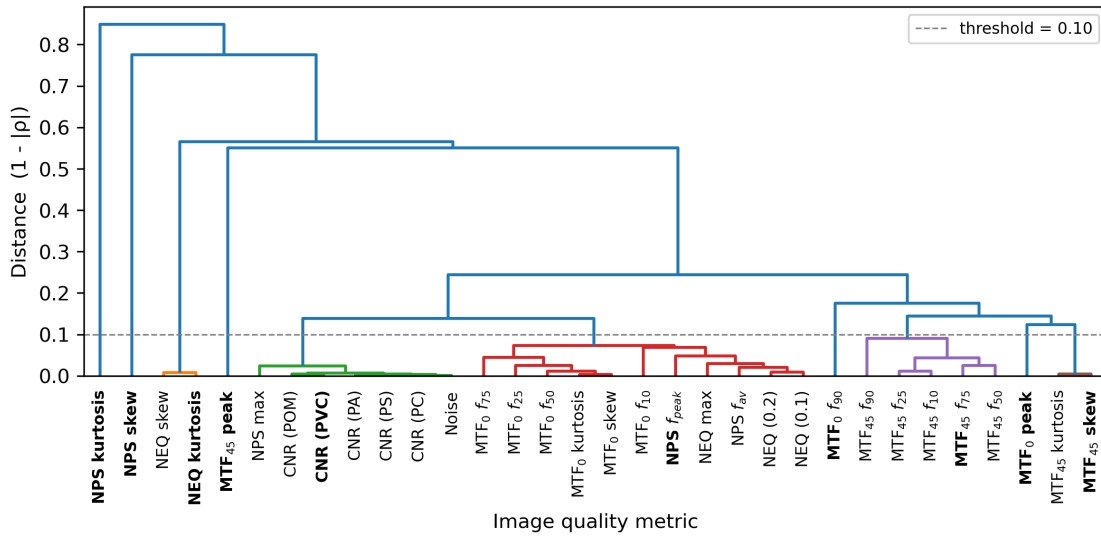


Figure 4.2: Dendrogram of the image quality metrics showing the clusters. The 10 selected IQMs are shown bolded.

Table 4.1: Representative image quality metrics selected after hierarchical clustering of the full 32-metric IQM set. One representative was selected from each cluster as the metric with the highest variance across all 170 protocols.

Category	Metric	Description
MTF axial (0°)	$MTF_0 f_{90}$	Spatial frequency at which the axial MTF falls to 90%
	$MTF_0 peak$	Peak value of the axial MTF
MTF oblique (45°)	$MTF_{45} f_{75}$	Spatial frequency at which the oblique MTF falls to 75%
	$MTF_{45} peak$	Peak value of the oblique MTF
	$MTF_{45} skew$	Skewness of the oblique MTF curve
NPS	$NPS f_{peak}$	Spatial frequency at which the NPS reaches its maximum
	$NPS skew$	Skewness of the NPS curve
	$NPS kurtosis$	Kurtosis of the NPS curve
NEQ	$NEQ kurtosis$	Kurtosis of the NEQ curve
CNR	$CNR (PVC)$	Contrast-to-noise ratio measured on the PVC insert

4.3.3 Effect of image quality on texture appearance

Figure 4.3 shows texture images from 12 randomly selected protocols, illustrating the substantial diversity in image appearance across the dataset despite imaging identical physical objects. Visual differences are apparent in both noise texture (granularity and magnitude) and spatial resolution (sharpness of texture pattern edges).

Figure 4.4 presents the same protocols sorted by $MTF_{45} f_{75}$ (top) and noise magnitude (bottom). When sorted by $MTF_{45} f_{75}$, a clear progression is visible from blurred texture patterns at low MTF values (left) to increasingly sharp, well-defined patterns at high MTF values (right). This is particularly evident in texture objects A and D, which contain high-frequency spatial patterns. An apparent exception can be seen when comparing the EVO BONE 17.0 mGy column to the neighboring Force Br40d 10.7 mGy column: despite both having an $MTF_{45} f_{75}$ value of

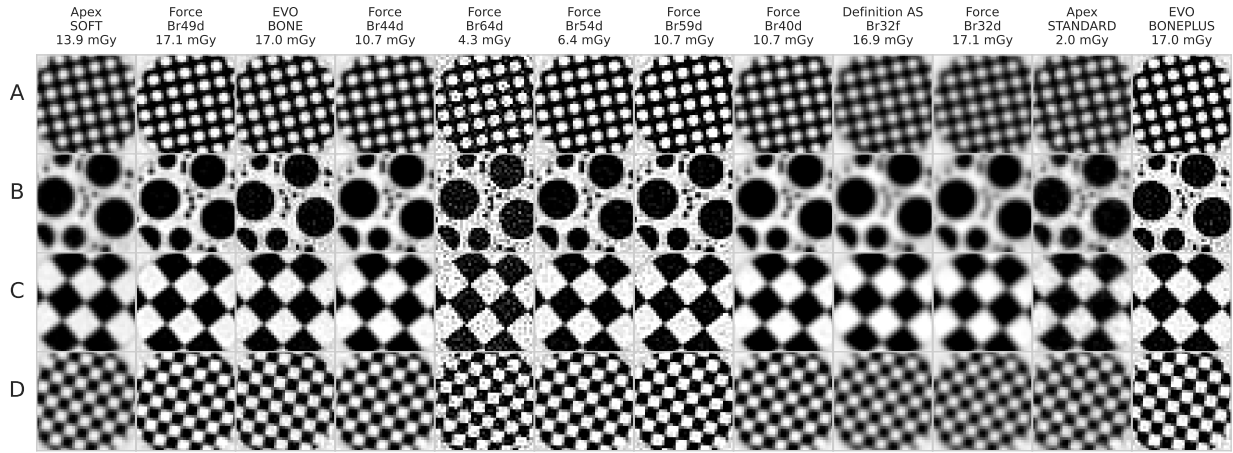


Figure 4.3: Close-up view of four of the texture objects (labeled A, B, C, D) on CT from 12 randomly selected protocols (selected from the 170 total).

0.30 mm^{-1} , the Force Br40d appears noticeably more blurred than the EVO BONE.

When sorted by noise magnitude, the progression shows increasing noise visibility from left to right, with smoother images at low noise values becoming progressively grainier at higher noise values. This is most noticeable in texture object C, which has relatively uniform regions that make noise more perceptible. Again, there is an apparent exception to the visual assessment: the Apex STANDARD 2.0 mGy column is much softer visually than its neighbors. This protocol corresponds to the lowest dose level included in the dataset, which happens to be the only one of the 12 protocols included at that dose level, resulting in a higher measured noise magnitude than the higher dose protocols it is being compared to, despite a relatively smoother kernel.

These visualizations demonstrate that measured image quality metrics (MTF, NPS) correspond to systematic changes in image texture, and that this correspondence captures relationships more subtle than visual inspection or protocol labels can easily resolve. While overall trends in

sharpness and noise can be visually distinguished, we cannot reliably order protocols whose resolution or noise differ only slightly, and visual assessment can conflate the two, as illustrated by the protocols that appear out of sequence when Figure 4.4 is sorted by a single metric. Manufacturer protocol labels provide only an approximate ordering: kernel and dose descriptors are categorical, vendor-specific, and combine several physical effects, so that protocols that share a nominal label can produce measurably different MTF and NPS. The measured metrics instead place every protocol on a common, continuous physical axis that captures these differences. Because radiomic features are mathematical quantifications of the same texture patterns, we expect them to respond systematically to these measured image quality characteristics.

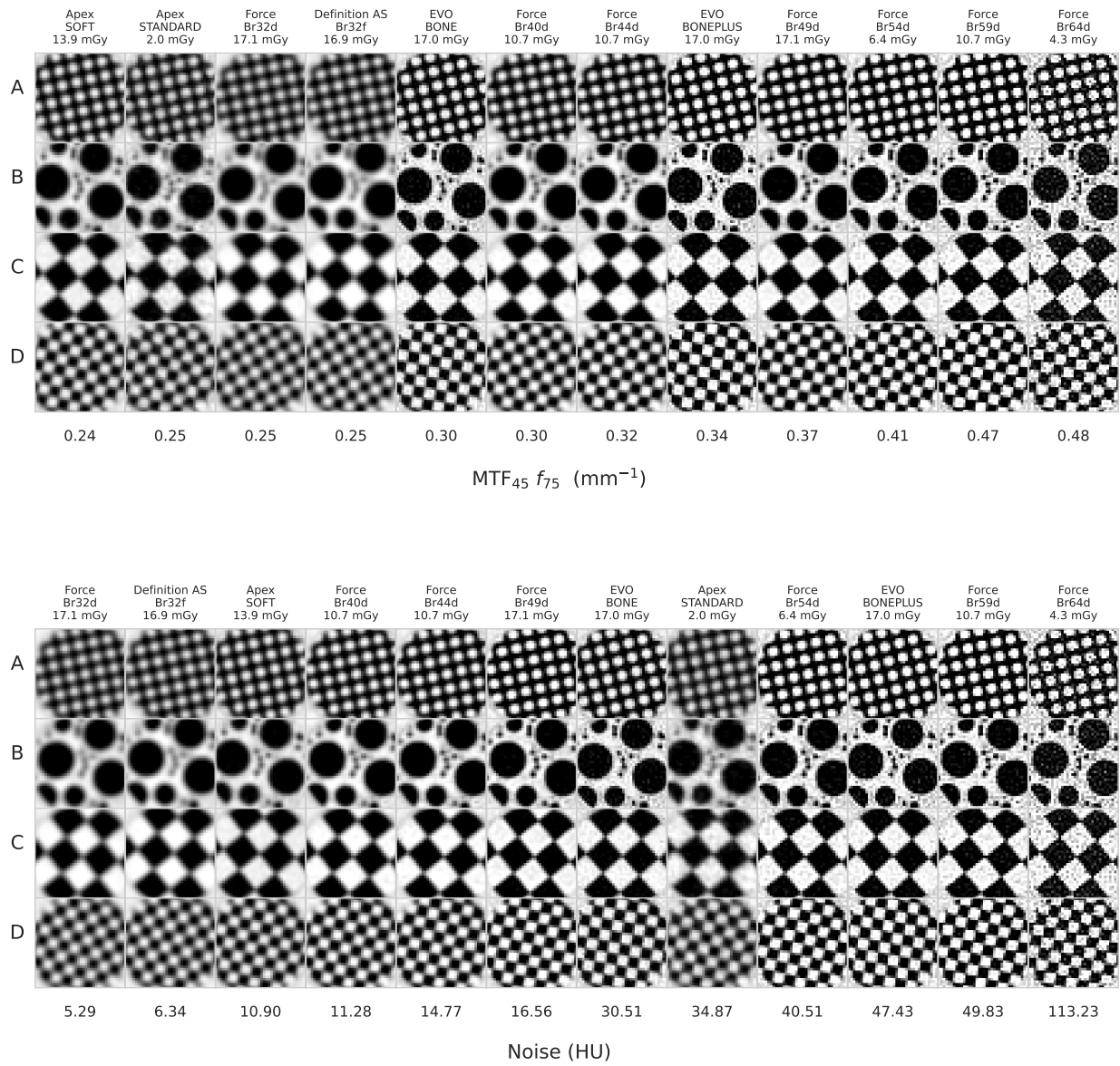


Figure 4.4: Close-up view of four of the texture objects (labeled A, B, C, D) on CT from a range of 12 protocols (selected from the 170 total). Images are shown sorted by $MTF_{45} f_{75}$ (top), and noise magnitude (bottom).

4.3.4 Radiomic feature dependence on image quality metrics

Figure 4.5 plots the four selected radiomic features against representative IQMs, and shows clear systematic relationships, although the form differs by feature. First order mean and GLCM autocorrelation vary monotonically with the resolution metrics, decreasing as MTF_{45} skew increases. GLCM IDN rises steeply with CNR and then saturates, approaching a plateau at high CNR. The MTF_{45} peak column shows a discretization, with protocols falling into a few vertical bands rather than spanning a continuous range—this is expected, as the peak MTF value is 1.0 for protocols reconstructed without an edge-enhancing kernel. Overall, the general form of the relationships between the radiomic features and IQMs is shared across the four textures, though first order kurtosis differs: textures A and D respond similarly to each other, as do B and C, but the two pairs respond differently, most notably to CNR (the first column).

Figure 4.6 summarizes the rank correlations between the four representative radiomic features and the selected IQMs, computed separately for each texture object. Overall, the correlations are strongest for the resolution metrics, CNR, and NPS f_{peak} , weaker NEQ kurtosis, and weakest for NPS skew and NPS kurtosis. GLCM autocorrelation and first order mean show the strongest correlations, followed by GLCM IDN, and all three respond to IQM variation consistently across texture objects, as shown by the visual uniformity down the four texture rows of each block. In contrast, the correlations for first order kurtosis are weaker overall and lack this consistency: textures A and D respond similarly to each other, while B and C respond differently from them.

Together, these results indicate that the selected features respond to the measured IQMs

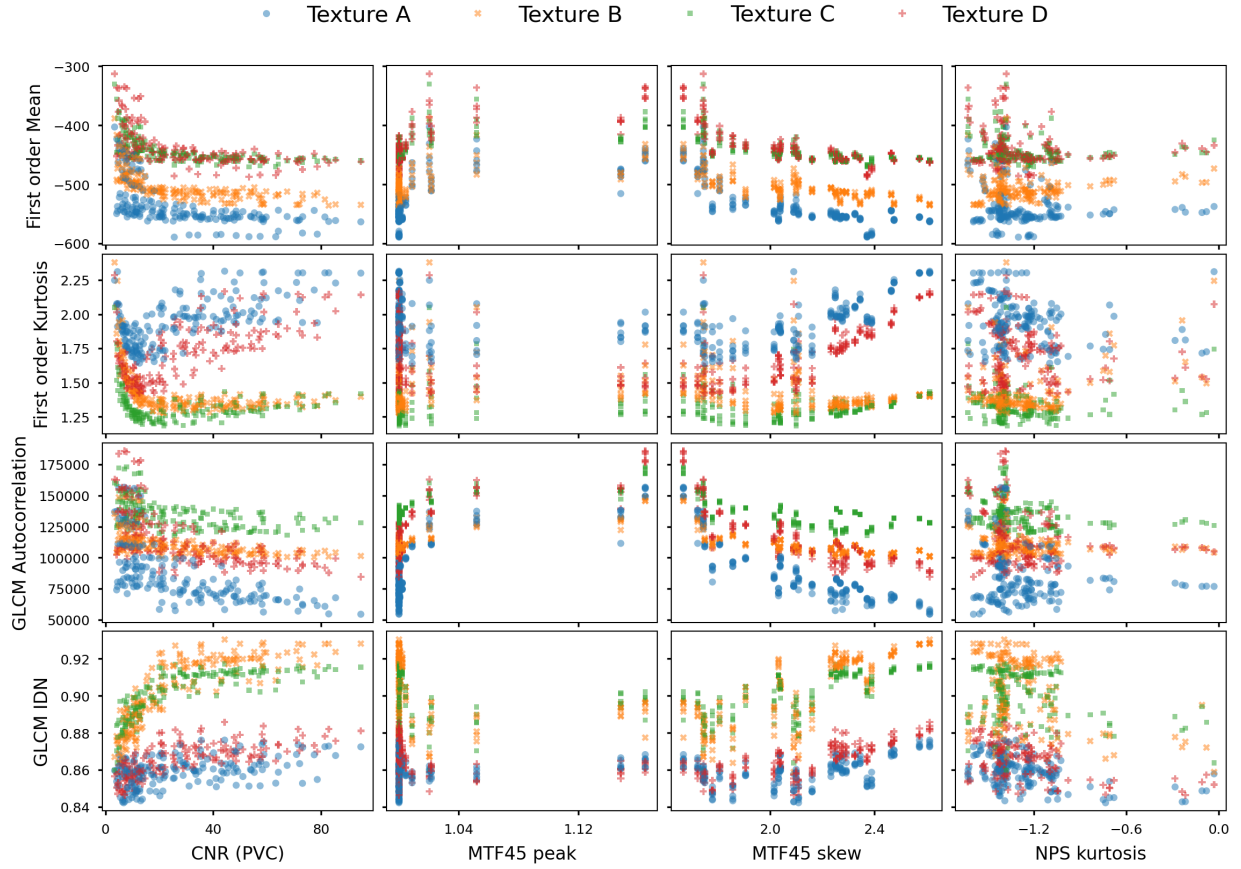


Figure 4.5: Scatter plots of the four selected radiomic features (subplot rows) vs example image quality metrics (subplot columns) to illustrate some of the relationships captured. Each point represents an individual CT protocol. Color represents the physical texture objects, with labeling corresponding to the ROI labeling in Figure 4.4.

in a structured way, that the resolution metrics (MTF), CNR, and the noise peak frequency (NPS f_{peak}) are the strongest correlates, and that this response is consistent across textures for first order mean, GLCM autocorrelation, and GLCM IDN, but texture-dependent for first order kurtosis. The saturating and non-linear forms observed here also indicate that a purely linear description would be insufficient, and that the model form may need to differ across radiomic features.

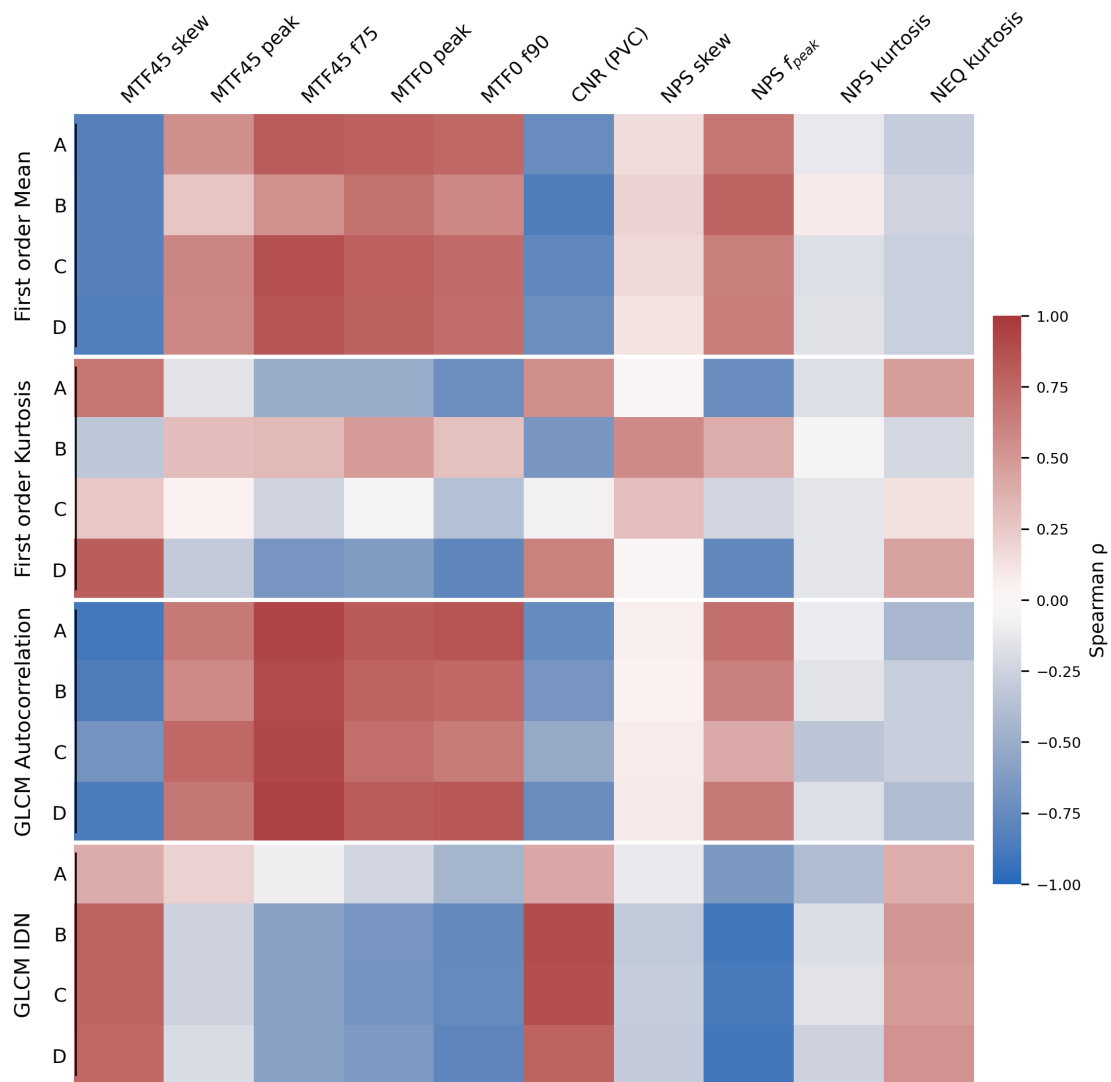


Figure 4.6: Spearman rank correlations between each of the four representative radiomic features and the selected IQMs, computed separately for the four texture objects (rows A–D within each feature block).

4.4 Discussion

This chapter establishes systematic, quantifiable relationships between physical image quality metrics and radiomic feature values across 170 protocols spanning four scanner models and two

vendors. Across the four selected features, the resolution metrics (MTF), CNR, and the noise peak frequency (NPS f_{peak}) were the dominant correlates, while the NPS shape descriptors (NPS skew, NPS kurtosis) contributed little. The functional forms were frequently non-linear, such as the saturating response of GLCM IDN to CNR, indicating that a linear description is insufficient. Results also indicate that the appropriate model form may differ across features.

We also illustrate how these measured IQMs resolve protocol differences more finely than either visual inspection or manufacturer protocol labels. Visual inspection orders the texture images only by gross trends in sharpness and noise, and conflates the two, while manufacturer labels are categorical, vendor-specific, and bundle several physical effects, so that protocols sharing a nominal label can produce measurably different MTF and NPS. By placing every protocol on a common, continuous, scanner-independent axis that the radiomic features are found to track, the measured metrics make a quantitative, physically interpretable description of radiomic variation possible, and provide the foundation for the predictive models of Chapter 5.

The image quality response was consistent across the four texture objects for first order mean, GLCM autocorrelation, and GLCM IDN, suggesting that for these features the dependence is a property of the feature definition rather than of the specific texture. First order kurtosis was the exception, responding differently between the A/D and B/C texture pairs, plausibly because as a higher-order moment of the intensity histogram it is sensitive to the tails and distribution shape particular to each texture. This consistency is encouraging for the prospect that image-quality-based descriptions transfer across textures, though whether it extends from these synthetic phantom textures to clinical tissue remains to be tested.

This analysis is descriptive and phantom-based. The functional forms are characterized qualitatively rather than fit, and each metric is examined univariately, whereas differences across CT protocols likely impact multiple combinations of resolution and noise that a univariate view cannot disentangle. The apparent exceptions in the sorted texture images, where protocols matched on a single metric still differ visually, are a direct symptom of this and motivate the multivariate treatment in Chapter 5.

Modeling these relationships is a natural next step, and is explored in the following chapter (Chapter 5). Other future directions could include alternative approaches to extracting scalar summary metrics from the image quality spectra, alternative approaches to feature and metric set reductions, and extending the analysis to additional image textures, including clinical images.

4.5 Conclusions

This chapter extends radiomic feature irreproducibility across CT protocols from an empirical observation to a quantifiable and physically interpretable phenomenon. By showing that radiomic features respond systematically to scanner-independent image quality measurements, and that the measured metrics resolve protocol differences more finely than visual inspection or manufacturer labels, this work establishes the foundation for physics-based approaches to characterizing and ultimately correcting protocol-induced radiomic variation in multi-center clinical applications.

5 Predictive modeling of radiomic feature variation due to CT protocol using physical image quality metrics

This Chapter is based on the following abstract: M. A. Daly, J. M. Hoffman, A. M. Hernandez, A. Uneri, J. Levy, B. Varghese, and M. F. McNitt-Gray, "Predicting Radiomic Feature Variation Due to CT Protocol Using Measured Image Quality Metrics," in 2026 Joint AAPM/ COMP Annual Meeting & Exhibition, Vancouver, BC: AAPM, 2026.

5.1 Introduction

In Chapter 4, radiomic feature values were shown to vary systematically with physical image quality metrics across a diverse range of CT protocols. Features responded to spatial resolution (MTF) and noise (NPS) in ways largely consistent across the four physical textures, with relationships that were often non-linear, some approximately linear (variance versus noise magnitude), others saturating or exponential (texture coarseness versus spatial resolution). These relationships were broadly consistent across scanner models and vendors, suggesting that image quality metrics may provide a scanner-independent basis for describing radiomic feature behavior.

This chapter investigates whether that relationship can be used predictively: whether the vari-

ation in radiomic feature values due to CT protocol can be quantitatively predicted from measured image quality metrics. A model mapping measurable, scanner-independent image quality metrics to radiomic feature values would allow the behavior of a feature under a given protocol to be predicted without measuring it on every scanner and protocol, supporting prospective assessment of radiomic model robustness and physics-based approaches to harmonization.

The primary objective of this chapter was to establish that radiomic feature variation due to CT protocol can be predicted from physical image quality metrics. Rather than modeling protocol-induced variation directly, each radiomic feature value was modeled as a function of image quality, so that variation across CT protocols is captured through its dependence on image quality. A separate model was fit for each feature, and its performance was reported on a held-out test set. A secondary objective was to evaluate the generalizability of this modeling approach, reporting within-vendor and cross-vendor generalization, and identifying the conditions under which predictive accuracy is retained or degraded.

