

The quest for “diagnostically lossless” medical image compression: A comparative study of objective quality metrics for compressed medical images

Ilona Kowalik-Urbaniak^a, Dominique Brunet^a, Jiheng Wang^b, David Koff^c, Nadine Smolarski-Koff^c, Edward R. Vrscay^a, Bill Wallace^d, Zhou Wang^b

^aDept. of Applied Mathematics, University of Waterloo, Waterloo, ON, Canada

^bDept. of Elect. and Comp. Eng, University of Waterloo, Waterloo, ON, Canada

^cDept. of Radiology, McMaster University, Hamilton, ON, Canada

^dAgfa Healthcare Inc., Waterloo, ON, Canada

ABSTRACT

Our study, involving a collaboration with radiologists (DK,NSK) as well as a leading international developer of medical imaging software (AGFA), is primarily concerned with improved methods of assessing the diagnostic quality of compressed medical images and the investigation of compression artifacts resulting from JPEG and JPEG2000. In this work, we compare the performances of the Structural Similarity quality measure (SSIM), MSE/PSNR, compression ratio CR and JPEG quality factor Q, based on experimental data collected in two experiments involving radiologists. An ROC and Kolmogorov-Smirnov analysis indicates that compression ratio is not always a good indicator of visual quality. Moreover, SSIM demonstrates the best performance, i.e., it provides the closest match to the radiologists' assessments. We also show that a weighted Youden index¹ and curve fitting method can provide SSIM and MSE thresholds for acceptable compression ratios.

Keywords: image quality assessment, JPEG, JPEG2000, medical images, medical image compression, SSIM, compression ratio, image quality.

1. INTRODUCTION

Given the explosive growth of digital image data being generated, medical communities worldwide have recognized the need for increasingly efficient methods of storage, display and transmission of medical images. There is also a general acknowledgement that lossless compression techniques, with low rates of data compression, are no longer adequate and that it is necessary to consider higher compression rates. Since the result, *lossy compression*, involves a loss of information and possibly visual quality, it is absolutely essential to be able to determine the degree to which a medical image can be compressed before its *diagnostic quality* is compromised.

The radiological community has not yet accepted a single objective assessment method for the quality of medical images. Recommended compression ratios (CR) for various modalities and anatomical regions have been published.^{2,3} To date, these recommendations have been based on experiments in which radiologists *subjectively* assess the diagnostic quality of compressed images. This study is primarily concerned with improved methods of *objectively assessing* the diagnostic quality of compressed medical images.

The “quality” of a compressed image can be characterized objectively in several ways. Radiologists most often employ the mean squared error (MSE) and its close relative, PSNR, even though they are known to correspond poorly to visual quality. The failure of MSE and PSNR is partially due to the fact that spatial relationships are ignored by the L^2 metric⁴ on which they are based. A more recent image fidelity measure, the *SSIM index*,⁵ measures the difference/similarity between two images by combining three components of the human visual system – luminance, contrast and structure. The result is a much improved assessment of visual quality.

In this paper, we examine whether compression ratio, mean squared error (MSE) and PSNR actually serve as reliable indicators of diagnostic quality. By this we mean “model the perception of trained radiologists in a satisfactory way.” We also investigate the *quality factor* (QF), the sole input parameter in the JPEG compression

algorithm, since it has also been employed as a reference for quality assessment. The performances of the above indicators are compared to that of the structural similarity index (SSIM), based on experimental data collected in two experiments involving radiologists. A second goal of this work has been to provide a method of determining of acceptable compression thresholds using the data from our experiments.

