Investigation of Data Modeling Strategies for Quantification of CT Image Quality

Maitham Naeemi

A thesis submitted in partial fulfillment of the
requirements for the degree of

Master of Science in Electrical Engineering

University of Washington

2017

Committee:

Sohini Roychowdhury, Chair

Pierre Mourad

Kaibao Nie

Program Authorized to Award Degree:
Engineering and Mathematics

Abstract

Computed Tomography (CT) is a widely used medical imaging technology that plays a crucial role in pathology diagnosis and treatment management. The growing use of CT imaging raises the risk of undue radiation exposure, particularly for patients imaged multiple times. Thus, there is a need to maintain radiation exposure *As Low As Reasonably Achievable* (ALARA) while ensuring diagnostic quality images. In this study, we investigate data modeling strategies to quantify CT image quality (IQ), in order to guide protocol selection and attain ALARA. Specifically, we present a novel, Windowed Fourier-domain Distance Metric (WFDM) that is used to select regions-of-interest (ROI) by their degree of spatial variability. By selecting regions of low variation (ROI-LV), an estimate of the noise in that region can be made. CT IQ is defined as the inverse of this additive noise. Against the phantom CT images, the WFDM model is shown to correlate strongly to image noise ($r > 0.76$ ($p \ll 0.001$)). As a CT IQ classifier, this model is comparatively analyzed against a fixed-size ROI (baseline) model and a Convolutional Neural Network (CNN), using phantom and patient CT images. The WFDM model and the CNN are shown to classify the phantom images accurately, with a mean accuracy of $\overline{\alpha}_{WFDM} \leq 100\%$, and $\overline{\alpha}_{CNN} = 93.8\%$, respectively. The baseline model manages a mean accuracy of $\overline{\alpha}_B = 73.6\%$ on the same phantom images. With the patient CT images, the baseline and WFDM accuracies drop to $\overline{\alpha}_B \leq 49.5\%$ and $\overline{\alpha}_{WFDM} \leq 66.1\%$, respectively. The CNN, however, performs at 100% accuracy when tested with images from the same CT stack as the training set, but below 1.9% otherwise. This indicates the CNN focus on structural rather than textural features. Finally, the WFDM model is used to predict high/low trends in 84 pairs of patient CT images. These trends are set against the trends in x-ray flux at the time of acquisition, $CTDI_{vol}$, which, for the same patient, directly correspond to CT IQ. The total percentage of image pairs with inverse trends is defined as the total percent error, which is found to be 30.95% and 21.43% for the baseline and WFDM models, respectively. However, this error drops to 0% for $CTDI_{vol}$ changes of at least 40.0%

for the baseline model and 27.5% for the WFDM model, respectively. Thus, for every patient that has been previously imaged, the WFDM model can be used to predict optimal parameters for adequate CT image acquisition. Future work will investigate the impact of WFDM parameters, such as the window sizes and transformation technique, on CT image quality assessment. In addition, the WFDM model can also be applied to pre-process CT images followed by CNN data models for CT image texture identification.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to thank Dr. Adam Alessio, from the Department of Radiology at the University of Washington's Seattle campus, for providing the CT images used in this study, and for his insights and guidance on analyzing them effectively. I would also like to thank Johnny Ren, a fellow Master's student at the University of Washington, Bothell, who performed the CNN analysis on the CT images.

DEDICATION

to my dear wife, Brittany, and my sweet Inara.

Chapter 1

INTRODUCTION

1.1 Background and Motivation

Computed Tomography (CT) is a widely used imaging technology that allows volumetric visualization of the internal micro-structure of a scanned object [1]. Various medical applications can be found in thoracic, cardiac, angiographic and colon imaging where CT images play a crucial role in pathology diagnosis and treatment management; thus it has become a vital diagnostic tool in modern health care with more than 70 million scans performed annually in 2007-2009 [2]. However, the growing number of CT image scans performed over the past decade has raised concerns regarding the adverse effects of patient exposure to radiation [3]. Studies have shown that the extent of exposure to radiation for patients has doubled from 1996 to 2010 [4].

CT image acquisition can operate at various combinations of machine settings, such as the x-ray tube voltage (kV_p) and current (mA), detector collimation, helical pitch, reconstruction filter, slice thickness, and so on. Associated with each scan setting, the resulting CT volumetric data will have different resolution and noise properties, which can subsequently impact the accuracy of pathology detection [5]. Contrast enhancement also plays a significant role in CT image quality and radiation exposure. The presence of contrast enhancement agents can improve contrast between different tissues, which allows the segmentation of a CT image into anatomical regions and the detection of pathologies. However, higher contrast enhanced dosages result in greater absorption of x-ray radiation, which can be harmful to the surrounding tissue. This encourages the implementation of optimized CT imaging protocols with sufficient diagnostic *image quality* (IQ) at the lowest achievable dosage levels to minimize radiation exposure. An image is the addition of signal and additive noise, hence

IQ is inversely proportional to additive noise. The goal is to maintain contrast enhanced dosages *As Low As Reasonably Achievable (ALARA)* while also ensuring that IQ is sufficient for accurate diagnoses [6]. This thesis presents and analyzes a novel CT image quality metric that can be used to give automated feedback regarding CT IQ in an effort to attain ALARA.

Before discussing the specifics of this system, it is necessary to define *image quality* as it pertains to this study. The quality of an image depends on the information desired from the image. Thus, a high quality image is one that provides the desired information with the smallest chance of error. In the case of CT images, decreasing x-ray exposure increases the additive noise in an image, thereby increasing the chance for error when analyzing the image. Furthermore, the addition of contrast enhancement agents increases the absorption of x-ray radiation. Thus, additive noise is inversely related to image quality, contrast enhancement dosage and x-ray exposure, a factor of interest in attaining ALARA. In this thesis, *image quality* is assessed in three ways:

1. Measuring scanning parameters proportional to radiation exposure.

   Since additive noise is inversely proportional to radiation exposure in a patient, parameters such as tube current or $CTDI_{vol}$, a CT dose index, would seem to serve as appropriate metrics. However, these are hardware settings at the time of acquisition and are independent of patient size and tissue density, two factors that contribute significantly to the resulting image quality. Larger patients, for example, require a greater amount of radiation than smaller ones to produce an image of diagnostic quality.

2. Subjectively analyzing of individual scans examined by a radiologist.

   Radiologists asses individual images based on their diagnostic quality. In this work, CT images have been subjected to subjective analysis where the specialist (radiologist) assigns a quality to individual scans. However, this subjective analysis, which incorporates factors pertaining to both scanner and patient by focusing on the resulting image, is difficult to standardize across different radiologists, instruments, and patients, thus

impeding ALARA. On the other hand, an automated method to estimate additive noise directly from an image could provide radiologists with a standardized tool to aid optimal protocol selection and ensure ALARA.

3. Automatically estimating the additive noise in individual CT images.

   Estimating additive noise from an image requires some understanding of the image and/or noise signal. The problem of separating signal from additive noise is fundamental to signal processing and has been investigated for a wide range of applications. In image processing, many image de-noising techniques have been developed to enhance an image post-acquisition. Such techniques, however, are not often used in clinical practices, to avoid image artifacts and distortions. Such effects can be difficult to distinguish from otherwise significant features in an image, and thus impede accurate pathology detection.

   One area of interest in medical image processing is image segmentation, which has spurred the development of various de-noising methods such as 1-D and 2-D correlations in multi-slice CT scans [5], noise estimation by conventional approaches [7], maximum-likelihood [8], Bayesian maximum-a-posteriori [8], linear estimator [9], or adaptive non-local means estimates [10]. These methods attempt to enhance existing CT images by improving signal-to-noise ratios (SNR) to improve segmentation. However, these methods can introduce image artifacts and distortions, such as false edges, and thus remain largely unused in clinical applications.

This thesis presents a novel approach to CT image modeling for additive noise estimation and image quality quantification. Such a study is unique in the field of CT imaging, representing a paradigm shift from image enhancement and segmentation post-acquisition, to image quality assessment during acquisition. Whereas prior approaches to CT image processing have focused on providing clinicians with tools to better interpret CT images, the goal of this work is to guide the technologist in taking diagnostic quality images, through

feedback on image quality, during the acquisition phase. Unlike image enhancement, image quality assessment does not modify the image but evaluates it. Thus, many image processing techniques that are avoided for CT image enhancement, due to the resulting artifacts and distortions, may function well for CT image quality assessment.

1.2 Contributions

The image quality quantification system presented in this thesis relies on estimating noise in an acquired CT image. By assuming noise to be additive to the signal, it is possible to estimate noise by measuring variations in pixel intensity of uniform regions in the image. These regions of interest (ROI) are not expected to vary due to signal, hence the measured variations provide an estimate of noise. The system functions by selecting ROI for noise estimation, and using that estimate to generate an image quality metric, as shown in Fig. 1.1.

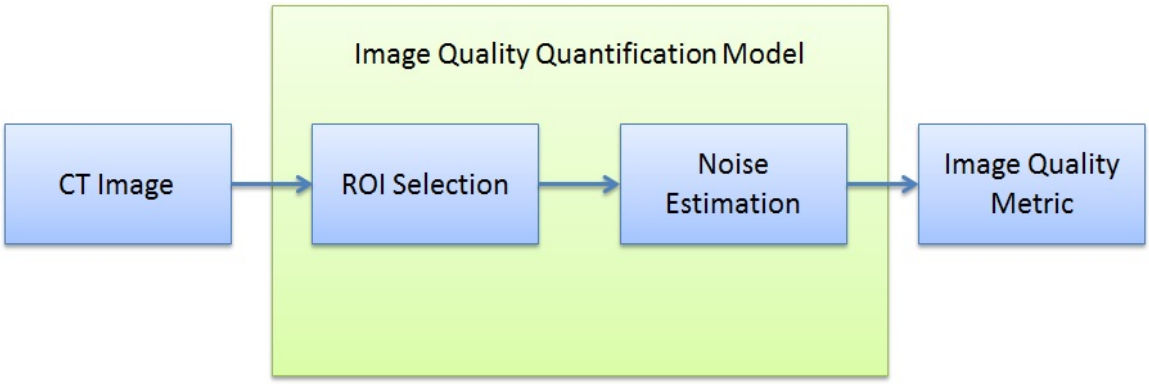


Figure 1.1: Overview of proposed image quality quantification system.

It is possible to propose different image quality models for this system based on the ROI selection criteria or mechanism, with the assumption that an optimal ROI exists within which any measured variation is due to noise. Thus, the performance of this system is limited by the ROI selection. This thesis presents two approaches to ROI selection:

1. Fixed-ROI based on domain knowledge.

   The Fixed-ROI approach attempts to select circular ROI within regions of uniform tissue. Initially, the heart was selected as a target region, but the resulting estimates were highly variable. Next, the paraspinal region was selected but the estimates remained variable. Finally, the fat was used as it proved less noisy than muscle, perhaps due to its location on the periphery of the body. However, the size of the fat region varies greatly across the patient population, posing challenges in patients with very little fat. In such a case, the paraspinal region can be considered as a secondary region for estimation.

   Once the ROI are selected, several metrics are viable for noise estimation, such as variance, standard deviation, coefficient of variation, and efficiency. By selecting multiple ROI, the variance between the ROI is also possible. Through testing, optimal metrics are identified to quantify image quality.

   In this thesis, the Fixed-ROI model is used as the baseline model against which other models are evaluated, due to its intuitiveness and reliance on domain knowledge.

2. Windowed Fourier-domain Distance Metric.

   This thesis also presents the *Windowed Fourier-domain Distance Metric* (WFDM) model as an approach to ROI selection by regions of signal uniformity. Specifically, the WFDM model subsamples a small and large region around each pixel of the original CT image , and compares the Fourier spectra of the two samples. Since both samples are centered on the same pixel, P, but vary in size, the resulting Fourier spectra show the variations around the same pixel affected by different sample sizes. The central region of the frequency spectra corresponds to the low frequency region. Measuring the distance between this central region in the frequency spectra of the two samples results in a distance metric corresponding to the spatial variation surrounding pixel P. The distance metric for each pixel in CT image I is then stored as a corresponding

pixel in a *distance image*, I_D. Finally, the ROI is selected by thresholding I_D, with the threshold q representing the maximum degree of variation in the selected region, as shown in Fig. 1.2. Once the ROI is selected, noise is estimated by various measures of the selected pixels, such as variance or standard deviation.

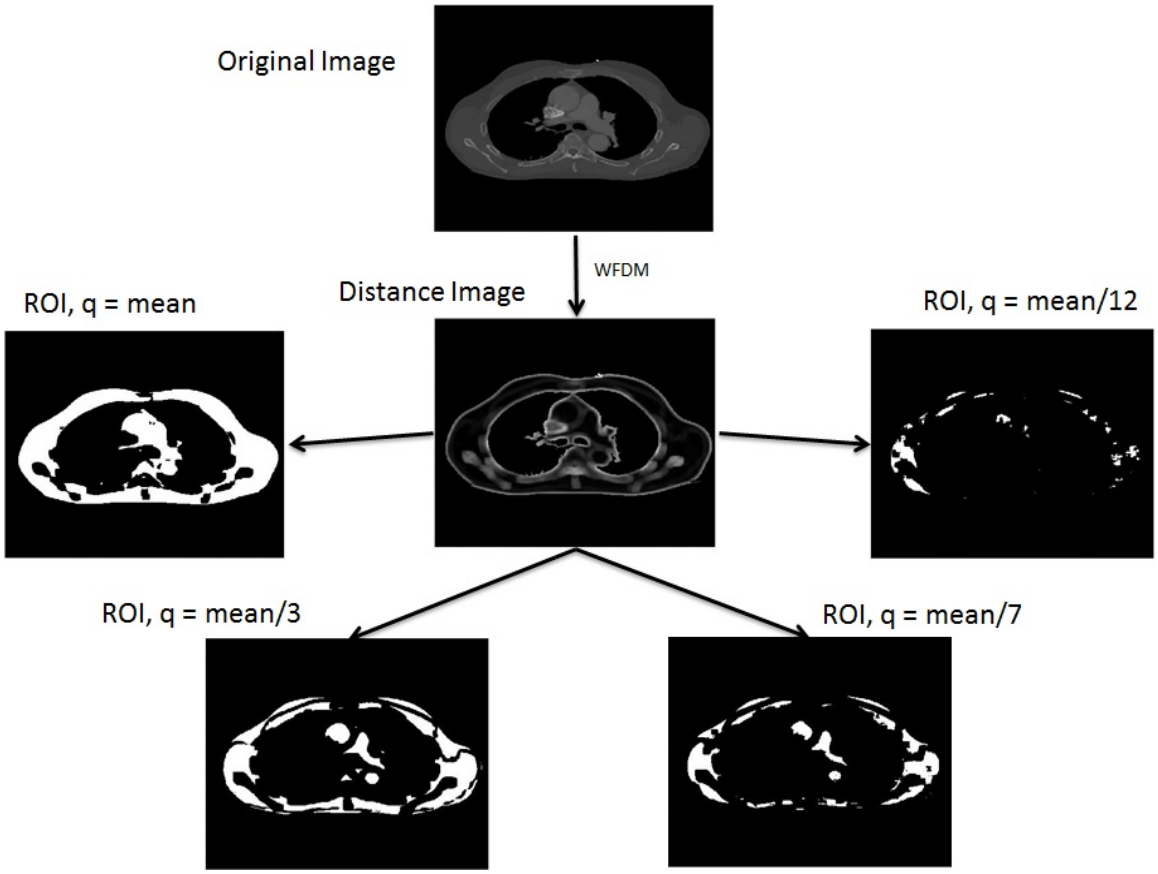


Figure 1.2: Region selection in the WFDM model.

One advantage of the WFDM region selection is that it does not rely on each patient's anatomy, thus avoiding the problem of selecting ROI in fat tissue for patients with little fat. Instead, the WFDM model selects the ROI in CT images by degree of local variation or, conversely, uniformity, suggesting a potential to compare similar regions across different images of the same patient or multiple patients.

In this study, the defined WFDM model is shown to perform well with phantom CT images and scale better than the alternative models with patient CT images. The relationship between the resulting image quality metric and scanning parameters, such as $CTDI_{vol}$, depends on many factors, most of which are uncontrolled. Thus, the WFDM model presented in this work is not yet a robust measure of absolute image quality and the resulting system does not yet meet the proposed goal of an automated image quality detector. However, the novel approach to image segmentation and ROI selection, as well as the observed trends in the image quality metrics, presents a promising step in that direction.

1.3 Thesis Outline

The body of this thesis is comprised of three chapters. Chapter 2 introduces the WFDM model and evaluates the performance of two resulting metrics on phantom CT images. Chapter 3 compares the performance of the WFDM model to that of the baseline model and a Convolutional Neural Network (CNN), using both phantom and patient datasets. Chapter 4 examines the performance of the baseline and WFDM models against changing input conditions, their scalability to large patient datasets, and provides an optimization for the WFDM threshold, q. Finally, results and future work are summarized in Chapter 5.

Chapter 2

THE PROPOSED WFDM METRIC: AUTOMATED SELECTION OF UNIFORM REGIONS

In this chapter, we propose a novel CT IQ metric that correlates the degree of spatial variation surrounding each pixel in an image. In uniform density (attenuation) regions, this variation is assumed to be an additive combination of signal and random noise. CT IQ is inversely proportional to the additive noise, such that low noise leads to high quality images and vice versa. Additionally, noise estimation, and therefore CT IQ classification, should be improved by extracting regions of low signal variations. This is achieved by relying on the observation that, in the absence of noise, the Fourier-domain estimate of a uniform spatial region of an image is independent of sample size. Thus, variations in the Fourier-domain spectrum of different sample sizes within a uniform spatial region can be indicative of additive noise, and the degree of variation in the Fourier-domain spectra can be used as a quantitative metric for IQ.

This chapter makes two key contributions. First, a novel windowed Fourier-domain based distance metric (WFDM) is introduced that is capable of estimating the variation in a relatively uniform spatial region due to additive noise. WFDM is found to identify thresholded spatial regions that correlate with the image acquisition parameters. Second, the WFDM is analyzed using phantom and real patient CT image data to ensure robustness and reliability of the proposed metric.

2.1 Proposed Method

For quantitative assessment of CT IQ, we extract spatial regions of interest (ROI) that can be used to estimate noise variance. CT images can be assumed to contain additive noise

along with the signal component and hence, segmentation of relatively uniform regions in every CT image is imperative for image quality estimation tasks.

A baseline method of ROI selection in CT images is performed by fixed thresholding, where all pixel values within a given range are set to 1 while values out of that range are set to 0. This baseline method is illustrated in Fig. 2.1, where a CT image with pixel values scaled to [0,1] is thresholded so that all pixels within [0.5,0.6] are set to 1 and the rest to 0. The threshold pixel intensity range is empirically determined. The thresholding results in the binary image shown in Fig.2.1(b) that can serve as a mask to extract a relatively uniform ROI shown in Fig.2.1(c). However, fixed thresholding selects ROI based on absolute pixel intensity, and hence cannot be generalized to extracting regions of low variation (ROI-LV) due to variations in pixel intensity across patients and different tissue types.

Prior work has shown that in the Fourier-domain, the central frequency spectrum structure (CFSS) of an image primarily corresponds to the signal rather than noise [11]. Therefore, the CFSS primarily correlates with the signal, and can be used to compare different regions of the image.

Specifically, for an abdominal CT image, the ROI-LV region is observed to have a Fourier-domain frequency spectrum that is more compact in the central, low frequency region than the in the higher frequencies. Further, the CFSS of the CT image data under analysis is observed to be mostly independent of the size of ROI-LV. Conversely, the spatial ROI with high variation (ROI-HV) is observed to have a Fourier-domain frequency spectrum with a less compact CFSS than the higher frequency regions and depend on sample size. Thus, changes in the CFSS due to changing sample size can be used to identify ROI-LV within an image or subset of an image.

Therefore, we propose a *windowed Fourier-domain distance metric (WFDM)* as a means of locating regions of low spatial variation that can be used to better correlate noise and CT IQ.

(a) (b) (c)

Figure 2.1: CT image thresholding for region of interest (ROI) extraction. (a) The original image of a phantom acquired at $10mA$ tube current. (b) A binary mask of the ROI generated by fixed thresholding after scaling the original pixel values to the range of [0,1] and setting [0.5,0.6] = 1, [0,0.5) & (0.6,1] = 0. (c) The extracted ROI resulting from pixel-wise multiplication of the binary mask and original image.

2.1.1 The Windowed Fourier-domain Distance Metric (WFDM)

The WFDM process follows the schematic shown in Fig. 2.2 and is detailed below.

We begin with a noisy image, I_n, of size $[n \times n]$ (2.1) where I is the ideal image and N is the additive noise. This image is passed through a high-pass filter (2.2), so that the local average intensities do not bias the measurement of low or high local spatial variation.

$$I_n = I + N \tag{2.1}$$

$$I_{HPF} = HPF(I_n). \tag{2.2}$$

At this point, two square windows centered at (x, y) are used to extract subsets of the image. The smaller window, w_s, has a dimension of $[d_s \times d_s]$ (2.3) while the larger window, w_l, has a dimension of $[d_l \times d_l]$ (2.4).

$$\forall(x, y), \quad w_s = I_{HPF}(x, y)[d_s \times d_s], \tag{2.3}$$

$$w_l = I_{HPF}(x, y)[d_l \times d_l], \quad d_l > d_s. \tag{2.4}$$

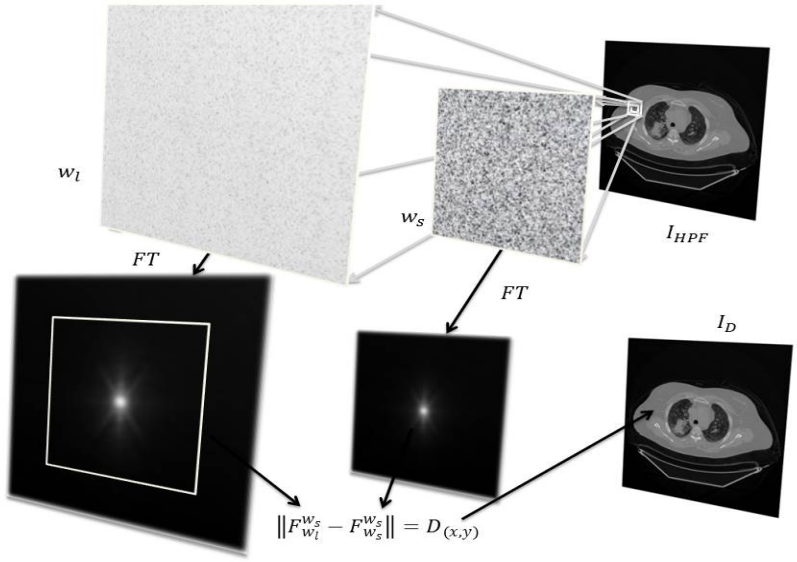


Figure 2.2: The high-pass filtered image, I_{HPF}, is analyzed in the Fourier-domain of two subsets of different size, w_l and w_s. The Euclidean distance between the two spectra, $D_{(x,y)}$, is found and stored at position (x, y) in the distance image, I_D.