We hypothesized that radiomic feature variation could be accurately predicted from physical image quality metrics, reflecting a dependence on measurable image quality rather than on proprietary scanner implementations. The extent to which this predictability transfers across scanners and vendors was treated as an open question, bounded by the image quality coverage of the dataset.

5.2 Methods

5.2.1 Dataset

The models in this chapter predict radiomic feature values from physical image quality metrics using the paired phantom dataset described in Chapter 3. The predictors were the 10 reproducible, non-redundant image quality metrics selected in Chapter 4, derived from the 3D MTF and NPS measurements. The prediction targets were the radiomic features retained after the inter-slice reproducibility filtering in Chapter 4, extracted from each of the four texture objects (A–D). A separate model was fit for each combination of radiomic feature and texture object.

5.2.2 Modeling approach

Each radiomic feature, for a given texture object, was predicted from the image quality metrics using a regression model with ElasticNet regularization. For a radiomic feature value y and the vector of image quality metrics $\mathbf{x}$ for a given protocol, the prediction was

$$\hat{y} = w_0 + \mathbf{w}^\top \phi(\mathbf{x}), \quad (5.1)$$

where $\phi(\mathbf{x})$ denotes a second-order polynomial expansion of the IQM vector, comprising the original metrics, their squares, and their pairwise interaction terms; w_0 is an intercept; and $\mathbf{w}$ is the vector of fitted coefficients. The second-order expansion allowed the model to represent the non-linear and combined dependencies on resolution and noise observed in Chapter 4.

The coefficients were estimated by minimizing the squared prediction error subject to an

ElasticNet penalty, which combines an $L1$ (LASSO) term that drives uninformative coefficients to zero, selecting a sparse subset of relevant terms, and an $L2$ (ridge) term that shrinks the remaining coefficients to stabilize the fit [103]. This was selected because the number of candidate terms was large relative to the 170 protocols. The penalty strength and the balance between the $L1$ and $L2$ terms were selected by cross-validation within the training data for each model. First-order (linear) models were additionally fit for comparison, to quantify how much of the predictive accuracy depended on the non-linear terms.

5.2.3 Model evaluation

Each model was fit and evaluated on a random split of the full dataset ($n = 170$ protocols) into training (80%) and test (20%) sets. Performance was evaluated using the coefficient of determination (R^2) as the primary measure, and the root mean squared error (RMSE), reported per feature in that feature's own units, as a measure of absolute error. The proportion of models exceeding $R^2 = 0.6$ and $R^2 = 0.8$ was also reported. This was repeated for each radiomic feature, for each of the four texture objects.

5.2.4 Generalization assessments

Establishing that radiomic feature variation can be modeled from image quality metrics establishes whether these relationships exist, but Section 5.2.3 does this by pooling the data from all four scanner models then splitting the train and test sets randomly. With that relationship established,

a secondary question is whether that predictive ability holds for scanner systems not included in the training set is more valuable but more challenging scenario. This analysis extends the in-distribution evaluation toward that more clinically realistic scenario, assessing how well predictions transferred to scanners and vendors held out from training.

To assess the models' generalizability beyond the conditions on which they were trained, two additional configurations were evaluated. The same fitting procedure was used; only the train-test splits of protocols differed. These splits are shown in Table 5.1. Results were aggregated and reported per vendor.

In the *within-vendor* configuration, each model was trained on one scanner and tested on the other scanner of the same manufacturer. All four pairings were evaluated: the two GE scanners (Revolution Apex and Revolution EVO), and the two Siemens scanners (SOMATOM Definition AS and SOMATOM Force). This tested transfer to an unseen scanner that shared a manufacturer with the training scanner.

In the *cross-vendor* configuration, each model was trained on both scanners of one manufacturer and tested on both scanners of the other, in both directions. This was the most challenging condition, as there are expected differences in hardware and software between manufacturers, as well as the known difference in slice thickness for this dataset. The two directions were evaluated separately to assess any asymmetry.

Table 5.1: Train–test splits for the generalization experiments, with the number of protocols n in each subset.

Configuration	Split	Training scanner(s)	n_{train}	Test scanner(s)	n_{test}
Within-vendor	1	GE Apex	48	GE EVO	42
	2	GE EVO	42	GE Apex	48
	3	Siemens Definition	40	Siemens Force	40
	4	Siemens Force	40	Siemens Definition	40
Cross-vendor	1	GE (Apex, EVO)	90	Siemens (Definition, Force)	80
	2	Siemens (Definition, Force)	80	GE (Apex, EVO)	90

5.2.5 Image quality across vendors

To interpret the cross-vendor results, the extent to which the two manufacturers overlapped in image quality space was assessed in two ways.

First, the overall structure of the image quality space was visualized using principal component analysis. The 10 metrics were standardized to zero mean and unit variance and projected onto their first two principal components. Protocols were labeled by scanner model and reconstruction kernel to reveal how the manufacturers and their kernel families were distributed across the space.

Second, the overlap was quantified per metric. For each image quality metric, the range spanned by the training manufacturer’s protocols was determined, and the proportion of the test manufacturer’s protocols falling outside that range was computed. Together these indicated whether poor cross-vendor transfer arose because the two manufacturers occupied different regions of image quality space (requiring extrapolation), or because the underlying relationship between image quality and radiomic features itself differed between them.

5.3 Results

5.3.1 Overall model performance

The physical image quality metrics modeled radiomic feature variation well and predicted the feature values accurately under the full-dataset configuration. Across the 52 radiomic features and four texture objects (208 models), the mean test R^2 was 0.88 ± 0.21 , with 98 % of models exceeding $R^2 = 0.6$ and 91 % exceeding $R^2 = 0.8$. Performance by feature class is summarized in Table 5.2.

Table 5.2: Full-dataset prediction performance by feature class. Models were fit per feature and texture object and evaluated on a held-out test set. Values were aggregated over all four texture objects and all features in the class. Mean test R^2 ($\pm$ standard deviation) is reported, along with median R^2 and the percent of models exceeding $R^2 = 0.6$ and $R^2 = 0.8$.

Feature class	n features	Mean R^2	Median R^2	$R^2 > 0.6$	$R^2 > 0.8$
First order	11	0.75 ± 0.43	0.92	89 %	75 %
GLCM	12	0.89 ± 0.06	0.91	100 %	92 %
GLDM	9	0.93 ± 0.01	0.93	100 %	100 %
GLRLM	10	0.92 ± 0.03	0.93	100 %	95 %
GLSZM	10	0.92 ± 0.03	0.93	100 %	95 %
All	42	0.88 ± 0.21	0.93	98 %	91 %

The results for each radiomic feature are provided in Figure 5.1. The texture feature classes (GLCM, GLDM, GLRLM, and GLSZM) were generally predicted both accurately and consistently, with mean R^2 between 0.89 and 0.93 per class, indicating that the variation in these features due to CT protocol is well described as a function of measured image quality. The first-order features had a lower mean R^2 (0.75) but a comparable median (0.92). This discrepancy was driven almost

entirely by first-order minimum. As minimum is an extreme-value (the single lowest-valued voxel in the ROI), it is dominated by noise rather than by systematic image quality effects. The other first-order features were predicted well.

Restricting the models to first-order terms reduced the overall mean test R^2 from 0.87 to 0.77, indicating that the underlying image quality–radiomics relationships are non-linear and are captured more fully by the second-order model, consistent with the functional forms observed in Chapter 4. Across-texture variation in R^2 was small for most features, indicating that a feature’s predictability from image quality was largely independent of the underlying physical texture. This is consistent with the texture-independence observed qualitatively in Chapter 4.

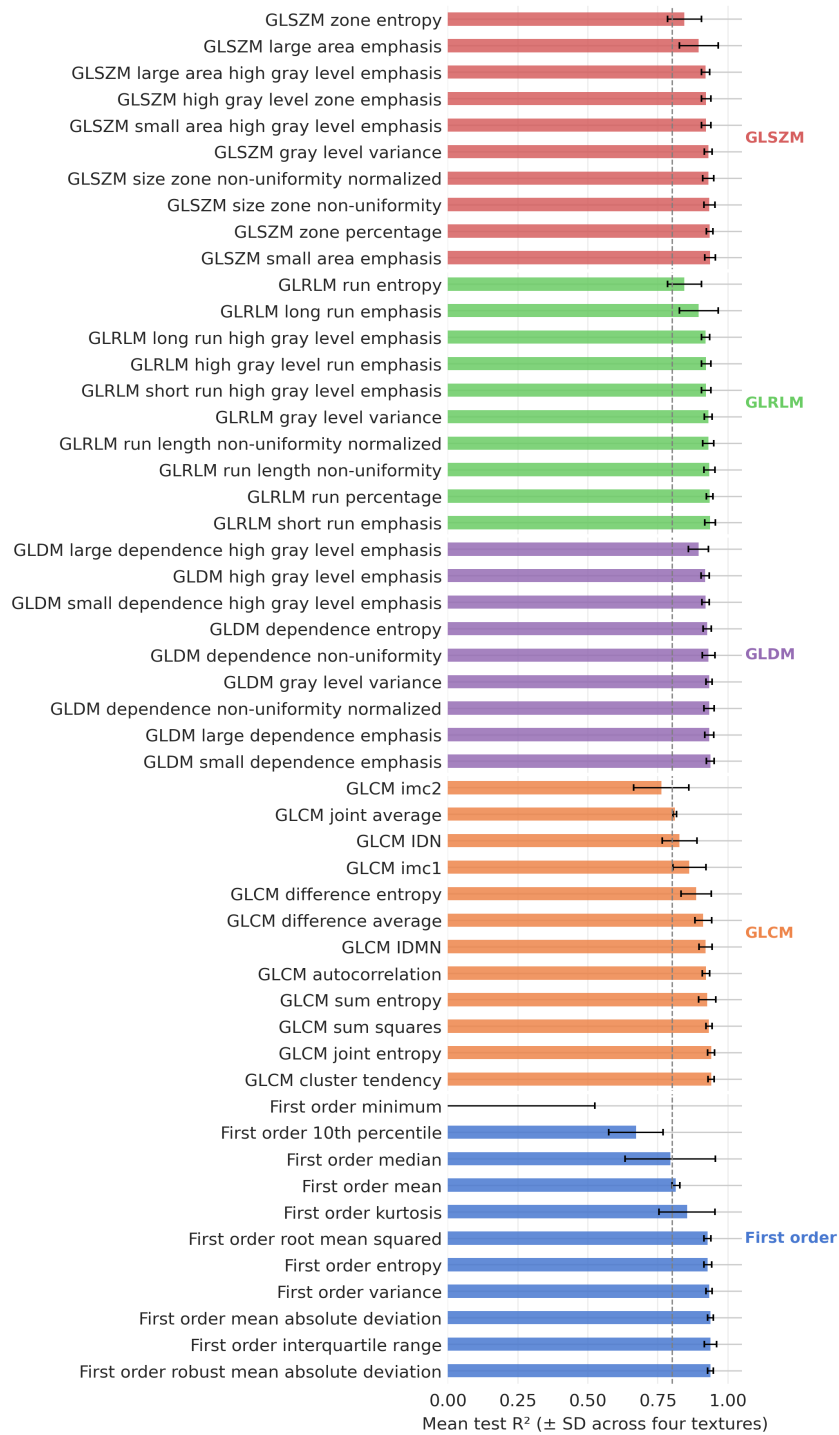


Figure 5.1: The mean test set R^2 value for each radiomic feature, averaged over the four texture objects. Error bars correspond to standard deviation.

5.3.2 Generalization across scanners and vendors

The generalization performance is summarized in Table 5.3. Within-vendor transfer, training on one scanner and predicting on the other scanner of the same manufacturer, retained much of the predictive accuracy: the median test R^2 was 0.67 for GE and 0.57 for Siemens, compared with 0.93 for the full-dataset configuration from Section 5.3.1. Prediction therefore transferred to an unseen scanner of the same manufacturer, but with a loss of accuracy relative to the full-dataset case.

Generalization across vendors, that is, training on one manufacturer’s scanners and predicting on the other’s, was markedly lower. The median test R^2 fell to 0.26 when testing on GE and 0.14 when testing on Siemens. Only 35 % and 22 % of models exceeded $R^2 = 0.6$ for GE and Siemens respectively, compared with 61 % and 46 % within-vendor.

The progression from full-dataset to within-vendor to cross-vendor produced a consistent decline in predictive accuracy as the test conditions moved further from those seen in training. Performance was also consistently higher for GE than for Siemens in both the within-vendor and cross-vendor configurations.

Table 5.3: Generalization performance across train–test configurations. Median test R^2 over all reproducible features and four texture objects, with the fraction of models exceeding $R^2 = 0.6$ and $R^2 = 0.8$.

Configuration	Test set	Median R^2	$R^2 > 0.6$	$R^2 > 0.8$
Within-vendor	GE	0.67	61 %	28 %
	Siemens	0.57	46 %	21 %
Cross-vendor	GE	0.26	35 %	13 %
	Siemens	0.14	22 %	11 %

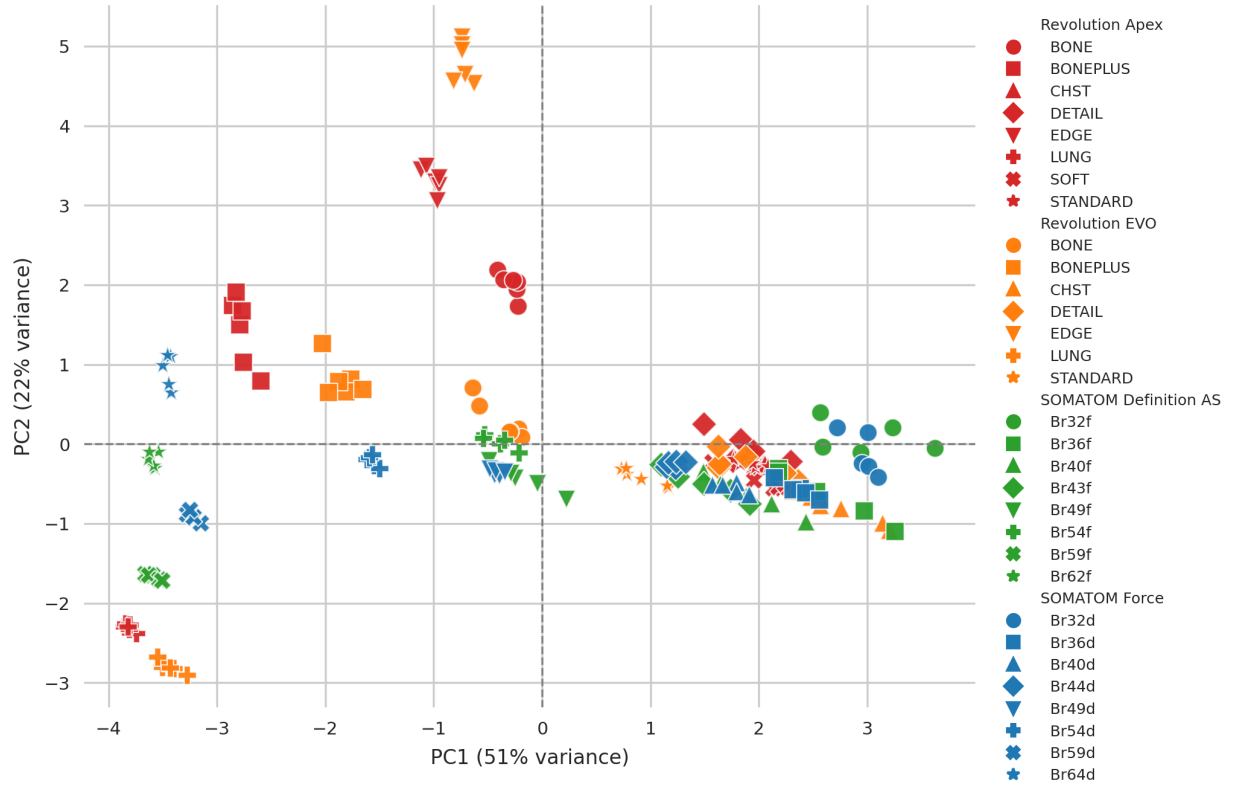


Figure 5.2: Principal component analysis of the 10 image quality metrics across all protocols. The first two components captured 73 % of the total variance (PC1: 51 %, PC2: 22 %). Marker color corresponds to the scanner model and marker shape to the reconstruction kernel.

5.3.3 Range in image quality across vendors

Two possible causes of the models' failure to generalize across vendors were: (1) the two manufacturers occupy different regions of image quality space, requiring models to extrapolate when trained on one and then predicting on the other, or (2) the underlying relationship between image quality and radiomic features itself differs between them.

To assess this, the structure of the image quality space was visualized using PCA of the 10 metrics across all protocols (shown in Figure 5.2). The first two components captured 73 % of

the total variance (PC1: 51 %, PC2: 22 %). Each scatter point corresponds to one protocol, with color corresponding to scanner model and marker shape corresponding to reconstruction kernel. Multiple points with the same color and marker shape are the different radiation dose levels.

The first principal component appears to act broadly as an axis of reconstruction kernel sharpness, with increasing values corresponding to smoother kernels (e.g., Br32f, CHST), and negative values corresponding to edge-enhancing kernels (e.g., Br62f, LUNG). While protocols are grouped within a small range about $PC2 = 0$ where $PC1 > 0$, the edge-enhancing kernels separate widely across $PC2$. In particular, the extrema of $PC2$ are distinctly occupied by GE protocols: LUNG at the minimum values of $PC2$, and EDGE at maximum values of $PC2$, consistent across both GE scanners. This suggests these are the most dissimilar protocols from the rest of the dataset in terms of IQMs measured.

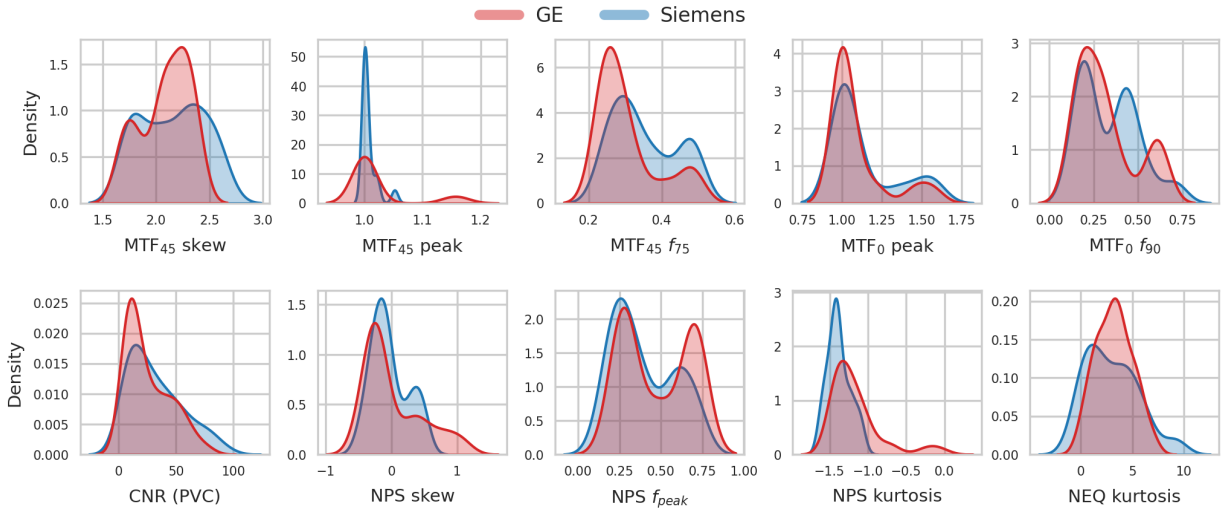


Figure 5.3: Distributions (kernel density estimation) of each of the IQMs used for modeling for each manufacturer.

However, this separation does not necessarily reflect non-overlapping ranges on the individual metrics. The distributions (as kernel density estimates) of each of the IQMs for the two vendors are shown in Figure 5.3. These show directly where the ranges overlap between vendors, and where they were “out-of-range”.

Table 5.4 reports, for each manufacturer, the fraction of the other’s protocols falling outside its measured range. Although most test protocols fell outside the training manufacturer’s range on at least one metric, the typical protocol did so on only about one of 10 metrics, remaining within range on the rest. Therefore, the two manufacturers overlap on most image quality metrics individually, and their separation in image quality space reflects a concentration of out-of-range behavior on a subset of metrics.

Table 5.4: Coverage of each test manufacturer’s protocols by the training manufacturer’s measured image quality metric ranges. “Out of range (≥ 1 metric)” is the fraction of test protocols outside the training range on at least one metric; “mean metrics out of range” is the average number of metrics (of 10 total) on which a test protocol falls outside the training range.

Train	Test	Out of range (≥ 1 metric)	Mean metrics out of range
GE	Siemens	57 %	1.2 / 10
Siemens	GE	63 %	1.0 / 10

Where protocols had out of range IQMs, they were the NPS shape descriptors and f_p , the oblique (45°) MTF (its skewness and f_{75}), while the axial (0°) MTF descriptors and the CNR remained largely in range. The leading metric differed by train-test direction: when predicting Siemens protocols from GE-trained ranges, the NPS kurtosis and f_p were most often out of range (each 29 % of protocols), followed by the oblique MTF skewness (25 %); when predicting

GE protocols from Siemens-trained ranges, the oblique MTF f_{75} led (33 %), while the NPS f_p fell entirely within range.

A model trained on one manufacturer must therefore extrapolate along this set of spectral shape and oblique-resolution metrics when predicting on the other, which may account for the poor cross-vendor generalization in Table 5.3.

5.4 Discussion

This investigation shows that CT-protocol induced variation in radiomic features, which has been established as an obstacle to clinical implementation of AI systems, can be predicted from measured physical image quality metrics, but that the strength of this prediction depends on how far the test conditions depart from those represented in training. Prediction was strong within the full sampled image quality space and substantially retained when generalizing to an unseen scanner of the same manufacturer, but degraded sharply across manufacturers. Rather than a uniform demonstration of scanner-independence, the results revealed conditions under which the physical image quality coordinate system supports prediction and the conditions under which it does not. The cross-vendor limit, in particular, has an identifiable physical origin, which we examine below.

A limitation of this work stems from the choice of modeling approach. ElasticNet with a second-order polynomial expansion was selected to provide a continuous, differentiable, and interpretable mapping of how physical terms interact. However, polynomial regression cannot

reliably extrapolate: outside the range of image quality spanned by the training protocols, low-order polynomial terms diverge rather than saturating to physically plausible values. The models therefore interpolate well across the sampled protocol space but degrade when predicting under image quality conditions not seen during training. This is likely a primary source of the poor cross-vendor transfer observed in Section 5.3.2: the metrics on which the manufacturers diverge place the test protocols outside the training range, where the polynomial terms are unconstrained.

Future work could consider alternative model families. Multivariate Adaptive Regression Splines (MARS) or cubic splines could provide flexibility to model complex, non-linear interactions within the sampled domain while enforcing linear constraints beyond the boundary, preventing exponential divergence. Alternatively, tree-based machine learning methods, such as random forests, could be evaluated; while they cannot extrapolate beyond their training targets, comparing these specific failure modes against spline-based extrapolation could inform the optimal framework for modeling radiomic features as a function of image quality metrics.

Another limitation is the reduction of each MTF and NPS curve to a small set of scalar descriptors. Two protocols can share these scalars while differing in the underlying curve shape, so any radiomic variation driven by that shape is not captured. Using the curves directly, either by sampling them densely or by decomposing them into a few dominant modes of variation, would retain this information. This was not done here because the curves are high-dimensional relative to the 170 protocols, and the scalar descriptors provide a tractable, interpretable summary. Using the full curves would require either a larger dataset or a functional representation, and is a natural extension.

5.5 Conclusions

Radiomic feature variation due to CT protocol can be predicted from physical image quality metrics. The texture feature classes were predicted with a mean R^2 of 0.89 to 0.93 within the sampled image quality space. While prediction transferred to unseen scanners of the same manufacturer, generalization across manufacturers was limited. Image quality metrics therefore provide a quantitative, scanner-independent basis for predicting protocol-induced radiomic variation, within the range of image quality a model has sampled.

6 Predicting CNN generalization across CT imaging conditions using physical image quality metrics

A version of this chapter is in preparation to be submitted as a manuscript to Medical Physics.

6.1 Introduction

Despite the transformational developments of the “deep learning revolution” of the 2010s, deep learning (DL) based artificial intelligence (AI) systems have so far achieved limited success in real-world medical imaging applications [104]. The application of these recent advances in DL to radiological imaging has been of great interest due to its potential to improve workflow efficiency by automating diagnostic and detection tasks, increase standardization through quantitative assessment of imaging data, and improve accuracy by leveraging imaging features beyond what can be perceived a human reader [5].

Although many studies have achieved state-of-the-art results by applying DL to medical imaging tasks, the reported performance is rarely replicated when models are tested on external datasets not encountered during training and development. For example, Zech et al. (2018) showed a pneumonia detection model that achieved an area under the receiver operating char-

acteristic curve of 0.931 on internal data, but decreased to 0.815 on an external test set from another institution [14]. Wang et al. (2020) tested three published CNN models for mammo-gram classification and found that the AUC scores decreased from between 0.88 and 0.95 on the reported validation data to between 0.44 and 0.65 on three external test sets [15]. A systematic review of radiologic DL models by Yu et al. (2022) found that 81% of algorithms demonstrate diminished performance when applied to external datasets, with 24% reporting a decrease in performance of ≥ 0.10 on the unit scale [16]. The observed performance decrease when applying DL models to external test sets is primarily attributed differences between the data the model was trained on and the diverse data encountered in real-world clinical practice.