2. METHODS: SUBJECTIVE EXPERIMENT DESIGN

Two subjective experiments were designed in order to assess both the local and global prediction of the image quality assessments being examined. The first experiment, designed for a global analysis, employed ten CT slices - five neurological and five upper body images - extracted from volumes stored in the Cancer Imaging Archive.⁶ These images were first windowed according to their default viewing parameters (window width and window centre) in order to reduce their bit-depth from 16 to 8 bits per pixel (bpp). Each of the resulting 512 x 512 pixel, 8 bpp, images were compressed at five different compression ratios using both the JPEG and JPEG2000 compression algorithms. (Since JPEG employs only the quality factor as input, it was adjusted in order to produce compression ratios as close as possible to those used for the JPEG2000 compression.) Preliminary visual observations were used to select the compression ratios employed in the experiment. The range of compression ratios was intended to represent a wide variety of visual qualities, from barely noticeable to fairly noticeable distortion.

An image viewer was constructed specifically for this study in order to provide an easy-to-use graphical interface for the radiologists. The viewer displayed a compressed image beside its uncompressed counterpart without zoom. (No zooming was permitted in this experiment.) The ten compressed images were presented randomly and independently to each subject. During the course of the experiment, each compressed image was presented twice to each radiologist, but without the radiologists' knowledge. The subjects were not made aware of the compression ratios or quality factors of the compressed images. Two buttons were placed at the bottom of the user interface: acceptable and unacceptable. A confirmation was requested before passing to the next stimulus. In the second experiment, designed for a local analysis, six CT slices - four brain images and two body images - were compressed with JPEG at five different levels. Ten regions - 35 x 35 pixel blocks - were manually selected from each of these images. The regions were chosen in order to obtain both a variety of features as well as local image quality. Ten pairs of buttons in the bottom of the interface allowed the subjects to rate each region as either acceptable or unacceptable. Once again, the experiment was repeated with the same medical images in a random order.

Two radiologist subjects participated in each of the experiments. (They were not specifically specialists for the sites and types of images presented.) The subjects were instructed to flag an image as unacceptable in the case of any noticeable distortion. The first experiment (global analysis) was held in an afternoon session while the second experiment (local/regional analysis) took place during the morning. Each experiment lasted about one hour.

3. OBJECTIVE QUALITY METRICS

Since some regions of images are of much less interest than others in terms of diagnosis, an automatic segmentation of the images employed in this study (that is, the uncompressed and all compressed images) was performed in order to remove the background and bony anatomy. In the 16-to-8 bit tone-mapping of the images, background pixels always assumed a zero value, while those corresponding to bony anatomy generally assumed values near or equal to 255. The following simple segmentation operation was therefore sufficient: First threshold the images to separate the foreground from the background and then perform a fill operation in order to include the black pixels within the body part being imaged. A similar operation of thresholding followed by filling was performed in order to remove the skull. The removal of objects such as the couch or the scalp was also automatically performed by selecting only the largest region according to an 8-neighbour connectivity. The couch and the skull are successfully removed, while the bony anatomy inside the body is preserved. Some examples of segmentation are presented in Figure 1. All computations reported below were performed on thresholded images.

In the following discussion, we let f denote an $M \times N$ digital image and g its compressed counterpart. The standard measure of error between f and g is the Mean Squared Error (MSE), defined as follows,

$$\text{MSE}(f, g) = \frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N (f(i, j) - g(i, j))^2.$$

The MSE essentially defines a distance between f and g . The more distortion there is in the compressed image g , the higher the $\text{MSE}(f, g)$ value. (If f and g are identical, then $\text{MSE}(f, g) = 0$.) Image quality is often expressed in terms of the Peak Signal-to-Noise ratio (PSNR), derived from MSE as follows,

$$\text{PSNR}(f, g) = 10 \log_{10} \left(\frac{R^2}{\text{MSE}(f, g)} \right).$$

Here R is the dynamic range, i.e. $0 \leq f(i, j) \leq R$. (For an N bit-per-pixel image, $R = 2^N - 1$. Note that if f and g are close to each other, i.e., $\text{MSE}(f, g)$ is low, then $\text{PSNR}(f, g)$ is high. On the other hand, if f and g are far from each other, i.e., $\text{MSE}(f, g)$ is high, then $\text{PSNR}(f, g)$ is low.