The Fourier-transform of these spatially windowed regions are computed as F^{w_s} and F^{w_l} in (2.5) and (2.6), respectively.

$$F^{w_s} = \mathcal{F}(w_s), \tag{2.5}$$

$$F^{w_l} = \mathcal{F}(w_l), \tag{2.6}$$

$$F^{w_s}_{w_s} = F^{w_s}[w_s], \tag{2.7}$$

$$F^{w_l}_{w_s} = F^{w_l}[w_s]. \tag{2.8}$$

A region of size w_s around the center of each spectrum, (F^{w_s}, F^{w_l}), is sub-sampled, in (2.7) and (2.8) respectively. The Euclidean distance (2.9) between the resulting spectra, $(F^{w_l}_{w_s}, F^{w_s}_{w_s})$, is then found, corresponding to changes in the CFSS between the two spectra. We previously observed that the CFSS of ROI-LV is independent of sample size, as opposed to that of ROI-HV. Thus, this distance, $D_{(x,y)}$, is a Fourier-domain based measure of spatial uniformity surrounding each pixel (x, y), and is stored in its respective location in a *distance*

image, I_D (2.10).

$$D_{(x,y)} = \|F_{w_s}^{w_l} - F_{w_s}^{w_s}\|_2, \tag{2.9}$$

$$I_D(x,y) = D_{(x,y)}. \tag{2.10}$$

The intensity of each pixel in I_D correlates the degree of spatial variation surrounding that pixel in the image I_n, hence regions with similar variations in pixel strength can be segmented by fixed thresholding I_D to find ROI-LV and ROI-HV, respectively.

Similar to the fixed thresholding used in the baseline approach shown in Fig.2.1, a threshold q can be applied to I_D to generate a mask, U_T^q (2.11). This mask can then be used to segment the original image I_n by regions of similar spatial variation and thus provides a generalizable approach to ROI-LV selection.

$$U_T^q(x,y) = \begin{cases} 1, & \text{if } I_D(x,y) \leq q \\ 0, & \text{if } I_D(x,y) > q \end{cases} \tag{2.11}$$

2.1.2 Data

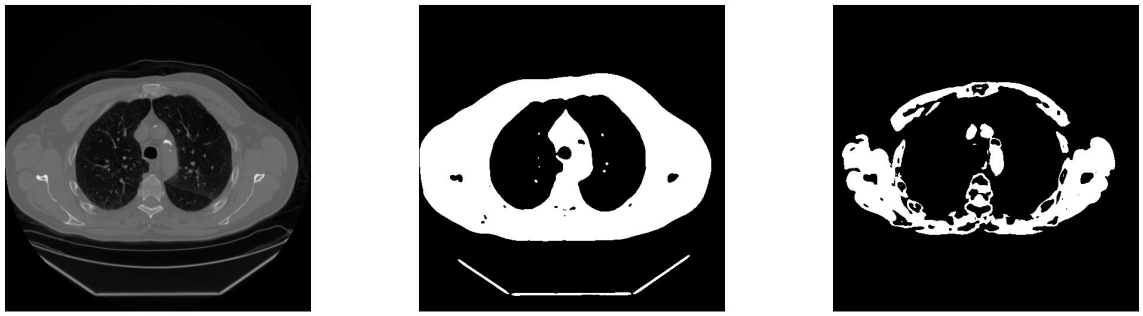
The data used for this work came from both phantom and patient scans. From an IRB approved study, CT images were acquired on two patients with a Lightspeed 16-slice scanner, where most settings were kept constant (helical, pitch 1.375, $20mm$ collimation, $120kVp$) and only the tube current was varied. For the patient whose scans were used in this study, the tube current values were $48mA$ and $5mA$, respectively. Both patient studies had 354 image slices from head to toe, but we extracted only the 101 image slices that correspond to the abdominal CT region. A dataset of phantom CT images was acquired using the tube current values of $10, 25, 75, 125, 175$, and $350mA$, respectively.

2.1.3 CT Image Visualization

We begin by visually comparing the ROI selected by the WFDM method to that of the baseline method. Fig. 2.3 demonstrates the baseline thresholding method from Fig. 2.1 now

applied to an image taken from a patient at $48mA$ tube current. Fig. 2.3(a) is the original image scaled to $[0, 1]$ and Fig. 2.3(b) is the mask acquired by thresholding pixel intensities such that $[0.4, 0.7] = 1$ and the remaining pixels are set to 0. The ROI shown in Fig. 2.3(c) results from further narrowing the threshold so that $[0.5, 0.6] = 1$ and the remaining pixels are set to 0. In addition, these threshold ranges were chosen empirically, hence the baseline method is not generalizable.

Conversely, applying WFDM to the same CT image shown in Fig. 2.3(a) and thresholding at $[q < \frac{\overline{\mathrm{I_D}}}{4}]$, where $\overline{\mathrm{I_D}}$ is the mean distance in I_D, results in the mask and corresponding ROI-LV shown in Fig. 2.4, respectively. Here, we observe the qualitative variations between the original image and the masked ROI-LV image. First, edges, as well as regions of high spatial variation such as the spine, have been removed. Second, the selected regions are not necessarily of the same absolute pixel intensity, as compared with the ROIs shown in Fig.2.3. Finally, the masked ROI-LV region in Fig. 2.4(b) is selected by relative thresholding as opposed to fixed intensity thresholding in the baseline method. Thus, the proposed WFDM based ROI-LV masking method is generalizable.



(a) (b) (c)

Figure 2.3: A CT image from the patient data imaged at $48mA$ tube current. (a) The original CT image, (b) a mask of ROI acquired by fixed thresholding $[0.5, 0.6] = 1$, and (c) a mask of ROI acquired by fixed thresholding $[0.525, 0.575] = 1$.

(a) (b)

Figure 2.4: The same CT image as in Fig. 2.3(a) after processing with WFDM. (a) A mask ROI-LV and (b) the masked ROI-LV are observed.

2.1.4 Performance Metrics

The CT images in these datasets are processed by WFDM. The selected ROI-LV are then used to calculate variance, standard deviation, and coefficient of variation of the noisy image, I_n, shown in (2.12), (2.13), and (2.14), respectively. Here, the symbol '$\circ$' represents pixel-wise multiplication.

$$\forall U_T^q = 1, \quad var : \sigma^2(U_T^q \circ I_n), \tag{2.12}$$

$$std : \sigma(U_T^q \circ I_n), \tag{2.13}$$

$$cov : \frac{\sigma}{\mu}(U_T^q \circ I_n). \tag{2.14}$$

Since the mask, U_T^q, used to extract the ROI-LV regions, depends on the threshold q, several values of q are comparatively analyzed. These q threshold values are defined as fractions of the mean of distances in the resulting distance image, I_D, for each CT image I_n.

The number of pixels below each q threshold value, ν, is also calculated (2.15).

$$\begin{aligned}\nu(q) &= \sum_{x=1}^{n}\sum_{y=1}^{n}(I_D \leq q)(x,y) \\ &= \sum_{x=1}^{n}\sum_{y=1}^{n} U_T^q(x,y).\end{aligned} \tag{2.15}$$

The abdominal ROI is defined as the tissue region of the CT image, hence the ROI-LV is a subset of the abdominal ROI. Thus, the pixels in the ROI-LV region, ν, are proportional to the number of pixels in the abdominal ROI for each patient CT image. The abdominal ROI is extracted by fixed thresholding, as shown in Fig. 2.1.

Defining the number of pixels in the abdominal ROI as ν_M, the number of pixels in the ROI-LV relative to ν_M can be calculated as in (2.16). This fraction of the abdominal ROI in ROI-LV, $\nu'(q)$, accounts for changes in abdominal ROI between patients and is therefore more generalizable.

$$\nu'(q) = \frac{\nu(q)}{\nu_M}. \tag{2.16}$$

The fraction of pixels, $\nu'(q)$, is representative of the relatively uniform ROI-LV region. Thus, it can be used to quantitatively analyze the correlation between CT image acquisition quality and the variation in spatial pixel intensities in the ROI-LV region.

2.2 Experiments & Results

The sensitivity analysis of the metrics *var* (2.12), *std* (2.13), *cov* (2.14), and $\nu'(q)$ (2.16) with respect to $q = [\overline{\mathrm{I_D}}, \frac{\overline{\mathrm{I_D}}}{2}, \frac{\overline{\mathrm{I_D}}}{3}, \frac{\overline{\mathrm{I_D}}}{4}, \frac{\overline{\mathrm{I_D}}}{5}, \frac{\overline{\mathrm{I_D}}}{6}, \frac{\overline{\mathrm{I_D}}}{7}, \frac{\overline{\mathrm{I_D}}}{8}, \frac{\overline{\mathrm{I_D}}}{9}, \frac{\overline{\mathrm{I_D}}}{10}]$ for different tube currents is shown below. Tube currents are directly proportional to CT image noise.

For the phantom CT images, the *var* measured in the baseline ROI is shown in Fig. 2.5(a). Here, *var* is observed to be inversely proportional to the CT image noise. A similar trend is observed for the ROI-LV regions obtained by the proposed WFDM method, such as in Fig. 2.5(b), where $q = \frac{\overline{\mathrm{I_D}}}{2}$. However, a higher variation in *var* is observed in the baseline

ROI than the ROI-LV. Thus, the ROI-LV of CT images acquired for each tube current are more uniform than the baseline method.

The sensitivity of *var* to the q threshold value is also observed by a comparative analysis of Fig. 2.5(b) - 2.5(d). The range of *var* values is observed to be inversely proportional to CT image noise, and increases with decreasing values of q, as shown in Fig. 2.5. This observation reflects the high variability with regards to sample size in ROI-LV regions obtained at lower thresholds.

Additionally, the *std* metric generates similar trends to *var*, with respect to CT image noise. However, the *cov* results in no significant trends, since q is a function of $\overline{\mathrm{I_D}}$.

The sensitivity of the ROI-LV to q is better illustrated by examining the size of the ROI-LV relative to the abdominal ROI, $\nu'(q)$, as shown in Fig. 2.6. Lower quality images show higher sensitivity to q than higher quality images. Thus, the correlation between $\nu'(q)$ and CT image noise for higher quality CT images is increased by decreasing the q threshold value.

For the phantom CT images, the overall sensitivity analysis for the correlation of each metric, $\nu'(q)$, *var*, *std*, and *cov*, to CT image noise, as a function of the q threshold value, is shown in Table 2.1. Here, $(*)$ represents $p \ll 0.001$.

The results in Table 2.1 indicate that the correlation between $\nu'(q)$ and CT image noise is highest at $q = \frac{\overline{\mathrm{I_D}}}{8}$, with a correlation coefficient (r-value) of 0.96 ($p \ll 0.001$). On the other hand, both *var* and *std* correlate better to CT image noise at $q = \frac{\overline{\mathrm{I_D}}}{3}$, with absolute correlation coefficient $|r| > 0.76$ ($p \ll 0.001$).

Next, the patient CT images are evaluated with respect to the metrics *var* and $\nu'(q)$, as shown in Fig. 2.7 for $q = \frac{\overline{\mathrm{I_D}}}{3}$. Here, the relationship of both *var* and $\nu'(q)$ to CT image noise reflects the trends observed in the phantom CT images, where *var* is observed to be inversely proportional to CT image noise while $\nu'(q)$ is observed to be directly proportional to CT image noise.

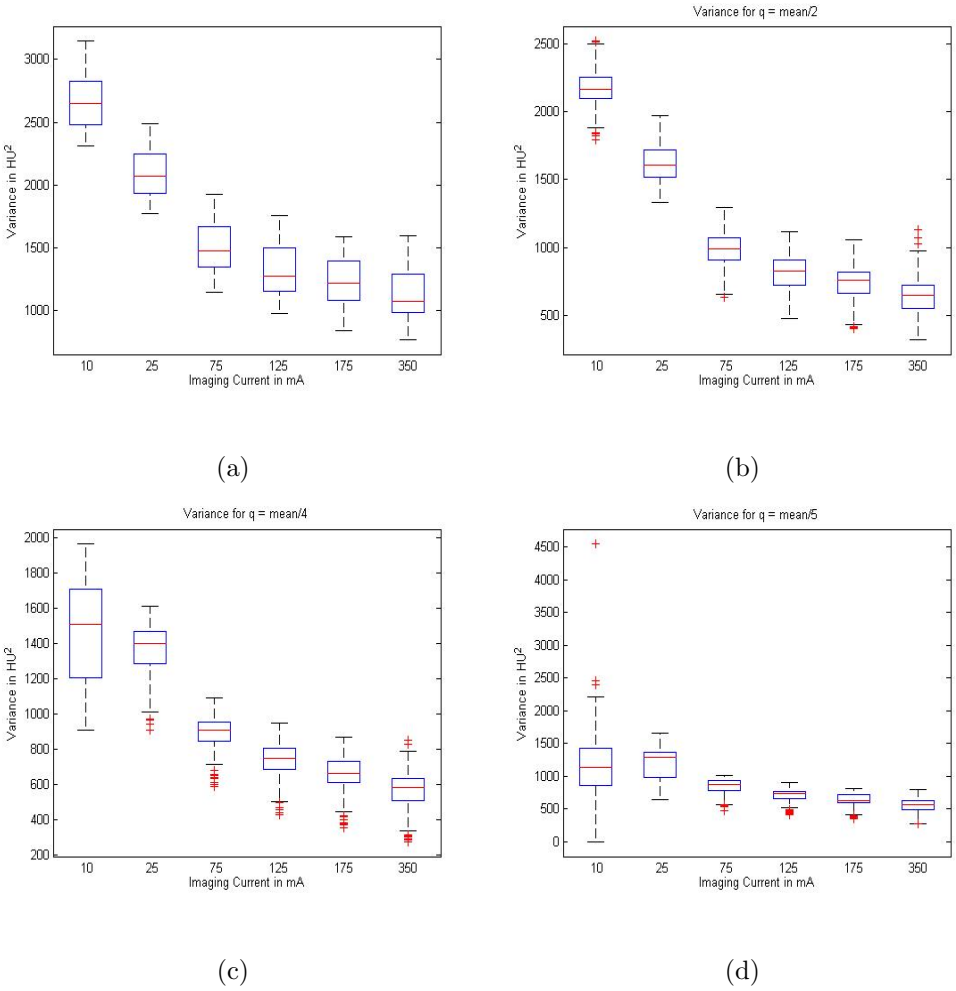


Figure 2.5: Analysis of *var* for the phantom CT images. (a) The *var* within ROI found by fixed thresholding. (b)-(c) The *var* within ROI-LV found by WFDM where $q = [\frac{\overline{I_D}}{2}, \frac{\overline{I_D}}{4}, \frac{\overline{I_D}}{5}]$, respectively. The x-axis is the tube current in mA. The y-axis is the variance in HU^2, where HU is the Hounsfield Unit.

In addition, $\nu'(q)$ for the combination of patient and phantom CT images at different q threshold values is shown in Fig. 2.8. Here, the patient CT images are observed to be significantly less sensitive to q than the phantom CT images. This is further supported by the results in Table 2.2, showing the correlation of each metric $\nu'(q)$, *var*, *std*, and *cov* to CT image noise at different q threshold values, for the combination of patient and phantom

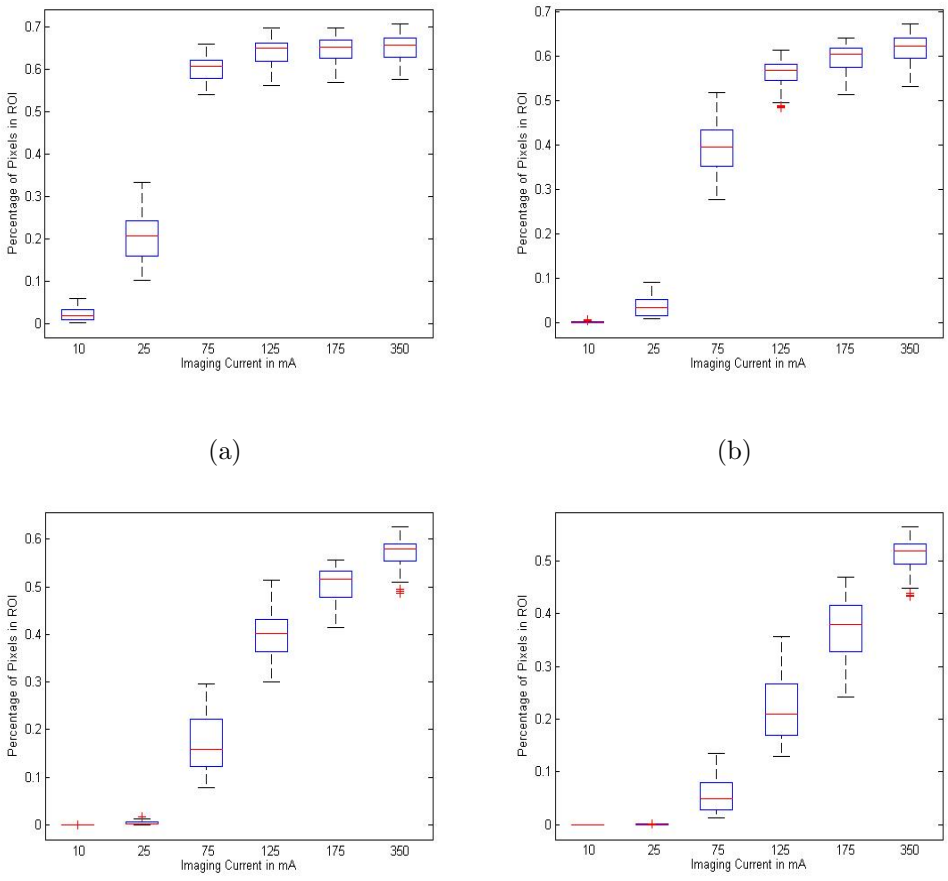


Figure 2.6: Analysis of $\nu'(q)$ for the phantom CT images. (a) $\nu'(q)$ at $q = \frac{\overline{I_D}}{4}$, (b) $q = \frac{\overline{I_D}}{5}$, (c) $q = \frac{\overline{I_D}}{6}$, and (d) $q = \frac{\overline{I_D}}{7}$, respectively. The x-axis is the tube current in mA. The y-axis is the fraction of ROI pixels in ROI-LV.

CT images. Here again, $(*)$ represents values of $p < 0.001$.

Table 2.2 shows that $\nu'(q)$ is best correlated to CT image noise in both patient and phantom CT images at $q = \frac{\overline{I_D}}{7}$ and $q = \frac{\overline{I_D}}{8}$, with a correlation coefficient of 0.95 ($p \ll 0.001$). On the other hand, the *var*, *std*, and *cov* do not correlate well to CT image noise in the

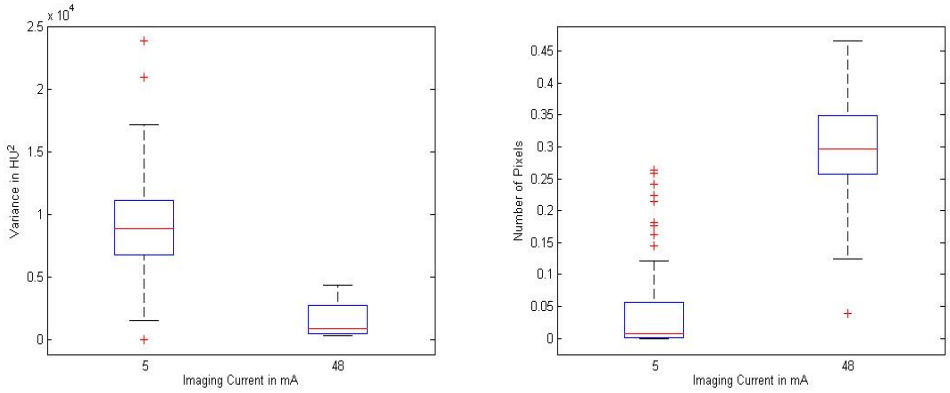
Table 2.1: Correlation of $\nu'(q)$, *var*, *std*, and *cov* with tube current in phantom CT images, at different q threshold values.

Threshold	Metric			
	$\nu'(q)$	*var*	*std*	*cov*
q	$r\,(p)$	$r\,(p)$	$r\,(p)$	$r\,(p)$
$\overline{\mathrm{I_D}}$	$-0.29\,(*)$	$-0.73\,(*)$	$-0.75\,(*)$	$-0.26\,(*)$
$\frac{\overline{\mathrm{I_D}}}{2}$	$0.21\,(*)$	$-0.74\,(*)$	$-0.76\,(*)$	$-0.28\,(*)$
$\frac{\overline{\mathrm{I_D}}}{3}$	$0.55\,(*)$	$-0.76\,(*)$	$-0.78\,(*)$	$-0.062\,(0.175)$
$\frac{\overline{\mathrm{I_D}}}{4}$	$0.69\,(*)$	$-0.75\,(*)$	$-0.77\,(*)$	$0.0044\,(0.924)$
$\frac{\overline{\mathrm{I_D}}}{5}$	$0.79\,(*)$	$-0.53\,(*)$	$-0.49\,(*)$	$-0.0022\,(0.960)$
$\frac{\overline{\mathrm{I_D}}}{6}$	$0.89\,(*)$	$-0.0010\,(0.826)$	$0.20\,(*)$	$-0.010\,(0.819)$
$\frac{\overline{\mathrm{I_D}}}{7}$	$0.95\,(*)$	$0.24\,(*)$	$0.41\,(*)$	$0.020\,(0.656)$
$\frac{\overline{\mathrm{I_D}}}{8}$	$0.96\,(*)$	$0.41\,(*)$	$0.56\,(*)$	$0.053\,(0.241)$
$\frac{\overline{\mathrm{I_D}}}{9}$	$0.95\,(*)$	$0.47\,(*)$	$0.599\,(*)$	$-0.00037\,(0.993)$
$\frac{\overline{\mathrm{I_D}}}{10}$	$0.91\,(*)$	$0.46\,(*)$	$0.611\,(*)$	$0.037\,(0.414)$

combined patient and phantom CT images, with absolute correlation coefficients below 0.46 ($p \ll 0.001$).

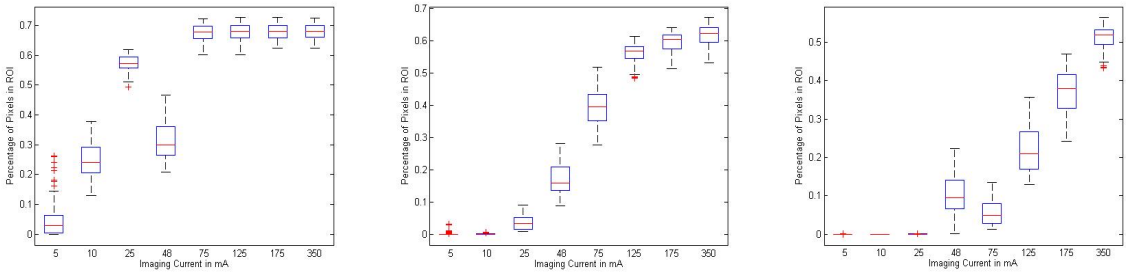
Although the trends observed in Fig. 2.5, 2.6, and 2.7, demonstrate that both *var* and $\nu'(q)$ correlate with CT image noise, the absolute values of *var* between the phantom and patient CT image datasets differ by an order of magnitude, whereas the value of $\nu'(q)$ is relative to the size of the abdominal ROI in each image. Thus, *var* does not correlate as well to CT image noise in the combined phantom and patient CT images as $\nu'(q)$.