Any difference between the distribution of the training dataset and the distribution of the data where the model will be applied is known as *data shift* [17]. In classic supervised learning, the training and test data are assumed to be independent and identically distributed, known as the *i.i.d. assumption* [105]. However, this assumption can be easily violated when models are deployed clinically, and the model will encounter samples that are *out-of-distribution* (OOD). Thus, the impact of data shift on model performance is innately related to the clinical utility of these AI systems.

One of the most prevalent and well-studied sources of dataset shift in medical imaging is known as *acquisition shift*. The term acquisition shift describes differences between the train and test data distributions caused by differences in scanners or imaging protocols [106]. In the context of CT imaging, acquisition shift can be caused by variation in the selection of technical parameters such as tube current, slice thickness, and reconstruction kernel, as well as by other

differences in scanner hardware and software. Thus, in this context, we consider sources of acquisition shift to include the full image processing pipeline, inclusive of image reconstruction (rather than referring to image acquisition and reconstruction as two independent processes, as is common in CT imaging). Scan protocol factors are well-known to influence the image’s spatial, contrast, and noise characteristics, partial volume averaging effects, and the ability to mitigate physiologic motion [20].

Although acquisition shift is not the only source of data shift in CT imaging applications, this work focuses specifically on the impact of data shift caused by CT protocol variation. The impact of CT protocol differences on downstream quantitative imaging assessments is well-established [8], [107], [108], and prior work has demonstrated that variation in physical imaging parameters is a primary driver of domain shift in medical imaging applications more broadly [109]. Importantly, protocol-induced shift differs from most other sources of domain shift in that many relevant technical parameters are recorded in the DICOM header of every CT image, guaranteeing their availability at test-time. This makes CT protocol variation a uniquely tractable form of domain shift.

Ideally, a clinical physicist deploying a trained model under a new CT protocol could obtain a quantitative estimate of expected performance degradation without needing to collect additional patient images. However, knowing the protocol information alone does not itself provide a mechanism for predicting its impact on model performance.

We propose that phantom-derived image quality metrics (IQMs), which translate protocol parameters into standardized physical measurements of spatial resolution and noise, provide a

mechanism for doing this. If the relationship between protocol-induced image quality shift and CNN performance degradation can be characterized, then this relationship could be used prospectively to predict performance under unseen protocols. In this study, we evaluate this hypothesis using classification of nodule vs non-nodule lung patches as a representative task, assessing whether the impact of CT protocol variation on CNN classifier performance can be characterized and predicted from phantom-based IQMs alone.

6.2 Methods

To predict how a CNN trained on one CT protocol will perform on another, two quantities must be measured for each protocol pair: (1) the difference in image quality, and (2) the difference in CNN performance. Phantom imaging provides (1) by acquiring scans of an image quality phantom under each protocol and extracting quantitative IQMs. Using the dataset developed in Chapter 3, we obtain a characterization of each protocol’s image resolution and noise. Obtaining (2) requires training and evaluating CNNs across a controlled set of imaging protocols.

The protocol on which a CNN was trained is the *source* condition p_s , and any protocol on which that model is subsequently evaluated as the *target* condition p_t . If the performance of that CNN is measured with a metric $\mathcal{M}$, then the quantity of interest is the change in CNN performance across a source-to-target protocol pair (p_s, p_t) , given by:

$$\Delta\mathcal{M}_{s,t} = \mathcal{M}_t - \mathcal{M}_s \tag{6.1}$$

where $\mathcal{M}_s$ is the performance of CNN evaluated on its source condition p_s , and $\mathcal{M}_t$ is the performance of the same CNN evaluated on target condition p_t .

After acquiring both quantities for a sufficient number of protocols, the relationship between the difference in image quality (ΔIQM), and the difference in CNN performance ($\Delta\mathcal{M}$) can be characterized with a regression model. That fitted model predicts the expected CNN performance degradation using only phantom measurements from an unseen target protocol. An overview of the approach is provided in Figure 6.1. In this work, both quantities are obtained under controlled conditions that isolate protocol shift as the sole source of variation. This controlled design is what makes the $\Delta\mathcal{M}_{s,t}$ measurements interpretable as a direct measure of protocol-induced performance shift.

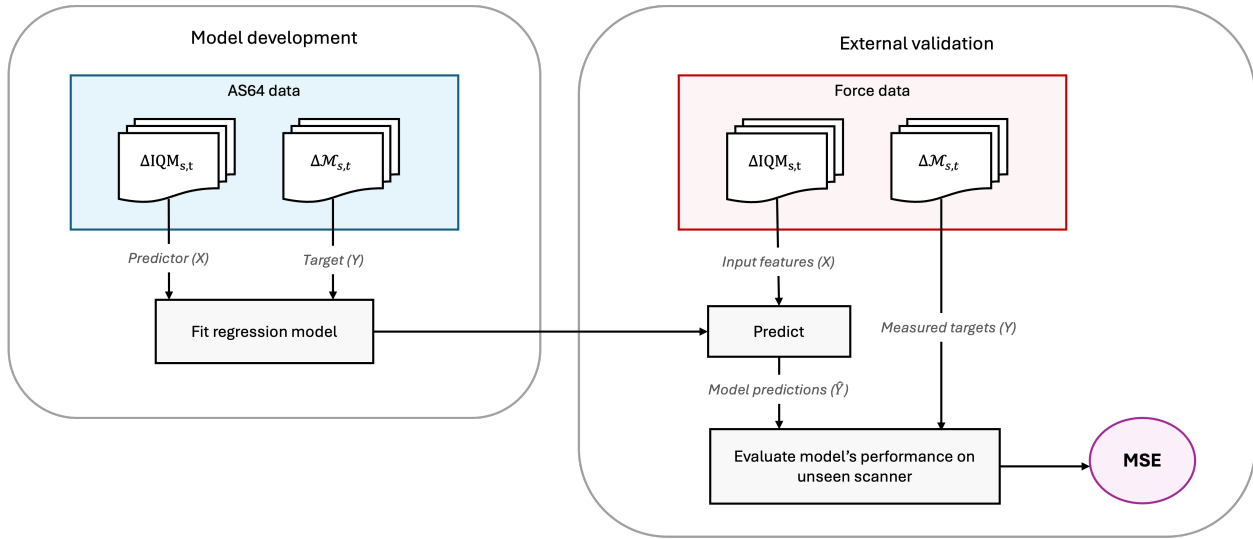


Figure 6.1: Overview of the experimental approach. The CNN performance prediction model is developed on $(\Delta\text{IQM}, \Delta\mathcal{M})$ measurements made from the AS64 dataset. The model is then tested on an external validation dataset from the Force scanner to assess its ability to generalize to an unseen scanner model (here, measured using mean square error (MSE)).

6.2.1 Representative task

As a representative DL task, a CNN was trained to classify 2D patches extracted from lung cancer screening CT images as “nodule” or “non-nodule,” as illustrated in Figure 6.2. The task was inspired by the false positive reduction step of many traditional lung nodule detection systems, which would often be applied after a nodule candidate detection step that operated with a high sensitivity resulting in a high false positive rate [110]. Tasks related to lung nodule detection are clinically relevant due to the the high volume of lung cancer screening [111] and well established as a deep learning applications through the LUNA16 challenge and its successors [110]. This application is also potentially sensitive to reconstruction parameters, making it a meaningful test case for the proposed IQM-based prediction framework [12], [112], [113].

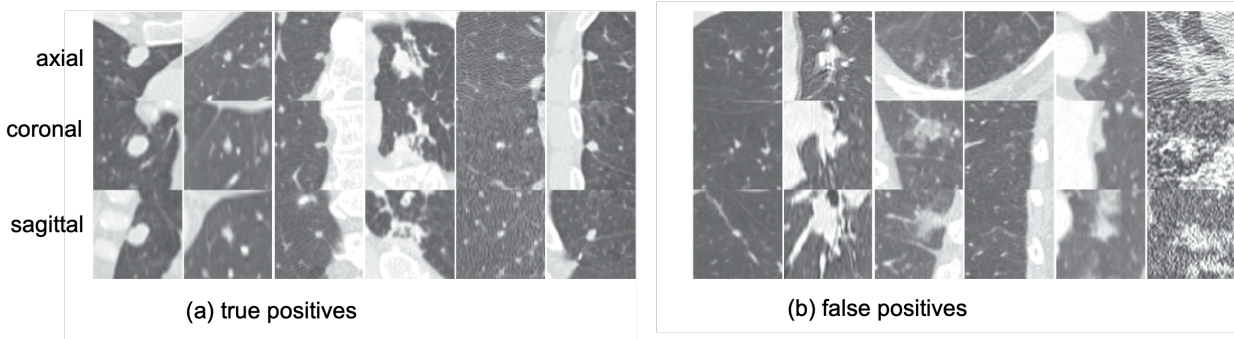


Figure 6.2: Illustration of the classification task using data from the LUNA16 challenge. (a) Shows patches containing lung nodules, (b) shows patches without true nodules. Modified from Setio et al. [110].

6.2.2 Imaging data and patient cohort

Our cohort consisted of 423 CT studies from 423 patients who had undergone lung cancer screening at our institution [93], [114]. Raw CT projection data was collected alongside standard image data with the approval of a local institutional review board, enabling retrospective reconstruction of each study under arbitrary acquisition and reconstruction settings.

Scans were acquired on two scanner models: Siemens Definition AS 64 and Siemens Force (Siemens Healthineers, Forchheim, Germany). All scans were acquired at 120 kV and a target CTDI_{vol} of 2.0 mGy. Image data was reconstructed with Br45f kernel, 1.0 mm slice thickness.

Lung nodule locations were provided by radiologists on the clinical scans and then manually annotated by expert readers; 232 scans contained nodules, and 191 did not contain nodules. Some scans contained more than one nodule. A total of 491 lung nodules were identified and annotated. A summary of the dataset is provided in Table 6.1.

Table 6.1: Patient cohort and nodule distribution by scanner model.

	Definition AS64	Force	Total
Patients (N)	392	31	423
With nodules	212	20	232
Without nodules	180	11	191
Nodules (n)	458	33	491

6.2.3 Dataset construction

6.2.3.1 Multi-reconstruction image dataset

For each study, the raw CT projection data was reconstructed with $K = 8$ different reconstruction kernels. The image processing was done using the program ReconCT (Siemens Healthineers, Forchheim, Germany, version 14.2). The settings used for each of the two scanner models are described in Table 6.2.

Reconstruction kernels were selected based on the imaging conditions used in Chapter 3 so that each reconstruction could be matched to phantom-based measurements of image quality, providing the paired $(\Delta\text{IQM}, \Delta\mathcal{M})$ observations required for model development. These kernels were all body kernels spanning from smooth (Br32f/Br32d) to sharp (Br62f/Br64d). Though Siemens reconstruction kernels are numbered in a way that roughly corresponds to image sharpness (with higher numbers being sharper), the algorithms used are proprietary and are not necessarily the same between scanner systems; this is reflected in the differing naming scheme (with the “f” suffix corresponding to the AS64 scanner and the “d” suffix corresponding to the Force scanner), and in that the numbers available on one system are not necessarily available on another system. For these reasons, the kernel names alone cannot be used as a way to match across scanner systems reliably. All kernels used were filtered backprojection (FBP).

Table 6.2: Image acquisition and reconstruction conditions for each scanner model.

Scanner model	CTDI _{vol}	Convolution kernels
Definition AS64	2.05 mGy	Br32f, Br36f, Br40f, Br43f, Br49f, Br54f, Br59f, Br62f
Force	2.14 mGy	Br32d, Br36d, Br40d, Br44d, Br49d, Br54d, Br59d, Br64d

6.2.3.2 Image quality measurements

Each reconstruction condition was paired with the corresponding image quality data, extracted from the dataset described in Chapter 3. This included metrics derived from axial and oblique MTF and NPS curves, as well as CNR measurements. A full description of the image quality measurements and extracted metrics can be found in Table 3.3.

6.2.3.3 Patch extraction

The inputs to the CNN consisted of 2D image patches extracted from lung cancer screening CT scans, with associated ground-truth labels indicating the presence of a target lesion (nodule).

Positive patches containing true nodules were centered on the annotated target coordinates. The patch bounding boxes were 60 mm squares, which were randomly jittered by up to 20 mm. All patches were resampled to 128×128 pixels and intensity-normalized to the range $[0, 1]$ using a clinical lung window (WW: 1600 HU, WL: -600 HU).

Negative patches, representing potential false positive (FP) candidates, were generated from CT studies that were free from nodules based on the radiologist’s report. Sourcing patches from these negative scans guaranteed the absence of true underlying pathology while maximizing the number of unique patients represented in the dataset. Because the studies originated from

the same cohort and therefore meets the lung cancer screening criteria, the background lung parenchyma is more likely to include clinical variations and pathologies expected in this population, such as fibrosis and emphysema (rather than being derived from “healthy” patients).

To simulate more challenging pseudo-FP nodule patches, negative patches were selected from regions of elevated attenuation within the lung. First, a lung mask was generated using the open-source LungMask tool applied to the smoothest reconstruction kernel available for each series [115]. Pixels within the lung mask with attenuation values > -600 HU were grouped into contiguous clusters, and the center of mass was calculated for each candidate group. Final patch center locations were then randomly sampled from these high-density candidate clusters to construct the desired class prevalence.

This sampling approach also allows for the construction of a dataset with a fixed class balance. As the focus of this investigation is on the impact of CT protocol on CNN performance, and prevalence directly impacts both the measured performance of a CNN as well as the interpretation of those performance metrics, ensuring consistent prevalence across data splits was considered highly important. The number of non-nodule patches was set to create a ratio of 3:1 negative to positive patches. This class imbalance reflects the clinical reality that a first-stage candidate generator would produce substantially more negative candidates than true nodules at a high sensitivity, while remaining moderate enough that (1) the CNNs could be expected to learn the task well with the relatively small dataset available, and (2) the performance metrics would remain interpretable and comparable across conditions.

6.2.3.4 Dataset splits

The data was split by scanner, such that studies acquired on the Definition AS64 ($N = 392$ patients) were used to develop the CNN performance prediction model, and studies acquired on the Force ($N = 31$ patients) were used as an external validation set to evaluate the final model.

Within the Definition development dataset, an 80/20 train/test split was applied, with a further 20% of the training set held out for validation. All splitting was done randomly, grouped by patient, and stratified by class. Splits were applied identically across all eight reconstruction conditions, ensuring that the same patients appeared in the same split regardless of reconstruction kernel. These data splits are shown in Figure 6.3.

A total of 1863 image patches were generated, comprising 475 true positive and 1388 false positive samples. Final nodules counts with exact prevalence are provide in Table 6.3.

Table 6.3: Patch counts and prevalence across dataset splits.

Scanner	Split	Nodule patches	Non-nodule patches	Total samples	Prevalence
Definition	Train	294	828	1122	26.20%
	Validation	61	214	275	22.18%
	Test	87	284	371	23.45%
Force	Test	33	99	132	25.00%
Combined	(total)	475	1388	1863	25.50%

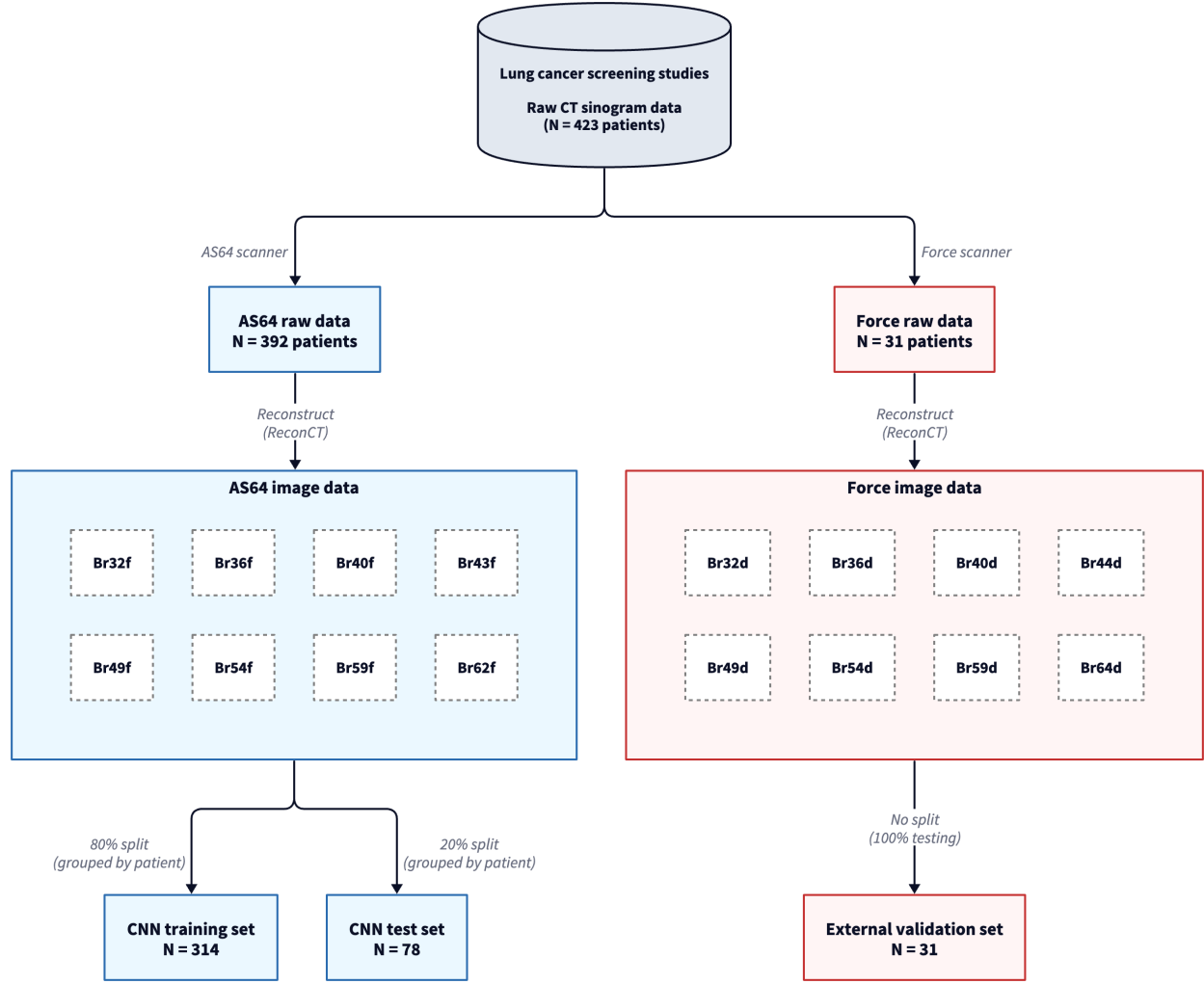


Figure 6.3: Dataset splits. Model development was done with the AS64 data (blue). The Force data (red) was used as an external test set.

6.2.4 Measuring CNN performance across CT imaging conditions

6.2.4.1 CNN training and evaluation

A binary classification network was trained independently on each of the $K = 8$ reconstruction conditions from the Definition AS64 development dataset, for $s = 1, \dots, K$, yielding eight separately trained CNNs. Each CNN was subsequently evaluated on all K conditions as target

protocols p_t , for $t = 1, \dots, K$, producing $K^2 = 64$ source-target pairings in total.

The primary evaluation metrics were the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC). Threshold-independent metrics were selected to characterize overall performance without defining a specific operating point; AUPRC was additionally favored given the class imbalance in the dataset, as it weights performance toward the minority (positive) class [116].

Each CNN used a ResNet-18 architecture [117] initialized with pretrained ImageNet weights [118], with the first layer frozen during training. Input patches were normalized to zero mean and unit variance using per-condition statistics computed from the training split. Data augmentation included random horizontal and vertical flipping and random affine transformations comprising rotations up to 180° , scaling between 0.8 and 1.2, and translations up to 10% of image dimensions. Training was performed with weighted cross-entropy loss with a class weight proportional to the negative-to-positive sample ratio, AdamW optimizer [119] with weight decay of 0.01, batch size of 32, and an initial learning rate of 0.001. Early stopping was employed based on validation AUROC with a patience of 10 epochs. The learning rate scheduler used ReduceLROnPlateau with a factor of 0.5 and a patience of 3 based on validation AUROC. The final model weights were selected from the earliest epoch within 0.5% of the peak validation AUPRC. All training was performed with an Nvidia GeForce RTX 2080 Ti GPU, implemented in PyTorch [120].

6.2.4.2 Change in CNN performance across CT imaging conditions

In this work, $\mathcal{M}$ is instantiated as AUPRC, such that Equation (6.1) becomes:

$$\Delta\text{AUPRC}_{s,t} = \text{AUPRC}_t - \text{AUPRC}_s \quad (6.2)$$

where AUPRC_s is the performance of CNN evaluated on its source condition p_s , and AUPRC_t is the performance of the same CNN evaluated on target condition p_t . In this work, the conditions were reconstruction kernels, and each condition corresponded to the set of images reconstructed using that kernel. A positive $\Delta\text{AUPRC}_{s,t}$ then corresponds to an increase in performance relative to the source condition, while a negative value indicates a decrease in performance relative to the source condition.

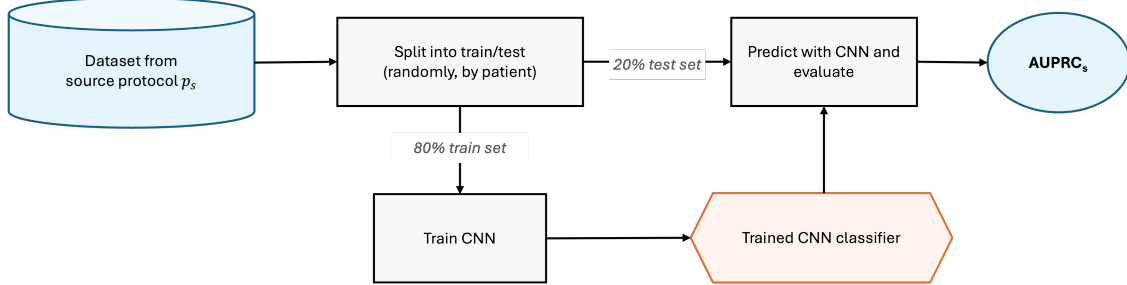


Figure 6.4: Schematic for training and evaluating a CNN within the same domain, source condition p_s .

Figure 6.4 illustrates the process of obtaining AUPRC_s by evaluating each CNN on its own source condition. Figure 6.5 illustrates the process of obtaining AUPRC_t by evaluating the same CNN on each target condition.

To obtain an estimate of the uncertainty in each $\Delta\text{AUPRC}_{s,t}$ value, both AUPRC_s and AUPRC_t were computed using bootstrap resampling of the test set with $B = 1000$ iterations.

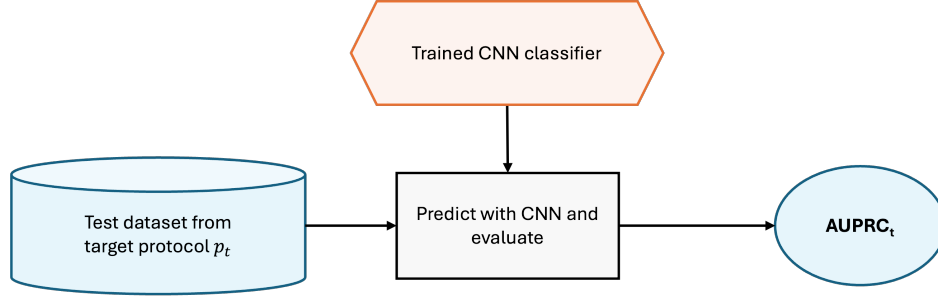


Figure 6.5: Schematic for evaluating the performance of a CNN trained on the source condition p_s when predicting on the target domain p_t .

The mean of the bootstrapped distribution was taken as the point estimate and the variance $\sigma_{s,t}^2$ as a measure of statistical uncertainty.

This yielded a dataset $\{(\Delta IQM_{s,t}, \Delta AUPRC_{s,t})\}$ of $N = K^2 = 64$ source-to-target pairs for use in model fitting.

6.2.5 Modeling CNN performance as a function of image quality

To systematically characterize the nature of CNN performance degradation under domain shift, we evaluated three distinct mathematical models. For each candidate model, the dependent variable $y = \Delta AUPRC_{s,t}$ represents the observed change in CNN performance, and the independent variable $x = \Delta IQM_{s,t}$ represents the scalar IQM shift, measured from the image quality assessment phantom in Chapter 3, between source and target conditions.

6.2.5.1 Candidate model forms

Three candidate model forms were considered:

1. **Linear model.** The linear model assumes that performance degrades at a constant rate as the image quality deviates from the training baseline:

$$y = \beta_0 + \beta_1|x| \quad (6.3)$$

This model uses the absolute magnitude of the shift, $|x|$, and therefore assumes that performance depends only on the size of the deviation from the training domain, not its direction.

