Title
Comparison of Traditional Image Quality Metrics with Human Observer Detection Performance in Ultrasound

Author
Mathew, Vineet Thomas

Publication Date
2020

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Comparison of Traditional Image Quality Metrics with Human Observer Detection Performance in Ultrasound

A thesis submitted in partial satisfaction

of the requirements for the degree of Master of Science

in Bioinformatics

by

Vineet Thomas Mathew

ABSTRACT OF THE THESIS

Comparison of Traditional Image Quality Metrics

with Human Observer Detection Performance in Ultrasound

by

Vineet Thomas Mathew

Master of Science in Bioinformatics

University of California, Los Angeles, 2020

Professor Van Maurice Savage, Chair

Given the critical role ultrasound imaging plays in medicine, it is important to have a reliable way to measure the quality of an ultrasound image. Image quality measurements allow engineers to design ultrasound imaging systems with configurations that allow users to perform the clinical tasks and diagnoses with the best accuracy, and also allow objective comparisons to be made between different ultrasound machines. The most meaningful way to measure image quality is to conduct a study on how well humans are able to perform the clinical task they will

be using the images for, but this is a resource intensive task. Therefore, in practice, ultrasound image quality is generally measured using basic mathematical metrics known as traditional image quality metrics. In this thesis, I explore how these metrics relate to human performance on a clinical task, and study whether human performance can be predicted from these metrics. To do so, I created a visual assessment which tests the ability of participants to detect artificial veins in an ultrasound image, at varying levels of noise and varying levels of undersampling (reduction in spatial resolution). I then compute the values of various traditional image quality metrics for the same ultrasound images and examine whether the trends of these traditional image quality metric values mirror the trends of human performance on our assessment, across varying levels of noise and undersampling. The results found that traditional image quality metrics had a relatively simple relationship with human performance across varying levels of noise, but a more complex relationship across varying levels of undersampling. In addition, I created regression models that were able to predict human performance from traditional image metrics for the cases studied. This work demonstrates a first step towards examining the relationship between traditional image quality metrics and task-based image quality assessments in the context of ultrasound images.

The thesis of Vineet Mathew is approved.

Eric Jameson Deeds

George Varghese

Van Maurice Savage, Committee Chair

University of California, Los Angeles

2020

To my parents Mathew and Gina,

and my sister Sneha:

Thank you for giving me your continued love and support,

and for working hard to give me the opportunities I've had in life.

Table of Contents

List of Figures

List of Tables

Acknowledgments

This thesis was completed during the COVID-19 pandemic. Given this, I want to give a special thanks to those who supported me in completing this thesis, because they did so while dealing with their own stresses caused by this time of uncertainty.

First, I am extremely grateful for Dr. Miles Wernick, who provided invaluable guidance, assistance and support throughout this project. I would also like to express my deepest appreciation to my thesis committee members Professor Savage, Professor Deeds and Professor Varghese. These individuals have helped shape my academic experience and have provided mentorship not only during my thesis, but throughout my college career. In addition, I would like to acknowledge Dr. Jovan Brankov for providing initial feedback on the study design.

Finally, I would like to thank all my friends, family, and loved ones who encouraged and helped me throughout the way. I would not be where I am today without the support my community gives me.

Comparison of Traditional Image Quality Metrics with Human Observer Detection Performance in Ultrasound

1. Introduction

Ultrasound imaging has grown to become the second most frequently used imaging modality in medicine[1]. Given its nonionizing nature, ability to provide real-time imaging and relative cost effectiveness, it has evolved into an indispensable tool for medical practitioners for a wide array of both diagnostic and therapeutic procedures[2].

Recent advances in technology have allowed for the development of portable ultrasound devices that can fit into a practitioner's pocket. This rise in ultrasound device accessibility, paired with the increasing realization of the benefit of routine and rapid ultrasound imaging in patient care, has contributed to the growth of point-of-care ultrasound (POCUS). POCUS involves the physician 'bringing ultrasound' to the patient, so that images are acquired and interpreted at the point-of-care. This differs from many traditional uses of ultrasound where a physician must first examine a patient, place an order for an ultrasound exam to be done by the radiology or imaging department, wait until the exam is complete, and then have the image interpreted by a specialized physician before treating the patient accordingly. On the other hand, POCUS is used by health care professionals to swiftly seek imaging information in order to guide immediate clinical decision making and treatment plans for the patient[3]. POCUS has already become adopted by fields such as emergency medicine and obstetrics and has shown to be effective in providing faster and more accurate diagnoses, as well as potential cost savings for patients[4,5,6]. One study showed that with just 18 hours of training, first year medical students

without any clinical experience were able to correctly identify 75% of cardiac pathologies when utilizing POCUS, compared to board-certified cardiologists who were only able to correctly identify 49% of the pathologies when using the standard practice bedside cardiovascular physical exam instead[7]. These benefits have begun to push POCUS into the general medical practice as a whole, leading to its growing adoption by primary care physicians in fields like family medicine and internal medicine[4]. As POCUS continues to become more accessible and its benefits become more evident, some physicians are calling for it to be added as another 'pillar' in the everyday standard practice of the bedside physical exam[8].

Given the growing importance of ultrasound imaging, it is critical that ultrasound devices produce images that are of high quality. Assurance of image quality becomes increasingly important as new POCUS ultrasound devices such as standalone portable probes that plug into a smartphone begin to enter the clinical space; these probes need to be able to produce images comparable to, if not better than, standard practice machines. To ensure this, the design and evaluation of ultrasound imaging devices require that image quality be measured in a reliable and quantitative fashion. Image quality assessments help guide initial optimizations of imaging system designs, and also demonstrate the performance of those designs in comparison to other ultrasound systems seen as the 'gold standard'[9].

In the field of image and signal processing, simple mathematical relationships are used to measure the 'quality' of an image or signal, such as Mean Squared Error (MSE) and Structural Similarity Index[10,11,12]. We will refer to these as "traditional image quality metrics". Traditional image quality metrics typically reflect some measure of correspondence between a measured image and a desired image (for example, a noisy image and a noise-free image). Ultrasound imaging engineers often use these traditional image quality metrics to guide device design and

measure the quality of captured images. However, in the medical community, there has been a recognition that a more fundamental definition of 'quality' is best. Specifically, "How effectively can the image be used for the clinical evaluation for which the image was acquired?"[13] Thus, the best measure of medical image quality is the user's performance on carrying out clinical tasks with the image, such as locating and distinguishing various anatomical structures.

Naturally, the most accurate way to measure quality in this way would be to have a series of physicians (or *human-observers*) take a 'test' and go through sets of images (e.g. images acquired by two competing devices). The ability to correctly diagnose disease or locate target anatomical structures would then be compared between both sets. However, this kind of task is costly in terms of time, money and logistics[14]. A simpler way would be to create a computerized method that is able to predict how well humans would perform on the test.

In the broader medical imaging field, especially in X-ray CT and nuclear medicine, computer algorithms known as *model observers* have been developed and have become an important tool in quantifying image quality[15]. These model observers act as a surrogate for human observers and perform the tasks human observers would do during an image quality evaluation test as described in the previous paragraph (say, locating an anatomical structure in an image), but do so in a way that the model observer's performance statistics mimic that of the human observers[16]. Good model observers can thus provide insight and predictions on how human observers would perform on the task used to evaluate image quality[17]. The efficacy of model observers in predicting human observer performance has had preliminary investigation in ultrasound as well[18]. But model observers have not found use in the ultrasound community, in spite of success in other communities, and further investigation in this direction would be worthwhile.

However, instead of creating a computerized model observer to predict human performance, we asked the following question: Since traditional imaging metrics are used as standard practice in the ultrasound field, what if we could simply use these traditional image metrics as a way to understand human-observer performance? Though engineers use traditional image metrics when optimizing and designing ultrasound imaging systems, the relationship between these metrics and quality defined by human clinical task performance does not appear to have been studied previously in the context of ultrasound.

As a first step towards understanding this relationship, I explore the following questions in this thesis:

- **What is the relationship between various traditional image quality metrics and human-observer performance in a defined clinical task?**
- **Can traditional image quality metrics be used to predict human-observer performance?**

Specifically, I evaluate the extent to which traditional image quality metrics can be used to understand and predict human-observer performance in identifying anatomical objects within images subject to various levels of degradation.

To study this relationship, I created a visual detection task by simulating anatomical structures consisting of image signals inserted artificially into clinical ultrasound images which were subject to varying levels of degradation. To assess human performance in detecting these signals, I recruited a group of volunteers to view these images and measured their accuracy in detecting the signals (simulated artificial blood vessels). Participants in such a study are known as *readers*. The readers viewed a series of pairs of ultrasound images in which one of the images

in each pair contained the inserted artificial blood vessel. Individuals would then make a binary choice as to which image in the pair they thought had the inserted artificial vessel, and their detection accuracy was computed. In other words, the assessment was measuring how visible these artificial vessels were to humans, at ten different levels of image degradation. Then, I calculated a collection of traditional image quality metrics for the images at these ten levels of image degradation, and examined whether these traditional image quality metric values followed patterns across these degradation levels that were similar to the patterns that human task performance followed across the degradation levels.

Human performance on this task was defined by the area under the receiver operating characteristic (ROC) curve (AUC), a well-established method of characterizing the accuracy of a binary decision-making agent[19]. In other words, the AUC measured how visible an artificial vessel was to a human, and the goal was to measure whether simple mathematical metrics (traditional image quality metrics) are actually predictive of AUC (Fig. 1).