The proposed quality measure is a variation of the Structural Similarity (SSIM) Index.⁵ First of all, the SSIM index between two images f and g is obtained by computing the following three terms which are important in the human visual system,

Luminance $l(f, g)$, estimated by the mean:

$$\mu_f = \frac{1}{NM} \sum_{i=1}^N \sum_{j=1}^M f(i, j)$$

Contrast $c(f, g)$, measured by variance:

$$\sigma_f^2 = \frac{1}{(NM - 1)} \sum_{i=1}^N \sum_{j=1}^M (f(i, j) - \mu_f)^2$$

Structure $s(f, g)$, measured by covariance:

$$\sigma_{fg} = \frac{1}{(NM - 1)} \sum_{i=1}^N \sum_{j=1}^M (f(i, j) - \mu_f)(g(i, j) - \mu_g)$$

The above terms are combined as follows in order to compute the SSIM index between images f and g :

$$\text{SSIM}(f, g) = \left(\frac{2\mu_f\mu_g + C_1}{\mu_f^2 + \mu_g^2 + C_1} \right) \left(\frac{2\sigma_f\sigma_g + C_2}{\sigma_f^2 + \sigma_g^2 + C_2} \right) \left(\frac{\sigma_{fg} + C_3}{\sigma_f\sigma_g + C_3} \right) \text{ over } m \times n \text{ pixel neighborhood.}$$

The SSIM index ranges between 0 and 1. As its name suggests, SSIM measures the similarity between f and g . The closer f and g are to each other, the closer $\text{SSIM}(f, g)$ is to the value 1. If f and g are identical, then $\text{SSIM}(f, g) = 1$.

The (non-negative) parameters C_1 , C_2 and C_3 are stability constants of relatively small magnitude, which are designed to avoid numerical “blowups” which could occur in the case of small denominators. For natural images, there are some recommended default values for these parameters.⁷ On the other hand, the question of optimal values for these stability constants for medical images is still an open one. The smaller the values of these constants, the more sensitive the SSIM index is to small image textures such as noise. In our study below, we shall examine a range of values for the stability constant (only one will be used, as explained below) in order to determine the value(s) which are optimal for the assessment of the diagnostic quality of medical images.

Note that in the special case $C_3 = C_2/2$, the following simplified, two-term version of the SSIM index is obtained:

$$SSIM(f, g) = \left(\frac{2\mu_f\mu_g + C_1}{\mu_f^2 + \mu_g^2 + C_1} \right) \left(\frac{2\sigma_{fg} + C_2}{\sigma_f^2 + \sigma_g^2 + C_2} \right).$$

In this work, we employ a variation of the above SSIM index by considering only the second term, i.e., the structure term. The first term, i.e., the luminance term, will not be considered since it does not make an impact on the quality score. This is because the luminance does not change visibly for the images and compression ratios encountered in this study. In summary, we consider the following SSIM index, which involves only a single stability parameter,

$$SSIM(f, g) = \left(\frac{2\sigma_{fg} + C}{\sigma_f^2 + \sigma_g^2 + C} \right).$$

The reader will note that there is a further complication because of the “apples-oranges” nature of MSE and SSIM, i.e., “error” vs. “similarity”: Recall that if f and g are “close”, then $MSE(f, g)$ is near 0 but $SSIM(f, g)$ is near 1. In order to be able to compare the quality assessments of both indices more conveniently, we define the following quantity,

$$SMSE(f, g) = 1 - MSE(f, g)/D$$

where D is a constant. In this paper D was chosen to be 255.

Using this definition, we now have that if f and g are “close”, then $SSIM(f, g)$ and $SMSE(f, g)$ are near 1.