Finally, the sensitivity of $\nu'(q)$ to q decreases at lower q threshold values, as observed in



(a) (b)

Figure 2.7: Analysis of (a) var and (b) ν' for the patient CT images at $q = \frac{\overline{I_D}}{3}$. The x-axis is the tube current in mA. The y-axes are (a) variance in HU^2, and (b) the fraction of ROI pixels in ROI-LV, respectively.



(a) (b) (c)

Figure 2.8: Analysis of $\nu'(q)$ for phantom CT images at 10, 25, 75, 125, 175, and $350mA$, and patient CT images at 5 and $48mA$ respectively. $\nu'(q)$ at (a) $q = \frac{\overline{I_D}}{3}$, (b) $q = \frac{\overline{I_D}}{5}$, and (c) $q = \frac{\overline{I_D}}{7}$, respectively.

Table 2.1 and Table 2.2. In addition, lower values of q result in larger variations in size of ROI-LV within CT images acquired at each tube current, as observed in Fig. 2.6 and Fig. 2.8. As a result, q can be optimized relative to both.

Table 2.2: Correlation of $\nu'(q)$, *var*, *std*, and *cov* with tube current in patient and phantom CT images, at different q threshold values.

Threshold	Metric			
	$\nu'(q)$	*var*	*std*	*cov*
q	$r\,(p)$	$r\,(p)$	$r\,(p)$	$r\,(p)$
$\overline{\mathrm{I_D}}$	$0.23\,(*)$	$-0.38\,(*)$	$-0.46\,(*)$	$0.021\,(0.601)$
$\frac{\overline{\mathrm{I_D}}}{2}$	$0.40\,(*)$	$-0.32\,(*)$	$-0.45\,(*)$	$0.013\,(0.737)$
$\frac{\overline{\mathrm{I_D}}}{3}$	$0.62\,(*)$	$-0.31\,(*)$	$-0.43\,(*)$	$0.0040\,(0.919)$
$\frac{\overline{\mathrm{I_D}}}{4}$	$0.75\,(*)$	$-0.23\,(*)$	$-0.34\,(*)$	$0.021\,(0.600)$
$\frac{\overline{\mathrm{I_D}}}{5}$	$0.83\,(*)$	$-0.14\,(*)$	$-0.21\,(*)$	$0.0042\,(0.916)$
$\frac{\overline{\mathrm{I_D}}}{6}$	$0.90\,(*)$	$-0.12\,(0.002)$	$-0.0057\,(0.886)$	$-0.017\,(0.670)$
$\frac{\overline{\mathrm{I_D}}}{7}$	$0.95\,(*)$	$-0.082\,(0.036)$	$0.084\,(0.033)$	$0.020\,(0.610)$
$\frac{\overline{\mathrm{I_D}}}{8}$	$0.95\,(*)$	$-0.058\,(0.138)$	$0.14\,(*)$	$0.046\,(0.245)$
$\frac{\overline{\mathrm{I_D}}}{9}$	$0.92\,(*)$	$-0.051\,(0.193)$	$0.16\,(*)$	$0.0050\,(0.900)$
$\frac{\overline{\mathrm{I_D}}}{10}$	$0.86\,(*)$	$-0.052\,(0.189)$	$0.16\,(*)$	$0.022\,(0.570)$

Chapter 3

COMPARATIVE ASSESSMENT OF SPATIAL SEGMENTATION METHODS ON PHANTOM AND PATIENT DATASETS

In the previous chapter, the WFDM model was presented as an alternative to the fixed ROI (baseline) model. For the baseline model, CT IQ is measured by estimating the variance in pixel intensities within manually-selected spatial regions-of interest (ROI) of a fixed circular size [12]. If an ROI is located in a region of spatially uniform signal intensity, then the measured variance in the ROI is an estimate of the noise added to the signal, which is inversely proportional to CT IQ. This method is dependent on the selection criterion for the fixed ROI that is guided primarily by domain knowledge. However, this fixed ROI-based CT IQ metric was not found to scale with inter-observer variabilities and large data sets. Additionally, this method suffers from incorrect ROI estimations in patients with minimal fat, or in poorly registered CT images.

In contrast, the WFDM model measures the degree of spatial variation in the regions surrounding each pixel. As demonstrated in the previous chapter, this can be used to select ROIs with low spatial variation (ROI-LV) to quantify CT IQ, regardless of the spatial orientation or location of these regions. This method is more generalizable than the prior fixed ROI approach for quantification of CT IQ. Also, the WFDM model demonstrated agree-ability with domain knowledge that specific regions, such as fat, are optimal for ROI selection, in abdominal CT images.

In this chapter, we comparatively analyze these existing models with cloud-based Big Data analytics tools using large numbers of abdominal CT images to realize the most generalizable CT IQ quantification metrics.

Convolutional Neural Networks (CNN) have been identified as a "valuable" Big Data analytics tool when raw image-based data is largely un-categorized [13]. In this study, we used CNN models developed using the cloud-based platform of Microsoft Azure machine learning studio (MAMLS) to classify CT images based on subjective CT IQ scores, thus bypassing the need for ROI selection and image segmentation. The CNN model learns the most significant hybrid features embedded in CT images from groups of pixels that are identified as feature maps by the convolutional and sub-sampling layers. The final neural network layer provides a probability score for the CT IQ based on the feature map regions. This cloud-based CNN work-flow overcomes CT image storage and processing limitations and leads to identification of structural patterns in CT images that can significantly benefit CT image acquisition parameter learning for each patient in personalized medicine applications.

The wide variety of CT image variabilities posed by the data analyzed in this chapter are representative of "Big" data analytics. Such rigorous analyses across CT image granularity, structural and textural variabilities from phantom to patient CT images has not been done before.

This chapter makes three key contributions. First, three data models with variable computational complexities are comparatively analyzed for predictive modeling of CT IQ. While structural ROI segmentation and pixel variance estimation methods are found to be generalizable models, CNN models are found to be useful for classification of image texture in the absence of structural variabilities. Second, a variety of classification parameters are analyzed for scalability in classification performance from phantom to patient CT images. Estimation of variance of pixels within a spatially uniform ROI is a robust classification parameter for segmentation data models. Third, the performance of data models is ranked based on average classification performance and consistency. We observe higher overall classification performances for *cross-sectional* CT IQ analysis when compared to *longitudinal* CT IQ classification. While ROI segmentation data models are found to be more consistent for CT IQ estimation and classification, the CNN model shows consistent CT IQ classification performance in the phantom image data set.

The organization of this chapter is as follows. Section II covers examples of prior work in automated CT image processing. In Section III, the data and mathematical models for CT IQ quantification are defined. Section IV describes the experiments and results followed by concluding remarks and discussion in Section V.

3.1 Prior Work

Major prior works on data analysis and implantation of machine learning for CT image processing include the following: automatic segmentation of different tissue types or organs in [14], identification of abnormalities or diseases in segmented regions in [15], and the suppression of imaging artifacts and noise, particularly in low-dose CT images in [16]. The significance of contributions in these works are discussed below.

Automatic segmentation of the lungs, as described in [14], is performed by thresholding pixel intensities. The intensity of pixels in a CT image region represents the attenuation of the x-ray through that region. Thus, tissue types that attenuate x-rays differently can be segmented by thresholding. Since pixel intensities are a function of imaging parameters and patient properties, the threshold pixels values can vary significantly across datasets. Thus, dynamic thresholding, which finds the threshold relative to the distribution of pixel intensities within a given image, is used to compensate for the variability between data sets. Once segmented, pixel intensities within the lung region are used to quantify lung volume. While the segmentation of the lungs by thresholding is shown to be consistent with manual segmentation performed by radiologists, the analysis of lung capacity largely relies on imaging protocol specifications. Further, noise due to low-dose protocols can significantly affect lung capacity estimations due to the significant variabilities in pixel intensities. However, these challenges can be addressed using CT IQ metrics to identify appropriate image processing methods for a given CT image and to group relevant data sets together for optimal analysis.

Artificial Neural Networks (ANN) have been applied to extract features for the identification of diseases, such as those in the lung as shown in [15]. In this work, the first step involves the segmentation and selection of the lungs by thresholding. The segmented lungs

are further processed to extract the relevant features used for disease identification, which in this case is the auto-correlation of the pixels in the lung. The ANN classifier is shown to identify various lung diseases by this method. However, the results of the auto-correlation are impacted by the noise and quality of the CT image, thereby making the ANN classifier dependent on the CT imaging parameters and the resulting CT IQ. In addition, the choice of features used in the ANN classifier may not be optimal for every patient. Thus, the development of quantifiable metrics for CT IQ, as well as the use of a CNN model to extract relevant features, can lead to generalizable data modeling.

Finally, the work in [16] demonstrates the application of machine learning to suppress image artifacts and noise in low-dose CT images, specifically by dictionary learning. Distinguishing artifacts from actual features in images is a traditionally challenging problem that is addressed in this work by expanding the dictionary to include artifact atoms trained along with the standard feature atoms. This discriminative dictionary can be used to identify both standard features and artifacts separately, allowing the latter to be suppressed. This method was tested with abdominal and mediastinum CT images and compared to the CT IQ scores provided by radiologists. However, the process of learning these discriminative dictionaries requires parameters and ROIs that are empirically set. Thus, this methods is applicable only to low-dose CT images with similar image acquisition parameters. Additionally, the discriminative dictionaries can significantly suppress artifacts, but random noise remains continues to impact the performance of the sparse coding model. The need for regularization terms in this method can be addressed through automated ROI selection methods and quantifiable CT IQ metrics, such as those discussed in this chapter.

3.2 Materials and Methods

This chapter presents comparative analysis of three distinct CT IQ quantification data models using abdominal CT images from a phantom and real patients. Two models rely on spatial segmentation of ROI that can best estimate noise in terms of pixel variability while the third CNN model relies learns hybrid features embedded in the image pixels that identify

CT IQ. Since additive noise in an image is observed as random variations in pixel intensities, CT IQ estimates must be limited to regions where the underlying signal is uniform. In all the three data models under analysis, domain knowledge and subjective CT IQ sores guide the optimal ROI selection strategies across a variety of image source variabilities.

Here, the three data models for CT IQ classification rely on different feature extraction methods. The first model uses the fixed size ROI segmentation by spatial orientation approach described in [12], and is used as the baseline model to compare the WFDM-based [17] and CNN classification models. Each model takes CT images as input and generates method-specific image quality scores, as shown in Fig. 3.1. The values Y_B, Y_{WFDM}, and Y_{CNN} represent the image quality scores for the baseline, WFDM, and CNN models, respectively. Prior to data modeling, each image is assigned a subjective CT IQ score (Y) based on its diagnostic quality. The goal is to identify a particular model that correlates highly with the subjective scores across data variabilities. The description of the datasets used for analysis, the data models for image classification and the output metrics for performance analysis are presented below.

3.2.1 CT Image Data

The data used for this work is obtained as abdominal scans from two distinct sources: phantoms and human patients. Each CT image comprises of [512x512] pixels. For the first set of phantom data, abdominal phantom CT scans are acquired by varying the image acquisition tube current values as $[10; 25; 75; 125; 175; 350]$ mA, respectively, using a third-generation dual-source multi-detector CT scanner (Siemens SOMATOM FORCE; Siemens Healthcare AG, Erlangen, Germany). During image acquisition, most system settings are kept constant (helical, pitch 1.375, 20 mm collimation, 120 kV p). Corresponding to each tube current setting, 6 sets of 81 abdominal CT images are thus acquired. Since tube current values are directly proportional to the CT IQ, the subjective CT IQ varies as $Y = [1; 2; 3; 4; 5; 6]$ for increasing tube current value, respectively. Thus, a multi-class classification task can be designed for the phantom data set.

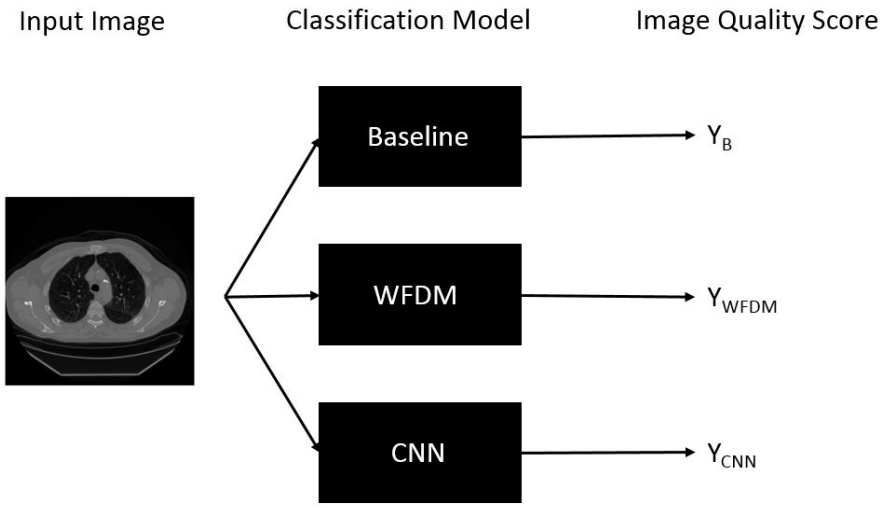


Figure 3.1: A CT image is passed as input to each image quality metric model, resulting in model-specific image quality scores Y_B, Y_{WFDM}, and Y_{CNN}.

For the second set of patient data, from an IRB approved study, CT scans are acquired from 6 individual patients who were imaged in 2 separate sessions with different image acquisition parameters. The subjective CT IQ for each of the 12 sets of CT scans were assigned in range $Y = [1 : 5]$, where $Y = 1$ implies poor quality image, $Y = 3$ implies optimal quality, and $Y = 5$ implies very high quality images, respectively. In this work, 100 images with [512x512] pixels acquired from the patient abdominal regions are analyzed from the 12 sets of patient image scans.

It is noteworthy that the patient data images are limited to $Y = [2, 3, 4]$ subjective CT IQ scores as opposed to the larger variations in CT IQ scores for the phantom data. This represents the lack of very poor and very high quality patient CT images, respectively. Thus, the following 3 classification tasks can be designed on the patient data set.

1. Multi-class Classification

This multi-class classification task, involves classification of each test image for $Y = [2, 3, or\ 4]$, separately.

2. Low by High Binary Classification

   This binary classification task, *low by high* (L/H), aims to distinguish CT images with low quality ($Y = 2$) from those with high quality ($Y = [3, 4]$).

3. Optimal by Sub-optimal Binary Classification

   This binary classification task, *optimal by sub-optimal* (O/S), aligns with the goal of limiting radiation dosages to comply with *ALARA* while maintaining diagnostic image quality. In this task based on domain knowledge, CT images with $Y > 3$ can be considered to have unnecessarily high quality and are therefore sub-optimal. Thus, in this binary classification task CT images that are of optimal quality ($Y = 3$) are classified from those with sub-optimal CT IQ ($Y = [2, 4]$). For each of the three classification tasks, the mean classification accuracy is for classification performance evaluation.

3.2.2 Data Models

The three data models analyzed for CT IQ estimation are described below. For predictive modeling of CT IQ, the input data samples are partitioned into training and test data sets, respectively. Using the training data set, the classification model parameters are estimated followed by classification of CT IQ for the test data samples.

1. Fixed Size ROI: Baseline Model

   This model [12] automatically locates a uniform ROI in each CT scan based on anatomical knowledge. An example of this model for an abdominal phantom image is shown in Fig. 3.2. First, the outer edge of the abdominal region is detected followed by estimation of two straight lines intersecting at the centroid of the abdominal region

and at angle θ from the horizontal axes, as shown in Fig. 3.2(b). At the intersection between these straight lines and the upper abdominal edge, the algorithm falls back by 1cm (8-12 pixels, depending on the image resolution), thereby avoiding the skin regions to detect pixel positions c. Using position c as the center, circles of area $1cm^2$ are identified as the fixed size patch ($Patch(\theta)$), as shown in Fig. 3.2(c). Next, the scaled mean ($\mu_\theta + 1000$) and standard deviations (σ_θ) of all the pixels lying within the patched regions are evaluated. Finally, the optimal ROI region (ROI_b) representative of the most uniform fat region is detected by varying $\theta = [5, 8, 11, 14, ...90]$ in (4.1)-(2). Here, the underlying assumption is that the most uniform region, that will be representative of most signal strength, will have the least coefficient of variation in pixel strengths.

$$\Theta = \arg_\theta \min\left(\frac{\sigma_\theta}{\mu_\theta + 1000}\right). \tag{3.1}$$

$$ROI_b \leftarrow Patch(\Theta). \tag{3.2}$$

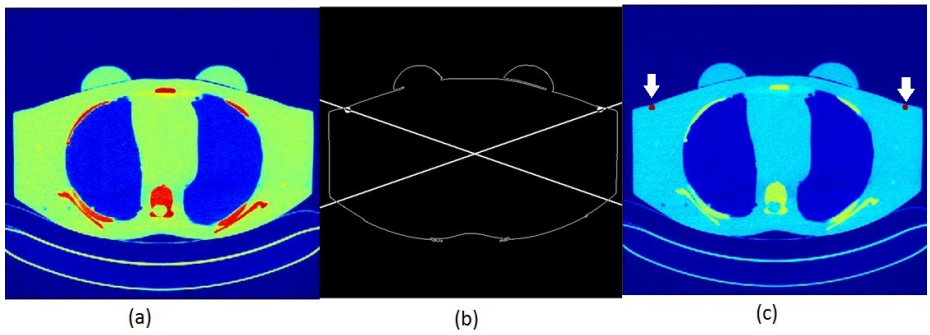
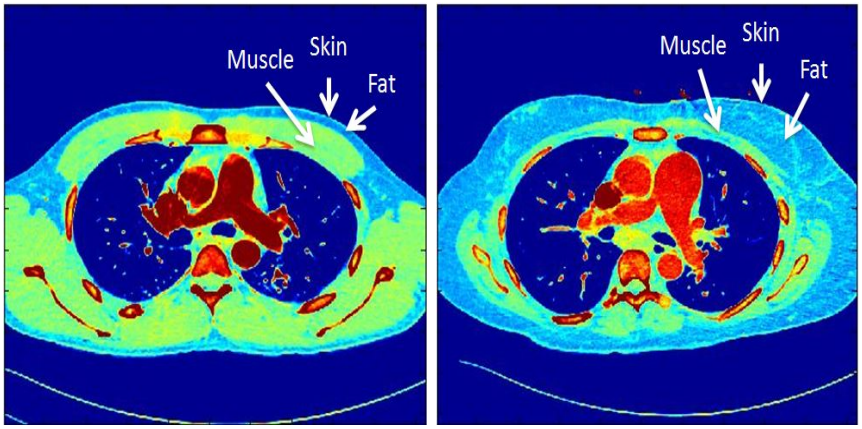


Figure 3.2: Steps for fixed-size ROI detection. (a) Phantom CT image. (b) Estimation of straight lines (horizontal angle θ) intersecting at the centroid of the abdominal region. (c) $Patch(\theta)$ regions detected as two circles with uniform area, indicated by the white arrows.

Although the fixed-size ROI model is independent of training and parameterization, it is heavily guided by the domain knowledge that the fat region is the most probable

uniform region for image signal strength estimation. However, in certain extreme situations, the estimated ROI is found to lie in fat, muscle, bone, or a combination of these regions. An example of these variations in patient CT images is shown in Fig. 3.3. Due to such variations in patient CT images, the fixed-size ROI model is not the most generalizable methods for CT IQ classification tasks.



(a) (b)

Figure 3.3: CT images of two patients showing the skin, fat, and muscle regions. (a) For a patient with low fat density between the skin and muscle, the fixed-size ROI region consists of non-fat or combination regions. (b) For a patient with high fat density between the skin and muscle, the fixed-size ROI is correctly identified in the fat region.

Since the fixed-size ROI detection model is minimally parameterized, it is the fastest CT IQ estimation method. For predictive modeling, the classification parameter learned from the training data includes median pixel variance , i.e., $med(\sigma^2_\Theta)$ or median coefficient of variation, i.e., $med(\frac{\sigma_\Theta}{\mu_\Theta+1000})$. Finally, the CT images in the test data set are classified for CT IQ corresponding to the closest matching classification parameter.

2. WFDM Model

The WFDM model [17] provides an alternative approach to ROI selection with the aim of avoiding uniform region estimation errors owing to patient variability. While

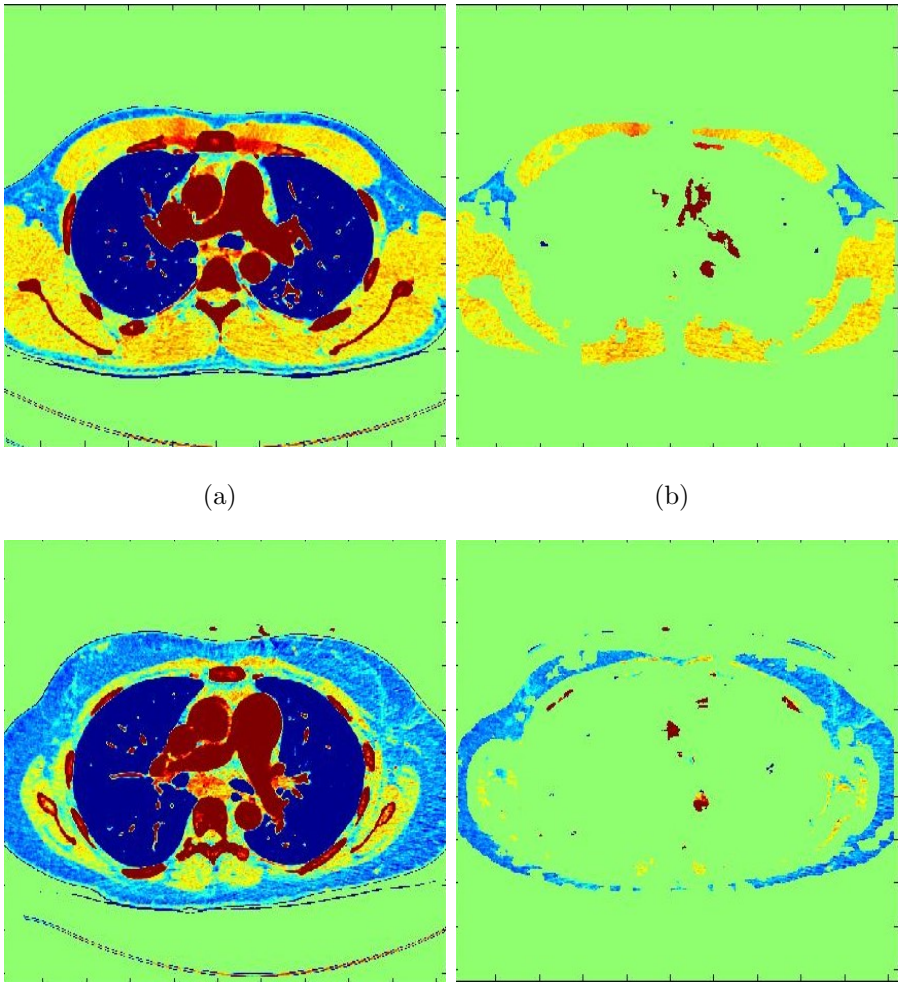
the baseline model uses spatial orientation guided by domain knowledge to select the best ROI, the WFDM approach involves image processing in the Fourier-domain. Each CT image is sub-sampled by two concentric moving square windows of different sizes, and the distances between the Fourier-spectra of the resulting sub-sampled images are stored in a *distance image* representation. The pixel intensities of the *distance image* correspond to orders of spatial variations in the sub-sampled region of the original CT image. Thus, the ROI mask generated by WFDM model (ROI_{WFDM}) is obtained by thresholding the *distance image*. An example this model for the images in Fig. 3.3 is shown in Fig. 3.4.