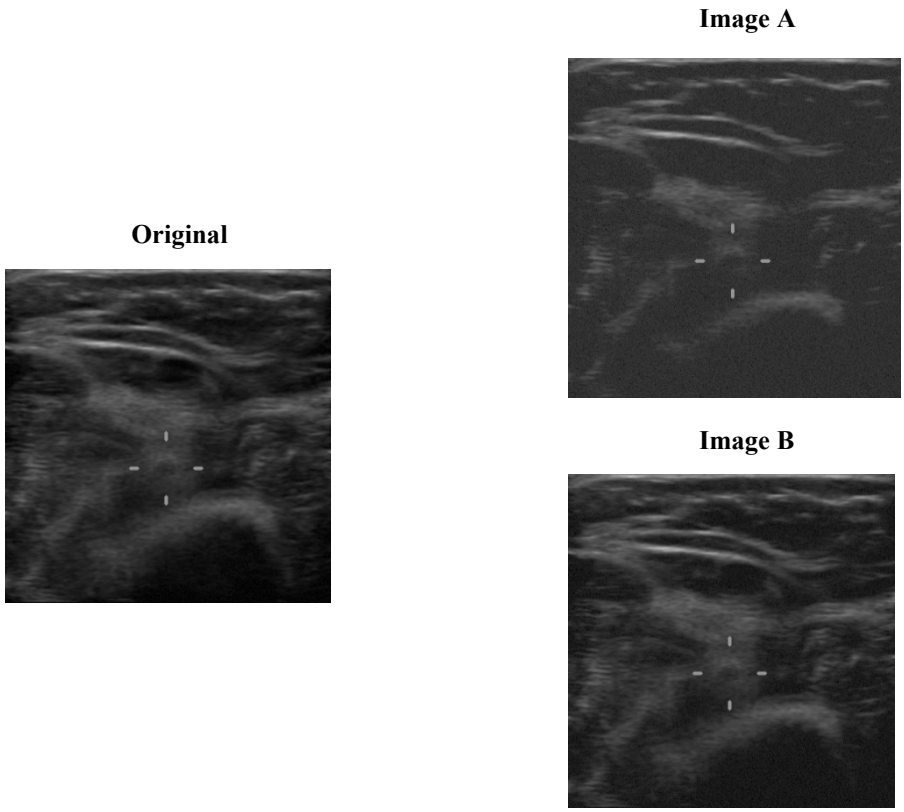


Figure 1: An image with artificial vessel inserted (left) at two levels of simulated degradation (right).

(Top right) **Image A**: This image is a degraded version of its original version (left). It has a Mean Squared Error (MSE) of 69.6 compared to its original version.

(Bottom right) **Image B**: This image is slightly degraded version of its original version (left). It has a Mean Squared Error (MSE) of 24.5 compared to its original version.

MSE is a traditional image quality metrics used to guide ultrasound imaging system design. However, are MSE values indicative of how well humans would be able to correctly detect the presence of the simulated blood vessel (located at the center of the crosshairs)? Since **Image A** has an MSE nearly three times higher than **Image B**, does this mean that humans will be three times less likely to correctly detect the presence of the simulated blood vessel in **Image A** than in **Image B**? This thesis seeks to identify the relationship between traditional image quality metrics and quality measured by human-observer detection performance.

2. Methods

2.1 Baseline image dataset

To measure the detection performance of humans in a clinical task it is necessary to have access to the ground truth for each image. It is common in medical imaging detection studies to use *hybrid* images, in which real clinical images are augmented with simulated anatomical

structures, with the presence and characteristics of these structures providing the ground truth[20]. It is also common in medical imaging studies to simulate anatomical structures with relatively simple structures, such as circular disks. Thus, I based my experiments on vascular ultrasound images, because veins and arteries (vessels) are in reality well approximated by dark circular disks within surrounding tissue.

A set of 20 vascular ultrasound images obtained by imaging with a Philips Lumify linear imaging probe formed the base dataset for the experiments. Each of these images was then modified by artificially introducing one dark circular disk at a location selected to appear visually realistic and not easy to detect. We will refer to these disks as simulated "vessels". The "vessel" location was chosen individually for each of the 20 images, so that no two were alike.

Each image containing a "vessel" was further modified by inserting crosshairs indicating the location of the "vessel". Crosshairs were also inserted at the same location in the corresponding original image (prior to insertion of the "vessel") to indicate the position where the "vessel" would be located if it were it to be present. We will refer to the images with and without a "vessel" as "object-present" and "object-absent" respectively. Thus, we had 20 object-absent/object-present pairs. Let us call these set of images in our baseline dataset S. Therefore,

$$S = \{(\mathbf{Y_1}, \mathbf{Y_1}^{obj}), (\mathbf{Y_2}, \mathbf{Y_2}^{obj}), \ldots, (\mathbf{Y_{20}}, \mathbf{Y_{20}}^{obj})\}$$

where $\mathbf{Y_n}^{obj}$ is the object-present version of $\mathbf{Y_n}$. The vessels that were inserted in each object-present image had one of three different radii (small:7 pixels, medium:10 pixels, large: 13 pixels), and one of three intensities (light: 90% brightness, medium: 86% brightness, and dark: 82% brightness) (Fig. 2). In a pilot study, these radius and intensity values were chosen such that any combination of radius and intensity would create a vessel that yielded approximately 90%

detection accuracy, so that once image degradations are added, the accuracy would drop into the range of "challenging" cases.

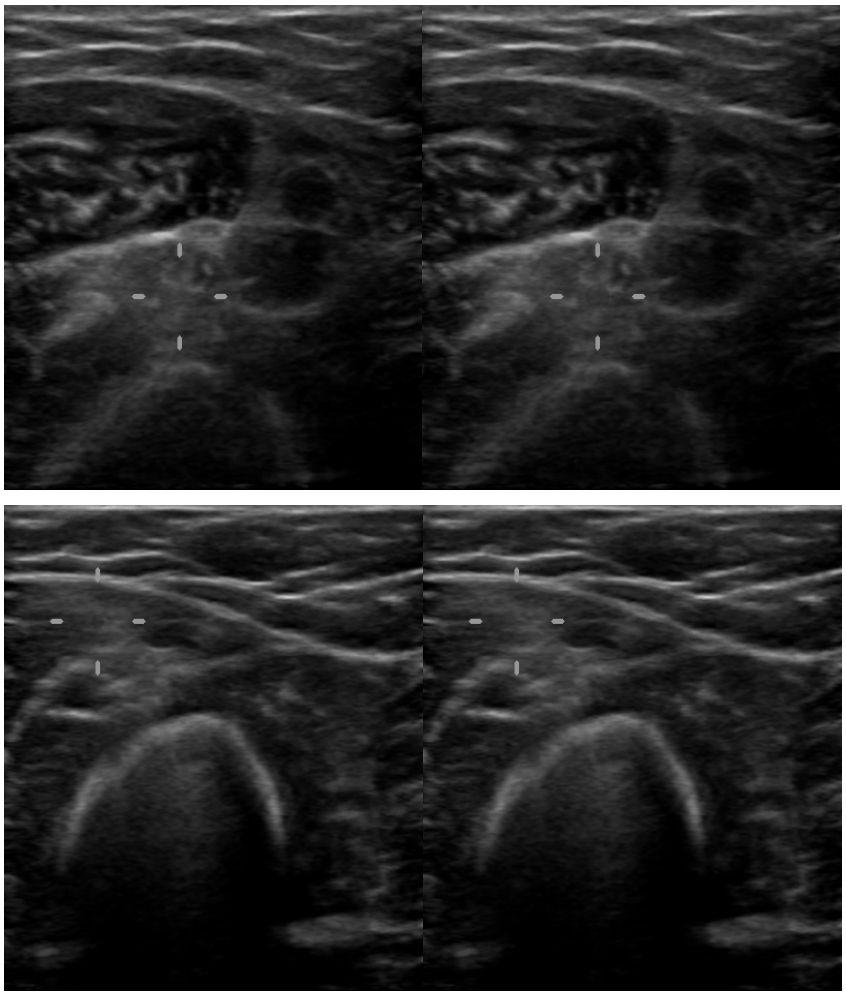


Figure 2: Examples of object-present/object-absent images in S

Crosshairs highlight the region of interest in the image in which an artificial vessel was inserted in the object present version of the image. Only one image in each pair contains an artificial vessel in the crosshairs. Participants' detection performance was evaluated based on their ability in correctly identifying which image in each pair contained the artificial vessel, on degraded versions of these pairs.

(Top) Here, the right side image is the object present image, since it contains the dark object at the center of the region with the crosshairs, unlike the left side image. The artificial vessel has a 'medium' intensity and a 'large' radius

(Bottom) Here, the left side image is the object present image, since it contains the dark object at the center of the region with the crosshairs, unlike the left side image. The artificial vessel has a 'dark' intensity and a 'small' radius

For each pair in *S*, I then created ten different versions of the pair, each at one of ten simulated levels of image degradation. I simulated five different levels of noise degradation, and five different levels of "undersampling" degradation (loss of information that occurs when insufficient imaging information is captured from the ultrasound probe, resulting in a lack of spatial resolution). These degradations will be more thoroughly explained in the following section.

Thus, I created five noised groups, ***noise_1***, ***noise_2***, ***noise_3***, ***noise_4***, and ***noise_5***, and five undersampled groups, ***undersample_1***, ***undersample_2***, ***undersample_3***, ***undersample_4***, and ***undersample_5***. Each group consisted of the 20 image pairs in *S*, subject to the group's given level of noise or undersampling degradation.

The 20 pairs of images in *S* at ten different levels of degradation yielded 200 images in the final dataset, which was used in the visual detection assessment. The following section explains the basis in which these degradation groups were created.

2.2 Simulated Image Degradation Groups: Noise and Undersampling

An important use of image quality metrics in ultrasound is to measure the effect of various imaging parameters on the performance of an imaging device or algorithm. This can be used, for example, to evaluate various design choices. Therefore, it is important to understand how these design choices affect the image quality metrics and—as I study in this thesis—to understand how these choices affect the performance of the user in conducting clinical tasks, which reflects the true real-world impact.

The imaging parameters that are most important to aspects of image quality are those that affect noise and spatial resolution. Noise level is a function of many factors, including the

manner of data acquisition and various characteristics of the imaging hardware. Spatial resolution is affected by various aspects of the data acquisition scheme as well, with an important parameter being the number of scanlines acquired to produce the image. Thus, to simulate the effect of imaging parameter choices that could lead to progressively greater levels of image degradation, I modified the original images by either introducing artificial noise or reducing the number of scanlines (undersampling).

Noise

During ultrasound image capture, the signal can encounter thermal noise, which obeys a Rician distribution (complex magnitude of i.i.d. Gaussian noise in real and imaginary parts of the signal)[21]. The captured signal is then transformed by a logarithmic transformation (*log compression*) to compress to greyscale values for purposes of improved visualization.

Thus, in order to insert simulated noise into the object-present and object-absent images, it was necessary to emulate a process of inserting the noise to the preprocessed ultrasound image information, prior to log compression.

The relationship between log compressed ultrasound image that is displayed, and the preprocessed image data is given by the following:

$$\mathbf{Y}_{\mathbf{n}_{ij}} = D\ln(\mathbf{X}_{\mathbf{n}_{ij}}) + G$$

where $\mathbf{Y}_{\mathbf{n}_{ij}}$ is the pixel in the *i*th row and *j*th column of the displayed image $\mathbf{Y}_{\mathbf{n}}$, and $\mathbf{X}_{\mathbf{n}_{ij}}$ is the pixel in the *i*th row and *j*th column of the preprocessed image. D and G are constants (in this project I used D=18.147 and G =0), which were inferred from the baseline image dataset by using the methodology described in a work by Paskaš [22].