Finally, we mention another variation in the computation of the SSIM, that is, the computation of the local SSIM. The above discussion of SSIM involved a computation of the similarity of two entire images f and g , in other words, the global similarity of the images. It is often useful to measure the local similarity of images, i.e., the similarity of corresponding regions or pixel subblocks between images. For this reason, one can employ any or all of the above formulas to compute SSIM values between corresponding $m \times n$ -pixel subblocks of two images.

One can proceed further and compute a SSIM quality map between two images f and g on a pixel-by-pixel basis as follows: At pixel (i, j) of each image, one constructs an $m \times n$ -pixel neighbourhood, or “window”, centered at (i, j) and then computes the SSIM index between the two neighbourhoods. This SSIM value is then assigned to location (i, j) . The result is a SSIM quality map which reveals local image similarities/differences between images f and g . A total SSIM score may then be computed by averaging over all the local SSIM values.

4. CLASSIFICATION PERFORMANCE METRICS

The Receiver Operating Characteristic (ROC) curve is a common tool for visually assessing the performance of a classifier in medical decision making. ROC curves illustrate the trade-off of benefit (true positives, TP) versus cost (false positives, FP) as the discriminating threshold is varied. For convenience, the contingency table is shown in Figure 2. At this point, we must qualify that due to the nature of the problem we are investigating, our definitions of FP and TP differ from those normally applied for the purposes of medical diagnosis. In this study, we wish to examine how well different “image quality indicators”, e.g., compression ratio, MSE, quality factor, SSIM, compare to the subjective assessments of image quality by radiologists. As such, we must assume that the “ground truth” for a particular experiment, i.e., whether or not a compressed image is acceptable or unacceptable, is defined by the radiologist(s). From this ground truth, we measure the effectiveness of each image quality indicator in terms of FP, TP, etc... This leads to the following definitions of P, N, TP, FP, etc.:

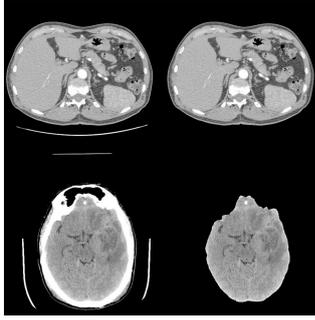


Figure 1. Examples of automatic image segmentation and object removal by thresholding and region growing.

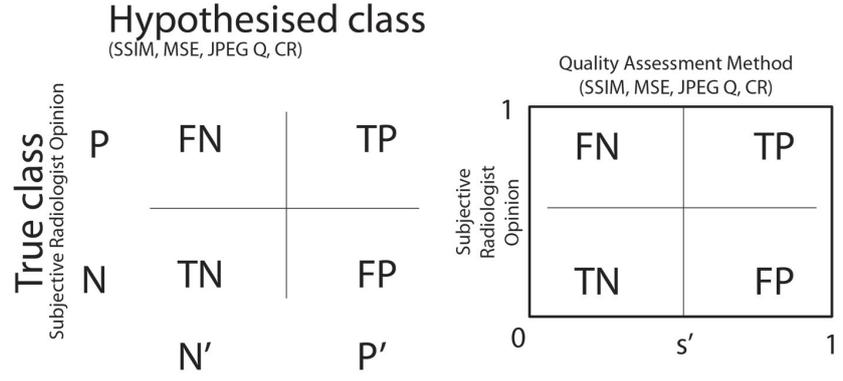


Figure 2. Contingency tables for Experiment 1.

1. P (or “1”) = $FP + TP$ total positives (acceptable) and N (or “0”) = $TN + FN$ total negatives (unacceptable): These refer to radiologists’ subjective opinions, which represent the True class. On the other hand, P' and N' belong to the Hypothesis class which, in our experiment, corresponds to a given quality assessment method, i.e., SSIM, MSE, quality factor, and compression ratio.
2. TP (true positives): images that are acceptable to both radiologists and a given quality assessment method.
3. TN (true negatives): images that are unacceptable to both radiologists and a given quality assessment method.
4. FN (false negatives): images that are acceptable to radiologists but unacceptable to a quality assessment method.
5. FP (false positives): images that are unacceptable to radiologists but acceptable to a given quality assessment algorithm.