In contrast to the baseline method, the WFDM model does not select a specific region, such as fat, based on domain knowledge, but rather finds the most uniform regions in a given CT image based on pixel frequency-based information. Thus, the CT image in Fig. 3.4(a) results in the ROI in Fig. 3.4(b) that is predominantly the muscle region, while the CT image in Fig. 3.4(c) results in the ROI in Fig. 3.4(d) that is predominantly the fat region. In both cases, however, the WFDM model selects the most spatially uniform region for that particular CT image.

For predictive modeling, two classification parameters need to be trained for the WFDM model. The first parameter ν refers to fraction of the abdominal region pixels in ROI_{WFDM} that fall below a certain given threshold [17]. The second parameter is the variance of pixels within the ROI_{WFDM} mask. Both these classification parameters are motivated by prior work in [17]. The WFDM method relies only on one threshold parameter for ROI estimation, thus its low computational complexity. The only difference in CT IQ classification performance between this model and the baseline model is the ROI segmentation procedure.

3. CNN Model

A CNN model is motivated by the prior work in [18], where each input image is



(a) (b)

(c) (d)

Figure 3.4: ROI selection in two different patients. (a) Patient with low subcutaneous fat density and the corresponding WFDM-based ROI in (b).(c) Patient with high fat density and the corresponding ROI in (d).

subjected to several hidden layers of feature learning to generate an output vector of probability scores, for each image to belong to a certain CT IQ. For this analysis, we implement CNN architecture with the following 7 hidden layers: convolutional (C)-subsampling (S)-activation (A)-convolutional (C)-subsampling (S)-activation (A)-neural network (NN). Each C-layer convolves the input image with a set of ker-

nels/filters, the S-layer performs pixel pooling, A-layer performs pixel scaling in the range [-1,1] and the final NN-layer implements classification using 200 hidden neurons. A detailed description is presented in [19].

The CNN model has high computational complexity since it involves intensive parameter training for the hidden layer feature maps and NN classifier weights and biases. To avoid computational bottlenecks, the CNN model is trained in the MAMLS platform followed by classification of the test sample images for CT IQ. For each test image, the CT IQ corresponds to the output class with maximum probability score assigned by the NN-layer. Prior works have shown that the finally trained feature maps from the hidden layers capture structural patterns embedded in images that aid image-based classification tasks [20]. In this work, we analyze these feature maps to gain new insights into the spatial configurations in CT images that aid CT IQ classification tasks. The CNN model is completely different from the other two models from the implementation standpoint, thereby allowing for potential identification of a more robust CT IQ estimation and classification method.

3.2.3 Performance Analysis and Metrics

To analyze the performance of each data model, the output metrics include: accuracy, α, and its mean, $\overline{\alpha}$ calculated using equations (3.3) and (3.4) respectively. Here TP and FN indicate true positives and false negatives for classification of test images, respectively.

$$\alpha_i = \frac{TP_i}{TP_i + FN_i}, \tag{3.3}$$

$$\overline{\alpha} = \sum \alpha_i. \tag{3.4}$$

The classification parameters used for each of the three data models under analysis and the corresponding mean accuracies are shown in Table 3.1.

For analysis of the phantom and patient data sets, the following 3 sets of experiments are performed:

Table 3.1: List of classification parameters for each data model and their mean accuracy.

Accuracy notation	Meaning
$\overline{\alpha}_B^{var}$	Mean accuracy of baseline model, by variance
$\overline{\alpha}_B^{cov}$	Mean accuracy of baseline model, by coefficient of variation.
$\overline{\alpha}_{WFDM}^{var}$	Mean accuracy of WFDM model, by variance
$\overline{\alpha}_{WFDM}^{\nu}$	Mean accuracy of WFDM model, by ν
$\overline{\alpha}_{CNN}^{p}$	Mean accuracy of CNN model, by probability score

1. Multi-class Classification of Phantom Data Set

   CT IQ classification for phantom data set is a multi-class classification task, where the data set is split such that 70% of all images are used for training and the remaining 30% images are used for model testing [19]. The sample class frequencies for the training and test data sets are same.

2. Cross-sectional Analysis and Classification of Patient Data Set

   In this experiment using the patient data set, 1200 patient images are split with 70% and 30% of the images in the training and test data sets, respectively. These two sets are then used to train and test the 3 data models for multi-class, L/H, and O/S classification tasks. This *cross-sectional* analysis allows for comparison of the scalability for each model from the phantom to the patient dataset.

3. Longitudinal Analysis and Classification of Patient Data Set

   In this experiment using the patient data set, one set of CT images from each of the 6 patients is used as the training set while the other set of CT images is the test set.

This corresponds to a 50/50 data split. Using this training and test data sets, the 3 data models are analyzed for multi-class, L/H, and O/S classification tasks. This *longitudinal* analysis allows for assessment of CT IQ predictability for "personalized medicine" applications.

The analysis of data models for CT IQ classification on phantom and patient data sets has never been performed before. Such comparative evaluation of data models provides a basis for ranking each data model by performance, i.e., classification accuracy. Also, the performances of the data models can be ranked for each data set separately or based on average classification performances. Further, by measuring the change in classification performance across datasets, the consistency of the data models can be assessed.

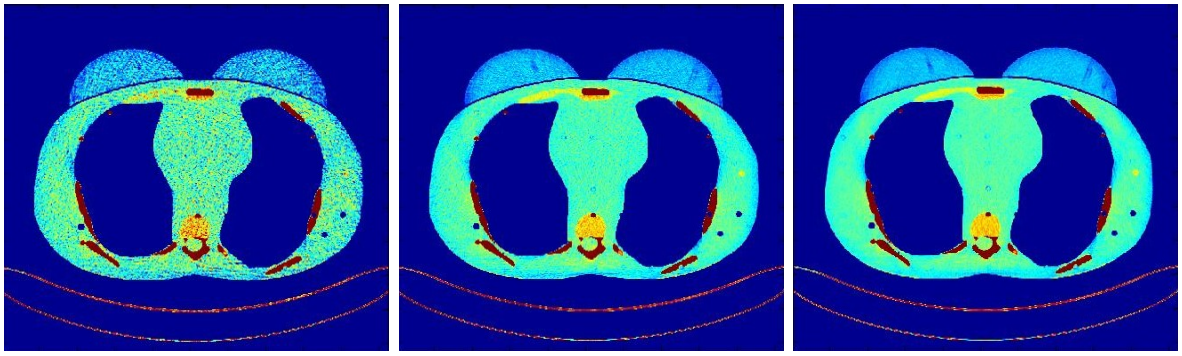
3.3 Experiments and Results

First, each of the 3 data models are trained and tested with the phantom dataset, which is demonstrative of the structural uniformity in CT images. While the phantom CT images were acquired at different tube currents, the size, orientation, and compositions of the phantom regions remain similar, as shown in Fig. 3.5. The CT images become significantly less grainy/noisy as tube current is increased from Fig. 3.5(a) - (c), but the structural aspects remain constant. Thus, the variabilities between the phantom CT images acquired at different tube currents, represent the effect of additive noise.

For images from the patient data in Fig. 3.3, the ROIs estimated by the baseline and WFDM methods are shown in Fig. 3.6(a). Fig. 3.6(b) shows the trained feature maps from the first C-layer in the CNN model. The range of variability in data modeling is illustrated in this example.

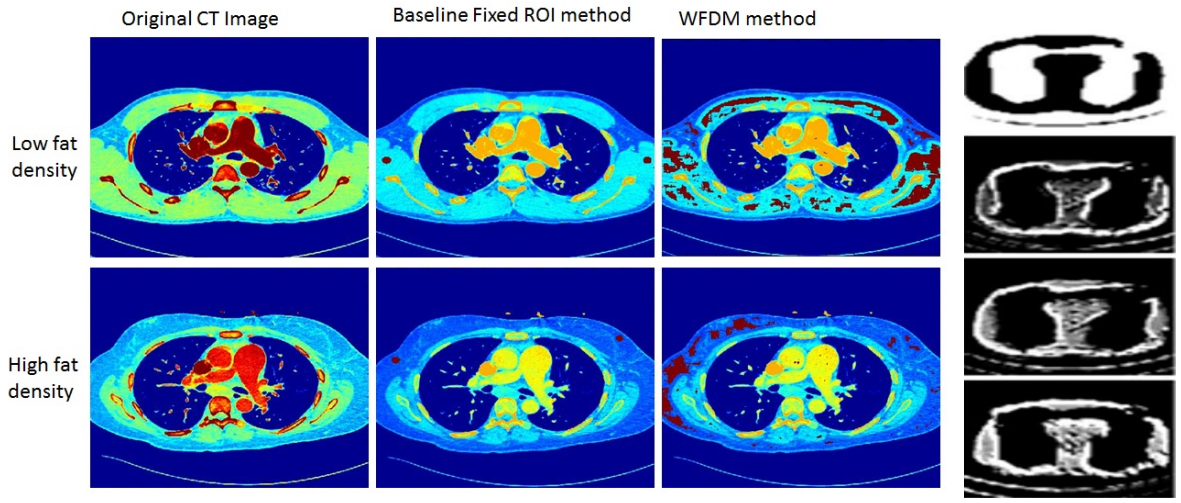
3.3.1 Multi-class Classification of Phantom Data Set

For the phantom CT images, the multi-class classification performances of all three data models are significantly better than random classification that would result in average accu-



(a) (b) (c)

Figure 3.5: Phantom CT images with (a) $Y = 1$, (b) $Y = 4$, and (c) $Y = 6$



(a) (b)

Figure 3.6: Example of data modeling variabilities for CT IQ classification tasks on patient data. (a) ROIs using the baseline and WFDM models. (b) ROIs learned from trained feature maps in the CNN model.

racy of 16.67%. These classification performances are reported in Table 3.2. The baseline model has the lowest mean accuracy of 73.6%, while the WFDM and CNN models both perform significantly better. We observe 100% classification accuracy for $\overline{\alpha}^{var}_{WFDM}$, which

indicates the feasibility of predictive modeling for phantom image CT IQ.

Table 3.2: Mean accuracy of phantom CT image classification using the 3 data models.

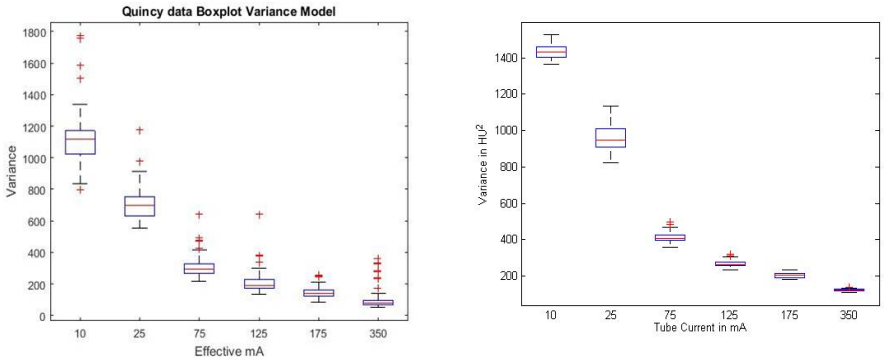
Subjective score (Y)	$\overline{\alpha}_B^{var}$	$\overline{\alpha}_B^{cov}$	$\overline{\alpha}_{WFDM}^{var}$	$\overline{\alpha}_{WFDM}^{\nu}$	$\overline{\alpha}_{CNN}^{p}$
Multi-class	0.736	0.736	1.000	0.882	0.938

Next, the performances of the baseline and WFDM data models are analyzed by comparing the distributions of pixel variances in both ROIs, respectively, as shown in Fig. 3.7. For CT images within each tube current, the spread of variances in the baseline model, shown in Fig. 3.7(a), is larger than that of the WFDM model, shown in Fig. 3.7(b). In particular, the outliers in the baseline model are likely due to imaging artifacts present in some of the CT images. The region with ROI_b is much smaller than that in ROI_{WFDM}. Thus, imaging artifacts disproportionately impact both the ROIs, thereby resulting in a greater spread of outliers, as seen in Fig. 3.7. While both plots show the expected inverse trend between pixel variance and CT IQ, the greater spread of outliers for the same training data set makes the WFDM model preferable over the baseline model.

We observe that the CNN data model performs significantly better than the baseline method, while both these methods have superior classification performances for low quality CT images than the high quality images. Besides, the WFDM model outperforms the CNN model with pixel variance as the classification parameter, but under-performs when using ν as the classification parameter.

3.3.2 Analysis and Classification of Patient Data Set

In contrast to the phantom dataset, the patient CT images vary structurally, as previously demonstrated in Fig 3.3 and 3.4. Thus, the classification performances of the data models are expected to decrease when transitioning from the structurally ideal CT phantoms to the real



(a) (b)

Figure 3.7: Phantom data multi-class classification for CT IQ with the (a) baseline and (b) WFDM models, using pixel variance as the classification parameter.

patient images. The resulting classification confusion matrices for multi-class classification of phantom and patient data set images are shown in Fig. 3.8.

Unlike the phantom dataset, the patient dataset is analyzed using *cross-sectional* and *longitudinal* experiments, as shown in Tables 3.3 and 3.4, respectively.

1. Cross-sectional Analysis and Classification

   The results of *cross-sectional* analysis for all 3 data models are shown in Table 3.3. The CNN model results in 100% accuracy for multi-class, L/H, and O/S classification tasks, suggesting the model is over-fitting the data. The WFDM model outperforms the baseline model for the multi-class, L/H, and O/S classification tasks.

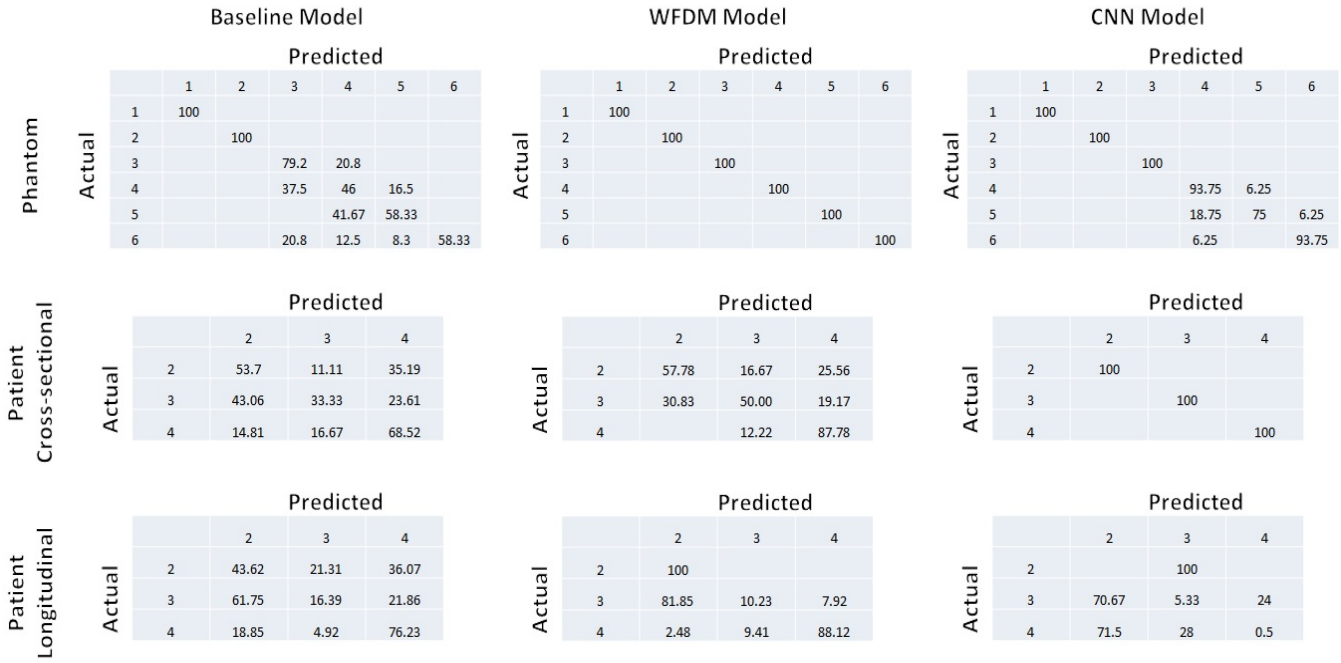


Figure 3.8: Confusion matrices for the 3 data models using the best classification parameter per model on the phantom and patient datasets, respectively. The phantom CT images are classified for CT IQ $Y = [1 : 6]$. The patient CT images are classified for CT IQ $Y = [2, 3, 4]$ using the *cross-sectional* and *longitudinal* analysis of patient data. The empty matrix locations imply a 0 value.

Table 3.3: Mean accuracy of patient CT image classification for *cross-sectional* analysis using the 3 data models.

Classification Task	$\overline{\alpha}_B^{var}$	$\overline{\alpha}_B^{cov}$	$\overline{\alpha}_{WFDM}^{var}$	$\overline{\alpha}_{WFDM}^{\nu}$	$\overline{\alpha}_{CNN}^{p}$
Multi-class	0.495	0.495	0.515	0.652	1
L/H	0.590	0.602	0.668	0.701	1
O/S	0.551	0.593	0.747	0.678	1

2. Longitudinal Analysis and Classification

   The results of the *longitudinal* analysis for all 3 data models are shown in Table 3.4. The WFDM data model outperforms both the baseline and CNN models for the multi-class, L/H, and O/S classification tasks. The CNN data model is most affected by the change from *cross-sectional* to *longitudinal* data splitting, as indicated by the significantly poor performance in multi-class classification.

Table 3.4: Mean accuracy of patient CT image classification for *longitudinal* analysis using the 3 data models.

Classification Task	$\overline{\alpha}_B^{var}$	$\overline{\alpha}_B^{cov}$	$\overline{\alpha}_{WFDM}^{var}$	$\overline{\alpha}_{WFDM}^{\nu}$	$\overline{\alpha}_{CNN}^{p}$
Multi-class	0.395	0.451	0.486	0.661	0.019
L/H	0.493	0.490	0.557	0.750	0.525
O/S	0.555	0.530	0.675	0.731	0.430

In the cross-sectional analysis, shown in Table 3.3, all classification parameters for the baseline model result in 49.5% accuracy in the multi-class classification, which is significantly lower than the baseline model's performance on the phantom dataset. The WFDM model results in multi-class classification accuracies up to 65.2%, while the CNN performs significantly better with an accuracy of 100%. Conversely, the CNN model results in multi-class classification accuracy of 1.9% for the *longitudinal* analysis of patient data, as shown in Table 3.4. Here, we observe that the CNN data model detects structural similarity in images for classification, thereby resulting in data over-fitting. However, the baseline and WFDM models have consistent classification performances in both the *cross-sectional* and *longitudinal* analyses.

For the two binary classification tasks (L/H and O/S) on the patient data, the baseline model performs slightly better than random classification with a mean accuracy $\overline{\alpha}_B^{var} \leq 0.590$

and $\overline{\alpha}_B^{cov} \leq 0.602$ for the L/H classification task in the *cross-sectional* analysis. Both *var* and *cov* classification parameters for the baseline model result in superior classification for L/H classification than O/S classification. The WFDM model performs significantly better than random classification, with $\overline{\alpha}_{WFDM}^{var} \leq 0.747$ for O/S and $\overline{\alpha}_{WFDM}^{\nu} \leq 0.701$ for L/H classifications in the *cross-sectional* analysis. Thus, for the WFDM model, the classification parameter *var* is preferable for O/S classification while the classification parameter ν is preferable for L/H classification. For the *cross-sectional* analysis, the CNN model performs with 100% accuracy for both binary classifications, as expected due to over-fitting.

In contrast, the performance of the baseline model is more random for the *longitudinal* analysis when compared to the *cross-sectional* analysis, with $\overline{\alpha}_B^{var} \leq 0.555$ and $\overline{\alpha}_B^{cov} \leq 0.530$. In this case, however, O/S classification results in the higher accuracy using classification parameters *var* and *cov* when compared to the L/H classification. On the other hand, the WFDM model again performs significantly better than random, with $\overline{\alpha}_{WFDM}^{var} \leq 0.675$ and $\overline{\alpha}_{WFDM}^{\nu} \leq 0.750$. Also, we observe that *var* and ν are preferable classification parameters for WFDM model for O/S classification and L/H classification, respectively, as was the case in the *cross-sectional analysis*. Here, the CNN model results in a near random performance, with $\overline{\alpha}_{CNN} \leq 0.525$.

We observe that the phantom dataset varies only in CT IQ, as shown in Fig. 3.5. Conversely, patient CT images have a higher degree of variability in image features and structures across patients and scans as shown in Fig. 3.9 and Fig. 3.10, respectively. These additional variabilities in patient data result in the deterioration of classification performances for patient data in *cross-sectional* and *longitudinal* experiments.

The specific differences between each patient's CT images also explain the classification results from the CNN model. CNN models are known to be more sensitive to detection of structural similarities in images when compared to variations in image texture and granularity [20]. Thus, the CNN identifies structural similarities in the CT anatomy per patient rather than detect CT IQ. Hence, the CNN model is not preferable for CT IQ classification in the presence of vast image structural variabilities.

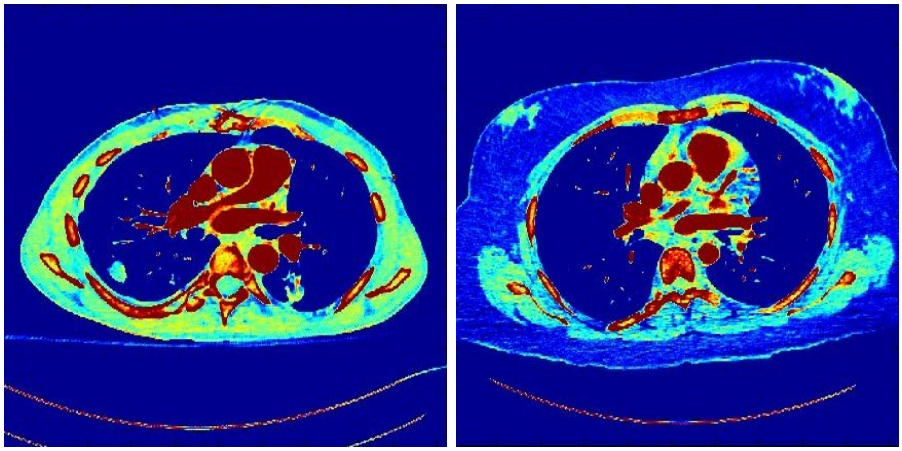


Figure 3.9: CT images of two different patients, with (a) low fat density, (b) high fat density.