2. **Quadratic model.** The quadratic model is the only model evaluated using the signed, directional IQM shift (x) rather than its absolute magnitude:

$$y = \beta_0 + \beta_1x + \beta_2x^2 \quad (6.4)$$

This allows the model to capture asymmetry with respect to zero domain shift. For example, the performance when the target condition is sharper than the source ($\Delta\text{IQM} > 0$) may differ from the response when it is smoother ($\Delta\text{IQM} < 0$). The quadratic is still symmetric about its own vertex, so it cannot capture asymmetric rates of change.

3. **Exponential model.** The exponential model tests whether the relationship between domain shift and performance degradation is nonlinear, with larger shifts associated with increasingly rapid declines in performance. The model is given by

$$y = \beta_0 + \beta_1e^{k|x|} \quad (6.5)$$

where β_0 is the intercept and k determines the rate at which the response changes with increasing domain shift. This model predicts relatively small changes at low levels of domain shift and progressively larger changes as the magnitude of the shift increases.

6.2.5.2 Model fitting and evaluation

The linear and quadratic models were estimated directly using weighted least squares. The exponential model was estimated using nonlinear least squares and required iterative optimization. Because reliable estimation of the quadratic and exponential models was not feasible using only the eight source-target evaluations available for a single trained CNN, observations were pooled across all eight independently trained CNNs ($n = 64$).

The weights were defined as the inverse variance of each bootstrapped $\Delta\text{AUPRC}_{s,t}$ estimate:

$$w_{s,t} = \frac{1}{\sigma_{s,t}^2} \quad (6.6)$$

where $\sigma_{s,t}^2$ denotes the variance of the bootstrapped $\Delta\text{AUPRC}_{s,t}$ estimate for a source-target pair (p_s, p_t) .

For a given image quality shift $\Delta IQM_{s,t}$, the fitted model predicts the corresponding change in performance:

$$\widehat{\Delta\text{AUPRC}}_{s,t} = g(\Delta IQM_{s,t}; \hat{\beta}) \quad (6.7)$$

where $g(\cdot)$ denotes the fitted regression function (linear, quadratic, or exponential; Equations (6.3) to (6.5)) and $\hat{\beta}$ are the estimated model parameters.

Model performance was evaluated by comparing predicted and observed $\Delta\text{AUPRC}_{s,t}$ values across all source-target pairs. Predictive accuracy was quantified using root mean squared error (RMSE) and the adjusted coefficient of determination (R_{adj}^2). Lower RMSE and higher R_{adj}^2 values indicate that image quality differences provide greater explanatory power for observed changes in CNN performance across CT acquisition protocols.

6.2.6 External validation on Force scanner

Model generalization was evaluated using the held-out multi-reconstruction dataset acquired on the Force scanner. Each of the $K = 8$ CNNs trained on the AS64 source conditions was evaluated on the test set reconstructed under all $K = 8$ Force target conditions, yielding $K^2 = 64$ source-target pairs. For each pair, the observed $\Delta\text{AUPRC}_{s,t}$ was computed as described in Equation (6.2), and the corresponding $\Delta\text{IQM}_{s,t}$ was used to generate a predicted performance change:

$$\widehat{\Delta\text{AUPRC}}_{s,t} = g(\Delta\text{IQM}_{s,t}; \hat{\beta}) \quad (6.8)$$

using the model parameters estimated from the AS64 development dataset. Generalization performance was assessed across all 64 pairs using RMSE and R_{adj}^2 between the predicted and observed $\Delta\text{AUPRC}_{s,t}$ values.

6.3 Results

6.3.1 CNN performance on OOD imaging conditions

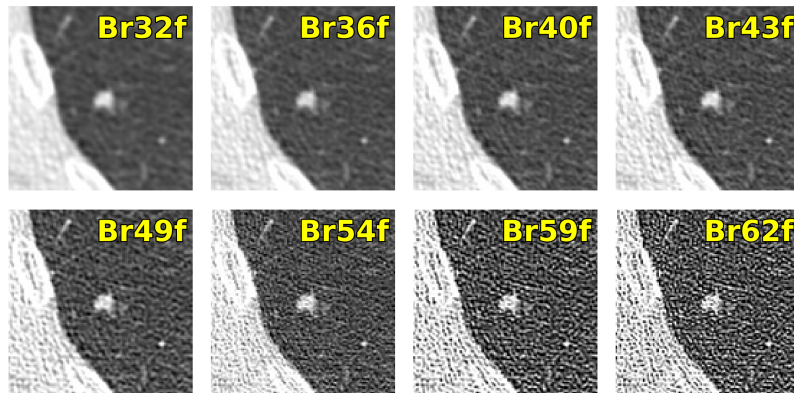


Figure 6.6: A sample patch containing a lung nodule (centered) reconstructed with each of the eight kernels from the Definition AS 64 dataset. As kernel sharpness increases, structured noise common in CT imaging can be seen emerging as streaks through the image.

Examples of lung patch images from each of the eight conditions used in the model development (Definition AS64) dataset are shown in Figure 6.6. The smoothest kernel, Br32f, shows visible smoothing of the image. The sharpest kernel, Br62f, produces a much sharper image with visible structured noise.

A summary of the performance of all 64 train-test combinations within the model development (Definition AS64) dataset is shown as a heatmap in Figure 6.7. The diagonal, outlined with black, shows the performance of the model when predicting on the source distribution ($p_t = p_s$). For the $p_t = p_s$ cases, AUPRC ranged from 0.790 (Br62f) to 0.850 (Br40f), indicating that all eight CNNs achieved reasonable classification performance when tested on their own training condition. As the prevalence for the datasets was 25%, the chance baseline for AUPRC was 0.25.

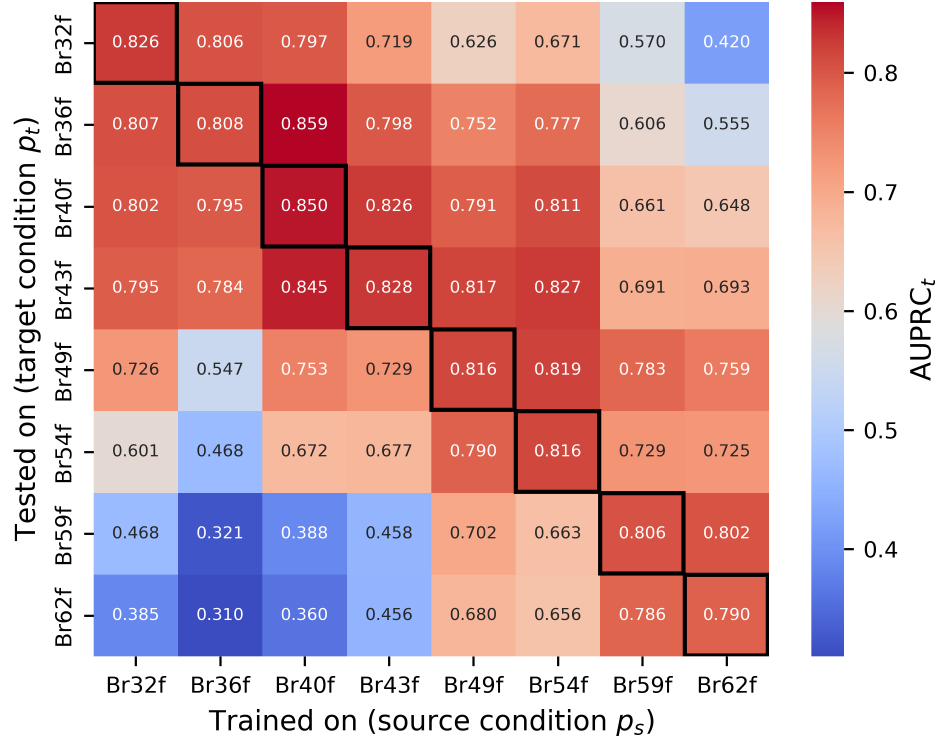


Figure 6.7: The CNN classification performance ($AUPRC_t$) for each train-test condition pairing for the Definition AS64 test set. The diagonal of corresponds to CNN performance on its source domain ($p_t = p_s$), while all non-diagonal elements are OOD target domains ($p_s \neq p_t$).

Performance drops as the target condition diverges from the source, with the most severe degradation occurring at the corners of the matrix. CNNs trained on smooth kernels (Br32f, Br36f) tested on the sharpest kernels (Br59f, Br62f) drop to as low as 0.310–0.385, while the reverse (sharp trained, smooth tested) drops to 0.420. There was a degree of asymmetry in the performance degradation: training on smooth kernels and testing on sharp conditions produced worse performance than the reverse, suggesting the CNNs were more sensitive to sharpening than to smoothing of the target condition relative to training.

While Figure 6.7 shows the raw AUPRC values, Figure 6.8 shows the performance differences relative to each CNN's in-distribution baseline that form the $\Delta\mathcal{M}_{s,t}$ observations used for model fitting. These $K^2 = 64$ $\Delta\text{AUPRC}_{s,t}$ values are the target variable data on which the regression models described in Section 6.2.5 are trained.

Both Figure 6.7 and Figure 6.8 show a region that appears somewhat anomalous compared to the rest of the heatmaps' generally smooth gradients: the CNN trained on Br36f shows a disproportionately large performance drop when tested on Br49f and Br54f relative to adjacent source conditions at the same target, where the degradation is considerably smaller.

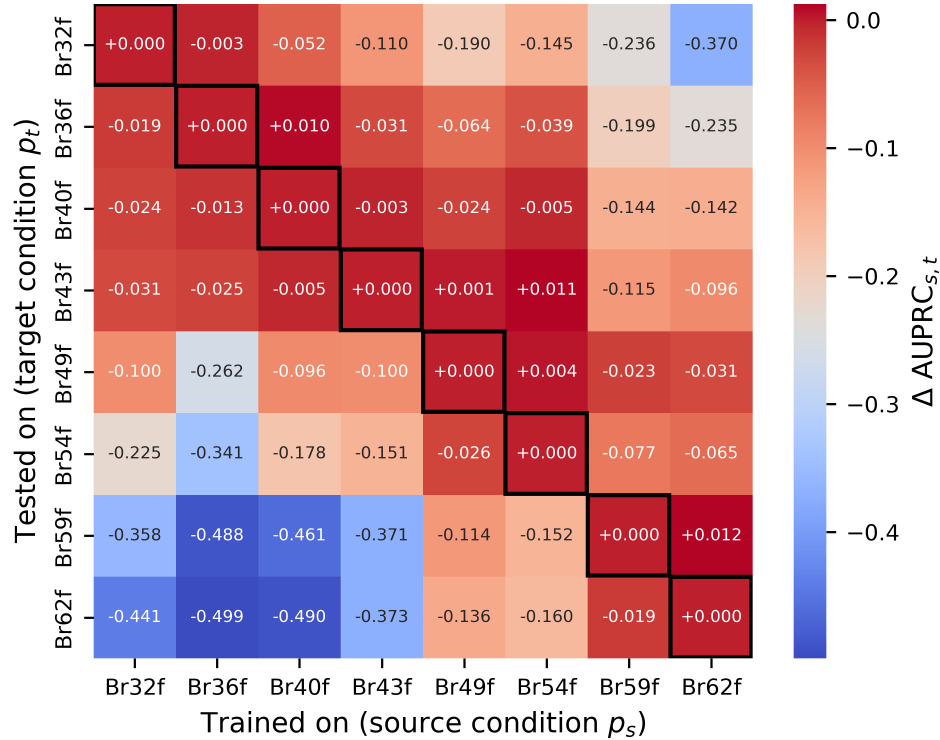


Figure 6.8: The change in CNN performance relative to the condition it was trained on ($\Delta\text{AUPRC}_{s,t}$) for each train-test condition pairing for the Definition AS64 test set. The top-left to bottom-right diagonal is 0, as it corresponds to $p_t = p_s$.

Additional results are provided in Appendix F, including Figure F.1 which shows barplots of the per-condition AUPRC for each of the eight trained CNNs with their bootstrapped 95% confidence intervals.

6.3.2 Modeling CNN performance as a function of image quality

The three candidate regression models (linear, quadratic, and exponential) were fit to the pooled AS64 (Δ IQM, Δ AUPRC) dataset for each image quality metric. Results across all IQMs and model forms are provided in Table F.1.

Of the three model forms, the quadratic model consistently achieved the highest training R_{adj}^2 across nearly all IQMs. MTF-derived metrics were the strongest individual predictors of CNN performance degradation, with $MTF_{0f_{75}}$, $MTF_{0f_{50}}$, and $MTF_{45f_{50}}$ achieving quadratic training R_{adj}^2 values of 0.895, 0.883, and 0.848, respectively. NPS and CNR-derived metrics yielded lower but still meaningful predictive power, with training R_{adj}^2 values generally in the range of 0.5–0.6. Based on training performance, $\Delta MTF_{0f_{75}}$ (the spatial frequency at the 75th centile of the axial MTF curve) was selected as the best single predictor of CNN performance degradation.

Figure 6.9 shows the three candidate model forms fit to the pooled AS64 data using $\Delta MTF_{0f_{75}}$ as the predictor. Each instance of $p_t = p_s$ corresponds to a (0, 0) datapoint. A positive $\Delta MTF_{0f_{75}}$ indicates that the target condition had a higher spatial resolution than the source condition ($MTF_{0f_{75},s} > MTF_{0f_{75},t}$), corresponding to a sharper target image. A negative

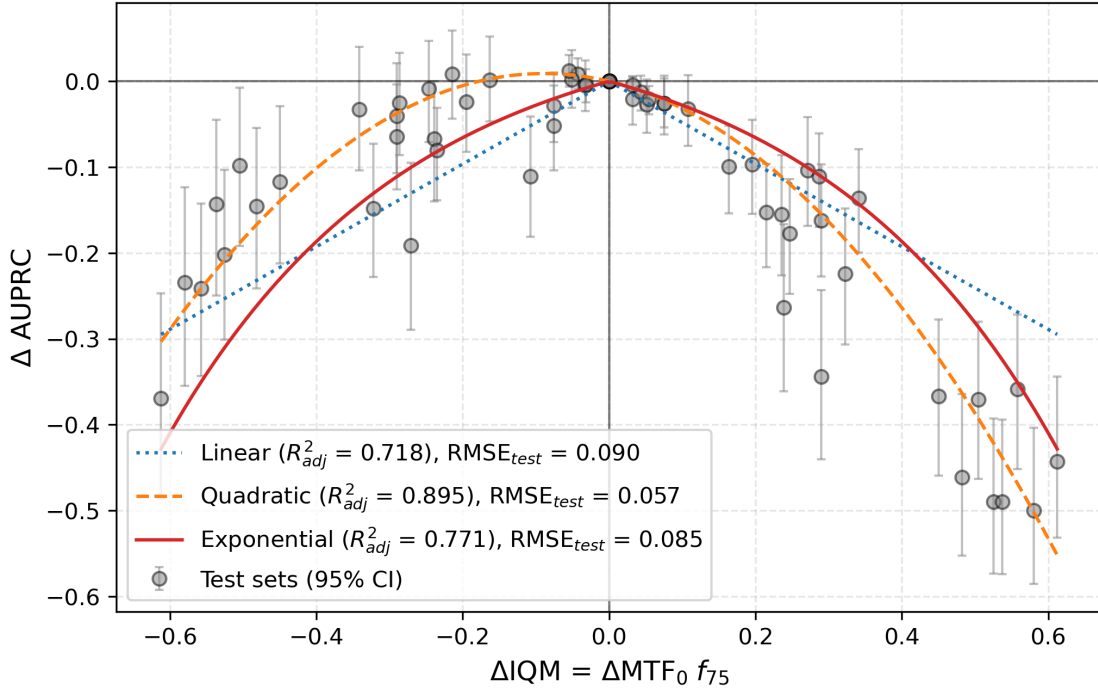


Figure 6.9: The best fit curves for each of the three model forms using $\text{MTF}_0 f_{75}$ as the predictor IQM. Scatter points correspond to measured $(\Delta\text{IQM}, \Delta\text{AUPRC})$ for all Definition AS64 train-test protocol with 95% confidence intervals. The legend indicates the R^2_{adj} for each best fit curve, as well as the RMSE of the test data.

$\Delta\text{MTF}_0 f_{75}$ indicates that the target condition had a lower spatial resolution than the source condition ($\text{MTF}_0 f_{75,s} > \text{MTF}_0 f_{75,t}$), corresponding to a smoother target image.

All three models capture the general trend of increasing performance degradation with increasing magnitude of domain shift. The best fit model is the quadratic (dashed orange, $R^2_{adj} = 0.895$), which is shifted left due to the asymmetry in the data: degradation is steeper for positive $\Delta\text{MTF}_0 f_{75}$ (target sharper than source) than for negative values (target smoother than source). The exponential model (solid red, $R^2_{adj} = 0.771$) captures the accelerating degradation at large shifts but is symmetric about 0. The linear model (dotted blue, $R^2_{adj} = 0.718$) assumes

symmetric degradation that increases at a constant rate with $|\Delta\text{IQM}|$ and is the least accurate.

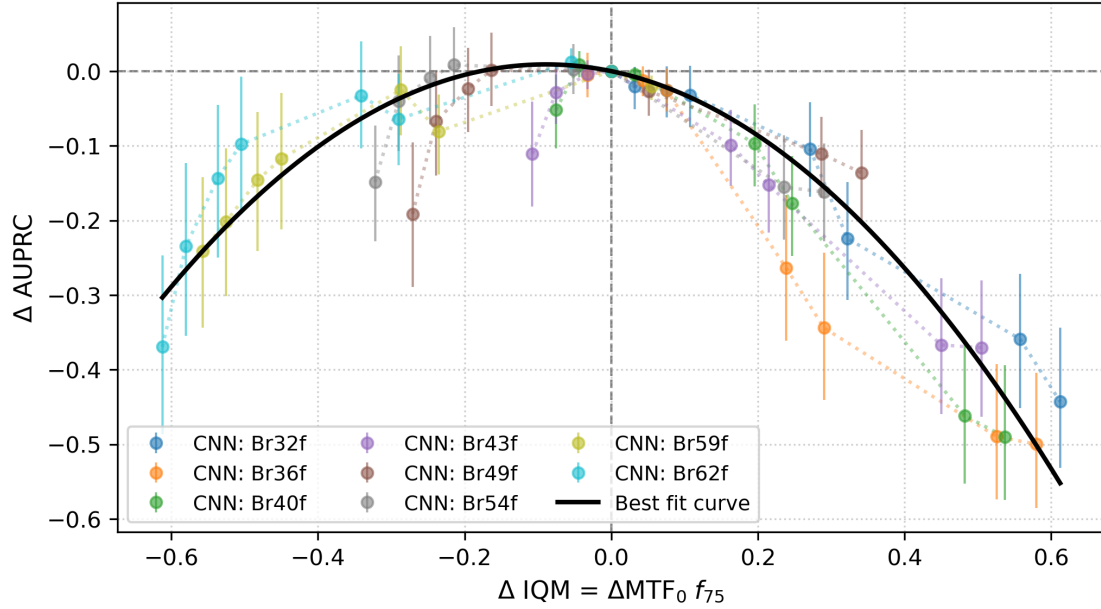


Figure 6.10: Evaluation of the CNN model pooling approach. The best fit curve (quadratic, $\text{MTF}_0 f_{75}$) with actual datapoints labeled by the protocol the CNN was trained on (p_s).

In Figure 6.10, the $(\Delta\text{MTF}_0 f_{75}, \Delta\text{AUPRC})$ datapoints are shown labeled by the CNN's p_s , along with the best fit model form (quadratic). This allows us to evaluate our pooling approach, where the data from the $K = 8$ trained CNNs was combined for modeling. It is apparent that individual CNNs do follow their own trajectories. However, the fit curve does accurately capture the overall behavior of performance degradation with increasing $|\Delta\text{IQM}|$, falling within the 95% CI of most of the datapoints. The Br36f (orange line) is a notable exception, where the performance decrease was underestimated for two datapoints (between ΔIQM of 0.2 and 0.3); this condition was discussed as potentially anomalous in Section 6.3.1. The performance decrease was also underestimated for Br43f, Br49f, and Br54f for the datapoint with the greatest negative change

in $\text{MTF}_0 f_{75}$; as this direction corresponds to a smoother target image compared to the source domain, we can infer that these datapoints are from $p_t = \text{Br32f}$ (the smoothest reconstruction kernel used).

6.3.3 External validation on Force scanner

Each model was evaluated on the held-out Force scanner dataset after being fit on the Definition AS64 dataset. Complete results across all IQMs and model forms are provided in Table F.1.

The model selected in Section 6.3.2 based on the highest R_{adj}^2 of 0.895 on the development data was the quadratic model. When applied to the external validation (Force) data, the quadratic model dropped to $R_{adj}^2 = 0.299$, while the exponential model achieved $R_{adj}^2 = 0.711$ and the linear model $R_{adj}^2 = 0.622$. The exponential model's RMSE on the Force data (0.082) was comparable to its training RMSE (0.085). The quadratic model's RMSE nearly doubled from 0.057 to 0.128. This discrepancy suggests that the quadratic model overfit to the training data.

Table 6.4: Predictive model performance for $\Delta\text{MTF}_0 f_{75}$ as the input feature, evaluated on the development dataset (Definition AS scanner) and held-out external validation dataset (Force scanner). Results are shown for each candidate model form with the best performing model form for each dataset bolded.

Model form	Definition AS64 (train)			Force (held-out)		
	R_{adj}^2	RMSE	MAE	R_{adj}^2	RMSE	MAE
Quadratic	0.895	0.057	0.037	0.299	0.128	0.099
Exponential	0.771	0.085	0.059	0.711	0.082	0.062
Linear	0.718	0.090	0.062	0.622	0.094	0.076

The best fit curves for the three model forms using the $\Delta\text{MTF}_0 f_{75}$ predictor are shown

with the held-out Force dataset overlaid in Figure 6.11. This visual shows clearly that the exponential model form better captures the data than the quadratic. However, all three model forms show a trend of overestimating the performance degradation for positive $\Delta\text{MTF}_0 f_{75}$, and underestimating the performance degradation for negative $\Delta\text{MTF}_0 f_{75}$. That is, on the Force data, overall the CNN's are performing better than predicted when the p_t is sharper than the CNN's training data, and worse than predicted when the p_t is smoother than the CNN's training data.

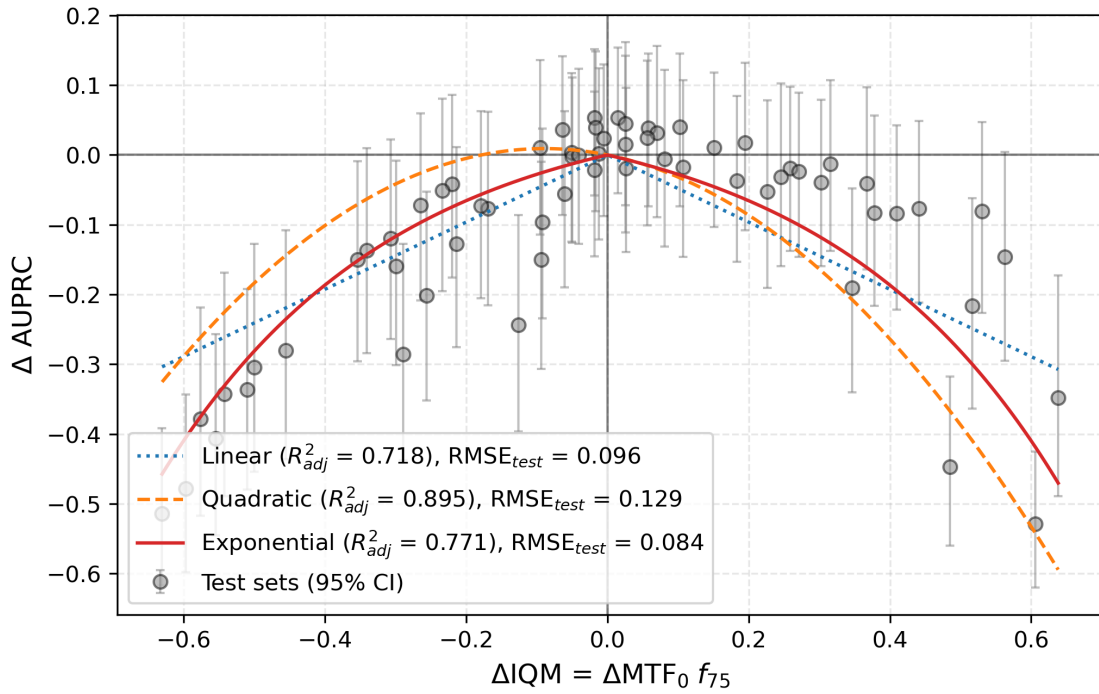


Figure 6.11: The best fit curves for each of the three model forms using $\text{MTF}_0 f_{75}$ as the predictor IQM. Scatter points correspond to measured $(\Delta\text{IQM}, \Delta\text{AUPRC})$ for all Definition AS64 train-test protocol with 95% confidence intervals. The legend indicates the R^2_{adj} for each best fit curve, as well as the RMSE of the test (Force) data.

6.4 Discussion and conclusions

This work investigated whether phantom-derived image quality metrics can predict the performance degradation of a CNN when applied to CT imaging conditions outside its training domain. The CNN performance results confirmed that reconstruction kernel has a substantial impact on classification performance, with AUPRC dropping from in-distribution baselines of 0.79–0.85 to as low as 0.31 for the most extreme source-target mismatches. This degradation was asymmetric: CNNs were more sensitive to smoothing of the target condition relative to their training domain than to sharpening. This may be due to smoothing effectively removing (blurring) discriminative features from the image that the CNN had fit to for classification. As sharpening preserves and amplifies existing features, it may have less impact despite introducing additional noise.