This, of course, leads to the question, “What constitutes acceptability/unacceptability to a given quality assessment method?” This is defined with respect to the discrimination threshold s' associated with the method, where $0 \leq s' \leq 1$. Each threshold value s' generates a point on the ROC curve which corresponds to the pair of values $(FPR, TPR) = (1 - SP, SE)$, where SP denotes specificity and SE denotes sensitivity, i.e.,

$$FPR \text{ (false positive rate)} = FP/N = 1 - SP \text{ (specificity)}$$

$$TPR \text{ (true positive rate)} = TP/P = SE \text{ (sensitivity)}.$$

The point $(1, 1)$ on the ROC curve corresponds to $FN = 0$, i.e., no false negatives, and $TN = 0$, i.e., and no true negatives. The opposite scenario occurs at the point $(0, 0)$ which corresponds to $FP = 0$, i.e., no false positives, and $TP = 0$, i.e., no true positives. The point $(0, 1)$ corresponds to perfect classification, since $FP = 0$, i.e., no false positives, and $TN = 0$, i.e., no true negatives.

Once again, we mention that the definitions and labels in the contingency table associated with our experiments differ from those associated with a general detection/diagnosis experiment. Here, by false negative we mean that an image with a low objective quality score (hence unacceptable according to the quality method) has actually received a positive subjective score (acceptable, according to the radiologists).

In ROC analysis, as is well known, the performance of a discriminant (here, the SSIM, MSE, JPEG quality factor and compression ratio quality assessment methods) is often characterized by (1) the area under the curve

(AUC) and/or (2) the Kolmogorov-Smirnov (KS) statistic or test.

AUC method: The AUC can be computed by numerical integration using the trapezoidal rule. Larger AUC values correspond to better performance of the classifier. It is possible that two ROC curves cross. In this special situation a given classifier might demonstrate better performance for some threshold values whereas another classifier behaves better for other threshold values. In this case, a single AUC may not be the best predictor of the performance of a classifier.

Kolmogorov-Smirnov (KS) test: Given two cumulative probability distributions $P1(x)$ and $P2(x)$, their KS statistic is defined as follows,

$$KS(P1, P2) = \sup_x |P1(x) - P2(x)|.$$

In our study, $P1$ and $P2$ are the cumulative distributions of positive (1's) and negative (0's) radiologists' responses (respectively). The larger the difference between the two distributions, the better the performance of a given model. A generic situation is illustrated in Figure 3, 4.

With reference to Figure 2, for a given threshold s' in $[0, 1]$, we have the following relations,

Cumulative Probability Distribution of 0's = $TN/(TN + FP) = 1 - FPR$
 Cumulative Probability Distribution of 1's = $FN/(FN + TP) = 1 - TPR$.

Thus, the KS statistic translates to

$$KS = \sup_x |TPR(x) - FPR(x)|.$$

This is shown graphically in Figure 4.

Let us first recall the idea of the Youden index.¹ With reference to Figure 4, in which the TPR and FPR curves are plotted as a function of the threshold s , the Youden index $Y(s)$ associated with the threshold value s is simply the difference,

$$Y(s) = TPR(s) - FPR(s).$$

Now suppose that the maximum value of $Y(s)$, the so-called maximum Youden index, occurs at the threshold value $s = s'$. This implies that the points $TPR(s')$ and $FPR(s')$ in Figure 4 lie farthest away from each other which, in turn, implies that the maximum Youden index is the Kolmogorov-Smirnov (KS) value associated with the cumulative probability distributions of 0's and 1's discussed in the previous section.