The pre-compression values $\mathbf{X_{n_{ij}}}$ can be written in terms of $\mathbf{Y_{n_{ij}}}$ as follows:

$$\mathbf{X_{n_{ij}}} = e^{(\mathbf{Y_{n_{ij}}}/D)+G}$$

The pre-processed pixel data $\mathbf{X_{n_{ij}}}$ was modified to create noisy data $\tilde{\mathbf{X}}_{\mathbf{n_{ij}}}$, as follows

$$\tilde{\mathbf{X}}_{\mathbf{n_{ij}}} = \sqrt{\left(\mathbf{X_{n_{ij}}} + A_1\right)^{\mathbf{2}} + \left(\mathbf{X_{n_{ij}}} + A_2\right)^{\mathbf{2}}}$$

with $A_1, A_2 \sim \mathbf{N}(\mathbf{0}, \sigma^2)$, which yielded the following after log compression:

$$\tilde{\mathbf{Y}}_{\mathbf{n_{ij}}} = D\ln\left(\tilde{\mathbf{X}}_{\mathbf{n_{ij}}}\right) + G.$$

So for a given object-absent / object-present image pair in S, $(\mathbf{Y_n}, \mathbf{Y_n^{obj}})$, I chose a value for σ and calculated the noisy version of each of the pixels, resulting in the noisy image pair $(\tilde{\mathbf{Y}}_{\mathbf{n}}, \tilde{\mathbf{Y}}_{\mathbf{n}}^{obj})$. Note that for a set of corresponding pixels in a given object-absent and object-present image pair $(\mathbf{Y_{n_{ij}}}, \mathbf{Y_{n_{ij}}^{obj}})$, the same noise values are used for each pair of $\tilde{\mathbf{X}}_{\mathbf{n_{ij}}}$ and $\tilde{\mathbf{X}}_{\mathbf{n_{ij}}}^{obj}$.

Using the noise simulation scheme above, I created five 'noise' groups, ***noise_1***, ***noise_2***, ***noise_3***, ***noise_4***, and ***noise_5***. Each group consisted of noisy versions of all 20 pairs of object-present and object-absent images in S, according to Table 1.

Table 1:Values of the standard deviation of the random noise components added to the preprocessed version of the images in each noise group.

Group Name	Value of σ for $A_1, A_2 \sim \mathbf{N}(\mathbf{0}, \sigma^2)$
noise_1	1
noise_2	e^2
noise_3	e^4
noise_4	e^6
noise_5	e^8

More specifically,

$$\textbf{\textit{noise\_k}} = \{(\mathbf{Y_1}_{\,noise\_k}, \mathbf{Y_1^{obj}}_{noise\_k}), (\mathbf{Y_2}_{\,noise\_k}, \mathbf{Y_2^{obj}}_{noise\_k}), \ldots, (\mathbf{Y_{20}}_{\,noise\_n}, \mathbf{Y_{20}^{obj}}_{noise\_k})\}$$

where $(\mathbf{Y_n}_{\,noise\_k}, \mathbf{Y_n^{obj}}_{noise\_k})$ refers to the *n*th object present/object-absent image pair in *S*, subject to the noise level as indicated by *k* as described in the above table.

Using the process described above, images became increasingly noisy moving from ***noise_1*** to ***noise_5*** (Fig. 3). Since each of the 20 object-absent/object-present image pairs were noised in each of the five groups at the groups respective noise level, a total of 100 noised object absent/object present image pairs were created.

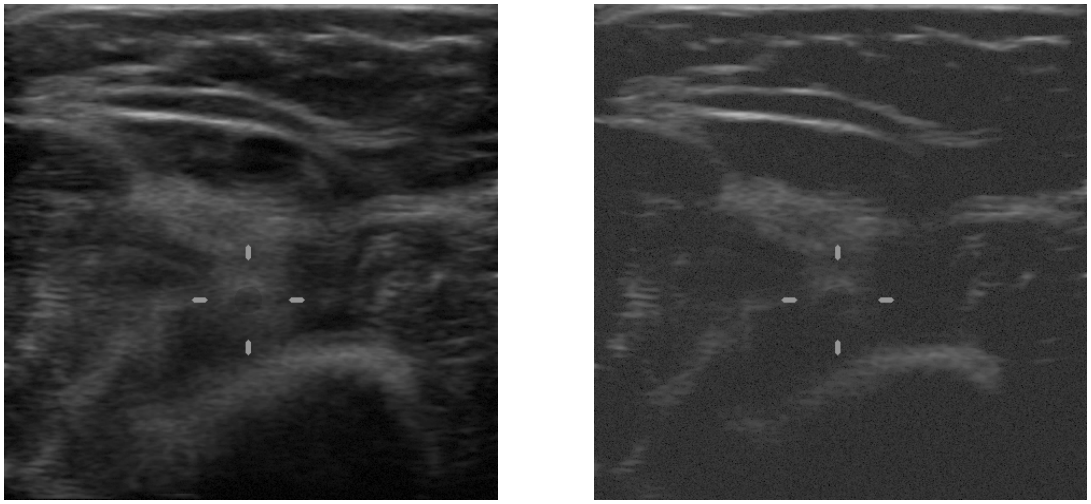


Figure 3: Examples of an object-present image $\mathbf{Y_1}^{\mathbf{obj}}$, with different levels of noise

Note: When displayed to the readers these images would never appear side by side during the visual assessment, but would rather appear side by side with their respective object-absent versions

(Left): $\mathbf{Y}^{\mathbf{obj}}_{\mathbf{1}_{noise\_1}}$, the **noise_1** version of the image $\boldsymbol{Y_1}^{\boldsymbol{obj}}$.

(Right): $\mathbf{Y}^{\mathbf{obj}}_{\mathbf{1}_{noise\_3}}$, the **noise_3** version of the image $\boldsymbol{Y_1}^{\boldsymbol{obj}}$.

Undersampling

Ultrasound images are formed by having multiple transducer elements capture analog acoustic echo signals that reflect off of tissues in the body from various directions. The analog signals captured by each transducer element are sampled at discrete spatial and temporal intervals to digitize the signal. The spatial sampling yields a collection of *scanlines* which depict reflections along a ray passing through the object[23]. In linear-mode ultrasound, which was the type of ultrasound used to capture the images in this study, the scanlines are displayed as columns of pixels in the image. If the acoustic signals are sampled at the Nyquist rate, a sufficient number of scanlines will be generated to fully characterize the signal[24]. However, if the signal is sampled below the Nyquist rate, fewer scanlines are generated, leading to a loss of visual information and reduced spatial resolution. This is called "undersampling."

The images acquired in this study were captured at the display resolution of 458×470 (width × height); however, much of the information in the horizontal direction is redundant. This is because the actual image information is determined not by the display resolution, but by the number of scanlines in the data before it is upsampled (stretched) to produce the geometrically correct aspect ratio which depicts the tissue in a geometrically correct way. Hardware manufacturers do not typically publish the number of scanlines acquired by the system, so I inferred this number empirically. I accomplished this by downsampling the display resolution of the images at increasingly lower rates, followed by upsampling back to the display resolution, until I obtained results that were just noticeably different than the original images. From this I inferred that the number of original scanlines is approximately 153.

Next, to emulate undersampling (diminished horizontal spatial resolution due to reducing the number of ultrasound scanlines), I downsampled each image to $w' \times 470$, and then upscaled the image back to 458×470 to retain the original aspect ratio. Here, w' was a number less than 153, the number of original scanlines that I estimated to be in the image. By doing this downsampling followed by upscaling, the images became more blurry, because the original image was attempted to be reconstructed by an 'undersampled' version.

Thus, in a given object-absent / object-present image pair ($\mathbf{Y_n}, \mathbf{Y_n}^{obj}$), prior to downsampling, each image in the pair represented an image sampled at the Nyquist rate that led to 153 ultrasound beams that were used to form the original 470×458 image. Using the undersampling simulation scheme above, I created five 'undersampling' groups, ***undersample_1***, ***undersample_2***, ***undersample_3***, ***undersample_4***, and ***undersample_5***. Each group consisted of undersampled versions of all 20 pairs of object-present/object-absent images, according to the following scheme in Table 2.

Table 2: Values of $\mathbf{w'}$ for images in each undersample group. Each image was downsampled to $w' \times 470$ and then upscaled back to 458×470 to simulate the effect of various levels of undersampling.

Group Name	w'
undersample_1	143
undersample_2	109
undersample_3	75
undersample_4	41
undersample_5	7

More specifically,

$$\boldsymbol{undersample\_k} = \{(\mathbf{Y_1}_{\,sample\_k}, \mathbf{Y}^{\mathbf{obj}}_{\mathbf{1}_{sample\_k}}), (\mathbf{Y_2}_{\,sample\_k}, \mathbf{Y}^{\mathbf{obj}}_{\mathbf{2}_{sample\_k}}), \ldots, (\mathbf{Y_{20}}_{\,sample\_k}, \mathbf{Y}^{\mathbf{obj}}_{\mathbf{20}_{sample\_k}})\}$$

where $(\mathbf{Y_n}_{\,sample\_k}, \mathbf{Y}^{\mathbf{obj}}_{\mathbf{n}_{sample\_k}})$) refers to the nth object present/object-absent image pair in *S*, subject to the undersampling level as indicated by *k* as described in the above table .

With this scheme, images became increasingly blurry moving from ***undersample_1*** to ***undersample_5*** (Fig. 4). Since each of the 20 object absent/object present image pairs had undersampled versions made in each of the five groups, a total of 100 undersampled object-absent/object-present image pairs were created.