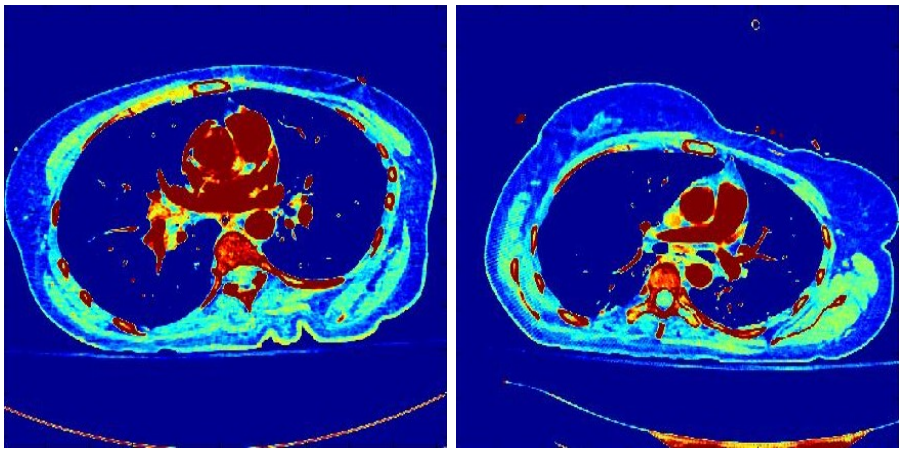


Figure 3.10: CT images of the same patient acquired at two different times, showing a change in alignment and sizes between the two imaging sessions.

The CT images shown in Fig. 3.9 present a challenge for the baseline and WFDM models as well as for the CNN model. The ROI selection is process is affected by the features available in that particular image. Thus, a patient with less body fat, or low fat density, will not have a comparable ROI to another patient with high fat density. For the WFDM model, one particular drawback in such situations is the use of classification parameter ν. Assuming fat

will predominantly be in ROI_{WFDM} regions, while bone regions will not, it is clear that a person with substantially more fat tissue will also have a higher percentage of all abdominal pixels in ROI_{WFDM} when compared to person with less fat tissue. Thus, the ν classification parameter is less generalizable for patient CT IQ classification when compared to var.

CT images contain complex and variable features that may yield useful information when analyzed in large quantities, hence the need for a Big Data approach to CT image analysis. However, such an approach relies on generalizable metrics and automated tools that are compromised by the complexity and variability of the CT images themselves. As a result, while the models attempt to produce noise estimation metrics useful for Big Data analysis, the performance of each model is significantly affected by various factors unrelated to noise. When these factors are considered, such as in the case of the phantom CT images where all image parameters are constant expect for the tube current, then these models consistently measure the noise. However, in the presence of other factors, such as the variations in the patient CT images, the predictive capabilities of the data models significantly deteriorate.

Additionally, the patient CT images are classified based on a subjective score (Y) that is provided for a single image in each patient CT image scan set. While the rest of the CT image set is assumed to be the same quality, this assumption may not always hold. The image acquisition parameters vary throughout a patient scan to account for anatomical changes in the regions being scanned. These changes may result in different amounts of noise in two images acquired from the same scan. Thus, lack of subjective data results in over-fitted CNN models.

Once the data models are comparatively analyzed for classification performances, the next step is to rank their performances. For both the phantom and patient CT images, the WFDM model is more accurate than the baseline model. This suggests that the ROIs selected by the WFDM model are better indicators of additive image noise than the baseline model. However, the comparable drop in classification performances between the two models, when processing patient CT images, is indicative of structural dissimilarity factors that affect the classification performances. Also, the lack of subjective scores on patient CT image scans

impact the CT IQ predictability. The CNN data model performs consistently on the phantom CT images and but it suffers from inconsistencies for the the patient CT images. However, with the phantom CT images, the CNN proves a powerful tool in classifying CT image noise so long as the the structural variabilities are low. Thus, the CNN-based classification performance may improve by masking the structural dissimilarities and by increasing the number of input images. Finally, the WFDM model outperforms the baseline model in both the multi-class and binary classification tasks and it is consistent in performance across *cross-sectional* and *longitudinal* data analyses. Thus, based on the data set and data models under analysis, the WFDM method is ranked highest followed by the CNN and the baseline model, respectively.

Chapter 4

COMPLETE AUTOMATED SYSTEM FOR CT IMAGE QUALITY DETECTION

In the previous chapters, we have discussed models that estimate additive noise in specific ROI to form IQ metrics. This chapter builds on those models, making two key contributions. First, a complete evaluation of the WFDM model is performed. A preliminary version of this model is presented in chapters 3 and 4. However, in this chapter, the model is applied to 6 stacks of phantom and 127 stacks of patient images, identifying exclusion criteria and WFDM threshold parameter for optimal IQ assessment. Second, the baseline and WFDM IQ metrics are used to predict trends in changing $CTDI_{vol}$ between 84 pairs of patient images. The sensitivity of these predictions to the percent change in $CTDI_{vol}$ are analyzed in Trend Error Curves (TEC). Including all 84 pairs of patient images, the total percent error is found to be 30.95% and 21.43% for the baseline and WFDM models, respectively. However, this error drops to 0% for $CTDI_{vol}$ changes of at least 40.0% for the baseline model and 27.5% for the WFDM model.

4.1 Prior Work

The goal of the baseline and WFDM models is to provide metrics for CT IQ. However, the accuracy of these metrics cannot be directly evaluated as there is no empirical measure for CT IQ to begin with. Indeed, the development of quantitative metrics for CT IQ is motivated by this lack of a simple measure. The quality of a CT image is a function of many variables including image acquisition parameters such as x-ray dosage and the size and condition of the patient being imaged. Thus, to evaluate the CT IQ metrics of the baseline and WFDM models, a variable proportional to IQ must be chosen for comparison, while controlling for

all other factors as much as possible.

In the previous chapter, the baseline and WFDM models were evaluated against a Subjective Quality Index (SQI) provided by a radiologist. Each image set was manually evaluated by the radiologist and labeled with the appropriate SQI score, from 1 through 5. A score of 3 is optimal, while scores below 3 are low quality and score above 3 fail to meet ALARA. While the baseline and WFDM models were able to classify some of the CT images correctly by SQI, the subjective nature of the scoring proved inconsistent for a thorough analysis of these models.

In this chapter, x-ray dosage as measured by $CTDI_{vol}$ is used for evaluation, and each patient is scanned multiple times with different dosages. By comparing slices of the same region from high and low dose scans of each patient, anatomical variations across scans are minimized and the x-ray dosage becomes the primary factor in CT IQ. Thus, differences in $CTDI_{vol}$ between the high and low scans should correspond with differences in CT IQ and be matched by the proposed IQ metrics. Two experiments are performed to evaluate these IQ metrics.

4.2 Materials and Methods

In this paper, three experiments are performed to analyze the performance of the baseline and WFDM models, using both phantom and patient datasets. First, the direct CT IQ estimation is analyzed for both the baseline and WFDM models applied to phantom and patient datasets, demonstrating the absolute performance of the proposed IQ metrics, as well as the limitations of $CTDI_{vol}$ as a reference value. Next, the performance of both models in estimating relative trends in IQ for pairs of high and low dose images is analyzed, as well as the correlation between changes in $CTDI_{vol}$ and the IQ metrics. Finally, the results of the low by high classification with the WFDM model are used to optimize the distance thresholding used in the ROI selection process, completing the design of the WFDM model.

Table 4.1: Notation Table

Notation	Description
Model-specific Metrics	
I	Input CT image.
I_{out}	Binary mask of the abdominal region.
I_{lungs}	Binary mask of the lungs.
I_{mask}	Binary masks of baseline ROI.
I_{reg}	Binary mask = $I_{out} - I_{lungs}$.
I_D	Distance image resulting from WFDM model.
q	Threshold used to select ROI from I_D.
I^q	Thresholded region ($I_D \leq q$).
r_{mask}	Radius of circular ROI in the baseline model.
θ	Angle from the vertical used to project straight lines in the baseline model.
$\hat{\theta}$	Optimal angle of baseline model.
d	Distance of each baseline ROI mask along the projected straight lines.
$\hat{d}$	Optimal distance of baseline model.
$cntr_{I_{out}}$	Centroid of abdominal region.
w_l	Size of large window in WFDM model.
w_s	Size of small window in WFDM model.
$I_{i,j}^{w_l}$	Sampled image of size w_l, centered at pixel (i, j).
$I_{i,j}^{w_s}$	Sampled image of size w_s, centered at pixel (i, j).
$F_{i,j}^{w_l}$	Fourier spectrum of $I_{i,j}^{w_l}$.
$F_{i,j}^{w_s}$	Fourier spectrum of $I_{i,j}^{w_s}$.
$V_{i,j}^{w_s}$	Variance of pixels in $I_{i,j}^{w_s}$.
Evaluation Metrics	
$\mu_{I'}$	Mean pixel intensity in image I'.
$\sigma_{I'}$	Standard deviation in image I'.
$\xi_{I'}$	Efficiency = $\frac{\sigma_{I'}^2}{\mu_{I'}^2}$
η_V^q	Mean of variances (V^{w_s}) across ROI in WFDM, at q.
TEC	The Trend Error Curve described in section 4.3.2.

4.2.1 CT Image Data

The experiments in this paper use both phantom and patient datasets. The phantom CT images were acquired at 6 different tube currents, corresponding to 6 different $CTDI_{vol}$ values. For this study, each set of 81 images is sub-sampled to exclude imaging artifacts, resulting in 5 phantom CT images at each $CTDI_{vol}$. The patient CT images were acquired from a group of patients at different times and $CTDI_{vol}$ values, including multiple abdominal scans of individual patients, as part of a Pulmonary Embolism study. Thus, comparative analysis of images from the same patient can be performed. For each scan of a patient, the image slice containing the pulmonary bifurcation is sampled, along with the image slices 3 and 6 cm above and blow it. This results in approximately the same 5 images per scan of each patient, which are then used to perform the analysis in this study.

4.2.2 Data Models

Fixed Size ROI: Baseline Model

The baseline model follows the steps outlined in Algorithm 1. Starting with CT image I, Fig. 4.1 (a), a *Highpass_filter* function is used to segment it into the body region, I_{out}, and the lungs, I_{lungs}, shown in Fig. 4.1 (b) and (c) respectively. This *Highpass_filter* function uses Gaussian blurring to reduce non-edge, high frequency signals such as noise and improving edge detection. This function also includes filling holes enclosed regions to produce the binary masks. Using the *Edge* function, I_{out} also provides the outer edge of the patient's body, Fig. 4.1 (d). Similarly, the *Edge* function applied to I_{lungs} to get the boundary of the lungs. The centroid, *cntr* is acquired from I_{out} and used to draw straight lines at θ angles from the vertical, Fig. 4.1 (e). The *Backtrack* function is used to place circular masks of radius r_{mask} at a distance, d, from the outer edge of the abdominal region. These masks are backtracked in r_{mask}-length intervals along a line, angled at *theta*, towards the centroid. Once 5 masks have been placed or the edge of I_{lungs} is reached, the *Backtrack* function switches to the next angle θ and begins the operation again. By this process, all possible $I_{mask}(\theta, d)$ are

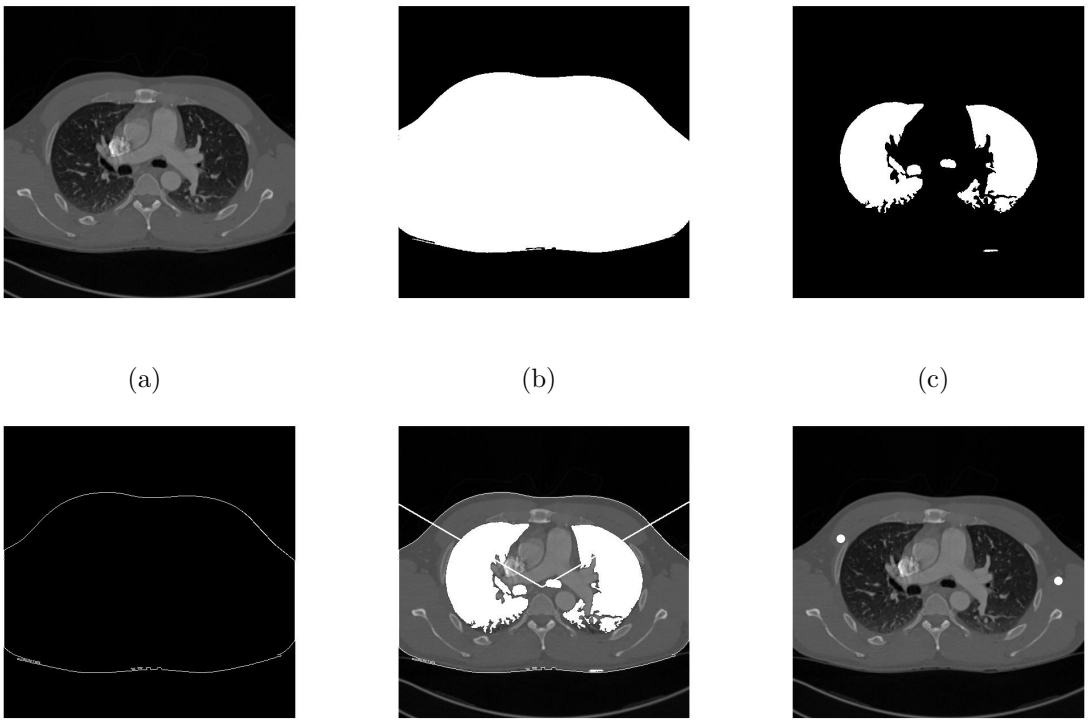
formed. The optimal mask is at the $\hat{\theta}$ and $\hat{d}$ that yields the minimum coefficient of variation, $\sigma^{\hat{\theta},\hat{d}}_{I_{mask}}/\mu^{\hat{\theta},\hat{d}}_{I_{mask}}$, Fig. 4.1 (f). The efficiency, ξ_I, is calculated for the optimal masks and the inverse is returned as the quality metric.

Algorithm 1 Baseline Model: Quality Estimation (I, r_{mask})

$[I_{out}, I_{lungs}] \leftarrow Highpass\_filter(I)$
$cntr \leftarrow Centroid(I_{out})$
for θ in $[0:3:90]$ **do**
  for d in $[Edge(I_{out}) : r_{mask} : Edge(I_{lungs})]$ **do**
    $I_{mask}(\theta, d) \leftarrow Backtrack(I, r_{mask}, \theta, d, cntr)$
  end for
end for
$[\hat{\theta}, \hat{d}] \leftarrow \arg_{(\theta,d)} \min(\sigma^{\theta,d}_{I_{mask}}/\mu^{\theta,d}_{I_{mask}})$
$\xi_I \leftarrow (\sigma^{\hat{\theta},\hat{d}}_{I_{mask}}/\mu^{\hat{\theta},\hat{d}}_{I_{mask}})^2$
return $1/\xi_I$

WFDM Model

The WFDM model follows the steps outlined in Algorithm 2. A *Highpass_filter* function is applied to the CT image I, Fig. 4.2 (a), extracting the overall body region, I_{out}, and the inner lungs, I_{lungs}. Subtracting I_{lungs} from I_{out} produces a binary mask of the abdominal tissue region, I_{reg}. For each pixel at (i, j) included in the mask I_{reg}, the *Sample* function is used to sample two windows of size w_s and w_l around pixel (i, j) of the original image I. This results in two new images, $I^{w_s}_{i,j}$ and $I^{w_l}_{i,j}$, respectively. Using the available 2-D Fast Fourier Transform function, *FFT2*, these windowed samples are transformed to their Fourier-domain spectra, $F^{w_s}_{i,j}$ and $F^{w_l}_{i,j}$, respectively. The *CFSS_distance* function is used to find the Euclidean distance between the CFSS of the two spectra, which is then stored as the corresponding pixel in the distance image, I_D, shown in Fig. 4.2 (b). The mean pixel intensity of I_D is



(a) (b) (c)

(d) (e) (f)

Figure 4.1: Baseline ROI selection process. (a) Patient CT image, I. (b) I_{out}, (c) I_{lungs}, and (d) outer edge of body. (e) Lines measuring θ angles from the vertical and passing through the centroid, $cntr$, used to backtrack I_{mask} from the outer edge to the lungs. (f) The I_{mask} that minimizes the coefficient of variation in those ROI.

found then found, resulting in μ_{I_D}. Since the optimal threshold, q, is unknown, a $[k = 1 : k']$ range of thresholds are calculated as $^1/_k$ factors of μ_{I_D}. The *Threshold* function finds all pixels in I_D at or below the threshold q, forming the mask I^q. For this study, k' was set to 7 so that the smallest threshold would be $q = \frac{\mu_{I_D}}{7}$. Below that threshold, the number of pixels in I^q for some of the images were too few. A selection of the masks for those ROI at $q = [\mu, \frac{\mu}{3}, \frac{\mu}{5}, and \frac{\mu}{7}]$ are shown in Fig. 4.2. Once I^q is formed, the variance is calculated in a w_s-sized region around each pixel (i, j) in I^q, resulting in $V_{i,j}^{w_s}$. These variances are averaged for all i, j, resulting in η_I^q. The CT IQ metric returned is the inverse of η_I^q.

Algorithm 2 WFDM Model: Quality Estimation (I)

$[I_{out}, I_{lungs}] \leftarrow Highpass\_filter(I)$

$I_{reg} \leftarrow I_{out} - I_{lungs}$

for (i, j) in $I_{reg}(i, j) > 0$ **do**

$[I - i, j^{w_l}, I_{i,j}^{w_s}] \leftarrow Sample(I(i, j), w_l, w_s)$

$[F_{i,j}^{w_l}, F_{i,j}^{w_s}] \leftarrow FFT2(I_{i,j}^{w_l}, I_{i,j}^{w_s})$

$I_D(i, j) \leftarrow CFSS\_distance(F_{i,j}^{w_l}, F_{i,j}^{w_s})$

end for

for $k = 1$ to k' **do**

$q \leftarrow \frac{\mu_{I_D}}{k}$

$I^q \leftarrow Threshold(I_D, q)$

for (i, j) in $I^q(i, j) > 0$ **do**

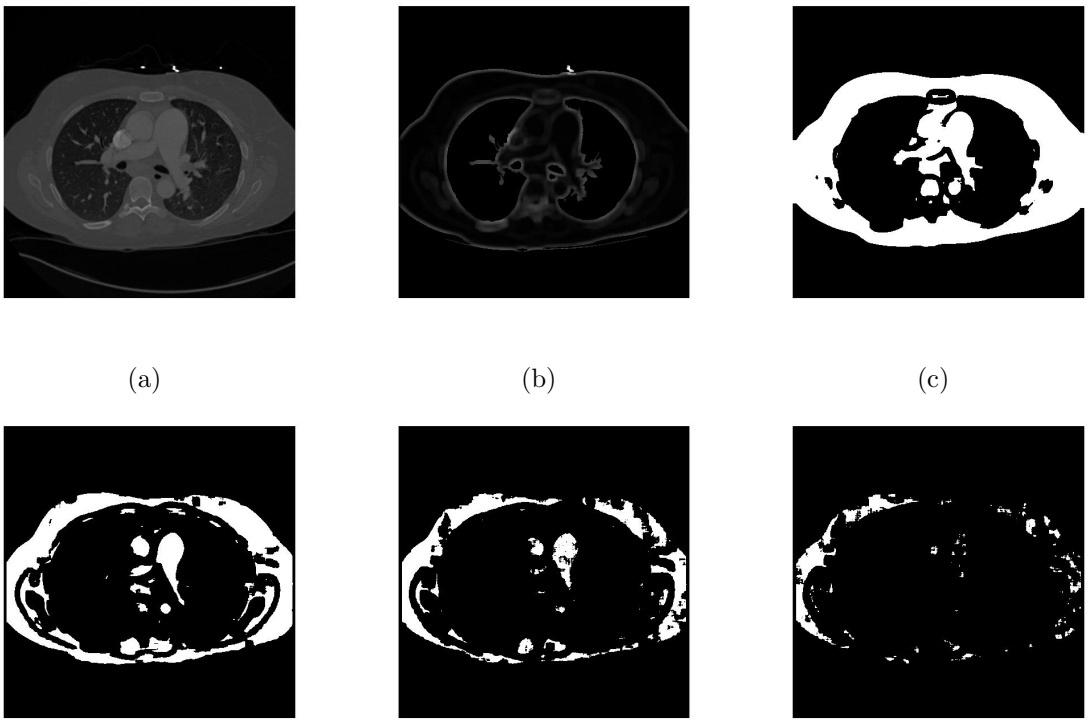
$V_{i,j}^{w_s} \leftarrow \sigma^2_{I_{i,j}^{w_s}}$

end for

$\eta_I^q \leftarrow \mu_{V^{w_s}}$

return $\frac{1}{\eta_I^q}$

end for



(a) (b) (c)

(d) (e) (f)

Figure 4.2: WFDM ROI selection process. (a) Patient CT image, I. (b) I_D, (c) Ψ_μ, (d) $\Psi_{\frac{\mu}{3}}$, (e) $\Psi_{\frac{\mu}{5}}$, and (f) $\Psi_{\frac{\mu}{7}}$.

From Fig. 4.2, it is clear that the q threshold value controls the number of pixels chosen as part of the ROI, as well as its sensitivity to regions of higher variation.

4.2.3 Experiments

To perform the experiments, both phantom and patient datasets are processed by the baseline and WFDM models to produce IQ metrics for each CT image. Using these calculated values and the $CTDI_{vol}$ of each image, the following experiments are performed.

1. Direct Image Quality Estimation

For each model, the IQ metric of each phantom and patient image is plotted against $CTDI_{vol}$, and a best fit line is found. The correlation coefficient provides a quantifiable measure of the observed trends, and serves as the basis of comparison between the results of the two models and datasets.

2. Low by High Classification

For the relative IQ estimation, the IQ metrics are used to classify pairs of CT images as belonging to the high or low dose scan, as measured by the $CTDI_{vol}$ of each image. This experiment serves as a qualitative test of the proposed metrics in distinguishing between two images of different quality, as well as the effectiveness of $CTDI_{vol}$ as a testing parameter. The latter is possible due to the differences in the high and low dose scans of each patient. For some patients, the difference in $CTDI_{vol}$ between the pair of scans is significantly greater than for others. Thus, the classification error in this experiment is evaluated as a function of the change in $CTDI_{vol}$, revealing the extent to which such a change is detectable by the IQ metrics.

In this experiment, 5 high and low dose pairs of CT images from each patient are processed using the baseline and WFDM models, and classified according to the resulting IQ metric. The image with the higher $CTDI_{vol}$ value has a higher effective x-ray dose, corresponding to a higher quality image, ceteris paribus. Both the baseline and WFDM models produce metrics based on noise estimates in their respective regions of interest. Since noise is inversely proportional to IQ, the metrics produced by both models should be inversely proportional to the $CTDI_{vol}$. Thus, for each pair of CT images, the one with the higher $CTDI_{vol}$ should correspond to the one with the lower IQ metric.

The experiment can be divided into four steps. First, 5 slices are used from each patients scan. The bifurcation of the pulmonary trunk is selected as a central slice, from which slices 3 cm and 6 cm before and after are selected. Patients with implants visible in these slices are excluded, as the implants produce additional imaging artifacts.

Each of these slices is then processed by the baseline and WFDM models to produce the corresponding IQ metric. Next, the metric from pairs of slices corresponding to the same patient are used to classify each image in a pair as high or low dose, by setting the image with the lower IQ metric as high dose and vice versa. Finally, this classification is compared with the actual $CTDI_{vol}$ of each image, and the classification error is measured as a function of change in $CTDI_{vol}$. A variable, representing the minimum change in $CTDI_{vol}$, is varied from 0% to 50% in 2.5% intervals. At each interval, pairs of images with differences in $CTDI_{vol}$ below this minimum are excluded from the count. Thus, the total count of image pairs decreases as the minimum threshold is increased, and the error count at each interval represents the number of mis-classifications among the image pairs that meet the threshold.