Of the image quality metrics evaluated, MTF-derived metrics were the strongest predictors of performance degradation, with $\Delta\text{MTF}_{0f_{75}}$ achieving the highest training R_{adj}^2 . This is expected given that the experimental design varied only the reconstruction kernel at a fixed dose level: kernel selection primarily affects in plane spatial resolution, which the axial MTF directly characterizes. In a study that also varied dose or slice thickness, noise-related metrics or oblique MTF would likely become more relevant predictors.

The quadratic model achieved the best fit on the development data ($R_{adj}^2 = 0.895$) but generalized poorly to the Force scanner ($R_{adj}^2 = 0.299$). The exponential model had lower training performance ($R_{adj}^2 = 0.771$), but retained most of its predictive accuracy on the external dataset ($R_{adj}^2 = 0.711$, RMSE = 0.082 vs 0.085 on training). This indicated that the quadratic

form overfit to the training data. The exponential model was constrained to be symmetric about $x = 0$, which may have had a regularization effect that prevented overfitting to the AS64 data. The model selection procedure, based on the training R_{adj}^2 , did not successfully produce a model that was able to generalize to the external validation data.

On the Force data, all three model forms showed a trend of overestimating degradation for positive $\Delta\text{MTF}_0 f_{75}$ and underestimating it for negative values. This directional bias likely reflects the fact that moving to a different scanner system introduces additional variables besides reconstruction kernel. Therefore, even when two protocols are matched on a single IQM such as $\text{MTF } f_{75}$, the Force scanner may differ from the AS64 in noise texture, dose efficiency, or other image characteristics not captured by that metric. A single-predictor model cannot account for these additional dimensions of variation. Future work should consider multivariate models that combine complementary IQMs to more fully characterize the domain shift between scanner platforms.

This study is an initial proof-of-concept with several limitations related to that scope. The experiment was restricted to a small number of imaging conditions across two scanner systems, producing a limited number of $(\Delta\text{AUPRC}, \Delta\text{IQM})$ data points. It was also limited to a single task (lung nodule classification), a single architecture (ResNet-18), and a single performance metric (AUPRC). Future work should investigate whether this approach generalizes to different architectures, tasks, and metrics.

The pooling of observations from all eight independently trained CNNs into a single regression dataset implicitly assumes that each CNN’s sensitivity to domain shift follows the same functional

relationship. The results show that this assumption is only approximately true. The individual CNNs followed their own trajectories, with the Br36f-trained CNN exhibiting a notably steeper rate of degradation than its neighbors. These differences may be due to some characteristic of the source condition's feature space, or from the stochasticity of the training process. These two sources cannot be distinguished with a single training run per condition. The pooled model averages over these CNN-specific sensitivities, which reduces predictive accuracy for individual CNNs that deviate from the group trend. With a larger dataset, future work could investigate per-source modeling, or incorporate the source condition's absolute IQM values as an additional predictor.

The source and target domains were each defined by a single reconstruction condition, so that the domain shift was fully described by a single protocol change. In practice, training datasets are often heterogeneous, and deployment environments may comprise a mixture of protocols. Under these conditions, the effective domain shift is no longer a single well-defined ΔIQM , and the single-predictor regression framework may not directly apply. Similarly, all conditions in this study varied only the reconstruction kernel at a fixed CTDI_{vol} , whereas real protocol differences can involve simultaneous changes in dose, kernel, slice thickness, and reconstruction algorithm. Extending the framework to accommodate these shifts as well as heterogeneous source domains remains an important direction for future work.

Finally, the differences in CT reconstruction protocol addressed here represent only one potential source of domain shift in clinical deep learning. Performance degradation can also stem from variations in patient demographics, disease prevalence, or institutional differences in clinical

practice [106]. This study deliberately isolates protocol-induced shift because it is controllable and quantifiable through phantom measurements, but the framework would need to be evaluated alongside other sources of variation before clinical deployment.

Despite these limitations, the exponential model's Force R_{adj}^2 of 0.711 demonstrates that a single phantom-derived image quality metric can explain a substantial portion of CNN performance degradation on an unseen scanner. This provides initial evidence that the IQM-based prediction framework can generalize beyond the system on which it was developed. With additional validation, this approach could provide a pathway for prospective performance prediction: before deploying a trained model under a new CT protocol, a physicist could acquire a phantom scan, compute the relevant IQMs, and obtain a quantitative estimate of expected performance change without requiring patient data. This has practical implications for protocol management, and model development and deployment in clinical AI systems.

7 Conclusions

7.1 Summary of work

This dissertation investigates whether physical image quality metrics can provide a quantitative, scanner-independent framework for characterizing and predicting variation in quantitative CT imaging across protocols.

Chapter 2 presented the design, manufacture, and initial evaluation of an open-source CT texture phantom specifically developed for radiomics research. The phantom features 3D-printed modular inserts with diverse texture patterns that elicit distinct radiomic signatures. This phantom served as the foundation for the controlled experimental studies in subsequent chapters.

Chapter 3 described the development of a comprehensive dataset acquired by scanning a texture phantom and an image quality phantom (Corgi, The Phantom Laboratory) using 170 diverse imaging protocols spanning four scanner models from two vendors across a range of radiation dose levels and reconstruction kernels. For each protocol, we measured 3D image quality (MTF, NPS) and extracted radiomic features from four unique texture objects. This paired measurement approach enabled direct investigation of relationships between image quality and radiomic features under controlled, identical imaging conditions. A version of the dataset was released publicly on Zenodo.

Chapter 4 characterized the relationships between physical image quality metrics and radiomic

feature values across the diverse protocol space using the dataset developed in Chapter 3. Through feature selection, visualization, and correlation analysis, we established that radiomic features respond systematically to fundamental image quality characteristics.

Chapter 5 developed regression models that predicted radiomic feature values from physical image quality metrics. Using polynomial regression with ElasticNet regularization, we demonstrated that image quality metrics can accurately predict radiomic variation within distribution, but with limited ability to generalize to unseen scanner platforms across vendors.

Chapter 6 extended the image quality-based prediction framework to deep learning models, investigating whether physical image quality metrics can predict out-of-distribution performance degradation of convolutional neural networks for lung nodule classification. Using clinical lung cancer screening CT data reconstructed with multiple kernels, we demonstrate that CNN performance changes across imaging conditions can be modeled as a function of measured image quality.

Through this body of work, we introduce a framework for characterizing and predicting CT protocol-induced variation in quantitative imaging using physical image quality as a basis.

7.2 Future directions

The work in this dissertation establishes that radiomic texture features respond systematically and, for most features, predictably to measured image quality, and that this relationship can be modeled within a given range of protocols. Several directions would extend these results toward

clinical use.

The motivation and long-term goal of this work was the definition of physical, vendor-neutral imaging inclusion criteria; this remains to be fully developed and tested. The predictive models developed here provide the necessary components: given a tolerance on acceptable radiomic feature variation or change in CNN performance, the region of image quality space over which a feature remains within that tolerance could be computed and used to accept or reject a candidate protocol. Demonstrating this prospectively, by confirming that protocols falling inside the defined region behave as predicted while those outside it do not, is a natural next step. This would require a specific task, and a clinically defined tolerance.

The relationships characterized in Chapters 4 and 5 were measured on simple geometric phantom textures. How they extend to real tissue is not established by this work and would need to be tested directly. The phantom textures are a deliberate simplification that isolates the protocol effect, and the degree to which feature behavior transfers to the heterogeneity of real tissue is an open question.

The cross-vendor results in Chapter 5 also motivate further investigation. Prediction degraded sharply when a model was applied to a vendor whose image quality range was not represented in training, a failure attributable to extrapolation beyond the training support. Future work should explore the performance on the within-range test set protocols, as well as alternative methods of IQM selection that may improve robustness. It could also explore alternative modeling approaches to better handle the case of out-of-range predictions. Relatedly, acquiring additional protocols to span the union of vendor image quality ranges would reduce the gaps that drive this extrapolation

in the first place, if such protocols are available.

The image quality metrics used here are well-posed for the linear, filtered backprojection reconstruction case, but clinical CT increasingly relies on iterative and deep-learning reconstruction, which are nonlinear. Extending the framework to these methods would require task-based metrics measured at representative contrast and dose, and it is not yet clear whether a single feature-to-image-quality relationship can be maintained when the underlying system response depends on the imaged object.

Another future direction is to reformulate the radiomic feature prediction target in terms of protocol *differences* rather than absolute feature values. While we took this approach in Chapter 6, we did not investigate the difference approach in Chapter 4 or Chapter 5. The models developed here predict a feature value from the image quality of a single protocol, and protocol comparisons are made only afterward; a delta formulation would instead predict the change in a feature directly from a measure of the difference between two protocols' image quality. An advantage of this approach is that it could leverage difference metrics defined on the full measured curves rather than on scalar summaries, such as the root-mean-square error between two MTFs or between two NPS curves. Because such metrics use the entire curve, they may distinguish protocols more efficiently than the metrics used here. The use of this type of difference metric could also be investigated for Chapter 6.

A Additional phantom manufacturing details for

Chapter 2

This appendix provides additional details for the manufacturing of the phantom described in Chapter 2.

A.1 Materials and manufacturing

The phantom platform utilizes a hybrid manufacturing approach, combining fused deposition modeling (FDM)-based 3D printing for the hollow outer shell and modular inserts with a liquid two-part epoxy resin cast for the core. This method dramatically reduces print time and mitigates the severe thermal warping that frequently causes dense FDM prints to fail.

The printable 3D models (STL files) are publicly available at [removed for peer review]. To ensure the interchangeable inserts slide smoothly into the phantom body without jamming, a 0.15 mm radial tolerance (0.3 mm total diametric clearance) is already incorporated into the provided design files.

The permanent outer shells and inserts were printed using a Bambu Lab X1-Carbon printer (0.4 mm nozzle, 0.28 mm layer height, 100% infill) and Jade White polylactic acid (PLA) filament (Bambu Lab, Shenzhen, China). The phantom body mold took approximately four hours to print using these settings.

To maximize metrological reproducibility across institutions, we strongly recommend utilizing un-dyed or pure white PLA; filaments containing heavy dyes or specialized additives (e.g., “glow-in-the-dark” blends) introduce unpredictable variations in CT attenuation. Furthermore, alternative FDM materials such as PET-G were found to be chemically incompatible with the epoxy resin and should be avoided.

The core of the phantom was cast using a two-part, clear “Deep Pour” epoxy resin system (Teexpert, <https://teexpert-resin.com>) to ensure a low-exotherm cure that does not melt the PLA shell. The resin and hardener were measured strictly by weight in a 2:1 ratio to ensure stoichiometric accuracy, and vigorously mixed for 10 minutes using a power drill equipped with a mixing paddle. Following mixing, the resin was placed inside a 3-gallon vacuum chamber powered by a 5 CFM dual-stage vacuum pump (Kozyvacu, Kozyard, Fontana, CA) to extract introduced air. Because epoxy foams and expands rapidly under vacuum, we strongly recommend calculating the volume required for a single phantom body, mixing it in an oversized disposable container, and degassing one batch at a time to prevent boiling over into the pump. For the phantoms manufactured in this study, the degassing phase was terminated once visual off-gassing subsided; however, to improve batch-to-batch reproducibility, we recommend future users adopt a strict 30-minute vacuum duration per batch to standardize air removal rather than relying on subjective visual cues.

The degassed epoxy was slowly poured over wooden stir stick resting inside the mold to safely break the surface tension and minimize the re-introduction of air. The liquid epoxy should be filled flush with the top rim of the PLA shell; the mold design intentionally includes 4 mm of

excess vertical depth to accommodate the approximate 10% volumetric shrinkage that occurs during the 48-hour room-temperature cure, ensuring a consistent final phantom depth of 40 mm. Minor bubbles should generally be left alone, as attempting to manually remove them (e.g., via toothpick) or utilizing a heat gun risks introducing more air or warping the PLA shell. The phantoms must be left undisturbed on a leveled surface during curing to prevent a slanted top surface that could introduce geometric misalignments.

For final image acquisition, the cylindrical phantom body can be stabilized using either the standard cradle provided with the ACR CT accreditation phantom or a custom 3D-printed stand, and aligned orthogonally to the scan plane utilizing the scanner's standard lasers and the physical alignment markings integrated directly into the phantom shell.

A.2 Texture definitions and print settings

All design work was done using Fusion 360 (version 2.0.17954; Autodesk, San Francisco, CA). This parametric computer-aided design (CAD) software was selected to ensure the precise, mathematically driven definition of the 3D structures, allowing for highly reproducible spatial dimensions prior to slicing and fabrication. Detailed design and print parameters are provided in Table A.1. All model slicing and print configuration was done in Bambu Studio (version 2.1.0.59; Bambu Lab, Shenzhen, China). Most textures (2-6, 8) were printed using a 0.4 mm nozzle with a layer height of 0.2 mm. Two textures (1 and 7) utilized a 0.2 mm print nozzle to achieve smaller details. The printed layer heights were chosen because they were the printer-optimized default

settings in our system and is below the minimum slice thickness of most CT scanners.

All inserts were manufactured using Jade White PLA (Bambu Lab, Shenzhen, China) at 100% infill, utilizing the default recommended infill pattern generated by the slicing software. Default print profiles were utilized for all inserts, with the exception of those requiring a 0.2 mm nozzle (Textures 1 and 7), for which the highest-quality print preset was selected.

Table A.1: Detailed geometric definitions and printing specifications for the eight modular texture inserts.

Texture ID	Geometry	Dimensions & pitch	Nozzle/Layer (mm)
1	Random circles (jittered)	base $r = 0.3$ mm, jitter ≤ 0.8 mm	0.2 / 0.08
2	Square rods grid	7.0 mm edge, 8.0 mm pitch	0.4 / 0.20
3	Checkerboard	3.0 mm squares	0.4 / 0.20
4	Circular rod grid	5.0 mm dia, 6.0 mm pitch	0.4 / 0.20
5	Circular rod grid	3.0 mm dia, 6.0 mm pitch	0.4 / 0.20
6	Circular rod grid	3.0 mm dia, 4.0 mm pitch	0.4 / 0.20
7	Inverted rods (air) grid	0.75 mm dia, 2.5 mm pitch	0.2 / 0.08
8	Solid block	N/A	0.4 / 0.20

A.3 Estimated production costs

The itemized manufacturing costs for the hardware and materials utilized in this study are provided in Table A.2. All listed values represent prices at the time of purchase, excluding taxes, tariffs, and shipping. In addition to the listed equipment, fabrication requires a standard power drill equipped with a mixing paddle for vigorous blending of the epoxy resin and hardener, as well as standard laboratory safety equipment (gloves, eye protection, and respiratory protection such as a ventilated fume hood).

Assuming a research group already possesses the necessary capital equipment (a 3D printer and vacuum degassing chamber), the marginal consumable cost to produce one complete phantom system (one epoxy body and one complete set of eight texture inserts) is highly economical. The material yields from this experiment were four phantom bodies and approximately four of the eight-texture inset sets (three sets described in the experiment, plus approximately one additional set in test prints of texture designs), produced from 3 gallons of resin and 3 spools of filament (with some filament remaining). With a total materials cost of \$202.31 to produce four phantom systems, the estimated material cost was about \$50 per complete system.

Even when factoring in the total capital equipment investment required to initiate this project from scratch, the amortized cost across the limited four-phantom manufacturing run was approximately \$401 per phantom (based on the estimated total cost of \$1,604.51). By comparison, standard proprietary quantitative imaging phantoms typically range from thousands to tens of thousands of dollars. Therefore, this open-source methodology provides an established research lab with a tool that is roughly two orders of magnitude cheaper (for marginal materials) than commercial alternatives, while lowering the initial barrier to entry for an unequipped lab by at least one order of magnitude. Furthermore, the lab now possesses the equipment required to both produce additional phantoms and extend the platform. Cost estimates exclude personnel time required for printing, casting, and post-processing.

Table A.2: Itemized cost breakdown of the capital equipment and consumable materials required to manufacture the four phantom bodies and three complete insert sets evaluated in this study. Costs reflect prices at the time of purchase and exclude local taxes, shipping, and tariffs.

Item	Source	Quantity	Cost (USD)
<i>Capital Equipment</i>			
X1 Carbon 3D Printer	Bambu Lab	1	1199.00
Dual-stage Vacuum Pump (5 CFM)	Kozyvacu (Amazon)	1	119.99
3-Gallon Vacuum Chamber	Vevor (Amazon)	1	83.21
Total equipment cost			1402.20
<i>Consumable Materials</i>			
Deep Pour Epoxy Resin	Teexpert (Amazon)	3 gallons	142.34
Jade White PLA Filament	Bambu Lab	3 spools	59.97
Total materials cost			202.31
Total cost (4 phantom systems)			1604.51

B Additional results for Chapter 2

This appendix provides additional results for Chapter 2.

B.1 Additional results for texture insert reproducibility

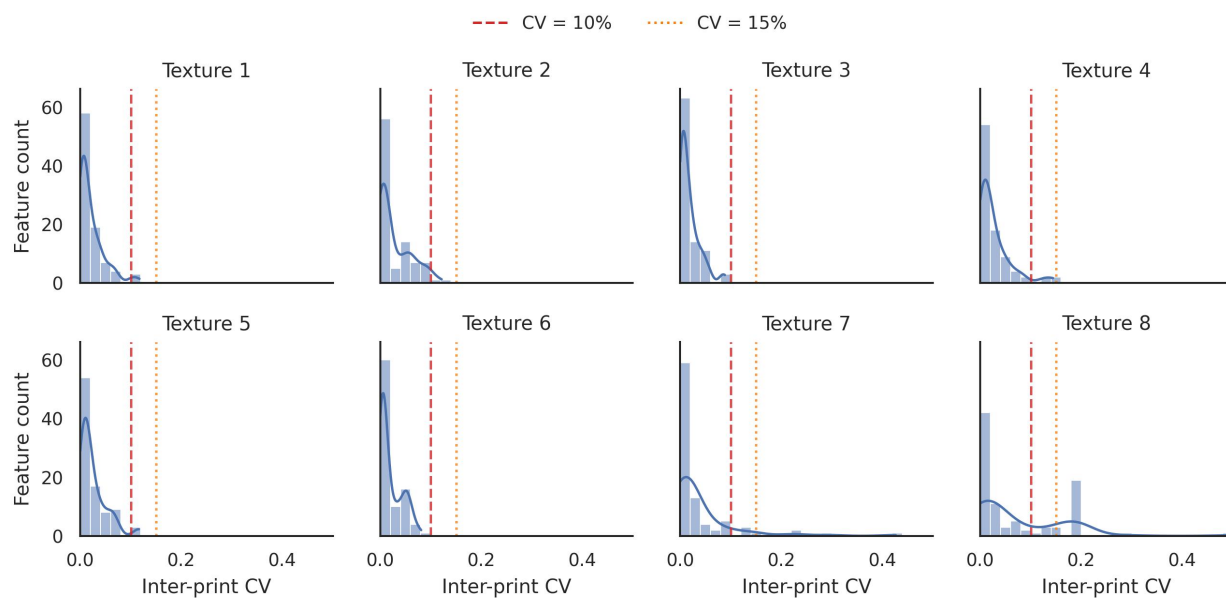


Figure B.1: Distributions of the inter-print CV values measured for each texture insert. Bars correspond to the number of radiomic features at a given CV value. The two thresholds used in the evaluation are indicated with vertical lines.

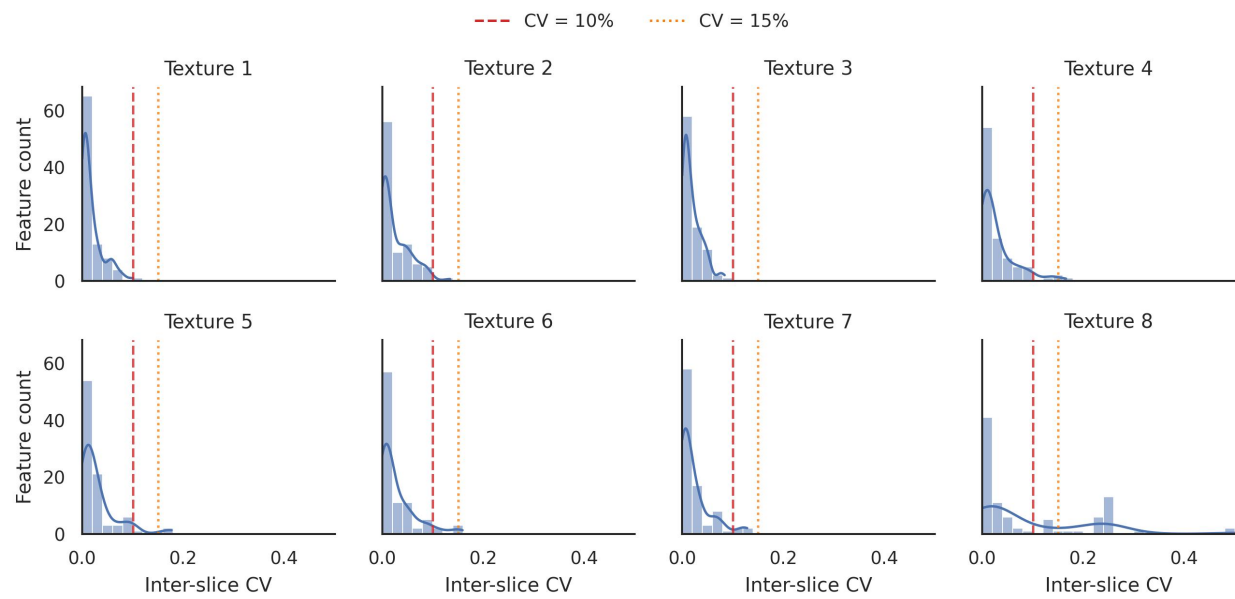


Figure B.2: Distributions of the inter-slice CV values measured for each texture insert. Bars correspond to the number of radiomic features at a given CV value. The two thresholds used in the evaluation are indicated with vertical lines.

Table B.1: Radiomic features with either inter-print or inter-slice CV above the stricter 0.1 threshold for Textures 1-7. The CV is provided where > 0.1 . Texture 8 was excluded, as it was solid insert rather than a heterogeneous texture pattern, making the reproducibility of radiomic texture features a less meaningful evaluation of manufacturing consistency.

Texture	Radiomic feature	Inter-print CV	Inter-slice CV
1	GLCM joint energy	0.117	0.100
1	GLRLM run variance	0.104	—
1	GLSZM zone variance	0.104	—
2	GLCM inverse variance	—	0.134
2	GLCM joint energy	0.121	—
2	GLCM maximum probability	0.103	—
4	First order skewness	0.127	0.126
4	GLCM joint energy	0.122	0.166
4	GLRLM run variance	0.144	0.141
4	GLSZM zone variance	0.144	0.141
5	GLCM joint energy	0.115	0.158
5	GLDM dependence variance	—	0.107
5	GLRLM run variance	0.116	0.177
5	GLSZM zone variance	0.116	0.177
6	GLCM joint energy	—	0.159
6	GLRLM long run low gray level emphasis	—	0.115
6	GLRLM run variance	—	0.146
6	GLSZM large area low gray level emphasis	—	0.115
6	GLSZM zone variance	—	0.146
7	GLCM maximum probability	0.426	0.116
7	GLDM dependence variance	0.139	0.115
7	GLDM large dependence low gray level emphasis	—	0.078
7	GLDM low gray level emphasis	0.147	—
7	GLRLM long run low gray level emphasis	0.224	—
7	GLRLM run variance	0.139	0.129
7	GLSZM large area low gray level emphasis	0.224	—
7	GLSZM zone variance	0.139	0.129

B.2 Additional results for sensitivity to imaging protocol variation

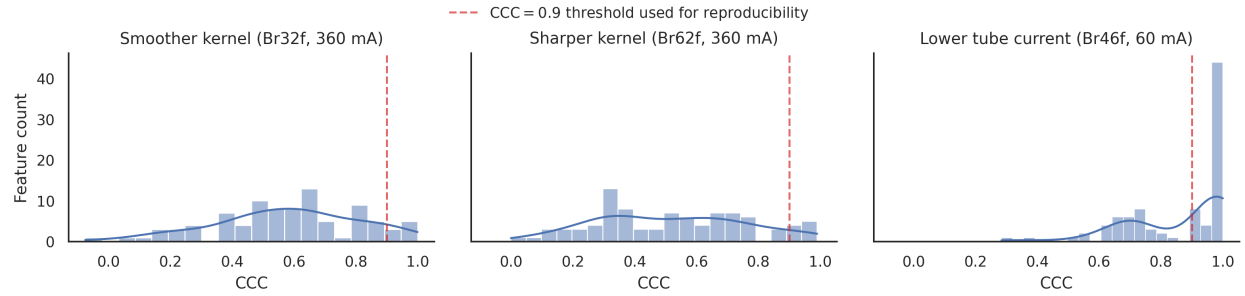


Figure B.3: Distributions of CCC values measured between radiomic feature sets for each CT protocol comparison (relative to the Br46f, 360 mA reference protocol). Bars correspond to the number of radiomic features at a given CCC value. The 0.9 threshold used in the evaluation is indicated with a vertical line.

C Table of radiomic features included in

Chapter 3

This appendix provides the complete list of radiomic features used in the construction of the dataset in Chapter 3.

Table C.1: Radiomics features and categories extracted from Pyradiomics [76].