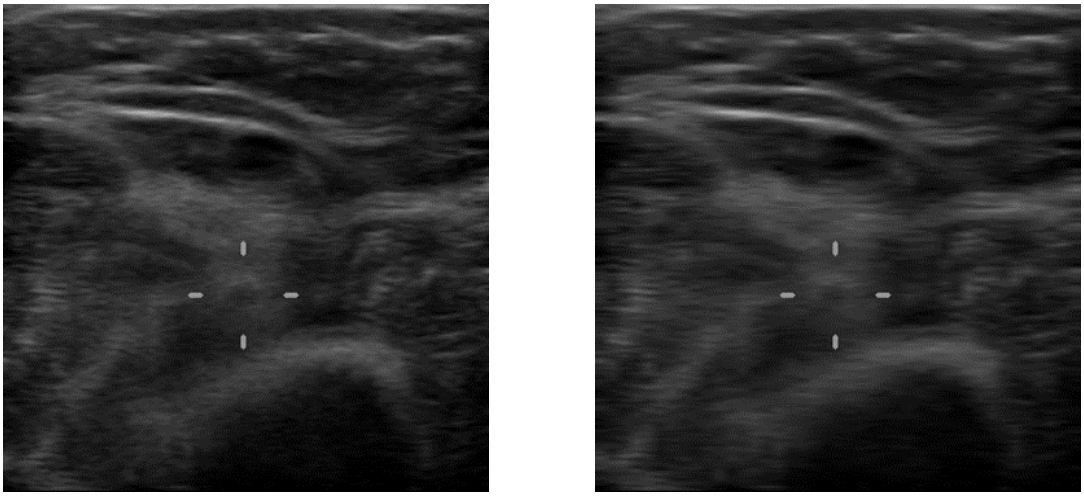


Figure 4: Examples of an object-present image $Y_1{}^{obj}$ with different levels of undersampling

Note: When displayed to the readers these images would never appear side by side during the visual assessment, but would rather appear side by side with their respective object--absent versions.

(Left): The ***undersample_1*** version of the image, $\mathbf{Y}^{\mathbf{obj}}_{\mathbf{1}_{sample\_1}}$.

(Right): The ***undersample_3*** version of the image $\mathbf{Y}^{\mathbf{obj}}_{\mathbf{1}_{sample\_3}}$.

In these undersampling groups, no simulated noise was introduced. Thus, our visual assessment consisted of two separate experiments: one to study only noise effects, and one to study only undersampling effects. Cross effects would be worthy of study, but would require a much larger-scale effort on the part of the participating readers.

Combining all noise and undersampling groups, our final dataset consisted of 200 images: 100 noisy pairs, and 100 undersampled pairs.

2.3 Visual Detection Assessment Overview

A classic approach to quantifying binary classification (detection) task-performance is the receiver-operating characteristic (ROC) curve, which depicts the trade-off of detection probability versus false alarm probability over a range of detection thresholds. To summarize this curve with a single numeric metric, it is common to use the area under the ROC curve, abbreviated as AUC.

One can measure the AUC in practice by first estimating the ROC curve from participants' reported confidence in the presence or absence of an object for each of a group of individual images presented to them (some object-present, some object-absent), and then calculating the true-positive and false-positive fractions. However, a more convenient approach that is mathematically equivalent is the *two-alternative forced-choice* (2AFC) test, in which participants are sequentially presented with matched pairs of images that are identical except that one contains the object and the other does not. A reader's ensemble average accuracy when guessing which of the two images they believe contains the object forms an estimate of the AUC value for that individual and image group[22]. The 2AFC is called *two-alternative* because the reader chooses which image contains the object; the term *forced-choice* refers to the fact the reader *must* choose one of the two images, even if they are unable to perceive the object in either of the two images.

In the literature of visual interpretation of medical images, participants in an experiment are known as *readers* or *human observers*. To measure human performance through AUC as described above, I recruited 13 *readers* and created a visual detection assessment measuring their performance in detecting simulated 'vessels' digitally embedded within genuine vascular ultrasound images. The assessment consisted of 200 2AFC tasks: readers were shown each of the 200 image pairs in our final dataset and were instructed to select which of the two images in the pair they perceived to contain the artificial vessel (Fig. 5), even if they were unsure. The 200 pairs were presented to the readers in random order to avoid ordering effects. For consistency, the readers performed the assessments under the following conditions, which were uniform across the readers. All readers were shown a small tutorial with examples of object present and object absent pairs before the test began. Readers were seated so that their eyes were 50cm away from the computer screen. Screen display was set to maximum brightness with a 90% zoom level on Google Chrome. All participants took the assessment on an HP Spectre 2-in-1 x360 laptop. No time limit was given.

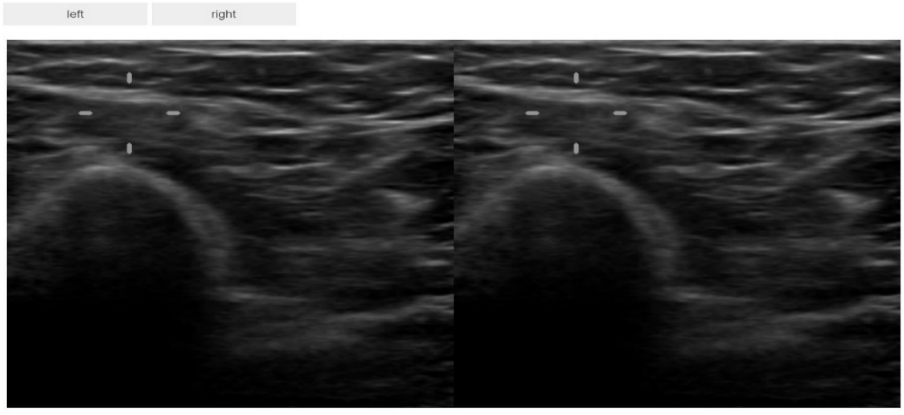


Figure 5: Example of 2AFC task

Readers had to choose which image in the pair they perceived to contain the object.. Here, the right image in the pair is the object present image, since it contains the dark object at the center of the region with the crosshairs, unlike the left side image.

2.4 Comparing Human Observer Performance with Traditional Metrics

After all visual assessments were complete, I then computed the average values of traditional image metrics within each noise and undersampling group and compared graphs of these metrics with graphs of AUC across noise and undersampling groups. AUC for each participant within each degradation group was calculated based on the percent of correct responses in that degradation group. In other words, for each set of 20 images in each of the 10 degradation groups, I calculated various traditional metric values (each averaged across all the images in a given degradation group), and compared it with each participant's percent of correct responses in that degradation group (AUC). I then used a resampling-based train/test machine-learning paradigm to develop a simple prediction model to predict human performance from the image metrics that had the simplest relationship to AUC. In this study, I compared the results of using linear, quadratic, and cubic polynomial models in predicting AUC for the noise and undersampling groups. In each iteration of the resampling scheme, 7 readers were randomly selected for training, and the remaining 6 readers for testing. Within each iteration, I performed a regression to minimize sum of squared errors between the prediction model and the AUC scores of all 7 readers, and then measured mean absolute error (MAE) and root mean square error (RMSE) with our models predictions and the average AUC for the 6 readers in the test. I then averaged the RMSE and MAE across 100 iterations as measures of generalization error (to measure generalizability of the model to unseen readers).

The following section will provide more detail on the creation of the dataset and the traditional image metrics that were calculated.

2.5 Metrics

Using the following traditional image quality metrics, I calculated average metric values across the images within each of the five noise groups, and each of the five undersampling groups. I then compared each metric's average value within each degradation group to the average AUC in each degradation group.

MSE (Mean Squared Error)

Given two $n \times m$ images $\mathbf{Y}$ and $\hat{\mathbf{Y}}$, the mean squared error is computed as follows:

$$MSE(\mathbf{Y}, \hat{\mathbf{Y}}) = \frac{1}{nm} \sum^{i} \sum^{j} (\mathbf{Y_{ij}} - \hat{\mathbf{Y}}_{\mathbf{ij}})^2$$

where $\mathbf{Y_{ij}}$ and $\hat{\mathbf{Y}}_{\mathbf{ij}}$ refer to the pixel in the ith row and jth column of each image respectively.

The MSE was computed between each degraded object present image and the original object-present counterpart.

MAE (Mean Absolute Error)

Given two $n \times m$ images $\mathbf{Y}$ and $\hat{\mathbf{Y}}$, the mean squared error is computed as follows:

$$MAE(\mathbf{Y}, \hat{\mathbf{Y}}) = \frac{1}{nm} \sum^{i} \sum^{j} |\mathbf{Y_{ij}} - \hat{\mathbf{Y}}_{\mathbf{ij}}|$$

where $\mathbf{Y_{ij}}$ and $\hat{\mathbf{Y}}_{\mathbf{ij}}$ refer to the pixel in the ith row and jth column of each image respectivley.

The MAE was computed between each degraded object present image and the original object-present counterpart.

CNR (Contrast to Noise Ratio)

The contrast to noise ratio (CNR) of an ultrasound image is given by

$$\mathrm{CNR} = \frac{|\mu_{\text{object}} - \mu_{\text{surrounding tissue}}|}{\sqrt{\sigma^2_{\text{object}} + \sigma^2_{\text{surrounding tissue}}}}$$

where μ_{object} and σ^2_{object} refer to the mean and variance of pixel values of the object within an image region of interest (in our case, an artificial 'vein') and $\mu_{\text{surrounding tissue}}$ and $\sigma^2_{\text{surrounding tissue}}$ refer to the mean and variance of pixel values in the concentric annulus surrounding the 'vein' that has the same area as the 'vein'.

CNR was measured on each object-present image and was then divided by the CNR values of each corresponding original object-present image (the version of the image before noise or undersampling degradation was added). This was done to measure CNR of the images relative to their degradation-free versions.

Contrast

The contrast of an ultrasound image is simply the numerator of the CNR expression, namely $|\mu_{\text{object}} - \mu_{\text{surrounding tissue}}|$.

Contrast was measured on each object-present and was then divided by the Contrast of each corresponding original object-present image (the version of the image before noise or undersampling degradation was added). This was done to measure Contrast of the images relative to their degradation-free versions.

SSIM

The Structural Similarity Index Measure defines the similarity between two images $\mathbf{Y}$and $\hat{\mathbf{Y}}$ as a function of comparisons between their similarities in luminance, contrast, and structure. In general,

$$SSIM(\mathbf{Y}, \hat{\mathbf{Y}}) = f(l(\mathbf{Y}, \hat{\mathbf{Y}}), c(\mathbf{Y}, \hat{\mathbf{Y}}, s(\mathbf{Y}, \hat{\mathbf{Y}}))$$

where $l, c,$ and s are functions that calculate the similarities between the luminance, contrast, and structure of the images respectively. More details of the function can be found in the paper by Wang et. al[11]. SSIM was measured on each object-present relative to each corresponding original object-present image (the version of the image before noise or undersampling degradation was added).