4.2.4 Change in Metric vs Change in $CTDI_{vol}$

Additionally, the percent change in CT IQ metrics for each pair of images is compared to the percent change in $CTDI_{vol}$ for those images. This serves as a quantitative test of the CT IQ metrics, as the amount of radiation exposure should correspond to IQ. However, a correlation between the magnitude change in an IQ metric and the magnitude change in $CTDI_{vol}$ depends on more than the effectiveness of the proposed metric. Other factors that affect IQ, such as differences in body types, size, and contrast enhancement, to name a few, can significantly impact the quality of a CT image or the resulting IQ metric, and some of these factors are unaccounted for in this study.

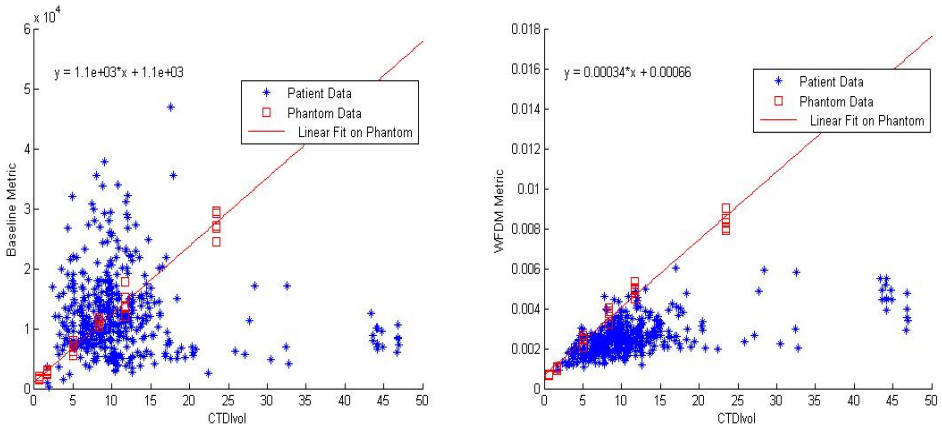
The difference in $CTDI_{vol}$ between the high and low dose scans varies across the CT image pairs, as does the difference in the IQ metric. In the previous experiment, only the relative value of the metric and $CTDI_{vol}$ values for each image pair is considered, defined as being either high or low. In this experiment, the correlation between the IQ metrics and $CTDI_{vol}$ is further investigated by comparing the magnitude difference of both values for each image pair. This is achieved by calculating the correlation coefficient of the percent difference in the metrics and the percent difference in $CTDI_{vol}$ for each image pair. Given

the direct relationship between $CTDI_{vol}$ and IQ, the IQ metrics are expected to be directly proportional to $CTDI_{vol}$.

4.3 Experiments and Results

4.3.1 Direct Image Quality Estimation

The baseline and WFDM models are applied to phantom and patient datasets to evaluate their effectiveness in estimating IQ. $CTDI_{vol}$ is used to evaluate the performance of the two models as it represents the x-ray dose applied to the body, and is thus a significant factor in the quality of a CT image. Specifically, increasing the x-ray dose reduces the noise in the image, corresponding to higher IQ at higher $CTDI_{vol}$. The baseline and WFDM models estimate noise and generate a metric for IQ, which should therefore be proportional to $CTDI_{vol}$. This is observed when applying the baseline and WFDM models to both phantom and patient datasets, as shown in Fig. 4.3.



(a) (b)

Figure 4.3: Correlation between the CT IQ metric from the (a) baseline and (b) WFDM models and the $CTDI_{vol}$ of 30 Phantom and 430 Patient images.

Both the baseline and WFDM models behave similarly when applied to the phantom

images, showing a strong linear correlation with R^2 of 0.9818 and 0.9866, respectively. However, the models behave differently when applied to the patient data, as observed by the wider scattering of points in Fig. 4.3 (a) than in Fig. 4.3 (b).

The decrease in performance from the phantom dataset to the patient dataset is likely due to other factors affecting the resulting IQ at a given $CTDI_{vol}$. A larger patient would require greater $CTDI_{vol}$ to produce an image of similar quality to a smaller patient. Thus, $CTDI_{vol}$ is a poor reference by which to evaluate the absolute performance of proposed CT IQ metrics. Additional differences in body features from patient to patient, the field of view or image cropping, the amount of contrast enhancement and the time since it was administered, and so on all have an effect on the resulting IQ. As shown in Fig. 4.4 (a) and (b), the phantom images taken at different axial locations differ in overall shape and size. However, the lack of complexity in the densities of the material ("tissue") inside the phantom results in comparable estimates of IQ from both images. On the other hand, Fig. 4.4 (c) and (d) show that images of the same patient can vary not just in overall shape and size, but also in the ratios of tissue densities and locations within. These differences significantly affect the location of ROI selected in each image and the resulting IQ estimate. To better control for these factors, pairs of images from the same location in the same patient, acquired at high and low doses, are compared instead. This is shown in Fig. 4.5.

Comparing identical slices in the phantom, at different $CTDI_{vol}$ values, is possible, as shown in Fig. 4.5 (a) and (b), since the size and structure of the "abdominal" region of the phantom remains identical. Thus, the only difference between the two images is the amount of noise, which allows for an accurate evaluation of their relative IQ. In contrast, the patient images shown in Fig. 4.5 (c) and (d) vary slightly in size, shape, angle, and contrast, all of which impact the resulting IQ metric. However, this variability is much less than that of Fig. 4.4 (c) and (d), suggesting that a pair-wise comparison of high and low dose images, from the same axial location in each patient, should be more effective.

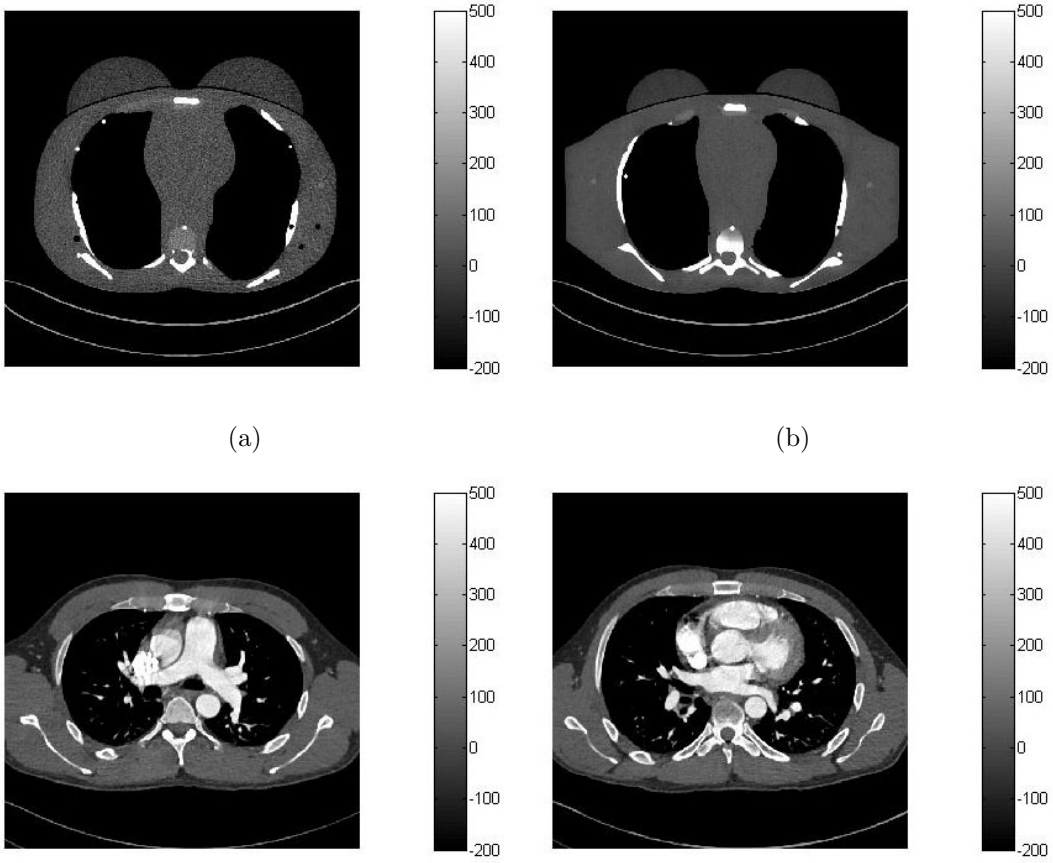


Figure 4.4: A comparison of variability in phantom and patient images from different locations along the z-axis. (a) A phantom image at one axial location and (b) a phantom image from a different axial location. Similarly, (c) a patient image at one axial location and (d) another of the same patient at a different axial location.

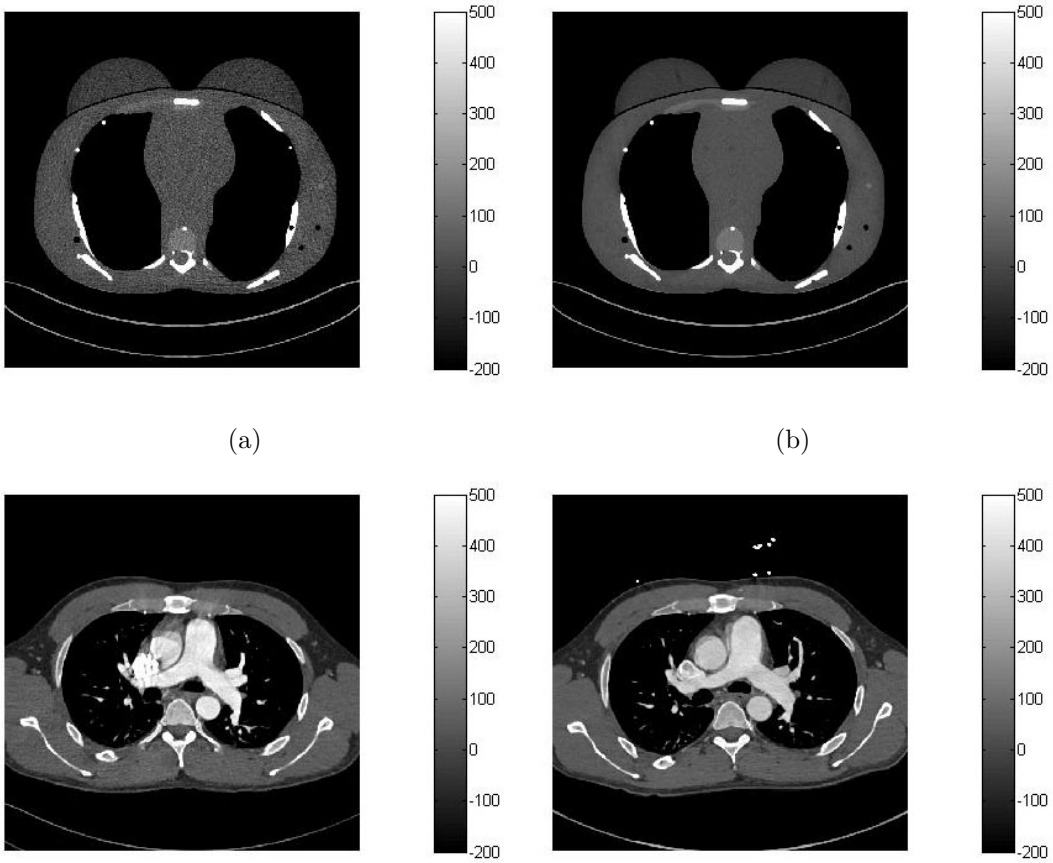


Figure 4.5: A comparison of phantom and patient images acquired from the same axial location at different $CTDI_{vol}$. (a) A low dose phantom image and (b) its corresponding high dose image. Similarly, (c) a low dose patient image and (d) its corresponding high dose image.

CT images of patients with implants are excluded from the study, as the implants cause streaking artifacts in the image, as show in Fig. 4.6. Severe clipping of images, such as in Fig. 4.6 (a), also adversely affects the accuracy of the IQ assessment. However, most of the images exhibit clipping to various extents, so only the extreme cases are excluded.

(a) (b) (c)

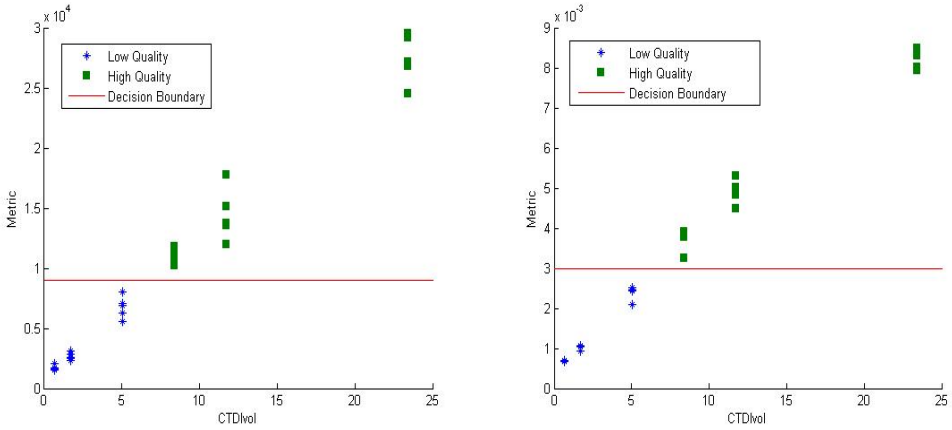
Figure 4.6: A comparison of patient images with implants (a) in the back, (b) in the chest, and (c) an object on the chest.

4.3.2 Low by High Binary Classification

CT IQ is directly correlated to radiation exposure during imaging, with high dose images resulting in higher IQ than low dose ones, as demonstrated by the linear trend of the phantom data in Fig. 4.3 of the previous section. Thus, two images acquired at different $CTDI_{vol}$ can be ranked by comparatively analyzing their respective IQ metrics. In the case of the phantom dataset, this is demonstrated by splitting the data into low and high dose halves, shown in Fig. 4.7. The horizontal line separating the two sets of data in both the baseline and WFDM model demonstrates the clear separation in IQ of phantom images acquired at different levels of exposure.

A similar test is performed with the patient data, as it consists of multiple scans of individual patients at different values of $CTDI_{vol}$. Some patients were imaged with the higher dose first and others with the lower dose first, so that image pairs have either an increasing or decreasing trend in $CTDI_{vol}$. These pairs of images are analyzed by both the baseline and WFDM models to produce increasing or decreasing trends in their IQ metrics. These trends are compared to the trends in $CTDI_{vol}$ and the percent error is calculated.

The magnitude change in $CTDI_{vol}$ varies across the image pairs, and thus the necessary



(a) (b)

Figure 4.7: Low by High classification of the phantom dataset for the (a) baseline and (b) WFDM models. The horizontal line in each plot represents a metric threshold separating the low and high dose cases.

sensitivity of the IQ metric to changes in $CTDI_{vol}$ for accurate prediction is examined. Sweeping through a range of cut-off $CTDI_{vol}$ values, below which image pairs are ignored, produces a curve matching the percent error in classification to minimum $CTDI_{vol}$ difference for each model. This Trend Error Curve (TEC) for the high and low dose binary classification, using metrics from the baseline and WFDM models is shown in Fig. 4.8.

Both the baseline and WFDM model show improving classification performance with increasing change in $CTDI_{vol}$. This is expected, as larger differences in dosage should result in more apparent differences in IQ. The WFDM performs better than the baseline overall, though both result in more than 20% error when images with lower changes in $CTDI_{vol}$ are included. These cases of misclassification may be due to other factors impacting IQ more significantly as the difference in $CTDI_{vol}$ decreases.

Changes in the shape, size, and angle of the abdominal region in an image can affect the placement of the baseline ROI, as shown in Fig. 4.9. For patient 17, an image acquired at low $CTDI_{vol}$, Fig. 4.9 (a) is compared to an image acquired at a high $CTDI_{vol}$, Fig. 4.9 (b).

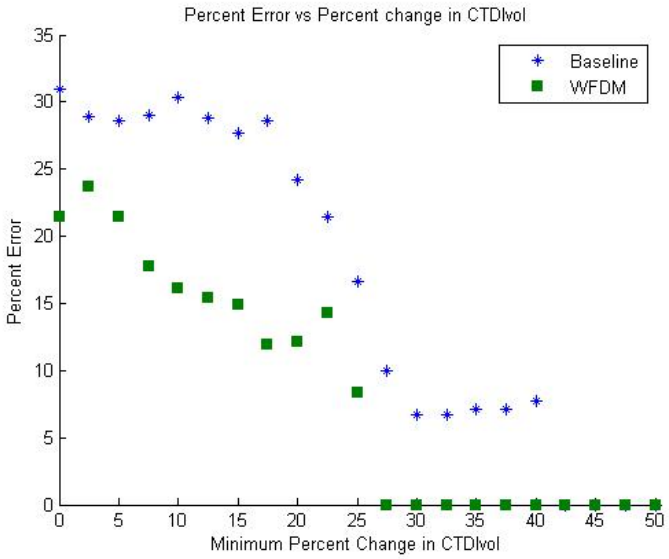


Figure 4.8: Trend Error Curve for the baseline and WFDM models.

The red circles in the images represent the ROI selected by the baseline model. There are slight differences in the shape and size of the abdominal regions between the two images, but these differences have little effect on the baseline ROI placement, as both images have ROI placed in similar locations. This results in baseline IQ metrics that trend similarly to the $CTDI_{vol}$ of the images, with the lower $CTDI_{vol}$ image having a lower baseline IQ metric than the higher $CTDI_{vol}$ image. On the other hand, patient 20 has more significant differences in shape and size of the abdominal region between the low $CTDI_{vol}$ and high $CTDI_{vol}$ images, shown in Fig. 4.9 (c) and (d), respectively. Thus, the baseline ROI are placed in significantly different locations, and the resulting baseline IQ metrics trend inversely to the $CTDI_{vol}$ of the images. This is an example of a trend error in the low by high classification analysis.

A similar analysis is performed for the WFDM model, shown in Fig. 4.10. In this case, the red regions represent the WFDM ROI. For patient 17, shown in Fig. 4.10 (a) and (b) the WFDM IQ metrics measured using these ROI trend similarly to the $CTDI_{vol}$ of the images, just as they did for the baseline model. However, for patient 20, Fig. 4.10 (c) and (d), the

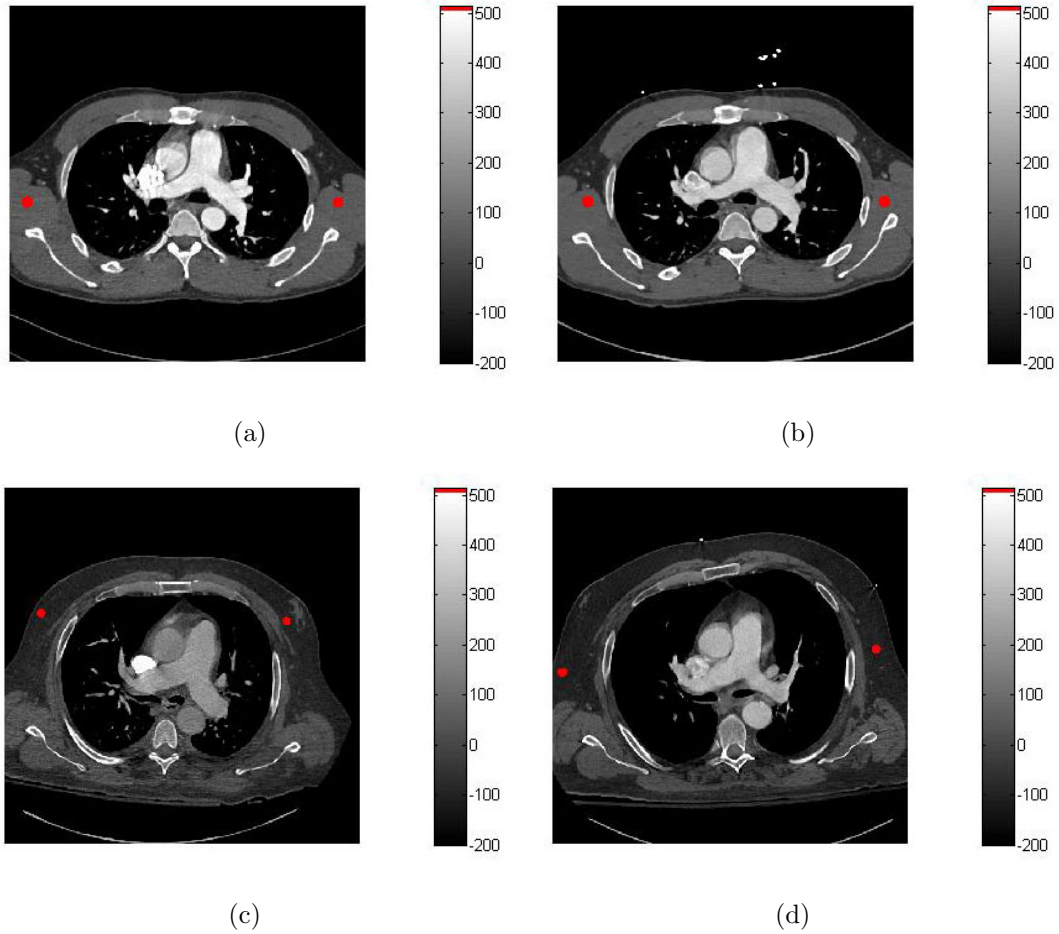


Figure 4.9: Baseline model comparison of image pairs acquired at two different $CTDI_{vol}$. Patient 17 has (a) a low $CTDI_{vol}$ image with a lower baseline IQ metric than (b) the corresponding high $CTDI_{vol}$ image. In contrast, patient 20 has (c) a low $CTDI_{vol}$ image with a higher baseline IQ metric than (d) the corresponding high $CTDI_{vol}$ image. The red circles are the baseline ROI regions used to find each IQ metric.

WFDM IQ metrics also trend similarly to the $CTDI_{vol}$ of the images, unlike the baseline model. Despite some of the differences in shape and size between the two images in Fig. 4.10 (c) and (d), the regions selected by the WFDM ROI are almost identical. This may explain the difference in performance between the WFDM and baseline models.

To better assess the response of each model to changes in $CTDI_{vol}$, the percent change in each metric is compared to the percent change in $CTDI_{vol}$.

4.3.3 Change in Metric vs Change in CTDIvol

Using the same CT image pairs as those used in the binary classification, the percent change in metric for each image pair is plotted against the percent change in $CTDI_{vol}$, shown in Fig. 4.11. The baseline model, shown in Fig. 4.11 (a), has a correlation coefficient of 0.2736, while the WFDM model, shown in Fig. 4.11 (b), has a correlation coefficient of 0.3710.

Although a best fit line does show the expected correlation between the IQ metric and $CTDI_{vol}$, this fit remains poor. The spread in the data appears to be larger at lower changes in $CTDI_{vol}$, matching the increase in classification error observed in the first experiment. However, excluding those data points does not improve the correlation.