Category	Feature Name
First Order Statistics (19 features)	
First Order	10th Percentile
First Order	90th Percentile
First Order	Energy
First Order	Entropy
First Order	Interquartile Range
First Order	Kurtosis
First Order	Maximum
First Order	Mean
First Order	Mean Absolute Deviation
First Order	Median
First Order	Minimum
First Order	Range
First Order	Robust Mean Absolute Deviation
First Order	Root Mean Squared
First Order	Skewness
First Order	Standard Deviation
First Order	Total Energy
First Order	Uniformity
First Order	Variance

Continued on next page

Table C.1 – Continued from previous page

Category	Feature Name
Gray Level Co-occurrence Matrix (24 features)	
GLCM	Autocorrelation
GLCM	Cluster Prominence
GLCM	Cluster Shade
GLCM	Cluster Tendency
GLCM	Contrast
GLCM	Correlation
GLCM	Difference Average
GLCM	Difference Entropy
GLCM	Difference Variance
GLCM	Informational Measure of Correlation 1
GLCM	Informational Measure of Correlation 2
GLCM	Inverse Difference
GLCM	Inverse Difference Moment
GLCM	Inverse Difference Moment Normalized
GLCM	Inverse Difference Normalized
GLCM	Inverse Variance
GLCM	Joint Average
GLCM	Joint Energy
GLCM	Joint Entropy
GLCM	Maximum Probability
GLCM	Sum Average
GLCM	Sum Entropy
GLCM	Sum of Squares
GLCM	Sum Variance
Gray Level Run Length Matrix (16 features)	
GLRLM	Gray Level Non-Uniformity
GLRLM	Gray Level Non-Uniformity Normalized
GLRLM	Gray Level Variance
GLRLM	High Gray Level Run Emphasis
GLRLM	Long Run Emphasis
GLRLM	Long Run High Gray Level Emphasis

Continued on next page

Table C.1 – Continued from previous page

Category	Feature Name
GLRLM	Long Run Low Gray Level Emphasis
GLRLM	Low Gray Level Run Emphasis
GLRLM	Run Entropy
GLRLM	Run Length Non-Uniformity
GLRLM	Run Length Non-Uniformity Normalized
GLRLM	Run Percentage
GLRLM	Run Variance
GLRLM	Short Run Emphasis
GLRLM	Short Run High Gray Level Emphasis
GLRLM	Short Run Low Gray Level Emphasis
Gray Level Size Zone Matrix (16 features)	
GLSZM	Gray Level Non-Uniformity
GLSZM	Gray Level Non-Uniformity Normalized
GLSZM	Gray Level Variance
GLSZM	High Gray Level Zone Emphasis
GLSZM	Large Area Emphasis
GLSZM	Large Area High Gray Level Emphasis
GLSZM	Large Area Low Gray Level Emphasis
GLSZM	Low Gray Level Zone Emphasis
GLSZM	Size Zone Non-Uniformity
GLSZM	Size Zone Non-Uniformity Normalized
GLSZM	Small Area Emphasis
GLSZM	Small Area High Gray Level Emphasis
GLSZM	Small Area Low Gray Level Emphasis
GLSZM	Zone Entropy
GLSZM	Zone Percentage
GLSZM	Zone Variance
Neighbouring Gray Tone Difference Matrix (5 features)	
NGTDM	Busyness
NGTDM	Coarseness
NGTDM	Complexity
NGTDM	Contrast

Continued on next page

Table C.1 – Continued from previous page

Category	Feature Name
NGTDM	Strength
Gray Level Dependence Matrix (14 features)	
GLDM	Dependence Entropy
GLDM	Dependence Non-Uniformity
GLDM	Dependence Non-Uniformity Normalized
GLDM	Dependence Variance
GLDM	Gray Level Non-Uniformity
GLDM	Gray Level Non-Uniformity Normalized
GLDM	Gray Level Variance
GLDM	High Gray Level Emphasis
GLDM	Large Dependence Emphasis
GLDM	Large Dependence High Gray Level Emphasis
GLDM	Large Dependence Low Gray Level Emphasis
GLDM	Low Gray Level Emphasis
GLDM	Small Dependence Emphasis
GLDM	Small Dependence High Gray Level Emphasis
GLDM	Small Dependence Low Gray Level Emphasis

D Distributions of radiomic features included in Chapter 3

This appendix provides the complete set distributions of radiomic features included in the dataset in Chapter 3.

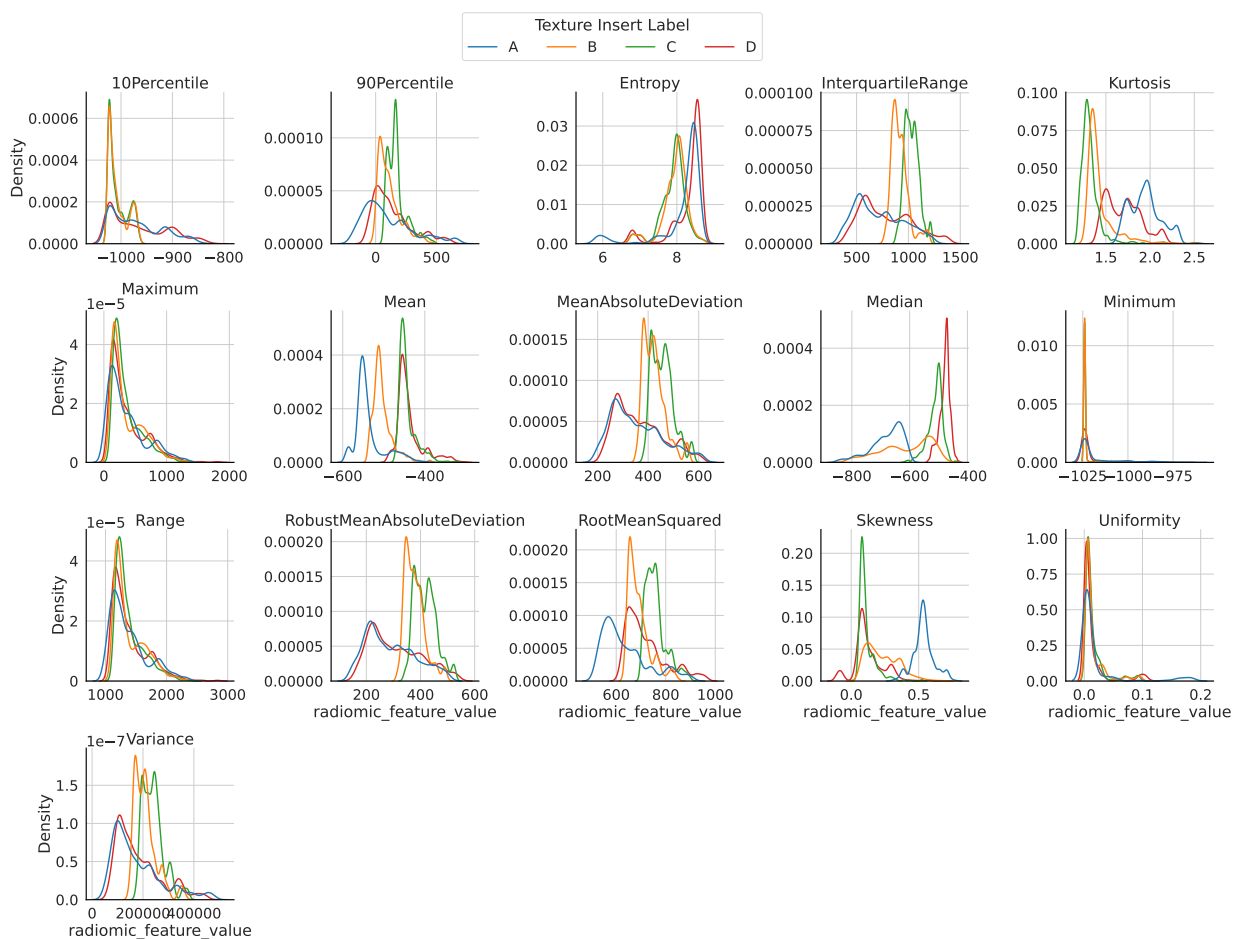


Figure D.1: Distributions for all first order features.

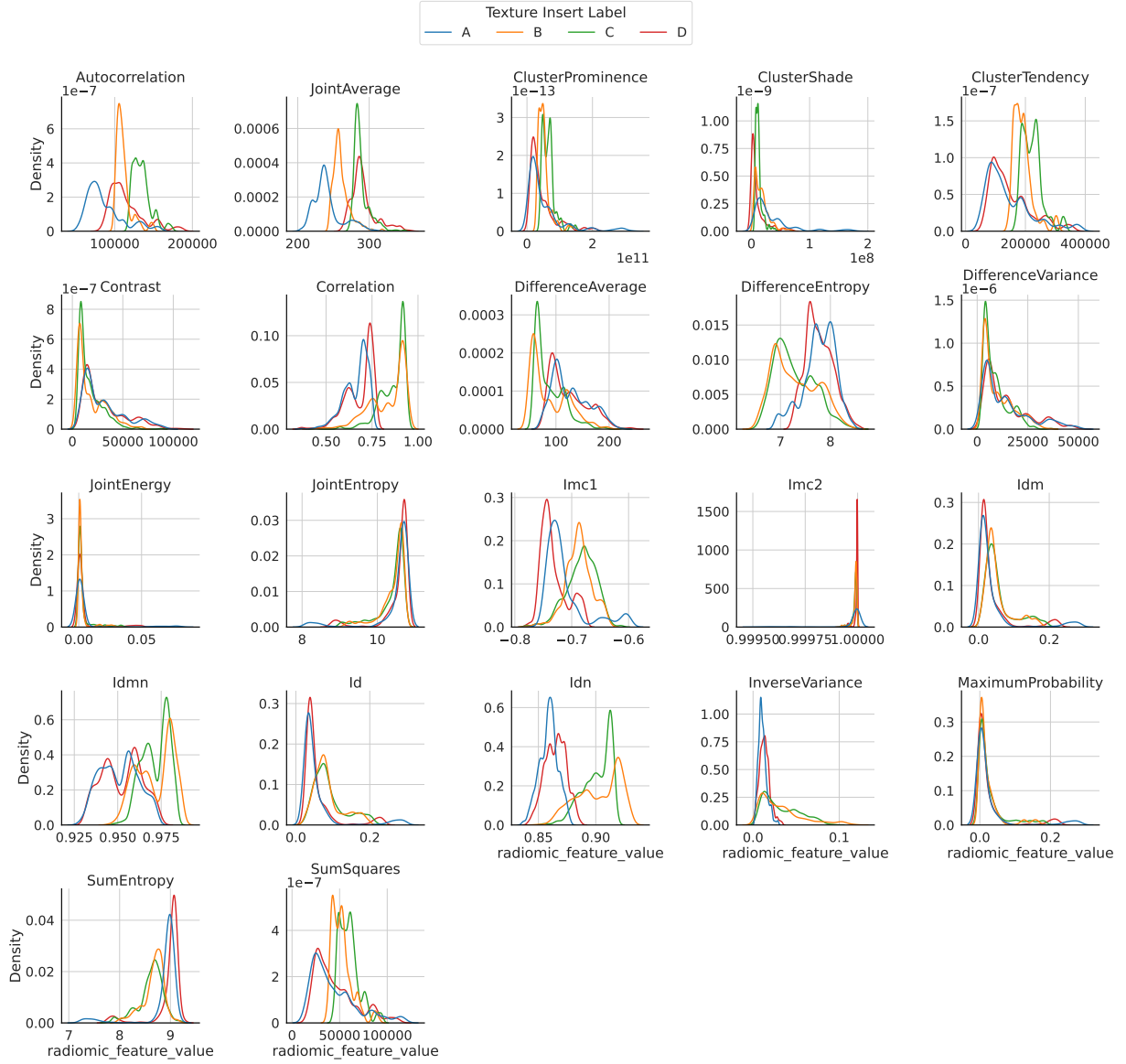


Figure D.2: Distributions for all GLCM features.

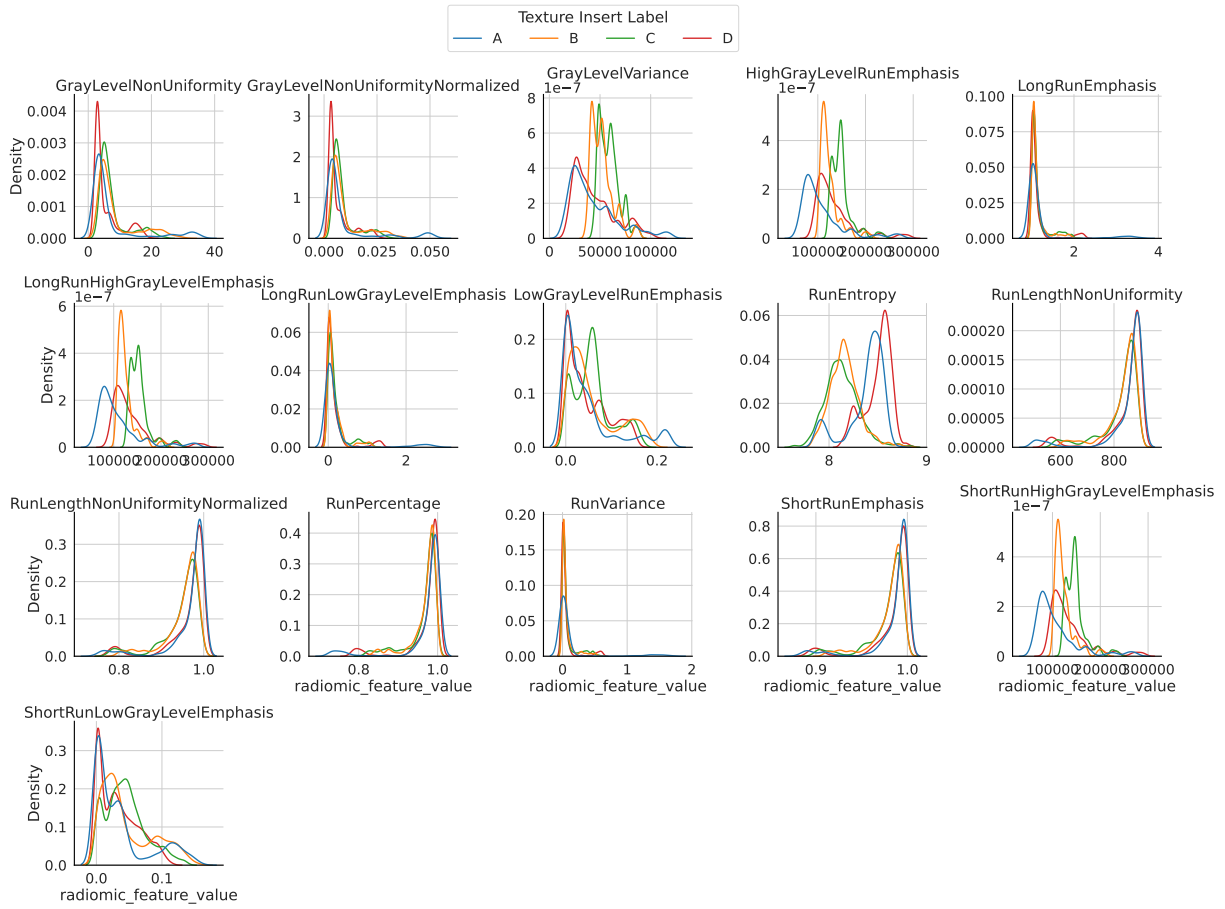


Figure D.3: Distributions for all GLRLM features.

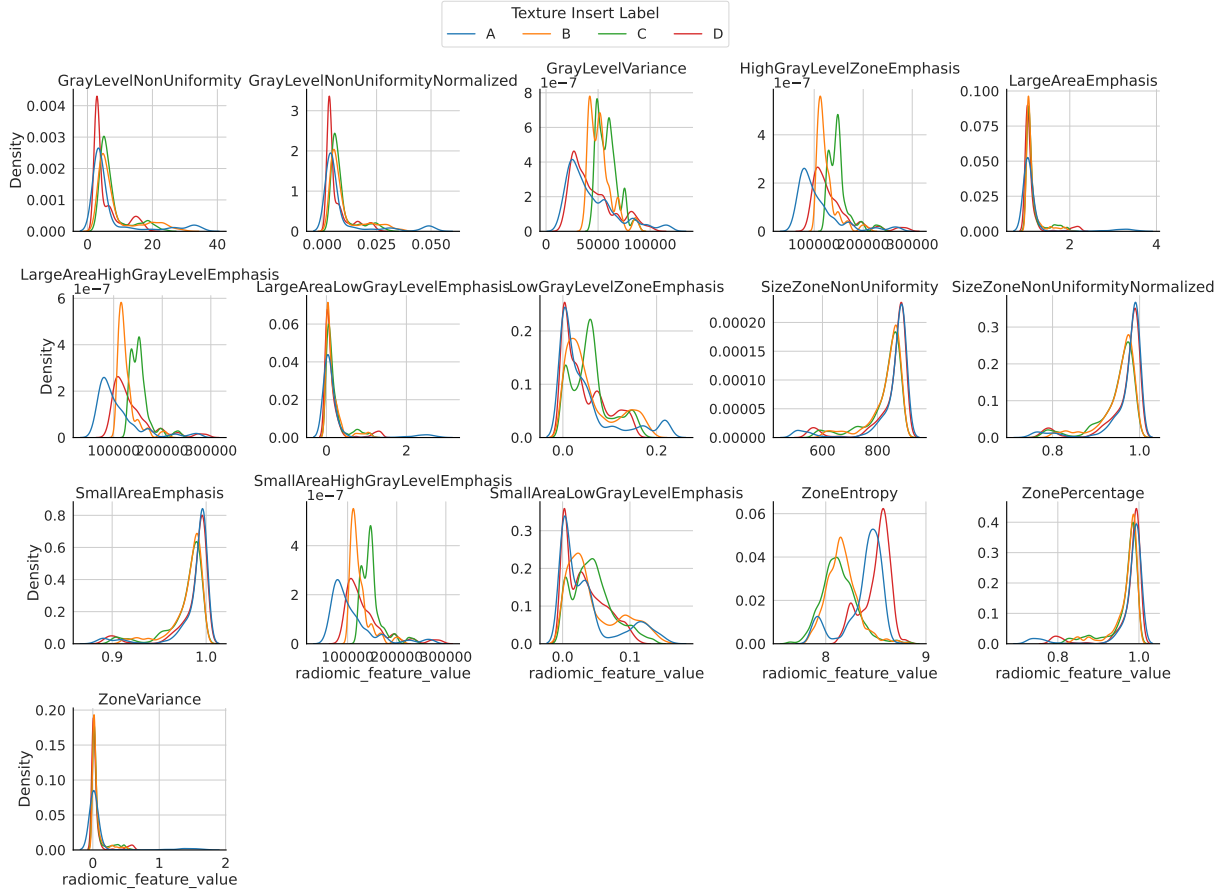


Figure D.4: Distributions for all GLSZM features.

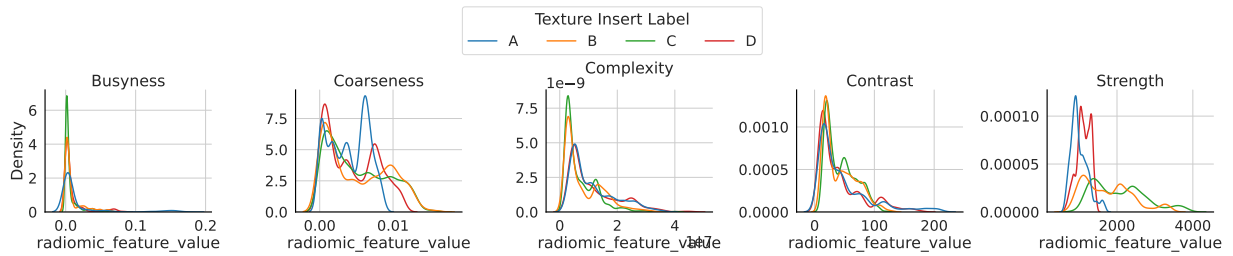


Figure D.5: Distributions for all NGTDM features.

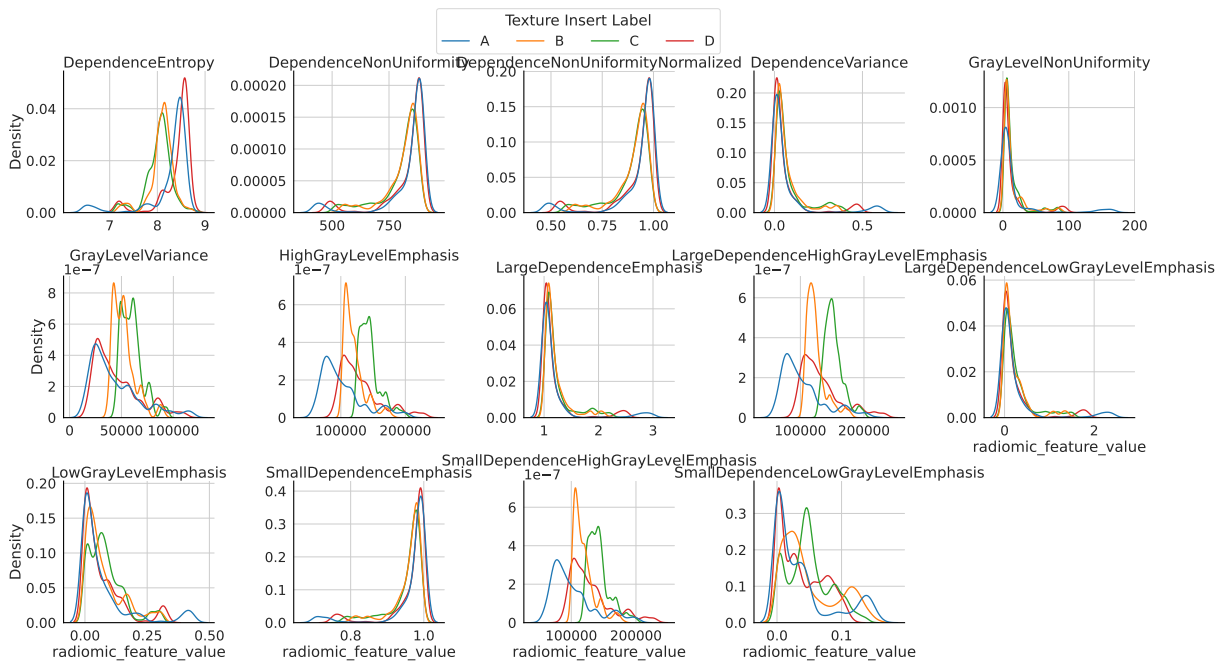


Figure D.6: Distributions for all GLDM features.

E Additional results for Chapter 4

This appendix provides additional results for Chapter 4.

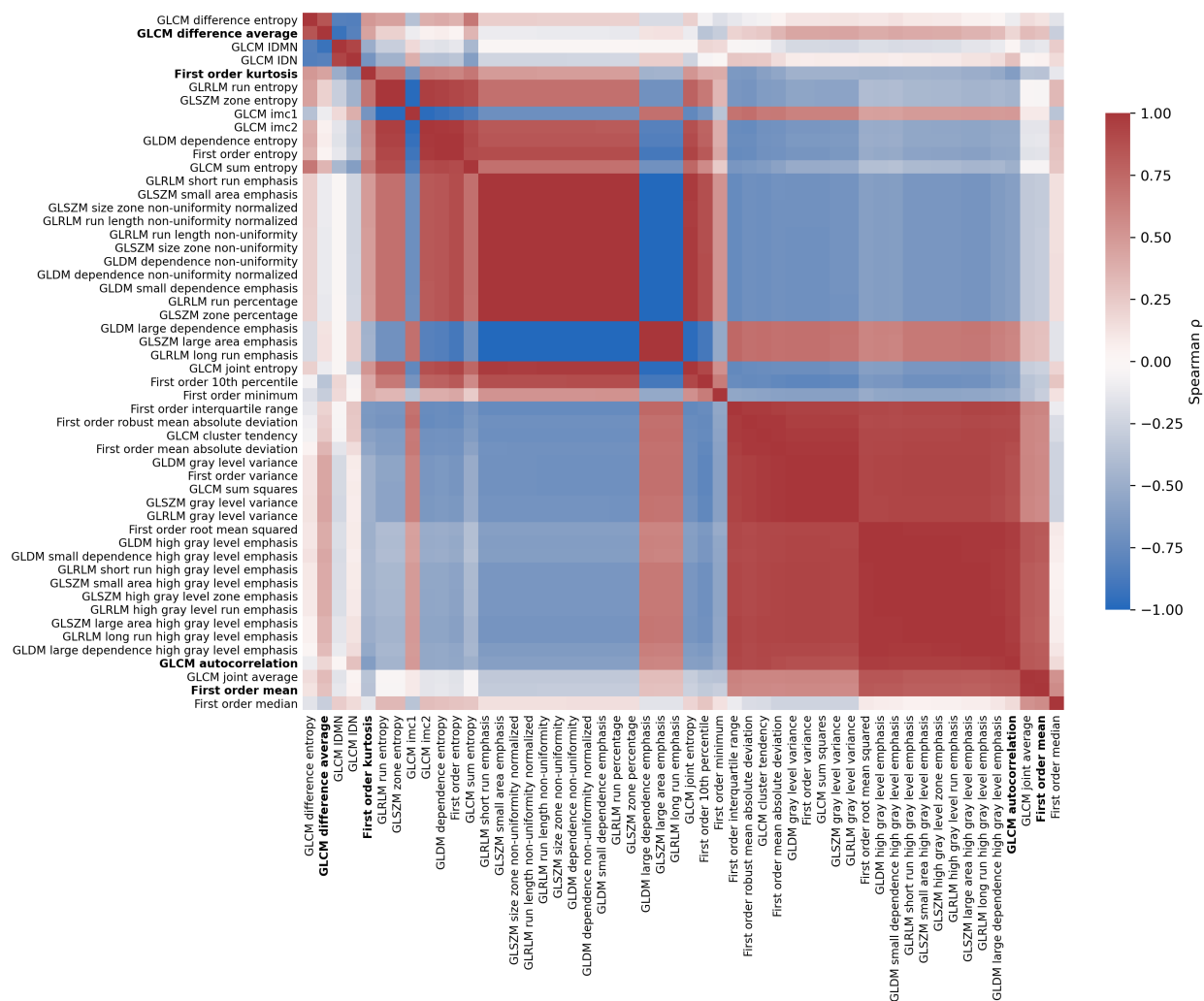


Figure E.1: Heatmap of the radiomic feature value Spearman correlations. The four representative radiomic features are shown bolded.