There is also a connection between the maximum Youden index and ROC curves. Since the values $FPR(s')$ and $TPR(s')$ lie farthest from each other, the corresponding point $(FPR(s'), TPR(s'))$ on the ROC curve (which contains all points $(FPR(s), TPR(s))$) lies farthest away from the diagonal line joining $(0, 0)$ and $(1, 1)$. The threshold value $s = s'$ for which the Youden index is maximized is considered to be the optimal threshold value.

It is possible to define a weighted version of the Youden index in the following way. First of all, note that the Youden index $Y(s)$ introduced above may be expressed as follows,

$$Y(s) = TPR(s) - FPR(s) = SP(s) + SE(s) - 1,$$

Where $SE(s)$ and $SP(s)$ denote, respectively, the sensitivity and specificity associated with the threshold value s . The Youden index $Y(s)$ may be viewed as employing an equal weighting of false positives (FP) and false negatives (FN). It may be desirable to employ a non-equal weighting of these statistics in order to alter

their relative importance. This can be accomplished by means of a parameter $\lambda, 0 \leq \lambda \leq 1$ so that the weighted Youden index is given by

$$WY(s) = \lambda SP(s) + (1 - \lambda)SE(s) - 1 = -(\lambda FPR(s) + (1 - \lambda)FNR(s)).$$

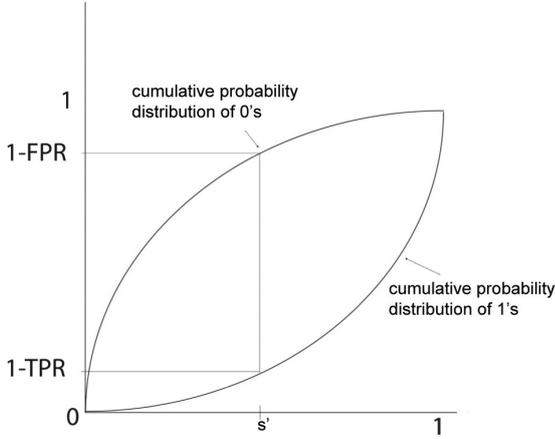


Figure 3. Kolmogorov-Smirnov distance.

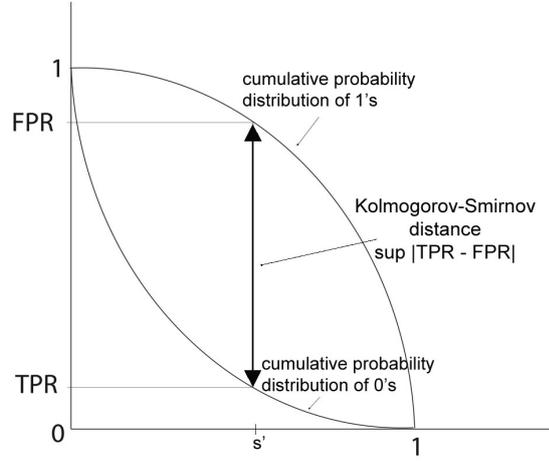


Figure 4. Kolmogorov-Smirnov distance.

5. THRESHOLD SELECTION

The selection of a threshold is accomplished by means of a curve-fitting model. In order to take into account the variability in the subjective quality assessment of compressed medical images, a logistic cumulative probability distribution is assumed to model the decision of a radiologist to either accept or not accept an image at a given objective score. A robust curve-fitting is then performed on plot of the average subjective score over all the radiologists and all the repetitions in function of the objective score. The threshold is selected so that the cumulative probability distribution model represents the desired level of confidence that the quality of the compressed image is diagnostically acceptable. For example, if one wants a 99% confidence, the threshold has to be selected at the value for which the fitted logistic curve is at 0.99.

The validation of the proposed metric is rigorously divided into three steps which employ data obtained from three disjoint sets. This is the correct method to perform statistical learning as presented in Ref. [8], although it is not always adhered to in the literature. First, in the training step the optimal parameters are determined using the weighted Youden index and the curve fitting model. Second, in the validation step the desired features as well as the threshold separating between acceptable and unacceptable compression are selected. Out of the 400 points (100 images, 2 radiologists, 1 repetition), 200 were used for training and 100 for validation in a 20-fold cross validation procedure. Finally, in the testing set the prediction accuracy is verified on fresh data (100 points).