Both plots show a very weak relationship between the magnitude difference in $CTDI_{vol}$ and the magnitude difference in each IQ metric, as indicated by R^2 values of 0.1651 and 0.1497 for the baseline and WFDM models, respectively. Although the binary classification showed an improvement at higher percent changes in $CTDI_{vol}$, those changes do not directly correspond to similar magnitude changes in the IQ metrics. For comparison, the same analysis applied to the phantom images is shown in Fig. 4.12

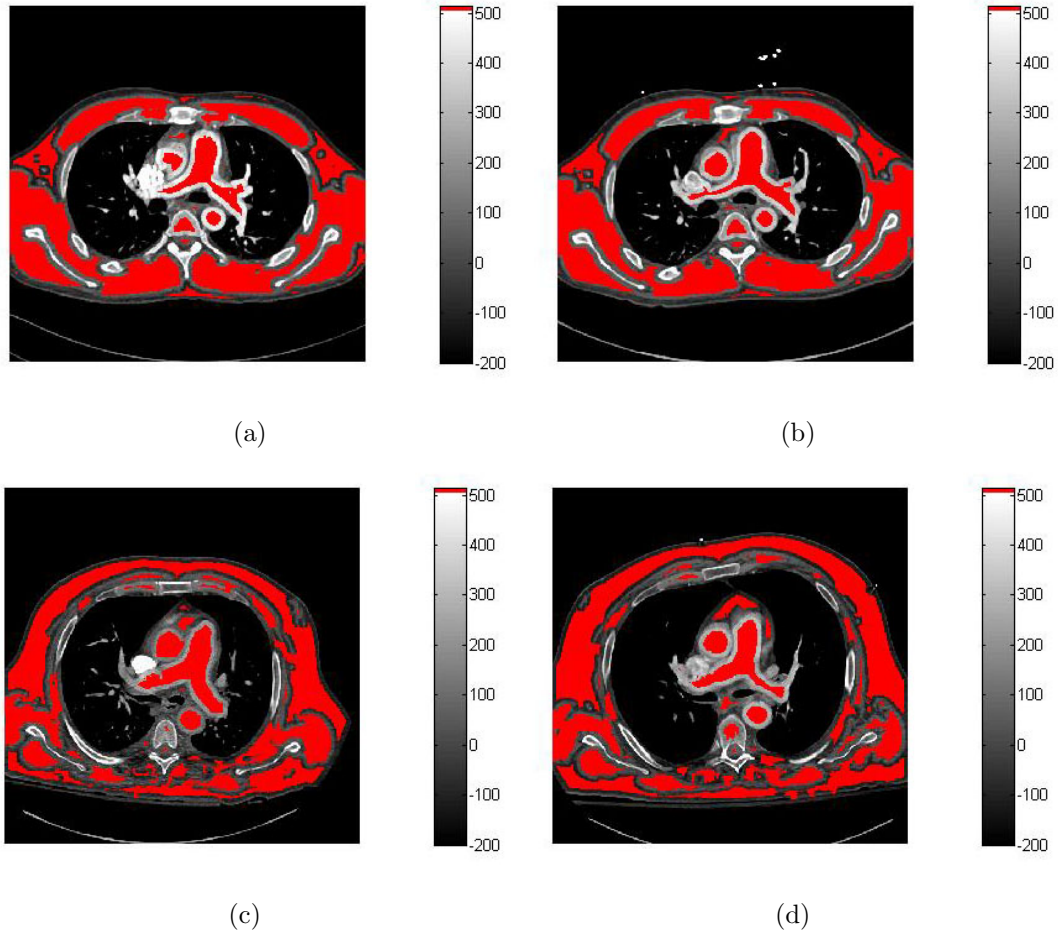
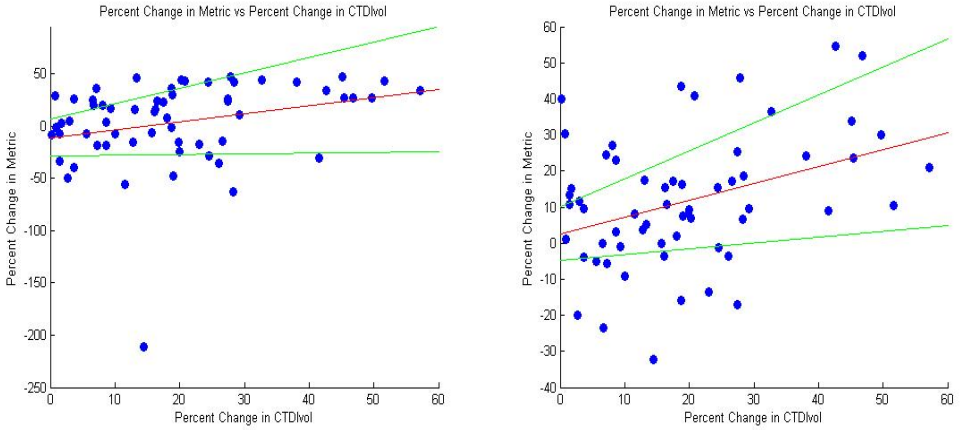
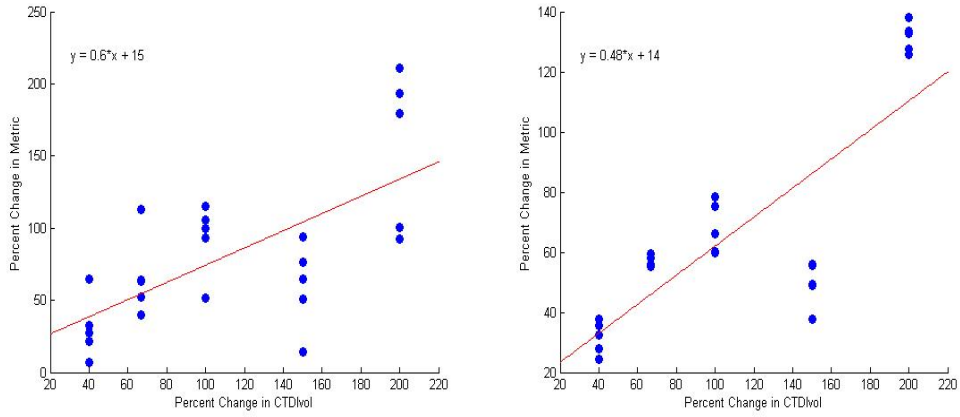


Figure 4.10: WFDM model comparison of image pairs acquired at two different $CTDI_{vol}$. Patient 17 has (a) a low $CTDI_{vol}$ image with a lower WFDM IQ metric than (b) the corresponding high $CTDI_{vol}$ image. Similarly, patient 20 has (c) a low $CTDI_{vol}$ image with a lower WFDM IQ metric than (d) the corresponding high $CTDI_{vol}$ image. The red regions are the WFDM ROI regions used to find each IQ metric.



(a) (b)

Figure 4.11: Percent change in metric vs percent change in $CTDI_{vol}$ for the (a) baseline and (b) WFDM models applied to patients with less than 20% change in size across their two scans. The red line in each plot is a linear fit and the green lines are the 95% confidence intervals.



(a) (b)

Figure 4.12: Percent change in metric vs percent change in $CTDI_{vol}$ for the (a) baseline and (b) WFDM models applied to phantoms.

The results in Fig. 4.12 are better correlated than for the patient dataset, with R^2 values of 0.4385 and 0.6489 for the baseline and WFDM models, respectively. Both models exhibit a similar deviation from the linear trend, suggesting this relationship may not necessarily be linear.

4.3.4 Optimal WFDM Threshold

The WFDM model produces the *distance image* I_D corresponding to a given CT image. The pixel values of the distance image represent the degree of spatial variation surrounding the corresponding pixels in the original image. Thus, thresholding the distance image by selecting only the pixels with intensities below a threshold value, q, produces a mask, I^q, of regions in the original image with spatial variation below the specified threshold. Ideally, regions below a certain threshold would prove optimal for image noise estimation. However, the selection of the optimal threshold, q, poses a challenge, especially for automated processes.

Here, the choice of threshold is experimentally analyzed by comparing its effect in the high vs low dose classification experiment. Specifically, the average percent error and the rate of change in the error of the TEC resulting from each threshold $q = [\mu, \frac{\mu}{2}, \frac{\mu}{3}, \frac{\mu}{4}, \frac{\mu}{5}, \frac{\mu}{6}, and \frac{\mu}{7}]$ is analyzed across the range of minimum $CTDI_{vol}$ values. The expectation is that the optimal threshold will correspond to the TEC with the steepest decline in error as a function of $CTDI_{vol}$ and the lowest average percent error, as shown in the following equations. First, the optimal threshold, q^*, is found by finding the q that results in a TEC with the steepest slope, as shown in equation 4.1. This is then compared to the threshold $\hat{q}$ corresponding to the lowest average percent error across all values of $CTDI_{vol}$.

$$q^* = \arg_q \max |\tfrac{\partial TEC(q)}{\partial CTDIvol}|. \tag{4.1}$$

$$\hat{q} = \arg_q \min(\mu_{TEC(q)}). \tag{4.2}$$

$$q^* = \hat{q}.$$

The two values, q^* and $\hat{q}$, are expected to be equal, indicating that the optimal threshold

selects an ROI that is the most responsive to changes in $CTDI_{vol}$ and the most accurate in classification. The plot for the average percent error at different values of q is shown in Fig. 4.13.

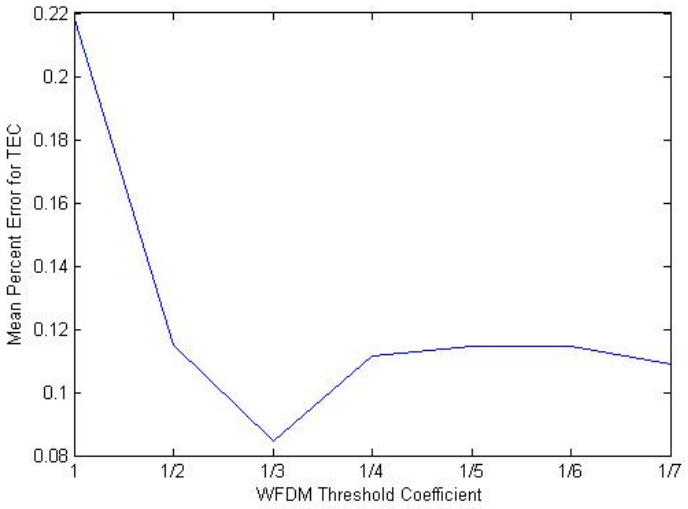


Figure 4.13: Average percent error across all values of $CTDI_{vol}$ at different thresholds $q = [\mu, \frac{\mu}{2}, \frac{\mu}{3}, \frac{\mu}{4}, \frac{\mu}{5}, \frac{\mu}{6}, and \frac{\mu}{7}]$

The optimal threshold based on average percent error, as shown by the minimum point in the curve, occurs at $q = \frac{\mu}{3}$. Comparing this result with the slopes of the TEC, shown in Fig. 4.14, suggests that the steepest decline does result from $q = \frac{\mu}{3}$.

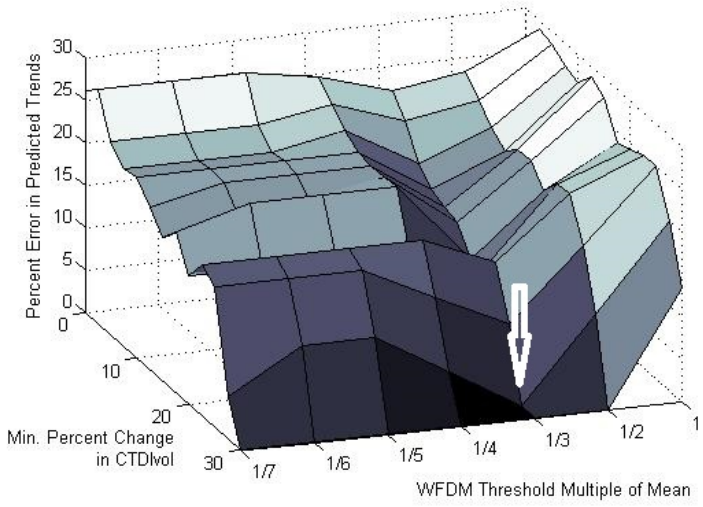


Figure 4.14: Slopes of the TEC at WFDM thresholds $q = [\mu, \frac{\mu}{2}, \frac{\mu}{3}, \frac{\mu}{4}, \frac{\mu}{5}, \frac{\mu}{6}, and \frac{\mu}{7}]$

One feature of the TEC is the impact of $CTDI_{vol}$ on classification error. Specifically, image pairs with $CTDI_{vol}$ differences below a given minimum are excluded, and the impact on classification error is observed. As image pairs are excluded, both errors and accurate classifications are lost. Thus, it is desirable to have a metric that is most sensitive to changing $CTDI_{vol}$. This analysis suggests that the threshold most sensitive to changing $CTDI_{vol}$ is also the threshold minimizing the percent error.

Chapter 5

CONCLUSIONS AND FUTURE WORK

5.1 Conclusions

This thesis presents a novel approach to improving CT image acquisition by providing metrics of CT IQ assessment. The appropriate CT IQ metric would allow a technician to use a patient's CT imaging history to guide future protocol selection and optimal acquisition parameters.

To achieve this goal, the WFDM model is presented in chapter 2. A comparative analysis between a baseline, WFDM, and CNN models is performed in chapter 3. Finally, the baseline and WFDM models are applied to large patient datasets in chapter 4, identifying exclusion and WFDM parameters for optimal CT IQ assessment.

1. Chapter 2: The Proposed WFDM Metric.

   The proposed metric involves automated segmentation of abdominal CT images by way of thresholding regions of low pixel variations and measuring metrics such as variance (var), standard deviation (std), coefficient of variation (cov), and relative size of ROI ($\nu'(q)$) that are found to be directly proportional to tube current.

   For the phantom CT images, a correlation between both var and std to CT image noise ($r > 0.76$ ($p \ll 0.001$)) and $\nu'(q)$ to CT image noise ($r = 0.96$ ($p \ll 0.001$)) is observed.

   However, after combining the phantom and patient CT images, var and std do not correlate well to CT image noise($r < 0.46$ ($p \ll 0.001$)) as compared to $\nu'(q)$ ($r = 0.95$ ($p \ll 0.001$)). This suggests that $\nu'(q)$ is a more generalizable metric for CT image noise detection than var and std.

2. Chapter 3: Comparative Assessment of Spatial Segmentation Methods.

   The WFDM model is also comparatively analyzed with the baseline and CNN models for inferencing CT image quality. The goal is predictive modeling of CT IQ across variabilities introduced by image acquisition parameters and imaged subjects. This would enable patients to be imaged with low contrast enhancement dosages, yet diagnostic quality images. The proposed analysis is directed towards capturing spatial and functional variabilities by analyzing phantom CT and patient CT images for *cross-sectional* and *longitudinal* analyses. The "Big" image-level variabilities are captured using the proposed analytics that can be fine tuned for personalized medicine and generalized CT imaging protocol guidance, respectively.

   The data models under analysis include computationally simple fixed-size and variable size ROI estimation strategies using a baseline and Fourier-domain based model, respectively. The baseline and WFDM-based models implemented using Matlab on a 3.10 GHz Intel core i5, 4GB RAM, PC require 5.77 and 9.45 seconds per image, respectively, for CT IQ estimation. However, the third data model is CNN model implemented on a cloud-based platform of Microsoft Azure. Although the CNN model is the most computationally intensive, it best detects subtle variations in image granularity, that impact CT IQ, for phantom images. The CNN models implemented on the MAMLS platform required 22-52 minutes of cloud computing time for phantom and patient image data sets for model training and testing operations. Further analysis of the CNN model feature maps demonstrates detection of significant edges as features for CT IQ classification.

   For the phantom CT image multi-class classification task, the baseline model has a mean accuracy $\overline{\alpha}_B = 73.6\%$, the WFDM has a mean accuracy $\overline{\alpha}_{WFDM} \leq 100\%$, and the CNN has a mean accuracy $\overline{\alpha}_{CNN} = 93.8\%$. For the patient CT image multi-class classification, the baseline model has a mean accuracy $\overline{\alpha}_B \leq 49.5\%$ for the *cross-sectional* analysis and $\overline{\alpha}_B \leq 45.1\%$ for the *longitudinal* analysis. The WFDM has

a mean accuracy $\overline{\alpha}_{WFDM} \leq 65.2\%$ for the *cross-sectional* analysis and $\overline{\alpha}_{WFDM} \leq 66.1\%$ for the *longitudinal* analysis. We observe that for the given set of CT images, the CNN model over-fits the data, resulting in an accuracy of 100% for the *cross-sectional* analysis and 1.9% for the *longitudinal* study. For the *cross-sectional* analysis of the baseline model, the *cov* classification parameter outperforms *var*. However, *var* outperforms *cov* for binary classification tasks in the baseline model's *longitudinal* analysis. The WFDM model outperforms the baseline model and is more consistent than the CNN model across all datasets, though the CNN model outperforms other data models in images with limited structural variabilities.

3. Chapter 3: Complete Automated System for CT Image Quality Detection

   In this chapter, two models for an automated system of CT IQ assessment have been presented, and their performance with respect to varying datasets and dosage conditions have been analyzed. As a direct measure of CT IQ, both the baseline and WFDM models show the expected linear relationship. However, the variability in factors such as patient size result in a non-linear relationship between IQ and $CTDI_{vol}$, underscoring the limitations of using it to assess CT IQ. Additional factors include image cropping and imaging artifacts present in some of the patient images, as well as a lack of information on the degree of contrast enhancement at the time of imaging.

   To avoid some of these issues, the relative IQ of high and low dose image pairs have been examined as a function of the magnitude change in $CTDI_{vol}$. While the phantom dataset separates effectively, as suggested by the strong linear relationship observed in Fig. 4.3, the patient dataset begins with at least 20% error. To address this, the minimum acceptable change in $CTDI_{vol}$ is increased. This is observed to significantly lower the classification error, until both models result in 0% error. In this case as well, the WFDM model outperformed the baseline model significantly. However, a comparative analysis of the change in metrics vs change in $CTDI_{vol}$ shows a general non-linearity in the relationship.

The optimization of the WFDM model's q threshold completes the work started on this model, and presents a complete system for further analysis and adjustments. While the optimal value of the threshold can vary depending on the CT images analyzed, the new threshold can readily be calculated by the same process.

5.2 Future Work

Some possible avenues for future work include further analysis and optimization of the WFDM model and improving overall image quality estimation.

5.2.1 WFDM

There are several areas that can be investigated with the WFDM model, regarding its ROI selection process and the resulting metric. The current region segmentation is based on the Fourier spectra of two square windows, one half the size of the other. The effect of the relative size of each window to the other can be investigated, as well as the shape of the windows used for subsampling. Additionally, the Fourier-domain was selected due to its well established history in signal processing, but other transforms such as the Discrete Cosine Transform (DCT) and the Wavelet Transform could be evaluated. Finally, metrics such as ν', as well as the relationship between image quality and optimal threshold q, suggest the ROI formed by the WFDM model could yield alternative metrics to image quality than a direct estimation of noise.

5.2.2 Image Quality Estimation

Image quality estimation can be enhanced by identifying other features to extract. The current work focused on extracting uniform regions for noise estimation, but other features may prove more robust. Using the WFDM model, it is possible to also select regions of high spatial variations, which often correspond to regions of high spatial variation. As an image becomes noisy, the distribution of regions of high and low spatial variation becomes

less distinct, so that the regions blend together. Hybrid metrics such as these may improve the accuracy of noise and IQ estimations.

In addition, several factors were observed to affect the models' performance but could not be accounted for in this work, such as the contrast enhancement dosage of each patient at the moment of imaging, the location of the scanner head during the scan, and the changing size and attributes of the patients being imaged. Assessing the impact of these factors is necessary to improve the scalability of the proposed image quality models.

Finally, access to additional data, including cadaver studies that allow for varying CT scan parameters akin to the phantom study, will address some of the uncontrolled variability in the study. Increasing the available data with good quality measurements can help shed light on the specific obstacles that remain. Additionally, access to larger samples of annotated data will allow for better evaluation of the image quality metrics, by allowing a thorough comparison between the automated system and the subjective scores provided by the radiologists.

BIBLIOGRAPHY

[1] CT Scan. https://en.wikipedia.org/wiki/CT_scan/.

[2] David B Larson, Lara W Johnson, Beverly M Schnell, Shelia R Salisbury, and Howard P Forman. National trends in ct use in the emergency department: 1995–2007 1. *Radiology*, 258(1):164–173, 2011.

[3] Rebecca Smith-Bindman, Diana L Miglioretti, and Eric B Larson. Rising use of diagnostic medical imaging in a large integrated health system. *Health Affairs*, 27(6):1491–1502, 2008.

[4] Rebecca Smith-Bindman, Diana L Miglioretti, Eric Johnson, Choonsik Lee, Heather Spencer Feigelson, Michael Flynn, Robert T Greenlee, Randell L Kruger, Mark C Hornbrook, Douglas Roblin, et al. Use of diagnostic imaging studies and associated radiation exposure for patients enrolled in large integrated health care systems, 1996-2010. *Jama*, 307(22):2400–2409, 2012.

[5] Rongping Zeng, Nicholas Petrick, Marios A Gavrielides, and Kyle J Myers. Approximations of noise covariance in multi-slice helical ct scans: impact on lung nodule size estimation. *Physics in medicine and biology*, 56(19):6223, 2011.

[6] John M Boone, William R Hendee, Michael F McNitt-Gray, and Steven E Seltzer. Radiation exposure from ct scans: how to close our knowledge gaps, monitor and safeguard exposureproceedings and recommendations of the radiation dose summit, sponsored by nibib, february 24–25, 2011. *Radiology*, 265(2):544–554, 2012.

[7] G McGibney and MR Smith. An unbiased signal-to-noise ratio measure for magnetic resonance images. *Medical physics*, 20(4):1077–1078, 1993.

[8] Santiago Aja-Fernández, Carlos Alberola-López, and Carl-Fredrik Westin. Noise and signal estimation in magnitude mri and rician distributed images: a lmmse approach. *Image Processing, IEEE Transactions on*, 17(8):1383–1398, 2008.

[9] José V Manjón, José Carbonell-Caballero, Juan J Lull, Gracián García-Martí, Luís Martí-Bonmatí, and Montserrat Robles. Mri denoising using non-local means. *Medical image analysis*, 12(4):514–523, 2008.

[10] Santiago Aja-Fernandez, Gonzalo Vegas-Sanchez-Ferrero, Rodrigo de Luis-Garcia, and Antonio Tristan-Vega. Noise estimation in magnetic resonance sense reconstructed data. In *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*, pages 1104–1107. IEEE, 2013.

[11] Sohini Roychowdhury, Dara Koozekanani, and Keshab Parhi. Automated denoising and segmentation of optical coherence tomography images. In *Signals, Systems and Computers, 2013 Asilomar Conference on*, pages 258–262. IEEE, 2013.

[12] Sohini Roychowdhury, Nathan Hollcraft, and Adam M Alessio. Blind analysis of ct image noise using residual denoised images. In *Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), 2015 IEEE*, pages 1–4. IEEE, 2015.

[13] Maryam M Najafabadi, Flavio Villanustre, Taghi M Khoshgoftaar, Naeem Seliya, Randall Wald, and Edin Muharemagic. Deep learning applications and challenges in big data analytics. *Journal of Big Data*, 2(1):1, 2015.