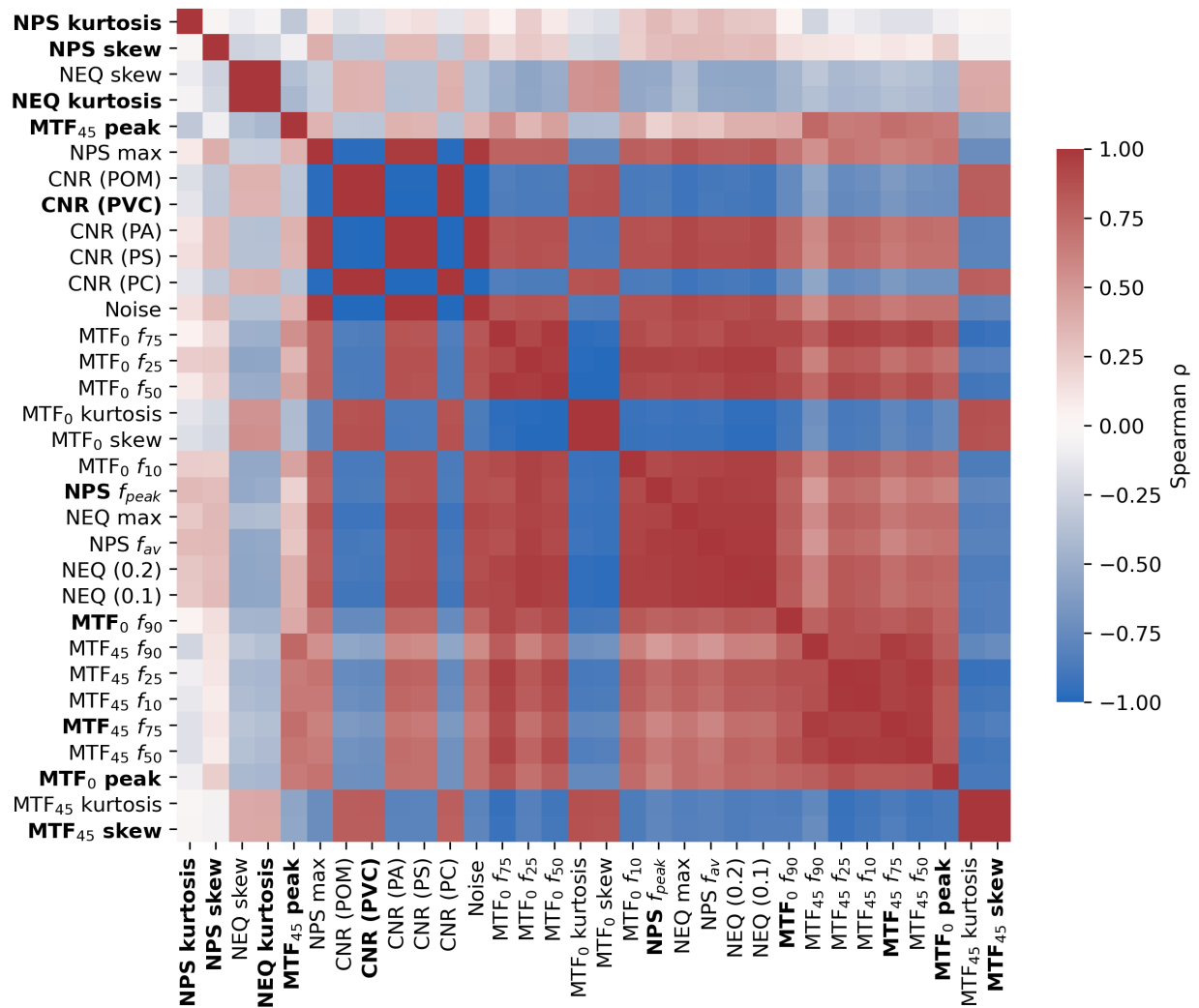


Figure E.2: Heatmap of the image quality metric Spearman correlations. The 10 selected IQMs are shown bolded.

F Additional results for Chapter 6

This appendix provides additional results for Chapter 6.

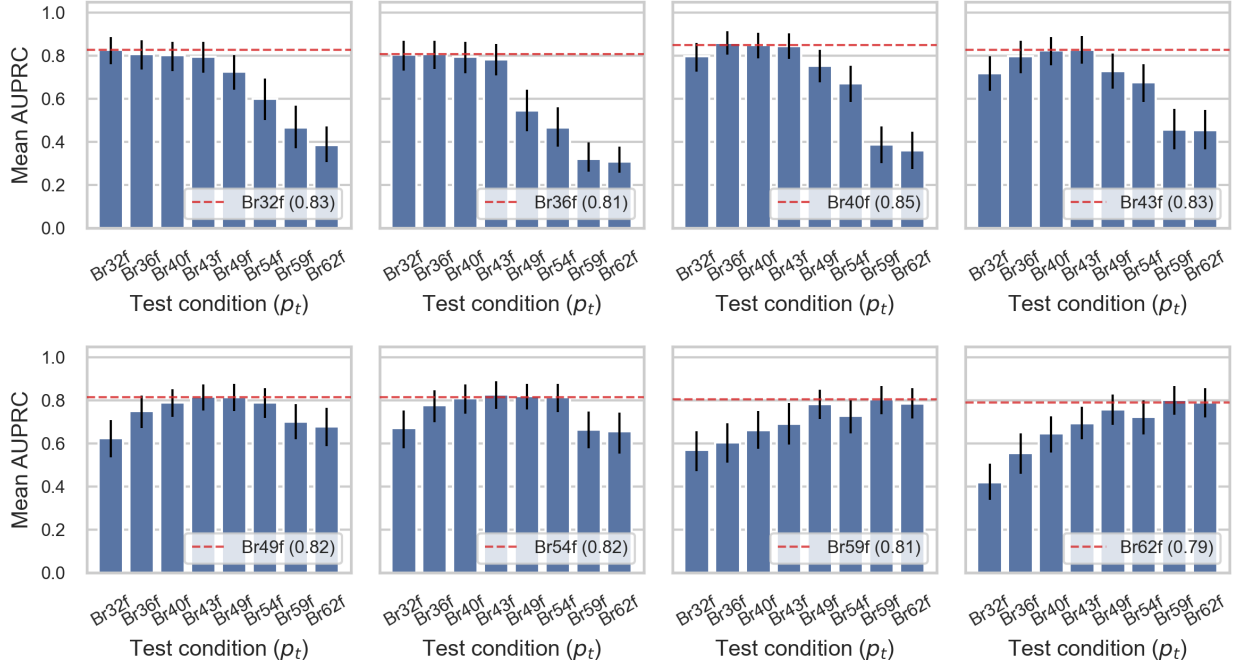


Figure F.1: The CNN classification performance (AUPRC) across all eight target conditions p_t for each of the eight trained CNNs, evaluated on the Definition AS64 dataset. Each subplot corresponds to a CNN trained on a different p_s ; the red dashed line indicates the performance of the CNN on the $p_t = p_s$ test set (corresponding to the diagonal of the heatmap in Figure 6.7). Error bars represent bootstrapped 95% confidence intervals.

Table F.1: Complete report of model fit results by IQM and fit shape, sorted by the quadratic model's train R_{adj}^2 . Train refers to the AS64 data used to fit the model, Force refers to the fit model's performance on the external test set from the Force scanner.

IQM	Quadratic		Exponential		Linear	
	Train R_{adj}^2	Force R_{adj}^2	Train R_{adj}^2	Force R_{adj}^2	Train R_{adj}^2	Force R_{adj}^2
MTF0 f75	0.895	0.299	0.771	0.711	0.718	0.622
MTF0 f50	0.883	0.327	0.752	0.703	0.675	0.612
MTF45 f50	0.848	0.420	0.748	0.728	0.681	0.609
MTF0 f40	0.847	0.508	0.752	0.725	0.699	0.627
MTF45 f25	0.845	0.457	0.744	0.718	0.680	0.618
MTF0 f60	0.836	0.378	0.728	0.674	0.704	0.583
MTF0 f20	0.828	0.486	0.736	0.685	0.674	0.592
MTF45 f60	0.827	0.337	0.729	0.681	0.702	0.588
MTF0 f25	0.820	0.310	0.689	0.644	0.605	0.558
NPS f_avg	0.820	0.468	0.699	0.700	0.622	0.580
MTF45 f10	0.817	0.453	0.721	0.671	0.652	0.570
MTF45 skew	0.814	0.439	0.720	0.696	0.639	0.585
MTF45 f40	0.806	0.463	0.718	0.692	0.649	0.586
NPS nyq/2 obliq	0.798	0.533	0.709	0.571	0.655	0.535
MTF45 f75	0.784	0.374	0.689	0.671	0.613	0.529
MTF45 kurtosis	0.777	0.438	0.689	0.669	0.602	0.554
NPS nyq/2 axial	0.768	0.360	0.680	0.493	0.671	0.497
MTF0 f10	0.766	0.349	0.654	0.594	0.591	0.501
Noise	0.759	0.423	0.649	0.587	0.608	0.516
MTF0 skew	0.733	0.373	0.626	0.611	0.564	0.535
NEQ max	0.715	0.356	0.603	0.499	0.562	0.447
NPS f_peak	0.703	0.333	0.483	0.443	0.500	0.443
NEQ kurtosis	0.698	-0.583	0.501	0.134	0.517	0.134
MTF45 f90	0.687	0.286	0.601	0.500	0.535	0.397
NEQ (0.1)	0.658	0.246	0.551	0.445	0.523	0.397
NEQ skew	0.652	-0.864	0.537	-0.422	0.496	0.092
MTF0 kurtosis	0.651	0.374	0.567	0.550	0.517	0.490
NPS max	0.609	0.157	0.506	0.305	0.489	0.304
MTF0 f90	0.601	-1.776	0.453	0.193	0.470	0.193
MTF0 peak	0.593	0.133	0.492	0.270	0.503	0.334
NEQ (0.2)	0.592	0.174	0.490	0.389	0.461	0.339
CNR (PVC)	0.579	0.547	0.523	0.474	0.472	0.511
CNR (PA)	0.570	0.469	0.518	0.420	0.476	0.499
CNR (PS)	0.560	0.428	0.509	0.529	0.468	0.454
CNR (PC)	0.538	0.470	0.493	0.439	0.457	0.480
MTF45 f20	0.533	0.160	0.364	0.172	0.385	0.172
CNR (POM)	0.532	0.481	0.490	0.441	0.455	0.483
MTF45 peak	0.221	-0.291	0.167	-0.182	0.194	-0.182
NPS skew	0.133	-0.298	0.129	-0.241	0.157	-0.234
NPS kurtosis	0.060	-0.338	0.070	-0.281	0.100	-0.281

References

- [1] J. P. B. O'Connor et al., "Imaging biomarker roadmap for cancer studies," *Nature Reviews Clinical Oncology*, vol. 14, no. 3, pp. 169–186, Mar. 2017. DOI: [10.1038/nrclinonc.2016.162](https://doi.org/10.1038/nrclinonc.2016.162).
- [2] H. J. W. L. Aerts et al., "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nature Communications*, vol. 5, no. 1, p. 4006, Jun. 2014. DOI: [10.1038/ncomms5006](https://doi.org/10.1038/ncomms5006).
- [3] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: Images Are More than Pictures, They Are Data," *Radiology*, vol. 278, no. 2, pp. 563–577, Feb. 2016. DOI: [10.1148/radiol.2015151169](https://doi.org/10.1148/radiol.2015151169).
- [4] X. Fave et al., "Delta-radiomics features for the prediction of patient outcomes in non-small cell lung cancer," *Scientific Reports*, vol. 7, no. 1, p. 588, Apr. 2017. DOI: [10.1038/s41598-017-00665-z](https://doi.org/10.1038/s41598-017-00665-z).
- [5] T. Davenport and R. Kalakota, "The potential for artificial intelligence in healthcare," *Future Healthcare Journal*, vol. 6, no. 2, pp. 94–98, Jun. 2019. DOI: [10.7861/futurehosp.6-2-94](https://doi.org/10.7861/futurehosp.6-2-94).
- [6] M. McNitt-Gray, "Quantitative Imaging in Computed Tomography," in *Quantitative Imaging in Medicine*, R. J. Nordstrom, Ed., AIP Publishing LLC Melville, New York, Nov. 2021, pp. 3–1–3–16. DOI: [10.1063/9780735423558_003](https://doi.org/10.1063/9780735423558_003).
- [7] R. Berenguer, M. d. R. Pastor-Juan, J. Canales-Vázquez, M. Castro-García, M. V. Villas, F. Mansilla Legorburo, and S. Sabater, "Radiomics of CT Features May Be Nonreproducible and Redundant: Influence of CT Acquisition Parameters," *Radiology*, vol. 288, no. 2, pp. 407–415, Aug. 2018. DOI: [10.1148/radiol.2018172361](https://doi.org/10.1148/radiol.2018172361).
- [8] N. Emaminejad, M. W. Wahi-Anwar, G. H. J. Kim, W. Hsu, M. Brown, and M. McNitt-Gray, "Reproducibility of lung nodule radiomic features: Multivariable and univariable investigations that account for interactions between CT acquisition and reconstruction

- parameters," *Medical Physics*, vol. 48, no. 6, pp. 2906–2919, 2021. DOI: [10.1002/mp.14830](https://doi.org/10.1002/mp.14830).
- [9] M. Shafiq-ul-Hassan et al., "Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels," *Medical Physics*, vol. 44, no. 3, pp. 1050–1062, Mar. 2017. DOI: [10.1002/mp.12123](https://doi.org/10.1002/mp.12123).
 - [10] B. Chen, H. Barnhart, S. Richard, J. Colsher, M. Amurao, and E. Samei, "Quantitative CT: Technique dependence of volume estimation on pulmonary nodules," *Physics in Medicine & Biology*, vol. 57, no. 5, p. 1335, Feb. 2012. DOI: [10.1088/0031-9155/57/5/1335](https://doi.org/10.1088/0031-9155/57/5/1335).
 - [11] P. Lo, S. Young, H. J. Kim, M. S. Brown, and M. F. McNitt-Gray, "Variability in CT lung-nodule quantification: Effects of dose reduction and reconstruction methods on density and texture based features," *Medical Physics*, vol. 43, no. 8, pp. 4854–4865, Aug. 2016. DOI: [10.1118/1.4954845](https://doi.org/10.1118/1.4954845).
 - [12] B. Zhao, Y. Tan, W.-Y. Tsai, J. Qi, C. Xie, L. Lu, and L. H. Schwartz, "Reproducibility of radiomics for deciphering tumor phenotype with imaging," *Scientific Reports*, vol. 6, 2016. DOI: [10.1038/srep23428](https://doi.org/10.1038/srep23428).
 - [13] D. Mackin et al., "Measuring Computed Tomography Scanner Variability of Radiomics Features," *Investigative Radiology*, vol. 50, no. 11, p. 757, Nov. 2015. DOI: [10.1097/RLI.000000000000180](https://doi.org/10.1097/RLI.000000000000180).
 - [14] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," *PLOS Medicine*, vol. 15, no. 11, e1002683, Nov. 2018. DOI: [10.1371/journal.pmed.1002683](https://doi.org/10.1371/journal.pmed.1002683).
 - [15] X. Wang, G. Liang, Y. Zhang, H. Blanton, Z. Bessinger, and N. Jacobs, "Inconsistent Performance of Deep Learning Models on Mammogram Classification," *Journal of the American College of Radiology*, vol. 17, no. 6, pp. 796–803, Jun. 2020. DOI: [10.1016/j.jacr.2020.01.006](https://doi.org/10.1016/j.jacr.2020.01.006).

- [16] A. C. Yu, B. Mohajer, and J. Eng, "External Validation of Deep Learning Algorithms for Radiologic Diagnosis: A Systematic Review," *Radiology: Artificial Intelligence*, vol. 4, no. 3, e210064, May 2022. DOI: [10.1148/ryai.210064](https://doi.org/10.1148/ryai.210064).
- [17] J. Quiñonero-Candela, Ed., *Dataset shift in machine learning* (Neural information processing series). Cambridge, Mass: MIT Press, 2009.
- [18] The National Lung Screening Trial Research Team, "The National Lung Screening Trial: Overview and Study Design," *Radiology*, vol. 258, no. 1, pp. 243–253, Jan. 2011. DOI: [10.1148/radiol.10091808](https://doi.org/10.1148/radiol.10091808).
- [19] J. P. Sieren et al., "SPIROMICS protocol for multicenter quantitative computed tomography to phenotype the lungs," *American Journal of Respiratory and Critical Care Medicine*, vol. 194, no. 7, pp. 794–806, 2016. DOI: [10.1164/rccm.201506-1208PP](https://doi.org/10.1164/rccm.201506-1208PP).
- [20] C. H. Cagnon, D. D. Cody, M. F. McNitt-Gray, J. A. Seibert, P. F. Judy, and D. R. Aberle, "Description and Implementation of a Quality Control Program in an Imaging-Based Clinical Trial," *Academic Radiology*, vol. 13, no. 11, pp. 1431–1441, Nov. 2006. DOI: [10.1016/j.acra.2006.08.015](https://doi.org/10.1016/j.acra.2006.08.015).
- [21] A. M. Hernandez et al., "Quality Assurance of the MIDRC Chest CT Database Via Surrogate Multi-Institutional Assessment of 3D Image Quality and Dose," *AAPM*, Jul. 2023.
- [22] A. M. Hernandez et al., "Assessing variations in 3D image quality in chest CT across sites and scanners," *Medical Physics*, vol. 53, no. 4, e70402, 2026. DOI: [10.1002/mp.70402](https://doi.org/10.1002/mp.70402).
- [23] L. G. Kessler et al., "The emerging science of quantitative imaging biomarkers terminology and definitions for scientific studies and regulatory submissions," *Statistical Methods in Medical Research*, vol. 24, no. 1, pp. 9–26, Feb. 2015. DOI: [10.1177/0962280214537333](https://doi.org/10.1177/0962280214537333).
- [24] J. T. Bushberg and J. M. Boone, *The Essential Physics of Medical Imaging*. Lippincott Williams & Wilkins, Dec. 2011.

- [25] E. L. Nickoloff and R. Riley, "A simplified approach for modulation transfer function determinations in computed tomography," *Medical Physics*, vol. 12, no. 4, pp. 437–442, 1985. DOI: [10.1118/1.595706](https://doi.org/10.1118/1.595706).
- [26] S. N. Friedman, G. S. K. Fung, J. H. Siewerdsen, and B. M. W. Tsui, "A simple approach to measure computed tomography (CT) modulation transfer function (MTF) and noise-power spectrum (NPS) using the American College of Radiology (ACR) accreditation phantom," *Medical Physics*, vol. 40, no. 5, p. 051 907, 2013. DOI: [10.1118/1.4800795](https://doi.org/10.1118/1.4800795).
- [27] J. Solomon, R. Saunders, and E. Samei, "Image Quality and Its Clinical Relevance," in *The Handbook of Medical Image Perception and Techniques*, E. Samei and E. A. Krupinski, Eds., 2nd ed., Cambridge: Cambridge University Press, 2018, pp. 197–212. DOI: [10.1017/9781108163781.013](https://doi.org/10.1017/9781108163781.013).
- [28] J. H. Siewerdsen, I. A. Cunningham, and D. A. Jaffray, "A framework for noise-power spectrum analysis of multidimensional images," *Medical Physics*, vol. 29, no. 11, pp. 2655–2671, Nov. 2002. DOI: [10.1118/1.1513158](https://doi.org/10.1118/1.1513158).
- [29] L. Yu, T. J. Vrieze, S. Leng, J. G. Fletcher, and C. H. McCollough, "Technical Note: Measuring contrast- and noise-dependent spatial resolution of an iterative reconstruction method in CT using ensemble averaging," *Medical Physics*, vol. 42, no. 5, pp. 2261–2267, 2015. DOI: [10.1118/1.4916802](https://doi.org/10.1118/1.4916802).
- [30] E. Samei et al., "Performance evaluation of computed tomography systems: Summary of AAPM Task Group 233," *Medical Physics*, vol. 46, no. 11, e735–e756, 2019. DOI: [10.1002/mp.13763](https://doi.org/10.1002/mp.13763).
- [31] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine Learning*, vol. 79, no. 1, pp. 151–175, May 2010. DOI: [10.1007/s10994-009-5152-4](https://doi.org/10.1007/s10994-009-5152-4).
- [32] G. Blanchard, G. Lee, and C. Scott, "Generalizing from Several Related Classification Tasks to a New Unlabeled Sample," in *Advances in Neural Information Processing Systems*, vol. 24, Curran Associates, Inc., 2011.

- [33] H. Ye, C. Xie, T. Cai, R. Li, Z. Li, and L. Wang, "Towards a Theoretical Framework of Out-of-Distribution Generalization," *Advances in Neural Information Processing Systems*, vol. 34, 2021.
- [34] K. Muandet, D. Balduzzi, and B. Scholkopf, "Domain Generalization via Invariant Feature Representation," *Proceedings of the 30th International Conference on Machine Learning*, vol. 28, 2013.
- [35] H. Bai, R. Sun, L. Hong, F. Zhou, N. Ye, H.-J. Ye, S.-H. G. Chan, and Z. Li, "DecAug: Out-of-Distribution Generalization via Decomposed Feature Representation and Semantic Augmentation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 8, pp. 6705–6713, May 2021. DOI: [10.1609/aaai.v35i8.16829](https://doi.org/10.1609/aaai.v35i8.16829).
- [36] J. Futoma, M. Simons, T. Panch, F. Doshi-Velez, and L. A. Celi, "The myth of generalisability in clinical research and machine learning in health care," *The Lancet Digital Health*, vol. 2, no. 9, e489–e492, Sep. 2020. DOI: [10.1016/S2589-7500\(20\)30186-2](https://doi.org/10.1016/S2589-7500(20)30186-2).
- [37] J.-P. Fortin et al., "Harmonization of multi-site diffusion tensor imaging data," *NeuroImage*, vol. 161, pp. 149–170, Nov. 2017. DOI: [10.1016/j.neuroimage.2017.08.047](https://doi.org/10.1016/j.neuroimage.2017.08.047).
- [38] F. Orlhac, F. Frouin, C. Nioche, N. Ayache, and I. Buvat, "Validation of A Method to Compensate Multicenter Effects Affecting CT Radiomics," *Radiology*, vol. 291, no. 1, pp. 53–59, Apr. 2019. DOI: [10.1148/radiol.2019182023](https://doi.org/10.1148/radiol.2019182023).
- [39] A. Ibrahim et al., "The Effects of In-Plane Spatial Resolution on CT-Based Radiomic Features' Stability with and without ComBat Harmonization," *Cancers*, vol. 13, no. 8, p. 1848, Jan. 2021. DOI: [10.3390/cancers13081848](https://doi.org/10.3390/cancers13081848).
- [40] L. Wei, Y. Lin, and W. Hsu, "Using a Generative Adversarial Network for CT Normalization and Its Impact on Radiomic Features," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, Apr. 2020, pp. 844–848. DOI: [10.1109/ISBI45749.2020.9098724](https://doi.org/10.1109/ISBI45749.2020.9098724).