3. Results

This thesis sought to explore the following questions:

- **What is the relationship between various traditional image quality metrics and human-observer performance in a defined clinical task?**
- **Can traditional image quality metrics be used to predict human-observer performance well?**

To answer these, we will look at the results from the methods in four steps:

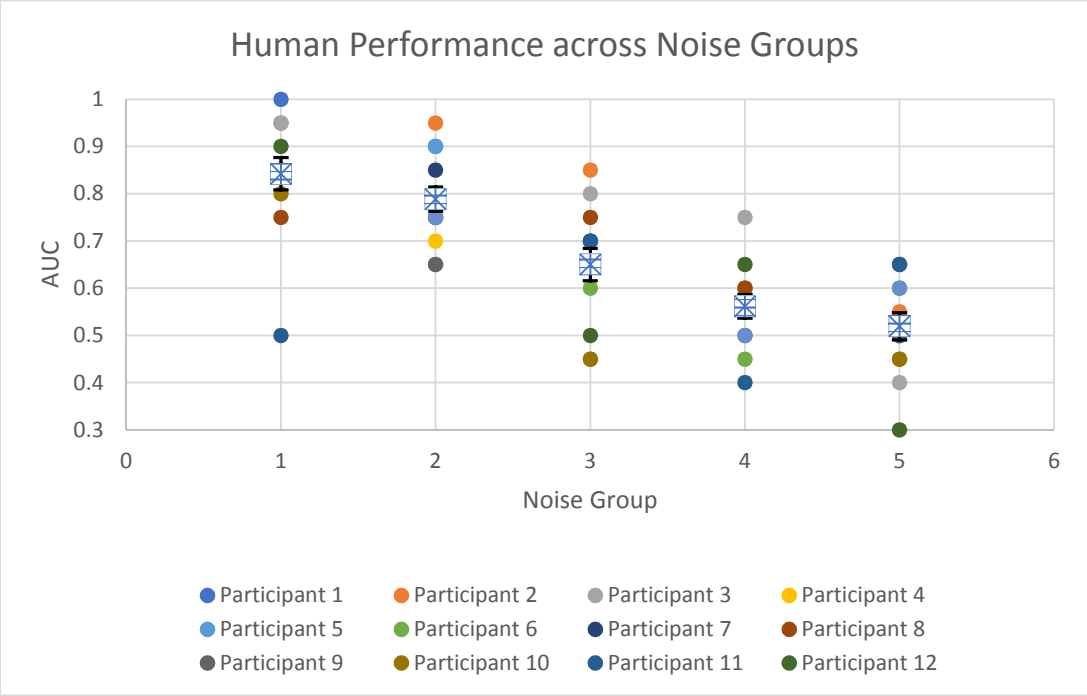
1. How does human detection performance respond to various levels of image degradation?
2. How do traditional image quality metrics respond to various levels of image degradation?

3. How do the traditional image metrics vary with respect to the AUC value (at each level of image degradation)?

4. How do regression-based machine learning models perform in predicting human AUC from the most-promising of the traditional image metrics?

3.1 Human Detection Performance Patterns

The detection performance of the readers at the noise and undersampling levels is summarized in the following graphs. Conducting an assessment like this is the most direct and accurate way to measure true image quality as we described in the introduction, though is resource intensive.



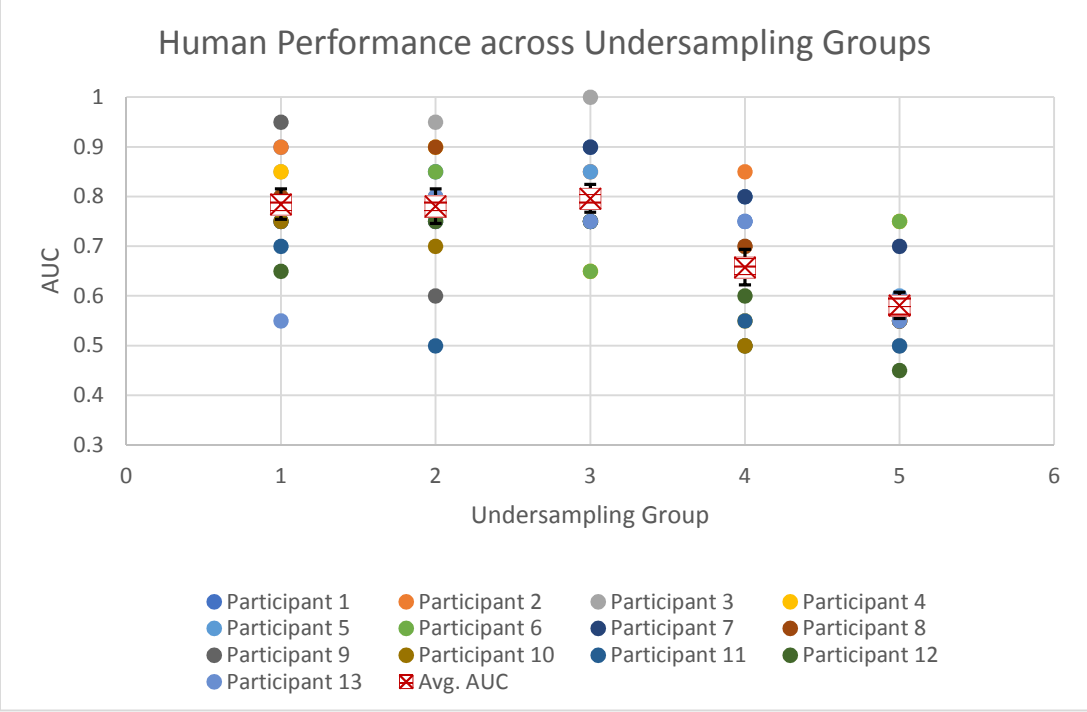


Figure 6: Plotted values of each participant's AUCs across all five levels of noise (Top) and all five levels of undersampling (Bottom).

Note: for the horizontal axis, group '1' refers to ***noise_1*** and ***undersample_1*** for each graph respectively, group '2' refers to ***noise_2*** and ***undersample_2***, and so on. Standard errors for the mean are depicted on the graph.

From these graphs, the first thing we observe is that human performance seems to steadily decrease as the amount of noise increases. However, human performance remains fairly steady across the first three undersampling levels, and then dips afterwards. If an ultrasound design team were able to directly measure human AUC by conducting an experiment as done in this study, and wanted to choose the image system parameters that would lead to the best AUC, they would choose the settings of ***noise_1***, though it would likely be the most resource intensive of all five noise configurations (it is the one that captures the least amount of noise in the signal). For a sampling setting, they would choose ***undersample_3***, as it provides similar AUC performance as ***undersample_1*** and ***undersample_2***, but is able to do so more efficiently due to its lower sampling rate as compared to ***undersample_1*** and ***undersample_2***. In fact, the readers exhibited a slightly higher AUC for detecting the vessels in ***undersample_3***, but this anomaly is likely due to inherent randomness in the readers' scores.

3.2 Traditional Image Quality Metric Patterns

If we accept that human AUC is the true measure of image quality, then traditional metrics may be able to approximate this, potentially offering the benefit of estimating the true assessment of image quality without the expensive and time-consuming step of conducting a reader study with human participants. But if a traditional image quality metric is to be used as a guide to design an imaging system or algorithm, it should ideally respond to the factors that degrade image quality in the same way as AUC responds to be useful

In this section, I compare the patterns of each of the traditional metrics with AUC patterns, as a function of levels of the different levels of noise and undersampling degradation. If the metrics and human AUC respond identically (up to a constant factor), this would imply that

human performance can be trivially predicted with these metrics. But as we can see, although some metrics show similar behavior to AUC, none of the metrics responds identically. The metric patterns for the noise groups are summarized in the graphs below and are shown along with the human AUC data for comparison where each point is the average (metric or AUC) value for images in the indicated noise level. Standard errors are also indicated on the graphs.

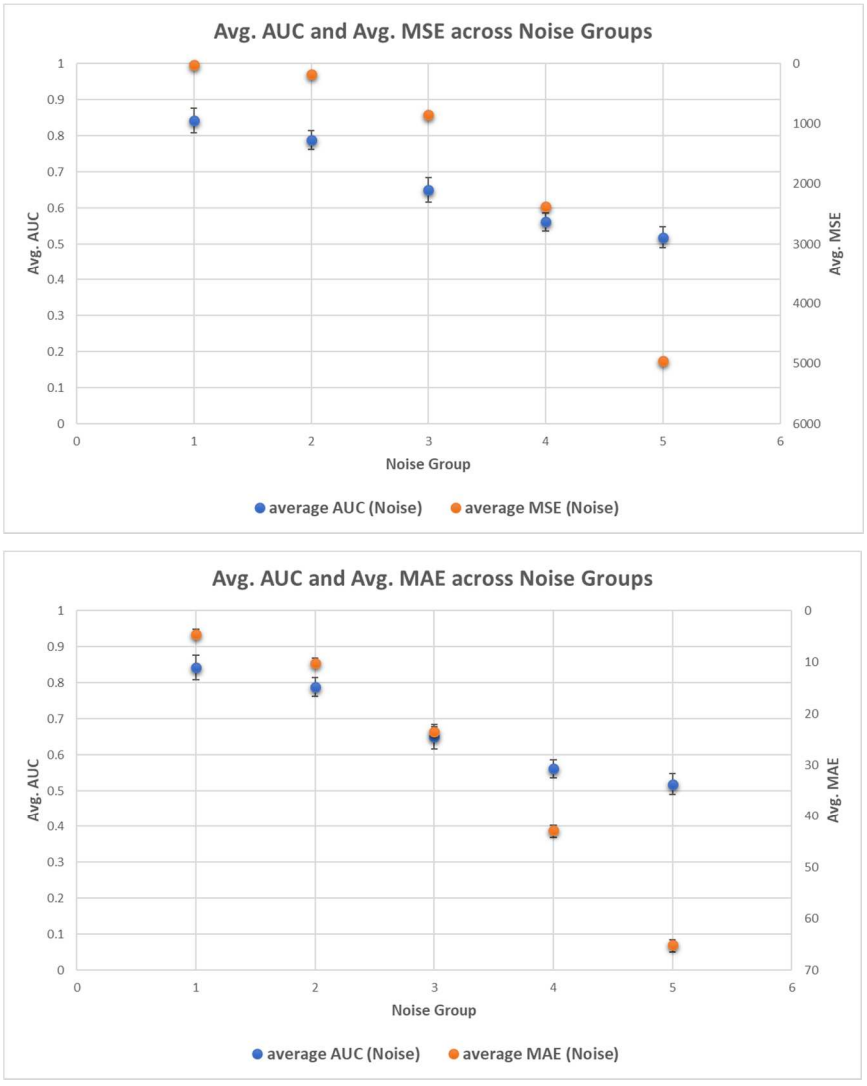


Figure 7: Plots of the average values of MSE and MAE across each noise group, compared to the average AUC values of participants

The blue points across each graph are identical because they are the trend of average AUC values in the noise groups, but the orange dots are different across each graph as they represents the trend of a specific image metric: average MSE (Top), and average MAE (Bottom). Standard errors are also denoted on the graphs.