[14] Shiying Hu, Eric A Hoffman, and Joseph M Reinhardt. Automatic lung segmentation for accurate quantitation of volumetric x-ray ct images. *IEEE transactions on medical imaging*, 20(6):490–498, 2001.

[15] S Gunasundari and S Baskar. Application of artificial neural network in identification of lung diseases. In *Nature & Biologically Inspired Computing, 2009. NaBIC 2009. World Congress on*, pages 1441–1444. IEEE, 2009.

[16] Yang Chen, Luyao Shi, Qianjing Feng, Jian Yang, Huazhong Shu, Limin Luo, Jean-Louis Coatrieux, and Wufan Chen. Artifact suppressed dictionary learning for low-dose ct image processing. *IEEE transactions on medical imaging*, 33(12):2271–2292, 2014.

[17] Maitham D Naeemi, Adam M Alessio, and Sohini Roychodhury. Automated selection of uniform regions for ct image quality detection. *arXiv preprint arXiv:1608.04381*, 2016.

[18] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[19] Johnny Ren. *Investigation of Convolutional Neural Network Architectures for Image-based Feature Learning and Classification*. PhD thesis, 2017.

[20] Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016.

[21] George C Kagadis, Christos Kloukinas, Kevin Moore, Jim Philbin, Panagiotis Papadimitroulas, Christos Alexakos, Paul G Nagy, Dimitris Visvikis, and William R Hendee. Cloud computing in medical imaging. *Medical physics*, 40(7):070901, 2013.

[22] S. Roychowdhury and M. Bihis. Ag-mic: Azure-based generalized flow for medical image classification. *IEEE Access*, PP(99):1–14, 2016.

[23] Yoshiharu Nakayama, Kazuo Awai, Yoshinori Funama, Masahiro Hatemura, Masanori Imuta, Takeshi Nakaura, Da Ryu, Shoji Morishita, Shamima Sultana, Natsuko Sato, and Yasuyuki Yamashita. Abdominal ct with low tube voltage: Preliminary observations about radiation dose, contrast enhancement, image quality, and noise. *Radiology*, 237(3):945–951, 2005.

[24] Z. S. Kelm, D. Blezek, B. Bartholmai, and B. J. Erickson. Optimizing non-local means for denoising low dose ct. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 662–665, June 2009.

[25] Diana L Miglioretti, Eric Johnson, Andrew Williams, Robert T Greenlee, Sheila Weinmann, Leif I Solberg, Heather Spencer Feigelson, Douglas Roblin, Michael J Flynn, Nicholas Vanneman, et al. The use of computed tomography in pediatrics and the associated radiation exposure and estimated cancer risk. *JAMA pediatrics*, 167(8):700–707, 2013.

[26] Rebecca Smith-Bindman. Is computed tomography safe. *N Engl j Med*, 363(1):1–4, 2010.

[27] Cynthia H McCollough, Shuai Leng, Lifeng Yu, Dianna D Cody, John M Boone, and Michael F McNitt-Gray. Ct dose index and patient dose: they are not the same thing. *Radiology*, 259(2):311–316, 2011.

[28] Rebecca Smith-Bindman, Jafi Lipson, Ralph Marcus, Kwang-Pyo Kim, Mahadevappa Mahesh, Robert Gould, Amy Berrington de González, and Diana L Miglioretti. Radiation dose associated with common computed tomography examinations and the associated lifetime attributable risk of cancer. *Archives of internal medicine*, 169(22):2078–2086, 2009.

[29] Y. Richard Wang, Mostapha Ziad, and Yang W. Lee. *Data Quality.* Number Vol. 23 in The Kluwer International Series on Advances in Database Systems. Springer, 2002.

[30] Shijun Wang and Ronald M Summers. Machine learning and radiology. *Medical image analysis*, 16(5):933–951, 2012.

[31] Cynthia H McCollough, Guang Hong Chen, Willi Kalender, Shuai Leng, Ehsan Samei, Katsuyuki Taguchi, Ge Wang, Lifeng Yu, and Roderic I Pettigrew. Achieving routine submillisievert ct scanning: report from the summit on management of radiation dose in ct. *Radiology*, 264(2):567–580, 2012.

Appendix A

WFDM ANALYSIS OF CHECKERBOARD IMAGE

The WFDM model is applied to a checkerboard image with added Gaussian noise to evaluate it's behavior. The checkerboard is divided into four quadrants, each with additive noise of $N(0, 0.01), N(0, 0.05), N(0, 0.1), and N(0, 0.5)$, respectively, as shown in Fig. A.1.

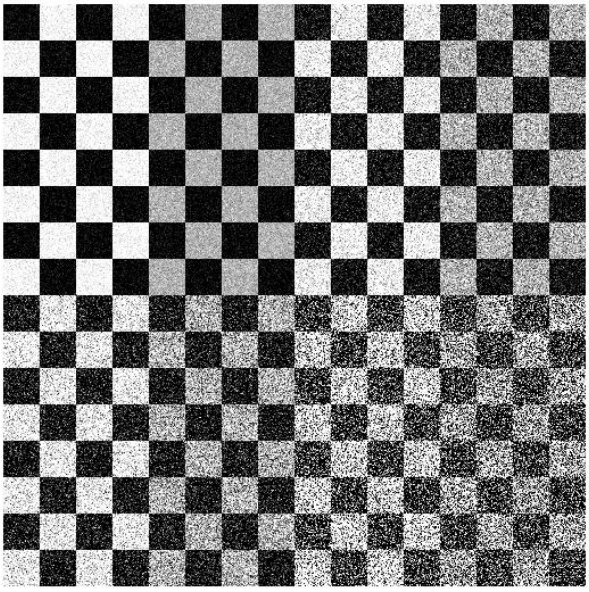


Figure A.1: Checkerboard image divided into four quadrants, each with additive noise of N(0,0.01), N(0,0.05), N(,0.1), and N(0,0.5), respectively.

A simple measurement of the variance of all pixels in each quadrant would include variations due to noise as well as due to the checkerboard itself. However, in this case, the relative trend in noise between the four quadrants can be distinguished in this manner, as shown in

Fig. A.2.

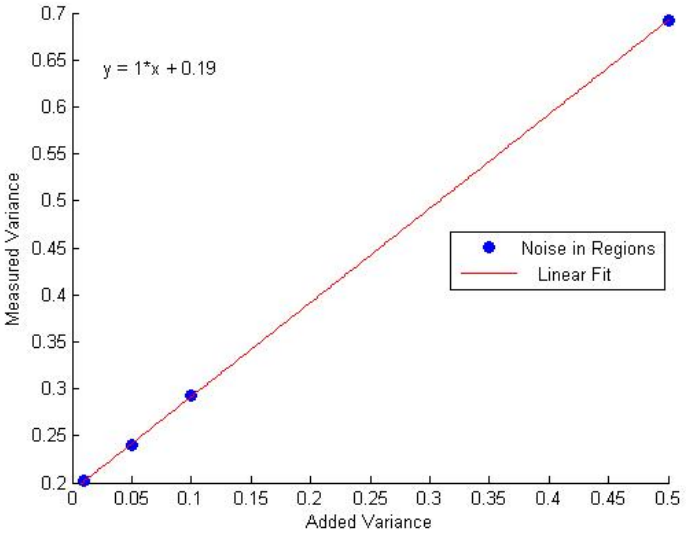


Figure A.2: Estimation of noise in each quadrant by direct measurement of overall variance.

Fitting a line to the results gives a slope of 1, suggesting an accurate assessment of relative noise is possible with this method. However, there is an offset of 0.19, which is significant for variances in the range of $[0.01 to 0.5]$. This offset represents the variance in the original checkerboard image.

Applying the WFDM method to the checkerboard produced a *distance image*, I_D, with pixels representing the degree of spatial variation around that pixel. In Fig. A.3, both the black and white squares of the checker board are detected as uniform regions, indicated by darker pixels (lower value) while the edges between the black and white squares are non-uniform regions and, hence, lighter. Similarly, the lower right quadrant with the highest additive noise variance appears lighter overall.

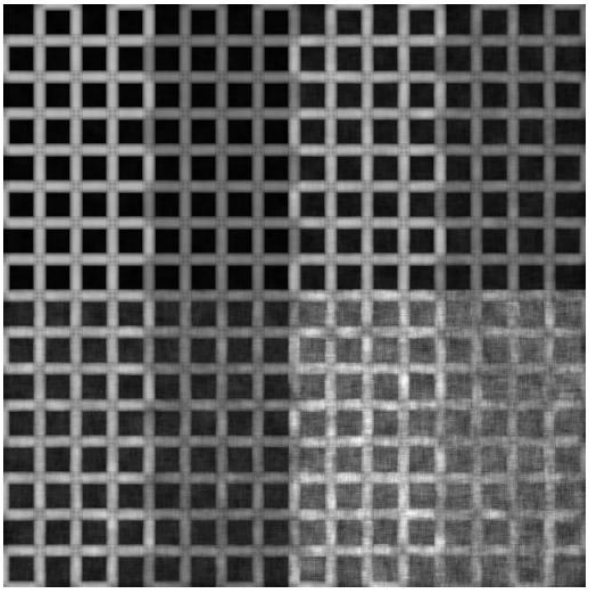


Figure A.3: The distance image, I_D, resulting from the WFDM process. Darker regions in the image are more spactially uniform than lighter ones.

Thresholding I_D to select the most uniform regions produces the mask that can be applied to the checkerboard, as shown in Fig. A.4. Here, all pixels in I_D below the average pixel intensity in I_D are selected, while those above are not.

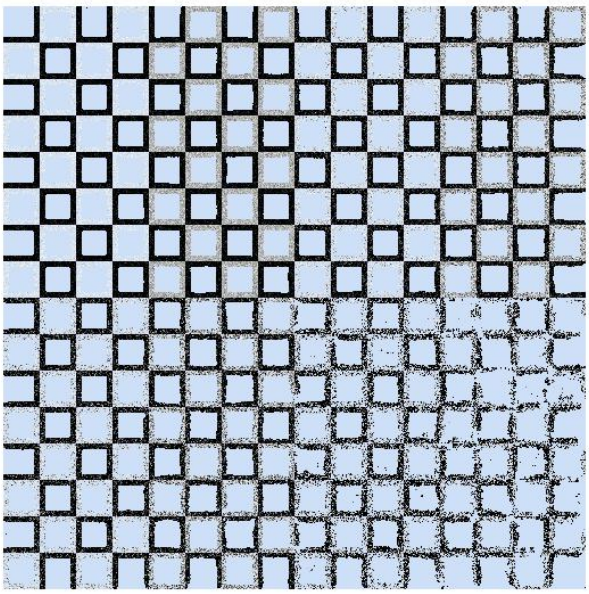


Figure A.4: The checkerboard masked by I_D, thresholded at $q = \mu$.

The edges in the original checkerboard image can clearly be seen, with only the inner protions of each square included. In the noisier regions, the distinction between edges and uniform regions become less clear, so the WFDM mask becomes less accurate. However, the WFDM model manages to exclude most regions that were causing the offset when measuring variance overall, as shown in Fig. A.5.

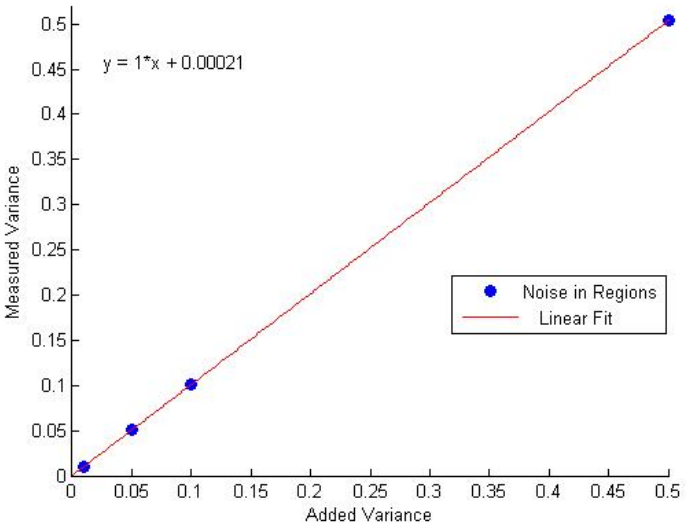


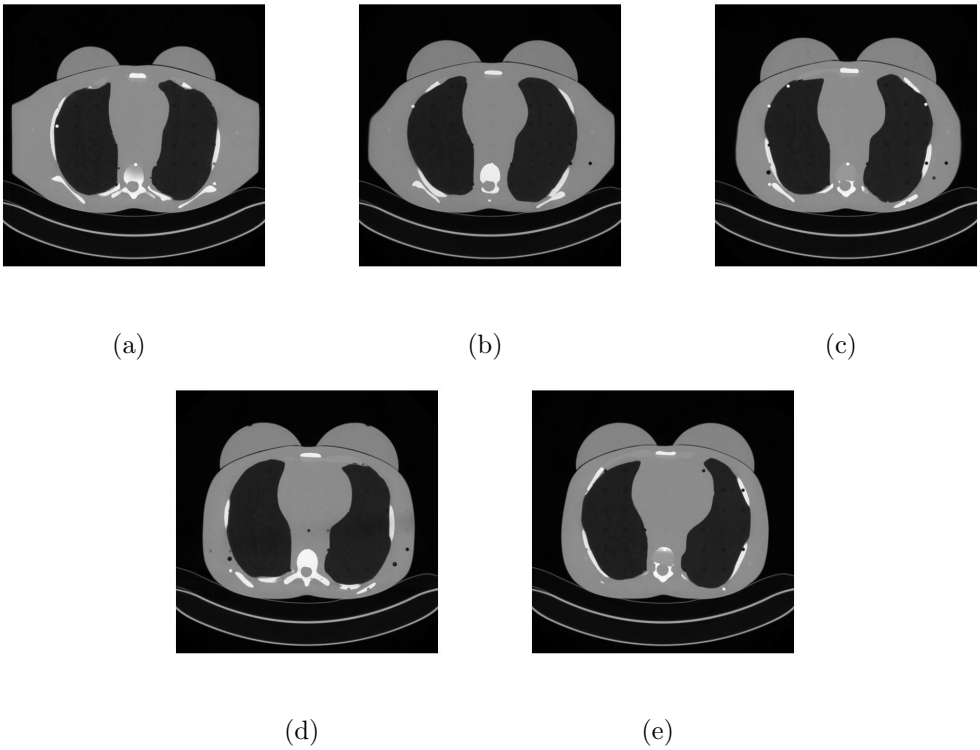
Figure A.5: Estimation of noise in each quadrant by measuring variance in the WFDM selected regions.

Once again, fitting a line to the data results in a slope of 1. However, the offset in this case is 0.00021, significantly less than both the overall estimation method of Fig. A.2 and the actual value of the variance. Thus, a comparison of the measured and actual values of variance shows them to be almost identical.

Appendix B

WFDM ANALYSIS OF NOISE IN PHANTOM IMAGE

Gaussian noise is added to 5 high quality phantom images and used to evaluate the behavior of the WFDM model. The phantom images are shown in Fig. B.1.



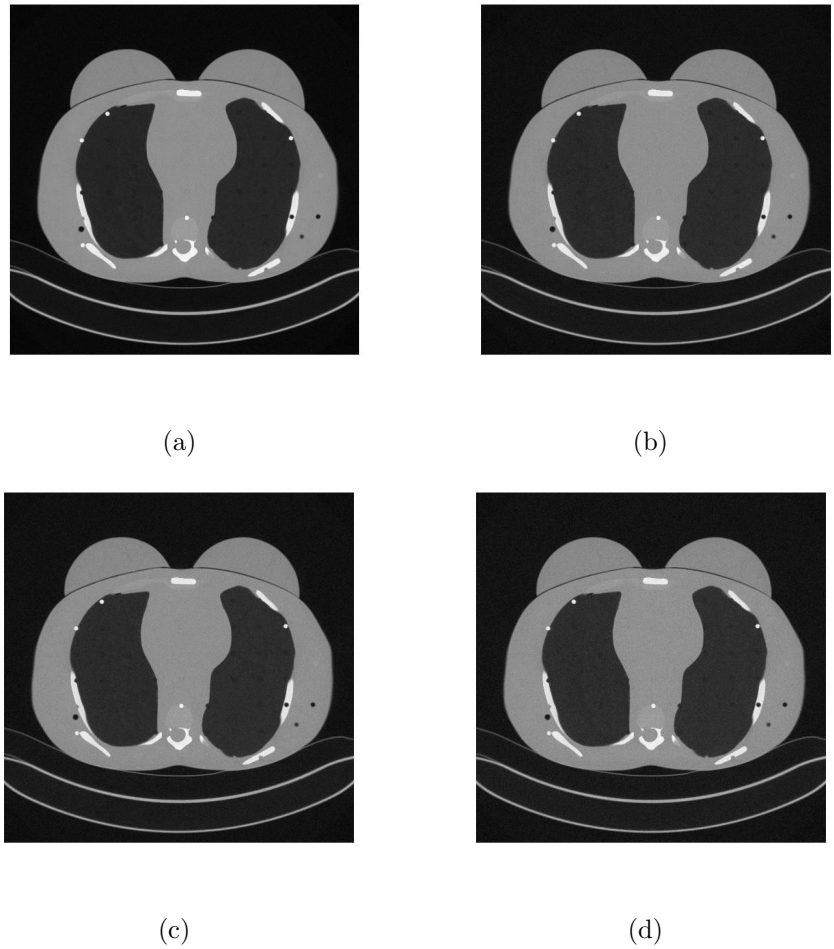
(a) (b) (c)

(d) (e)

Figure B.1: Five high quality phantom images, acquired from the same scan, showing variations in shape and size between slices.

Each image is degraded by noise with distributions of $N(0, 0.0001)$, $N(0, 0.0002)$, $N(0, 0.0003)$, and $N(0, 0.0004)$, respectively, resulting in 20 total images. Examples of each

of the 4 noise levels for one phantom image are shown in Fig. B.2.



(a) (b)

(c) (d)

Figure B.2: Phantom images degraded by noise with distributions of $N(0, 0.0001)$, $N(0, 0.0002)$, $N(0, 0.0003)$, and $N(0, 0.0004)$, respectively.

As before, an overall estimate is performed by taking the variance of the images. Although the relative trend between noise levels, for each phantom image, is observed, the offset once again affects the results. In this case, however, that offset varies depending on the underlying features of the image itself. Given the variations in shape and size of each phantom image, the resulting noise estimates are difficult to assess, as shown in Fig. B.3.

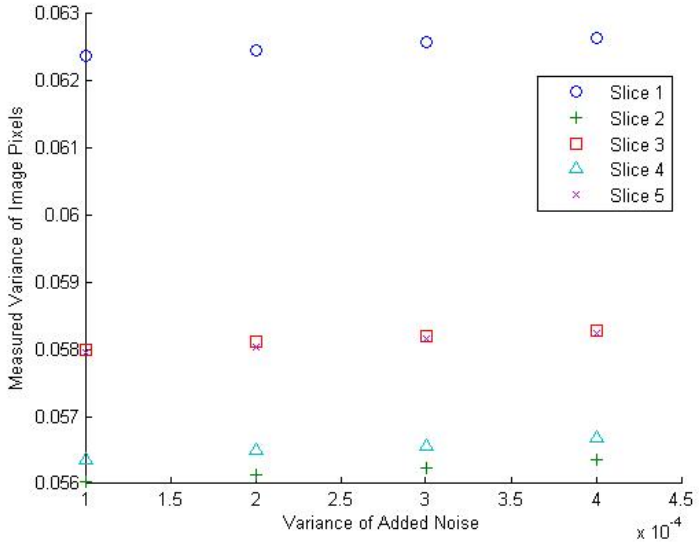


Figure B.3: Estimation of noise in each phantom image by measuring overall variance of pixels.

Performing the same analysis with the WFDM model results in a significantly lower offset, and, specifically, one that is consistent across the different phantom images, as shown in Fig. B.4.

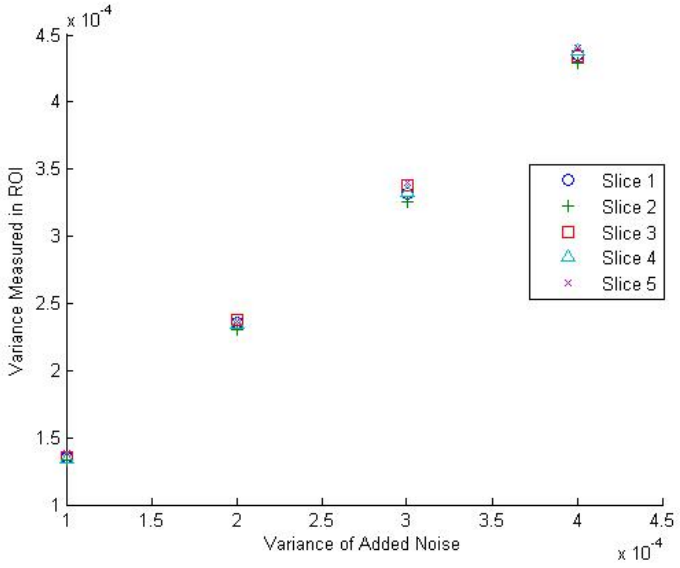


Figure B.4: Estimation of noise in each phantom image by measuring variance in the WFDM selected regions.

The linear trend and the close grouping of images with similar noise levels underscores the importance of feature selection in CT image analysis. By selecting the regions relevant to noise estimation, such as spatially uniform regions, the estimate is vastly improved. A line is fit to the results, shown in Fig. B.5, to find the slope and offset. Again, the slope is 1, indicating a direct relationship between the added variance and the amount measured by this method. The offset, in this case, indicates the amount of noise originally in the phantom images used for this experiment.

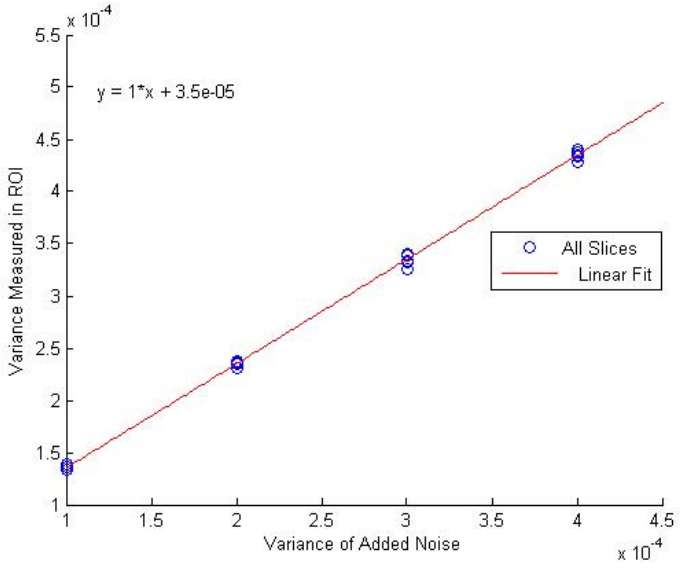


Figure B.5: Linear fit to WFDM noise estimation. The slope is 1, as expected, and the offset here can be an indicator of the noise originally in the phantom images.

The simple consistency of the phantom allows for such consistent results with the WFDM model, and the same response is not expected from patient images. However, if equivalent regions from different patients can be compared, such as the regions selected by the WFDM model, it would significantly reduce the error in noise estimations.