6. RESULTS

6.1 First Experiment: global quality

The first part of our analysis involves all data points accumulated in Experiment 1. These data points include the two image types (brain CT and body CT) and the two compression methods (JPEG and JPEG2000). Figure 5 shows the ROC curves that correspond to the two quality measures SSIM and MSE. Figures 6 and 7 show the ROC curves corresponding to JPEG and JPEG2000 compressed images and the four quality measures SSIM,

MSE, JPEG quality factor Q and compression ratio CR. We observe that the ROC curve corresponding to CR demonstrates the worst performance, i.e., the lowest area under curve (AUC). Figures 8 and 9 show ROC curves associated with each of the two image types, i.e., brain CT and body CT. Such an analysis in terms of image types is particularly important since these two classes of images possess different characteristics (e.g. texture) which may yield different compression artifacts. Note that the ROC curves with the largest AUC correspond to the SSIM index quality measure.

In Figures 10, 11, 12 and 13 are presented the individual ROC curves corresponding to JPEG and JPEG2000 compression methods, the four quality measures and the two image types. Once again, the ROC curves associated with the SSIM index quality measure yield the largest AUCs. This suggests that of the four quality measures, SSIM performs the best in modeling the radiologists' subjective assessments of compressed images when the AUC is used as a performance indicator.

From Figures 10, 12 and 15, the AUCs associated with JPEG-compressed brain images are seen to be significantly lower than those associated with JPEG-compressed body images. This indicates that assessments of compressed CT brain images agree the least with the subjective assessments of radiologists. A closer examination of the data provides an explanation of this disparity. The radiologists perceived JPEG-compressed brain images as almost always acceptable, even in the cases when the quality of these images was deemed unsatisfactory in terms of SSIM or MSE. As a result, the ratios FP/TN, which are the false positive rates (FPR) plotted along the horizontal axis of ROC space, assume values of only 0 or 1. This explains why the two JPEG-compressed ROC curves are not only linear but almost horizontal.

Furthermore, Figures 10 and 12 show that in the case of compressed brain images, JPEG2000 demonstrates better agreement with the radiologists' opinions than JPEG. However, at the same compression ratios, more JPEG images were judged as acceptable by the radiologists. How do we explain this oddity? It is generally accepted that JPEG2000 "performs better" than JPEG for most classes of images - in other words, at a given compression ratio, the error (both visual as well as quantitative) between uncompressed and compressed images is less for JPEG2000 than JPEG. That being said, Koff et al. noted that the opposite was often observed in the case of brain images, i.e., JPEG performed better than JPEG2000. This anomaly is due to the bony skull in brain images.⁹ The sharp edges between the skull bone and neighboring regions (both background and interior regions) are difficult to approximate by any method. This is particularly the case for JPEG2000 since a larger number of wavelet coefficients are required to approximate these edges. JPEG can better accommodate these strong edges because of the 8x8 pixel block structure employed in the algorithm. As a result, the visual quality in the interior regions, which contain the most diagnostic information, can be lower for JPEG2000 than for JPEG.

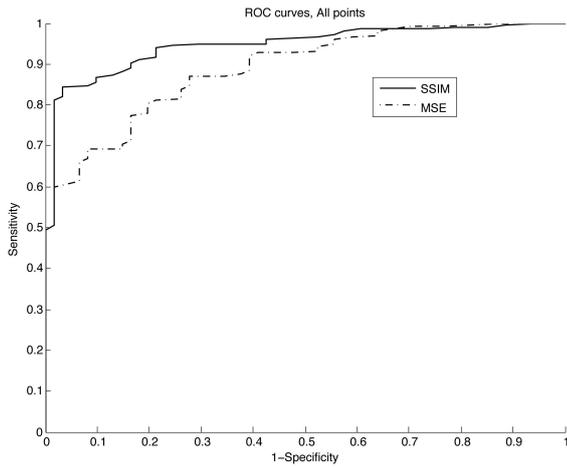


Figure 5. ROC curves corresponding to all data points. AUC SSIM = 0.9471 AUC MSE = 0.8900.