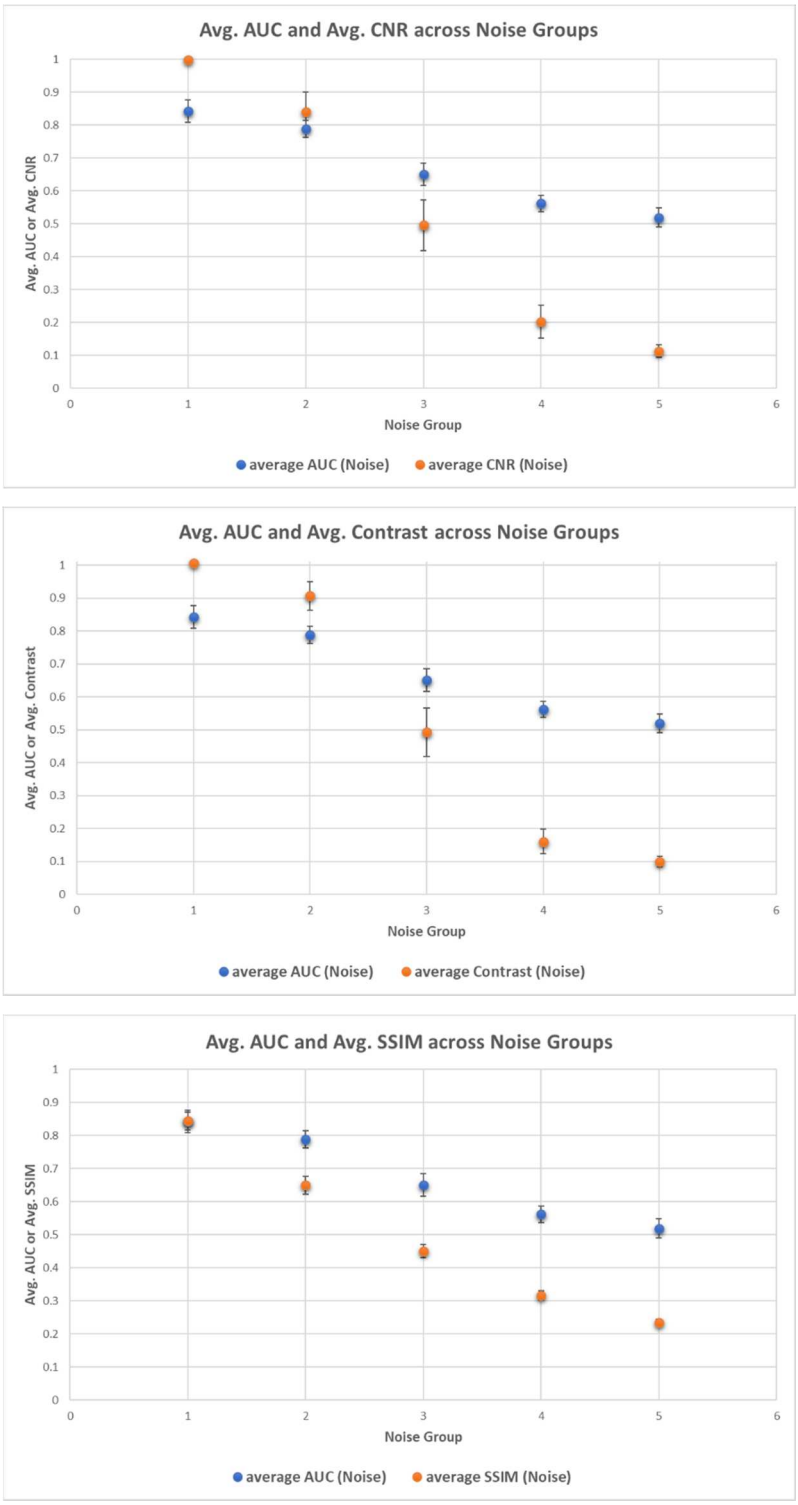


Figure 8: Plots of the average values of CNR, Contrast and SSIM across each noise group, compared to the average AUC values of participants

The blue points across each graph are identical because they are the trend of average AUC values in the noise groups, but the orange dots are different across each graph as they represents the trend of a specific image metric: average CNR (Top), and average Contrast (Middle) and average SSIM (Bottom). Standard errors are also denoted on the graphs.

The metric patterns for undersampling groups are summarized in the graphs below and are shown along with the human AUC data for comparison, where each point is the average (metric or AUC) value for images in the indicated undersampling group. Standard errors are also indicated on the graphs.

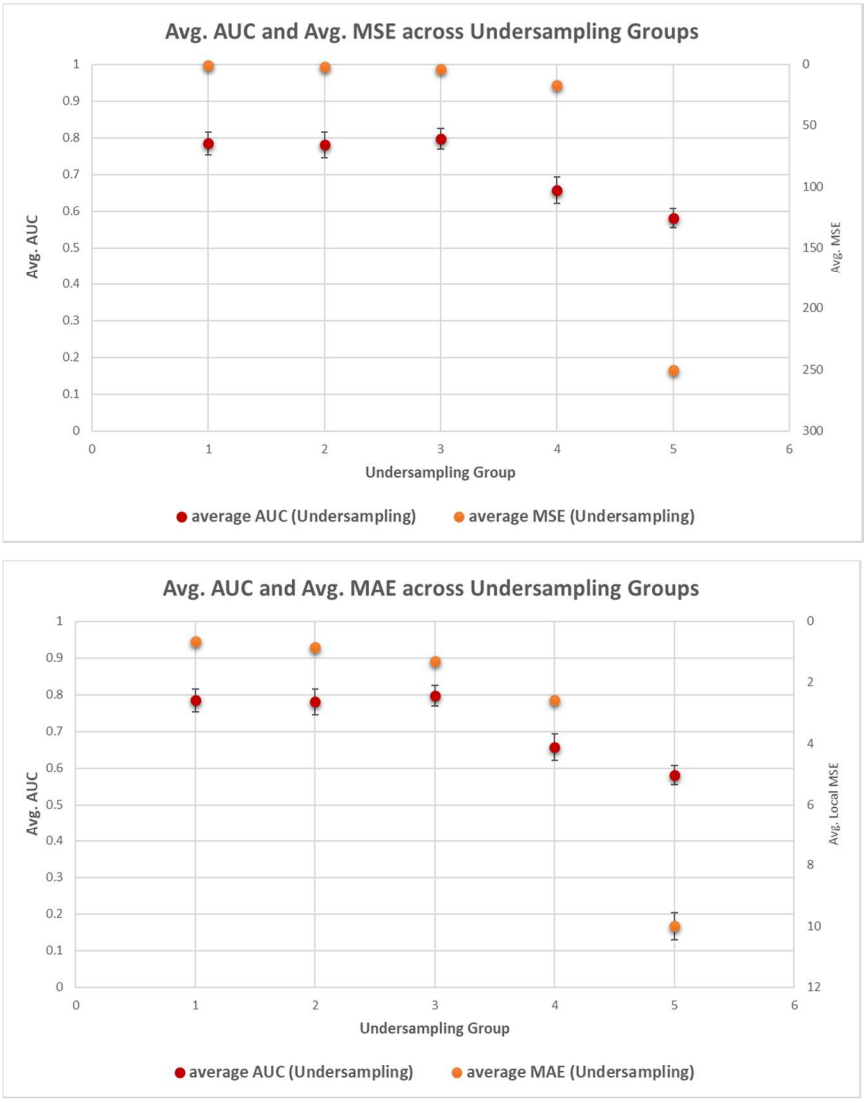
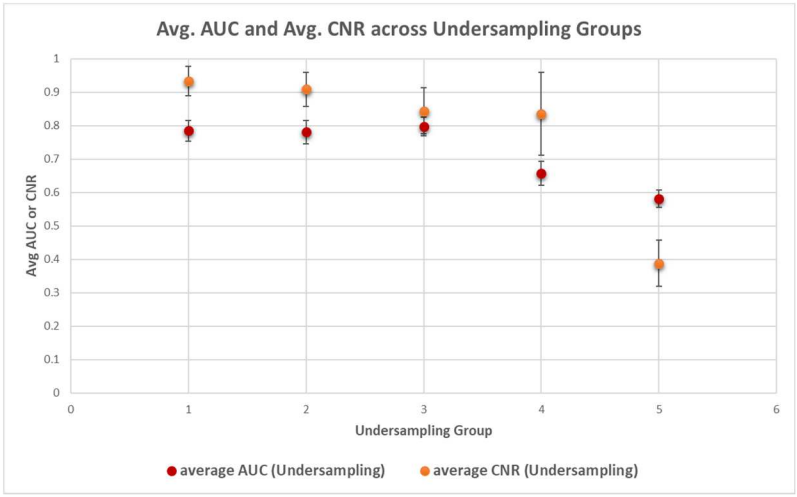
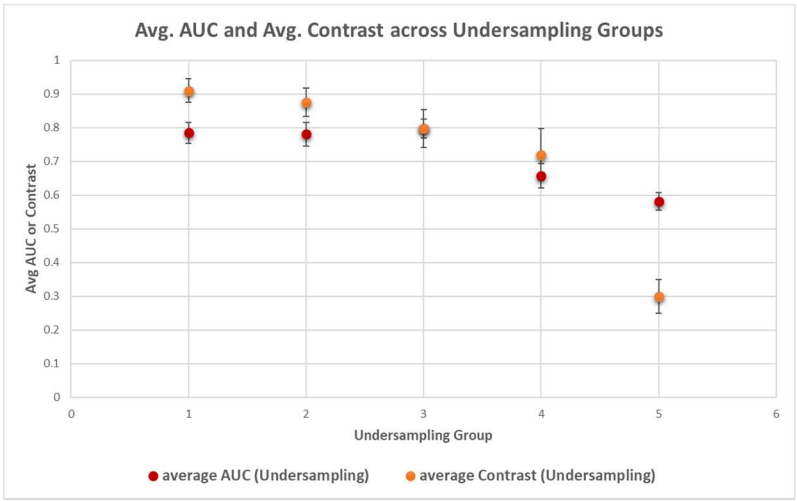


Figure 9: Plots of the average values of MSE and MAE across each undersampling group, compared to the average AUC values of participants

The red points across each graph are identical because they are the trend of average AUC values in the sampling groups, but the orange dots are different across each graph as they represents the trend of a specific image metric: average MSE (Top), and average MAE (Bottom). Standard errors are also denoted on the graphs.