- [41] S. Yang, E. Y. Kim, and J. C. Ye, "Continuous Conversion of CT Kernel Using Switchable CycleGAN With AdaIN," *IEEE Transactions on Medical Imaging*, vol. 40, no. 11, pp. 3015–3029, Nov. 2021. DOI: [10.1109/TMI.2021.3077615](https://doi.org/10.1109/TMI.2021.3077615).
- [42] M. Daly et al., "Harmonizing quantitative imaging feature values in CT using image quality metrics as a basis," in *Medical Imaging 2024: Computer-Aided Diagnosis*, S. M. Astley and W. Chen, Eds., San Diego, United States: SPIE, Apr. 2024, p. 44. DOI: [10.1117/12.3006885](https://doi.org/10.1117/12.3006885).
- [43] M. A. Daly, B. Varghese, A. M. Hernandez, A. Uneri, J. Levy, D. J. Goodenough, J. M. Hoffman, and M. F. McNitt-Gray, "Reduction of Inter-Scanner Quantitative Imaging Feature Variability in CT through Image Quality Based Protocol Harmonization," in *AAPM 66th Annual Meeting & Exhibition*, AAPM, 2024.
- [44] G. J. Gang, R. Deshpande, and J. W. Stayman, "Standardization of histogram- and gray-level co-occurrence matrices-based radiomics in the presence of blur and noise," *Physics in Medicine and Biology*, vol. 66, no. 7, p. 074 004, Apr. 2021. DOI: [10.1088/1361-6560/abeea5](https://doi.org/10.1088/1361-6560/abeea5).
- [45] H. Din, Y. Yuan, G. H. Kim, M. McNitt-Gray, J. W. Stayman, and G. J. Gang, "Using a physics-based approach to standardize radiomics values: Experimental validation in an anthropomorphic phantom on a clinical CT scanner using a range of dose levels and reconstruction kernels," in *Medical Imaging 2026: Physics of Medical Imaging*, vol. 13924, SPIE, Apr. 2026, p. 32. DOI: [10.1117/12.3087679](https://doi.org/10.1117/12.3087679).
- [46] M. Zarei, E. Abadi, L. Vancoillie, and E. Samei, "Protocol selection formalism for minimizing detectable differences in morphological radiomics features of lung lesions in repeated CT acquisitions," *Journal of Medical Imaging*, vol. 11, no. 2, p. 025 501, Apr. 2024. DOI: [10.1117/1.JMI.11.2.025501](https://doi.org/10.1117/1.JMI.11.2.025501).
- [47] D. V. Fried, S. L. Tucker, S. Zhou, Z. Liao, O. Mawlawi, G. Ibbott, and L. E. Court, "Prognostic Value and Reproducibility of Pretreatment CT Texture Features in Stage III Non-Small Cell Lung Cancer," *International Journal of Radiation Oncology*Biophysics*Physics*, vol. 90, no. 4, pp. 834–842, Nov. 2014. DOI: [10.1016/j.ijrobp.2014.07.020](https://doi.org/10.1016/j.ijrobp.2014.07.020).

- [48] G. Wu et al., "Diagnosis of Invasive Lung Adenocarcinoma Based on Chest CT Radiomic Features of Part-Solid Pulmonary Nodules: A Multicenter Study," *Radiology*, vol. 297, no. 2, pp. 451–458, Nov. 2020. DOI: [10.1148/radiol.2020192431](https://doi.org/10.1148/radiol.2020192431).
- [49] M. Selvam, A. Chandrasekharan, A. Sadanandan, V. K. Anand, A. Murali, and G. Krishnamurthi, "Radiomics as a non-invasive adjunct to Chest CT in distinguishing benign and malignant lung nodules," *Scientific Reports*, vol. 13, no. 1, p. 19062, Nov. 2023. DOI: [10.1038/s41598-023-46391-7](https://doi.org/10.1038/s41598-023-46391-7).
- [50] H. J. Park, S. M. Lee, J. W. Song, S. M. Lee, S. Y. Oh, N. Kim, and J. B. Seo, "Texture-Based Automated Quantitative Assessment of Regional Patterns on Initial CT in Patients With Idiopathic Pulmonary Fibrosis: Relationship to Decline in Forced Vital Capacity," *American Journal of Roentgenology*, vol. 207, no. 5, pp. 976–983, Nov. 2016. DOI: [10.2214/AJR.16.16054](https://doi.org/10.2214/AJR.16.16054).
- [51] J. D. Budzikowski, J. J. Foy, A. A. Rashid, J. H. Chung, I. Noth, and S. G. Armato, "Radiomics-based assessment of idiopathic pulmonary fibrosis is associated with genetic mutations and patient survival," *Journal of Medical Imaging*, vol. 8, no. 3, p. 031903, May 2021. DOI: [10.1117/1.JMI.8.3.031903](https://doi.org/10.1117/1.JMI.8.3.031903).
- [52] V. Nardone et al., "3D bone texture analysis as a potential predictor of radiation-induced insufficiency fractures," *Quantitative Imaging in Medicine and Surgery*, vol. 8, no. 1, pp. 14–24, Feb. 2018. DOI: [10.21037/qims.2018.02.01](https://doi.org/10.21037/qims.2018.02.01).
- [53] Y.-C. Chen, Y.-T. Li, P.-C. Kuo, S.-J. Cheng, Y.-H. Chung, D.-P. Kuo, and C.-Y. Chen, "Automatic segmentation and radiomic texture analysis for osteoporosis screening using chest low-dose computed tomography," *European Radiology*, vol. 33, no. 7, pp. 5097–5106, Jul. 2023. DOI: [10.1007/s00330-023-09421-6](https://doi.org/10.1007/s00330-023-09421-6).
- [54] Y. Lee et al., "Radiomics identifies distinct cortical bone texture alterations in patients with CKD using HR-pQCT," *Bone Research*, vol. 14, no. 1, p. 36, Apr. 2026. DOI: [10.1038/s41413-026-00515-7](https://doi.org/10.1038/s41413-026-00515-7).
- [55] A. Midya, J. Chakraborty, M. Gönen, R. K. G. Do, and A. L. Simpson, "Influence of CT acquisition and reconstruction parameters on radiomic feature reproducibility," *Journal of Medical Imaging*, vol. 5, no. 1, p. 011020, Jan. 2018. DOI: [10.1117/1.JMI.5.1.011020](https://doi.org/10.1117/1.JMI.5.1.011020).

- [56] J. Hoyer, J. Solomon, T. J. Sauer, M. Robins, and E. Samei, "Systematic analysis of bias and variability of morphologic features for lung lesions in computed tomography," *Journal of Medical Imaging*, vol. 6, no. 1, p. 013504, Jan. 2019. DOI: [10.1117/1.JMI.6.1.013504](https://doi.org/10.1117/1.JMI.6.1.013504).
- [57] D. C. Sullivan et al., "Metrology Standards for Quantitative Imaging Biomarkers," *Radiology*, vol. 277, no. 3, pp. 813–825, Dec. 2015. DOI: [10.1148/radiol.2015142202](https://doi.org/10.1148/radiol.2015142202).
- [58] M. McNitt-Gray et al., "Standardization in Quantitative Imaging: A Multicenter Comparison of Radiomic Features from Different Software Packages on Digital Reference Objects and Patient Data Sets," *Tomography*, vol. 6, no. 2, pp. 118–128, Jun. 2020. DOI: [10.18383/j.tom.2019.00031](https://doi.org/10.18383/j.tom.2019.00031).
- [59] A. Zwanenburg et al., "The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping," *Radiology*, vol. 295, no. 2, pp. 328–338, May 2020. DOI: [10.1148/radiol.2020191145](https://doi.org/10.1148/radiol.2020191145).
- [60] C. H. McCollough, M. R. Bruesewitz, M. F. McNitt-Gray, K. Bush, T. Ruckdeschel, J. T. Payne, J. A. Brink, and R. K. Zeman, "The phantom portion of the American College of Radiology (ACR) Computed Tomography (CT) accreditation program: Practical tips, artifact examples, and pitfalls to avoid," *Medical Physics*, vol. 31, no. 9, pp. 2423–2442, Sep. 2004. DOI: [10.1118/1.1769632](https://doi.org/10.1118/1.1769632).
- [61] K. Gulliksrud, C. Stokke, and A. C. Trægde Martinsen, "How to measure CT image quality: Variations in CT-numbers, uniformity and low contrast resolution for a CT quality assurance phantom," *Physica Medica*, vol. 30, no. 4, pp. 521–526, Jun. 2014. DOI: [10.1016/j.ejmp.2014.01.006](https://doi.org/10.1016/j.ejmp.2014.01.006).
- [62] P. Kalendralis et al., "Multicenter CT phantoms public dataset for radiomics reproducibility tests," *Medical Physics*, vol. 46, no. 3, pp. 1512–1518, Mar. 2019. DOI: [10.1002/mp.13385](https://doi.org/10.1002/mp.13385).
- [63] K. Buch, B. Li, M. M. Qureshi, H. Kuno, S. W. Anderson, and O. Sakai, "Quantitative Assessment of Variation in CT Parameters on Texture Features: Pilot Study Using a Nonanatomic Phantom," *American Journal of Neuroradiology*, vol. 38, no. 5, pp. 981–985, May 2017. DOI: [10.3174/ajnr.A5139](https://doi.org/10.3174/ajnr.A5139).

- [64] B. A. Varghese et al., "Technical and clinical considerations of a physical liver phantom for CT radiomics analysis," *Journal of Applied Clinical Medical Physics*, vol. 25, no. 4, e14309, Apr. 2024. DOI: [10.1002/acm2.14309](https://doi.org/10.1002/acm2.14309).
- [65] Y. Li, M. Reyhan, Y. Zhang, X. Wang, J. Zhou, Y. Zhang, N. J. Yue, and K. Nie, "The impact of phantom design and material-dependence on repeatability and reproducibility of CT-based radiomics features," *Medical Physics*, vol. 49, no. 3, pp. 1648–1659, 2022. DOI: [10.1002/mp.15491](https://doi.org/10.1002/mp.15491).
- [66] B. A. Varghese, D. Hwang, S. Y. Cen, X. Lei, J. Levy, B. Desai, D. J. Goodenough, and V. A. Duddalwar, "Identification of robust and reproducible CT-texture metrics using a customized 3D-printed texture phantom," *Journal of Applied Clinical Medical Physics*, vol. 22, no. 2, pp. 98–107, Feb. 2021. DOI: [10.1002/acm2.13162](https://doi.org/10.1002/acm2.13162).
- [67] J. Solomon, A. Ba, F. Bochud, and E. Samei, "Comparison of low-contrast detectability between two CT reconstruction algorithms using voxel-based 3D printed textured phantoms," *Medical Physics*, vol. 43, no. 12, pp. 6497–6506, Dec. 2016. DOI: [10.1118/1.4967478](https://doi.org/10.1118/1.4967478).
- [68] E. Ehler, D. Craft, and Y. Rong, "3D printing technology will eventually eliminate the need of purchasing commercial phantoms for clinical medical physics QA procedures," *Journal of Applied Clinical Medical Physics*, vol. 19, no. 4, pp. 8–12, 2018. DOI: [10.1002/acm2.12392](https://doi.org/10.1002/acm2.12392).
- [69] E. Samei, J. Hoyer, Y. Zheng, J. B. Solomon, and D. Marin, "Design and fabrication of heterogeneous lung nodule phantoms for assessing the accuracy and variability of measured texture radiomics features in CT," *Journal of Medical Imaging*, vol. 6, no. 2, p. 021606, Apr. 2019. DOI: [10.1117/1.JMI.6.2.021606](https://doi.org/10.1117/1.JMI.6.2.021606).
- [70] K. Mei, M. Geagan, L. Roshkovan, H. I. Litt, G. J. Gang, N. Shapira, J. W. Stayman, and P. B. Noël, "Three-dimensional printing of patient-specific lung phantoms for CT imaging: Emulating lung tissue with accurate attenuation profiles and textures," *Medical Physics*, vol. 49, no. 2, pp. 825–835, Feb. 2022. DOI: [10.1002/mp.15407](https://doi.org/10.1002/mp.15407).
- [71] D. F. Craft, S. F. Kry, P. Balter, M. Salehpour, W. Woodward, and R. M. Howell, "Material matters: Analysis of density uncertainty in 3D printing and its consequences for radiation

- oncology," *Medical Physics*, vol. 45, no. 4, pp. 1614–1621, 2018. DOI: [10.1002/mp.12839](https://doi.org/10.1002/mp.12839).
- [72] J. Solc, T. Vrba, and L. Burianova, "Tissue-equivalence of 3D-printed plastics for medical phantoms in radiology," *Journal of Instrumentation*, vol. 13, no. 09, P09018–P09018, Sep. 2018. DOI: [10.1088/1748-0221/13/09/P09018](https://doi.org/10.1088/1748-0221/13/09/P09018).
- [73] S. R. Badiuk, D. K. Sasaki, and D. W. Rickey, "An anthropomorphic maxillofacial phantom using 3-dimensional printing, polyurethane rubber and epoxy resin for dental imaging and dosimetry," *Dentomaxillofacial Radiology*, vol. 51, no. 1, p. 20200323, Jan. 2022. DOI: [10.1259/dmfr.20200323](https://doi.org/10.1259/dmfr.20200323).
- [74] C. Dillon et al., "Computed Tomography Quality Control Manual," American College of Radiology, Tech. Rep., 2017.
- [75] D. J. Schuirmann, "A comparison of the Two One-Sided Tests Procedure and the Power Approach for assessing the equivalence of average bioavailability," *Journal of Pharmacokinetics and Biopharmaceutics*, vol. 15, no. 6, pp. 657–680, Dec. 1987. DOI: [10.1007/BF01068419](https://doi.org/10.1007/BF01068419).
- [76] J. J. Van Griethuysen et al., "Computational Radiomics System to Decode the Radiographic Phenotype," *Cancer Research*, vol. 77, no. 21, e104–e107, Nov. 2017. DOI: [10.1158/0008-5472.CAN-17-0339](https://doi.org/10.1158/0008-5472.CAN-17-0339).
- [77] T. K. Koo and M. Y. Li, "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research," *Journal of Chiropractic Medicine*, vol. 15, no. 2, pp. 155–163, Jun. 2016. DOI: [10.1016/j.jcm.2016.02.012](https://doi.org/10.1016/j.jcm.2016.02.012).
- [78] J. J. Bartko, "The Intraclass Correlation Coefficient as a Measure of Reliability," *Psychological Reports*, vol. 19, no. 1, pp. 3–11, Aug. 1966. DOI: [10.2466/pr0.1966.19.1.3](https://doi.org/10.2466/pr0.1966.19.1.3).
- [79] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011.

- [80] L. I. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, vol. 45, no. 1, pp. 255–268, Mar. 1989.
- [81] G. B. McBride, "A proposal for strength-of-agreement criteria for Lin's concordance correlation coefficient," *NIWA client report: HAM2005-062*, vol. 45, pp. 307–310, 2005.
- [82] P. Lambin et al., "Radiomics: The bridge between medical imaging and personalized medicine," *Nature Reviews Clinical Oncology*, vol. 14, no. 12, pp. 749–762, Dec. 2017. DOI: [10.1038/nrclinonc.2017.141](https://doi.org/10.1038/nrclinonc.2017.141).
- [83] A. Traverso, L. Wee, A. Dekker, and R. Gillies, "Repeatability and Reproducibility of Radiomic Features: A Systematic Review," *International Journal of Radiation Oncology*Biophysics*, vol. 102, no. 4, pp. 1143–1158, Nov. 2018. DOI: [10.1016/j.ijrobp.2018.05.053](https://doi.org/10.1016/j.ijrobp.2018.05.053).
- [84] M. Ravanelli, D. Farina, M. Morassi, E. Roca, G. Cavalleri, G. Tassi, and R. Maroldi, "Texture analysis of advanced non-small cell lung cancer (NSCLC) on contrast-enhanced computed tomography: Prediction of the response to the first-line chemotherapy," *European Radiology*, vol. 23, no. 12, pp. 3450–3455, Dec. 2013. DOI: [10.1007/s00330-013-2965-0](https://doi.org/10.1007/s00330-013-2965-0).
- [85] J. Solomon, A. Mileto, R. C. Nelson, K. Roy Choudhury, and E. Samei, "Quantitative Features of Liver Lesions, Lung Nodules, and Renal Stones at Multi-Detector Row CT Examinations: Dependency on Radiation Dose and Reconstruction Algorithm," *Radiology*, vol. 279, no. 1, pp. 185–194, Apr. 2016. DOI: [10.1148/radiol.2015150892](https://doi.org/10.1148/radiol.2015150892).
- [86] A. M. Hernandez, P. Wu, M. Mahesh, J. H. Siewerdsen, and J. M. Boone, "Location and direction dependence in the 3D MTF for a high-resolution CT system," *Medical Physics*, vol. 48, no. 6, pp. 2760–2771, 2021. DOI: [10.1002/mp.14789](https://doi.org/10.1002/mp.14789).
- [87] J. H. Siewerdsen, A. Uneri, A. M. Hernandez, G. W. Burkett, and J. M. Boone, "Cone-beam CT dose and imaging performance evaluation with a modular, multipurpose phantom," *Medical Physics*, vol. 47, no. 2, pp. 467–479, Feb. 2020. DOI: [10.1002/mp.13952](https://doi.org/10.1002/mp.13952).
- [88] Phantom Laboratory, *Corgi® Phantom Product Guide*, 2025.

- [89] J. P. Cruz-Bastida, D. Gomez-Cardona, K. Li, H. Sun, J. Hsieh, T. P. Szczekutowicz, and G.-H. Chen, "Hi-Res scan mode in clinical MDCT systems: Experimental assessment of spatial resolution performance," *Medical Physics*, vol. 43, no. 5, p. 2399, May 2016. DOI: [10.1118/1.4946816](https://doi.org/10.1118/1.4946816).
- [90] M. A. Daly, M. F. McNitt-Gray, and J. M. Hoffman, "An open-source CT texture phantom for radiomic imaging studies: Design, manufacture, and initial evaluation," *Medical Physics*, vol. 53, no. 8, e70613, 2026. DOI: [10.1002/mp.70613](https://doi.org/10.1002/mp.70613).
- [91] J. Hoffman and M. Daly, *HD phantom body and radiomics texture inserts for CT imaging*, Feb. 2026. DOI: [10.5281/zenodo.15733372](https://doi.org/10.5281/zenodo.15733372).
- [92] K. M. Kanal, P. F. Butler, D. Sengupta, M. Bhargavan-Chatfield, L. P. Coombs, and R. L. Morin, "U.S. Diagnostic Reference Levels and Achievable Doses for 10 Adult CT Examinations," *Radiology*, vol. 284, no. 1, pp. 120–133, Jul. 2017. DOI: [10.1148/radiol.2017161911](https://doi.org/10.1148/radiol.2017161911).
- [93] AAPM, "Lung Cancer Screening CT Protocols Version 5.1," American Association of Physicists in Medicine, Tech. Rep., Sep. 2019.
- [94] F. R. Verdun et al., "Image quality in CT: From physical measurements to model observers," *Physica Medica*, vol. 31, no. 8, pp. 823–843, Dec. 2015. DOI: [10.1016/j.ejmp.2015.08.007](https://doi.org/10.1016/j.ejmp.2015.08.007).
- [95] M. A. Daly, J. M. Hoffman, A. M. Hernandez, A. Uneri, B. A. Varghese, J. Levy, and M. F. McNitt-Gray, "CT texture phantom dataset with paired image quality assessments for quantitative imaging," *In press at Medical Physics*, 2026. DOI: [10.1002/mp.70633](https://doi.org/10.1002/mp.70633).
- [96] M. Daly, J. Hoffman, A. Hernandez, A. Uneri, B. Varghese, J. Levy, and M. McNitt-Gray, *PTP-IQ: Paired texture phantom & image quality data for CT*, Feb. 2026. DOI: [10.5281/zenodo.16820899](https://doi.org/10.5281/zenodo.16820899).
- [97] M. D. Wilkinson et al., "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific Data*, vol. 3, no. 1, p. 160018, Mar. 2016. DOI: [10.1038/sdata.2016.18](https://doi.org/10.1038/sdata.2016.18).

- [98] D. Mackin, X. Fave, L. Zhang, J. Yang, A. K. Jones, C. S. Ng, and L. Court, "Harmonizing the pixel size in retrospective computed tomography radiomics studies," *PLOS ONE*, vol. 12, no. 9, e0178524, Sep. 2017. DOI: [10.1371/journal.pone.0178524](https://doi.org/10.1371/journal.pone.0178524).
- [99] F. Orlhac, J. J. Eertink, A.-S. Cottureau, J. M. Zijlstra, C. Thieblemont, M. Meignan, R. Boellaard, and I. Buvat, "A Guide to ComBat Harmonization of Imaging Biomarkers in Multicenter Studies," *Journal of Nuclear Medicine*, vol. 63, no. 2, pp. 172–179, Feb. 2022. DOI: [10.2967/jnumed.121.262464](https://doi.org/10.2967/jnumed.121.262464).
- [100] A. Yadav et al., "A comparative analysis of image harmonization techniques in mitigating differences in CT acquisition and reconstruction," *Physics in Medicine and Biology*, vol. 70, no. 5, p. 055 015, Mar. 2025. DOI: [10.1088/1361-6560/adabad](https://doi.org/10.1088/1361-6560/adabad).
- [101] G. F. Reed, F. Lynn, and B. D. Meade, "Use of Coefficient of Variation in Assessing Variability of Quantitative Assays," *Clinical and Diagnostic Laboratory Immunology*, vol. 9, no. 6, pp. 1235–1239, Nov. 2002. DOI: [10.1128/CDLI.9.6.1235-1239.2002](https://doi.org/10.1128/CDLI.9.6.1235-1239.2002).
- [102] C. Spearman, "The Proof and Measurement of Association between Two Things," *The American Journal of Psychology*, vol. 15, no. 1, p. 72, Jan. 1904. DOI: [10.2307/1412159](https://doi.org/10.2307/1412159).
- [103] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005. DOI: [10.1111/j.1467-9868.2005.00503.x](https://doi.org/10.1111/j.1467-9868.2005.00503.x).
- [104] T. J. Sejnowski, *The deep learning revolution*. Cambridge, Massachusetts: The MIT Press, 2018.
- [105] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, vol. 25, Curran Associates, Inc., 2012.
- [106] D. C. Castro, I. Walker, and B. Glocker, "Causality matters in medical imaging," *Nature Communications*, vol. 11, no. 1, p. 3673, Jul. 2020. DOI: [10.1038/s41467-020-17478-w](https://doi.org/10.1038/s41467-020-17478-w).

- [107] B. M. Trotta, A. V. Stolin, M. B. Williams, S. B. Gay, A. S. Brody, and T. A. Altes, "Characterization of the Relation Between CT Technical Parameters and Accuracy of Quantification of Lung Attenuation on Quantitative Chest CT," *American Journal of Roentgenology*, vol. 188, no. 6, pp. 1683–1690, Jun. 2007. DOI: [10.2214/AJR.06.1153](https://doi.org/10.2214/AJR.06.1153).
- [108] M. W. Wahi-Anwar, N. Emaminejad, Y. Choi, H. G. Kim, W. Hsu, M. S. Brown, and M. F. McNitt-Gray, "A novel physics-based data augmentation approach for improved robust deep learning in medical imaging: Lung nodule CAD false positive reduction in low-dose CT environments," in *Medical Imaging 2021: Physics of Medical Imaging*, vol. 11595, SPIE, Feb. 2021, pp. 89–104. DOI: [10.1117/12.2582126](https://doi.org/10.1117/12.2582126).
- [109] O. Kilim, A. Olar, T. Joó, T. Palicz, P. Pollner, and I. Csabai, "Physical imaging parameter variation drives domain shift," *Scientific Reports*, vol. 12, p. 21 302, Dec. 2022. DOI: [10.1038/s41598-022-23990-4](https://doi.org/10.1038/s41598-022-23990-4).
- [110] A. A. A. Setio et al., "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge," *Medical Image Analysis*, vol. 42, pp. 1–13, Dec. 2017. DOI: [10.1016/j.media.2017.06.015](https://doi.org/10.1016/j.media.2017.06.015).
- [111] A. Leiter, R. R. Veluswamy, and J. P. Wisnivesky, "The global burden of lung cancer: Current status and future trends," *Nature Reviews Clinical Oncology*, vol. 20, no. 9, pp. 624–639, Sep. 2023. DOI: [10.1038/s41571-023-00798-3](https://doi.org/10.1038/s41571-023-00798-3).
- [112] S. Young et al., "The effect of radiation dose reduction on computer-aided detection (CAD) performance in a low-dose lung cancer screening population," *Medical Physics*, vol. 44, no. 4, pp. 1337–1346, 2017. DOI: [10.1002/mp.12128](https://doi.org/10.1002/mp.12128).
- [113] A. O. Omigbodun, F. Noo, M. McNitt-Gray, W. Hsu, and S. S. Hsieh, "The effects of physics-based data augmentation on the generalizability of deep neural networks: Demonstration on nodule false-positive reduction," *Medical Physics*, vol. 46, no. 10, pp. 4563–4574, Oct. 2019. DOI: [10.1002/mp.13755](https://doi.org/10.1002/mp.13755).
- [114] The National Lung Screening Trial Research Team, "Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening," *New England Journal of Medicine*, vol. 365, no. 5, pp. 395–409, Aug. 2011. DOI: [10.1056/NEJMoa1102873](https://doi.org/10.1056/NEJMoa1102873).

- [115] J. Hofmanninger, F. Prayer, J. Pan, S. Röhrich, H. Prosch, and G. Langs, "Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem," *European Radiology Experimental*, vol. 4, no. 1, p. 50, Dec. 2020. DOI: [10.1186/s41747-020-00173-2](https://doi.org/10.1186/s41747-020-00173-2).
- [116] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, Jul. 1997. DOI: [10.1016/S0031-3203\(96\)00142-2](https://doi.org/10.1016/S0031-3203(96)00142-2).
- [117] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 770–778. DOI: [10.1109/CVPR.2016.90](https://doi.org/10.1109/CVPR.2016.90).
- [118] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2009, pp. 248–255. DOI: [10.1109/CVPR.2009.5206848](https://doi.org/10.1109/CVPR.2009.5206848).
- [119] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *International Conference on Learning Representations*, 2019.
- [120] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA: Curran Associates Inc., Dec. 2019, pp. 8026–8037.