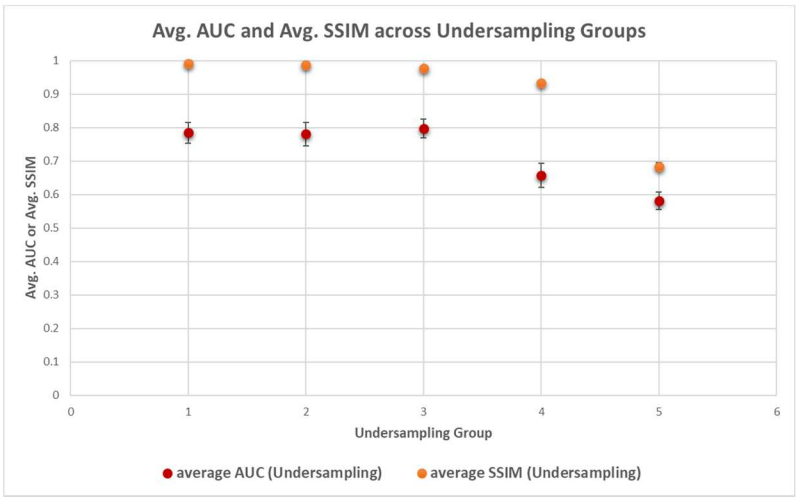


Figure 10: Plots of the average values of CNR, Contrast and SSIM across each undersampling group, compared to the average AUC values of participants

The red points across each graph are identical because they are the trend of average AUC values in the sampling groups, but the orange dots are different across each graph as they represents the trend of a specific image metric: average CNR (Top), and average Contrast (Middle) and average SSIM (Bottom). Standard errors are also denoted on the graphs.

When we study the metric patterns for the noise groups in Fig. 7 and Fig.8, we see that all of them decrease at each increasing noise level, similar to the human AUC pattern. Thus, if an ultrasound design team were to use these metrics as a measure of human AUC, they would make the choice of image system parameters of ***noise_1*****,** since it provides the best metric scores. And in doing this, they would make the correct choice, since we saw from the human experiment that ***noise_1*** indeed led to the best human AUC. All the metrics capture the decreasing pattern of human AUC performance across noise levels, albeit with varying complexity of relationship than others.

When we study the metric patterns for the undersampling groups in Fig. 9 and Fig. 10, we see a slightly different story. Though the AUC points remain flat and only begin dropping after the third undersampling group, the points for MSE, CNR, and SSIM remain relatively flat up through the fourth undersampling group. But the other two metrics (MAE and Contrast) instead of remaining flat, steadily decrease through the fourth undersampling group. All metrics show a relatively large dip between the fourth and fifth undersampling group. If an ultrasound design team were to consider all these metrics to gauge human AUC, they may be inclined make the choice of image system parameters of ***undersample_4*****,** since it provides metric scores very similar to ***undersample_1*****, *undersample_2*****, *undersample_3*****,** and ***undersample_4*** but with a more efficient sampling rate. Yet in doing this, they would make a misguided choice, since we saw from the human experiment that it was ***undersample_3*** that led to the best human AUC in terms of efficiency of sampling. Due to the differences in the metric patterns and AUC patterns for undersampling groups, we would expect that conclusions about the pattern of human performance from traditional image quality metrics may be a more difficult task than in the noise case, and that these metrics may misguide us at certain undersampling levels.

3.3 Comparing Traditional Image Quality Metrics Against Human AUC

A more direct way to compare the patterns of traditional image quality metrics and human-observer AUC is to plot them against one another. These give us a better idea if any of the metrics can be used to predict the AUC value at the different degradation levels. Below are the trends of AUC vs the various metrics, at various levels of noise (blue points) and undersampling (red points).

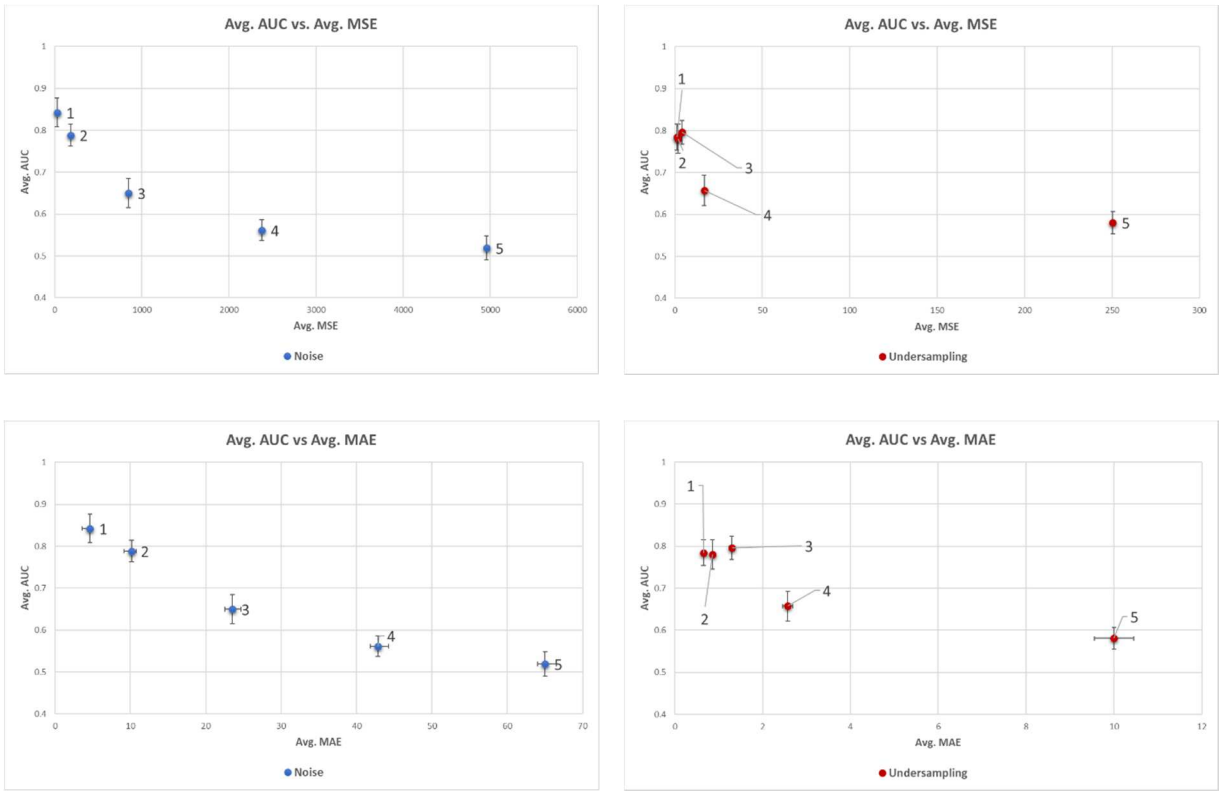


Figure 11: Plots of the average AUC values vs. average metric values of MSE (top row) and MAE (bottom row) metrics, across the noise (left column) and undersampling groups (right column)

Note: The number label next to each point represents the respective noise or undersampling group number, such as ***noise_1, undersampling_1*****,** etc.

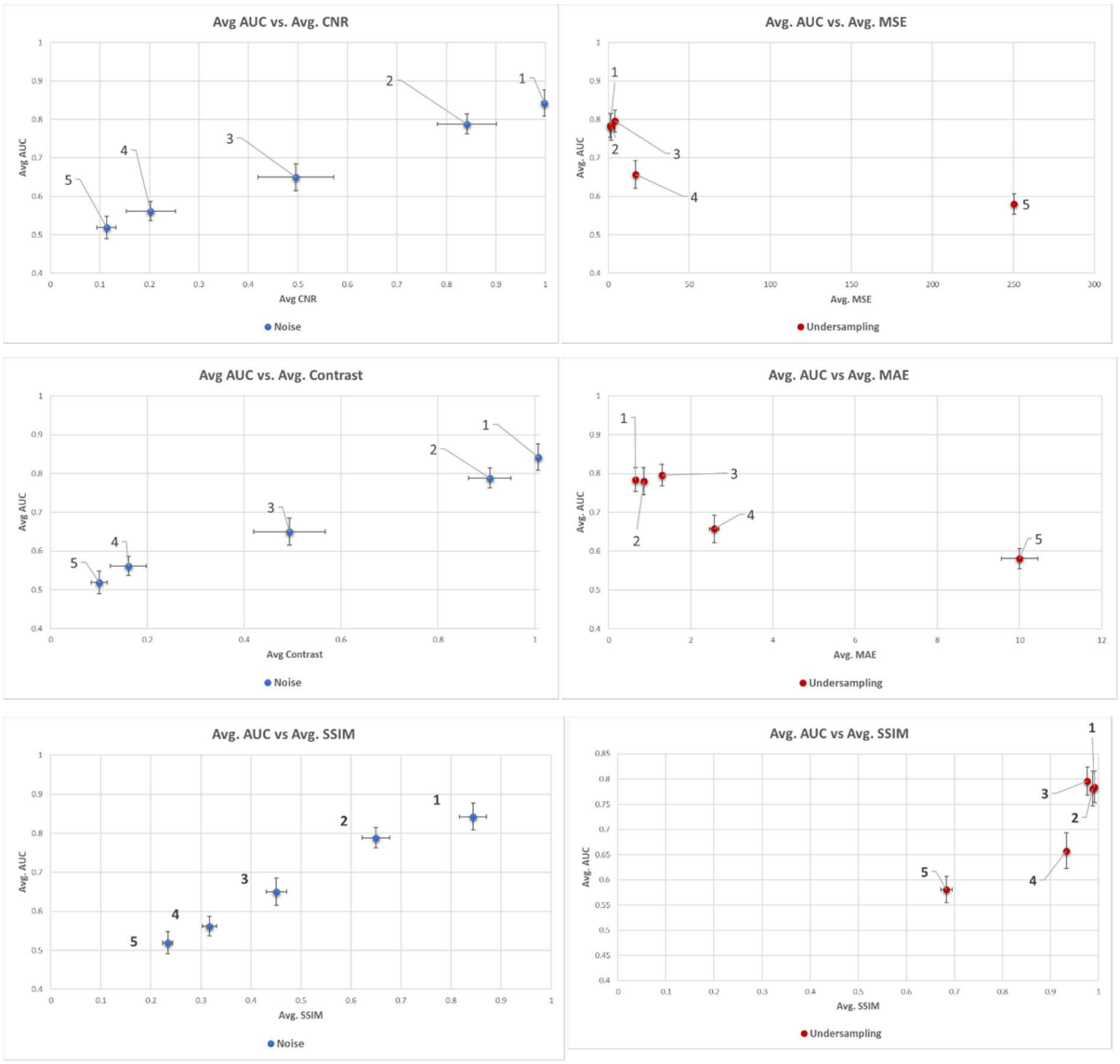


Figure 12: Plots of the average AUC values vs. average metric values of CNR (top row), Contrast (middle row) and SSIM (bottom row)metrics, across the noise (left column) and undersampling groups (right column)

Note: The number label next to each point represents the respective noise or undersampling group number, such as ***noise_1, undersampling_1*,** etc.

When examining the trends of AUC verses the individual metrics, and comparing across noise and undersampling groups, we observe that all the trends for the noise groups are monotonic. In addition, for the noise groups, the AUC trends across three of the metrics (CNR, Contrast, and SSIM) look nearly linear. For MSE and MAE, the AUC trend for the noise groups

looks somewhat exponential. However, in the AUC vs metric graph for the undersampling groups, none of the trends are monotonic, and the relationships look rather complex. From exploring these patterns, we would expect that predicting AUC from image metrics from noise groups may be possible with a simple model, but for undersampling groups, a more complex model may be needed.

3.4 Predicting Human AUC from Traditional Image Metrics

To explore whether or not a traditional image metric could be used to predict AUC well, I chose one metric for the noise groups and one metric for the undersampling groups that I expected to give the best prediction results using regression models, based on our graphs above. After inspecting Figures 9 and 10, I decided that the metrics with the most straightforward relationships with AUC (and thus were most likely to result in successful prediction models) were CNR for the noise groups, and Contrast for the undersampling groups. These metrics were used to predict human AUC within each group across 100 resampling iterations as described in the methods section. The results of linear, quadratic, and cubic models are given in Table 3.

Table 3: Average performance of regression models in predicting AUC across 100 train/test split iterations

(Top): Results of various models using CNR to predict AUC across noise groups. Here, the linear model showed the best performance.

(Bottom): Results of various models using Contrast to predict AUC across undersampling groups. Here, the cubic model showed best performance

Models for Noise Groups	Average MAE	Average RMSE
Linear	**0.038**	**0.044**
Quadratic	0.04	0.047
Cubic	0.044	0.052

Models for Undersampling Groups	Average MAE	Average RMSE
Linear	0.05	0.06
Quadratic	0.053	0.064
Cubic	**0.048**	**0.057**

A linear model performed best on the noise groups, with an average test set MAE and RMSE of 0.038 and 0.044 respectively. The cubic polynomial performed best on the undersampling groups, with an average test set MAE and RMSE of 0.048 and 0.057 respectively. Both of these MAE values are less than 5% of the total range of values AUC can take (0-1). These prediction errors are small enough to draw useful conclusions about the effect of noise and sampling on human observer performance.

4. Discussion

This study showed that for ultrasound images, traditional image metrics have a variety of relationships with human performance, depending on each metric, and the type of degradation the image is subject to. For most metrics, there seems to be a liner relationship between the metric value and human performance in noise groups. However, all metrics have nonlinear relationships with human performance with respect to undersampling.

Predictions of human performance can be made from metrics with limited accuracy. In this study, I chose to predict AUC in the noise and undersampling groups with CNR and Contrast respectively. This was based on the fact that in Figures 9 and 10, these metrics seemed to exhibit the simplest relationship with AUC amongst other metrics. However, further study is needed to see if other metrics can be used to predict AUC with comparable or increased performance.

Because the dataset and number of participates was rather limited, this study's results need to be interpreted with caution. More data collection through testing more readers, as well as including more noise and undersampling groups (such as 10 groups of noise and undersampling each, rather than 5) will likely give a more accurate picture of the relationship between these metrics and human performance. In addition, since prediction models were fit on a limited dataset, the degree to which the results generalize to other undersampling rates or noise levels is unknown.

Further model optimizations should also be done, such as combining multiple metrics to predict AUC, or using more sophisticated models. For example, neural networks might be able to capture the nonlinearities shown in the relationship between metrics and human performance on undersampling groups. In addition, this thesis studied the effect of noise and undersampling on human AUC in isolation, though in the real world, images may be subject to *both* noise and undersampling. A further study should be done to examine these cross effects, and whether or not the relationships between the traditional image metrics and human-observor AUC that was found when varying noise and undersampling carry over when varying noise and sampling together., The methods and results of this thesis prove to be a first step towards a potentially

simpler way of predicting human observer performance – through traditional image metrics rather than model observers.

References

1. Bhidé, A., Datar, S., & Stebbins, K. (2019). Case Histories of Significant Medical Advances: Ultrasound. Harvard Business School. https://www.hbs.edu/faculty/Publication%20Files/20-003_3e85b83a-e765-4f93-9bbc-651517b6738e.pdf
2. Klibanov, A. L., & Hossack, J. A. (2015). Ultrasound in Radiology: From Anatomic, Functional, Molecular Imaging to Drug Delivery and Image-Guided Therapy. Investigative radiology, 50(9), 657–670. *https://doi.org/10.1097/RLI.0000000000000188*
3. Alonso, J. V., Turpie, J., Farhad, I., & Ruffino, G. (2019). Protocols for Point-of-Care-Ultrasound (POCUS) in a Patient with Sepsis; An Algorithmic Approach. Bulletin of emergency and trauma, 7(1), 67–71. *https://doi.org/10.29252/beat-0701010*
4. Arnold, M. J., Jonas, C. E., & Carter, R. E. (2020). Point-of-Care Ultrasonography. American family physician, 101(5), 275–285.
5. Melgarejo, S., Schaub, A., & Noble, V. E. (2017, October 31). Point of Care Ultrasound: An Overview. American College of Cardiology. https://www.acc.org/latest-in-cardiology/articles/2017/10/31/09/57/point-of-care-ultrasound
6. Van Schaik, G., Van Schaik, K. D., & Murphy, M. C. (2019). Point-of-Care Ultrasonography (POCUS) in a Community Emergency Department: An Analysis of Decision Making and Cost Savings Associated With POCUS. Journal of ultrasound in medicine : official journal of the American Institute of Ultrasound in Medicine, 38(8), 2133–2140. *https://doi.org/10.1002/jum.14910*
7. Kobal, S. L., Trento, L., Baharami, S., Tolstrup, K., Naqvi, T. Z., Cercek, B., Neuman, Y., Mirocha, J., Kar, S., Forrester, J. S., & Siegel, R. J. (2005). Comparison of

effectiveness of hand-carried ultrasound to bedside cardiovascular physical examination. The American journal of cardiology, 96(7), 1002–1006. *https://doi.org/10.1016/j.amjcard.2005.05.060*

8. Narula, J., Chandrashekhar, Y., & Braunwald, E. (2018). Time to Add a Fifth Pillar to Bedside Physical Examination: Inspection, Palpation, Percussion, Auscultation, and Insonation. JAMA cardiology, 3(4), 346–350. *https://doi.org/10.1001/jamacardio.2018.0001*
9. Sassaroli, E., Crake, C., Scorza, A., Kim, D. S., & Park, M. A. (2019). Image quality evaluation of ultrasound imaging systems: advanced B-modes. Journal of applied clinical medical physics, 20(3), 115–124. *https://doi.org/10.1002/acm2.12544*
10. Wang, Z., & Bovik, A. C. (2009, January). Mean squared error: Love it or leave it? A new look at Signal Fidelity Measures. IEEE Signal Processing Magazine. 26(1), pp. 98-117. doi: 10.1109/MSP.2008.930649.
11. Wang, Z., & Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004, April). Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing. 13(4), pp. 600-612.
12. Thung, K., & Raveendran, P. (2009, December, 14-15). A survey of image quality measures. 2009 International Conference for Technical Postgraduates (TECHPOS), Kuala Lumpur. https://ieeexplore.ieee.org/document/5412098
13. International Commission on Radiation Units & Measurements. (1995). Medical Imaging - The Assessment of Image Quality (Report 54). https://www.icru.org/home/reports/medical-imaging-the-assessment-of-image-quality-report-54

14. Solomon, J., & Samei, E. (2016). Correlation between human detection accuracy and observer model-based image quality metrics in computed tomography. Journal of medical imaging (Bellingham, Wash.), 3(3), 035506. *https://doi.org/10.1117/1.JMI.3.3.035506*

15. *Yu, L., Leng, S., Chen, L., Kofler, J. M., Carter, R. E., & McCollough, C. H. (2013). Prediction of human observer performance in a 2-alternative forced choice low-contrast detection task using channelized Hotelling observer: impact of radiation dose and reconstruction algorithms. Medical physics, 40(4), 041908. https://doi.org/10.1118/1.4794498*

16. He, X., & Park, S. (2013). Model observers in medical imaging research. Theranostics, 3(10), 774–786. *https://doi.org/10.7150/thno.5138*

17. Zhang, Y., Leng, S., Yu, L., Carter, R. E., & McCollough, C. H. (2014). Correlation between human and model observer performance for discrimination task in CT. Physics in medicine and biology, 59(13), 3389–3404. *https://doi.org/10.1088/0031-9155/59/13/3389*

18. Insana, M. F., & Hall, T. J. (1994). Visual detection efficiency in ultrasonic imaging: A framework for objective assessment of image quality. *Journal of the Acoustical Society of America, 95*(4), 2081–2090. *https://doi.org/10.1121/1.408669*

19. Hajian-Tilaki K. (2013). Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. Caspian journal of internal medicine, 4(2), 627–635.

20. Krupinski E. A. (2010). Current perspectives in medical image perception. Attention, perception & psychophysics, 72(5), 1205–1217. https://doi.org/10.3758/APP.72.5.1205

21. Li, H., Wu, J., Miao, A., Yu, P., Chen, J., & Zhang, Y. (2017). Rayleigh-maximum-likelihood bilateral filter for ultrasound image enhancement. Biomedical engineering online, 16(1), 46. *https://doi.org/10.1186/s12938-017-0336-9*

22. Green, D.M. (1964). General prediction relating yes-no and forced-choice results. *Journal of the Acoustical Society of America,* 36(5). https://doi.org/10.1121/1.2143339

23. Paskaš, M. (2009). Two approaches for log-compression parameter estimation: Comparative study. Serbian Journal of Electrical Engineering, 6(3), 419-425. *https://doi.org/10.2298/SJEE0903419P*

24. Ma, J., Karadayi, K., Ali, M. & Kim Y. (2011, October, 18-21). Software-based ultrasound phase rotation beamforming on multi-core DSP. *2011 IEEE International Ultrasonics Symposium*, Orlando, FL. doi: 10.1109/ULTSYM.2011.0121.

25. Springer Link. (2000). Nyquist theorem. In Computer Science and Communications Dictionary. Springer, Boston, MA.