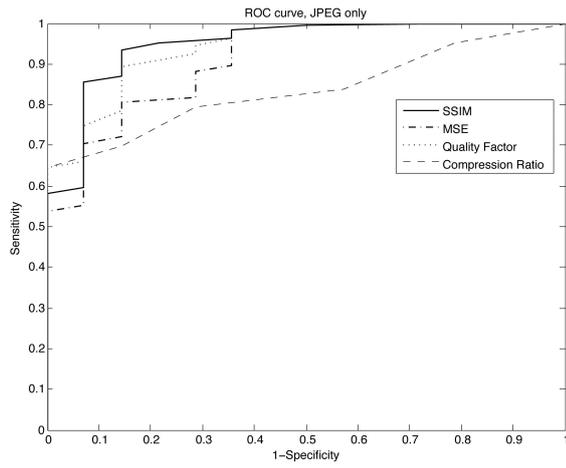


Figure 6. ROC curves corresponding to JPEG images and SSIM, MSE, Quality Factor, Compression Ratio. AUC JPEG SSIM = 0.9485 AUC JPEG MSE = 0.9101 AUC JPEG Quality Factor = 0.9401 AUC JPEG Compression Ratio = 0.8372.

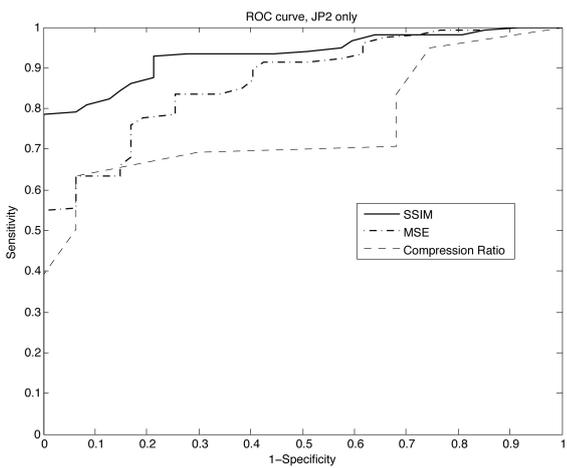


Figure 7. ROC curves corresponding to JPEG2000 images and SSIM, MSE, Compression Ratio. AUC JPEG2000 SSIM = 0.9330 AUC JPEG2000 MSE = 0.8691 AUC JPEG2000 Compression Ratio = 0.7573.

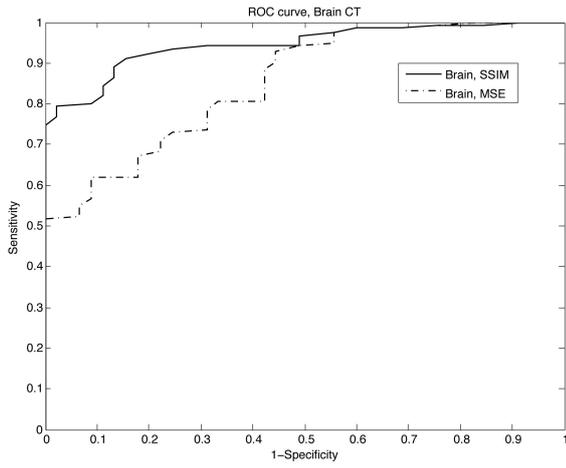


Figure 8. ROC curves corresponding to Brain CT. The area under the curve for each of the types is: Brain SSIM: AUC = 0.9447 Brain MSE: AUC = 0.8524.

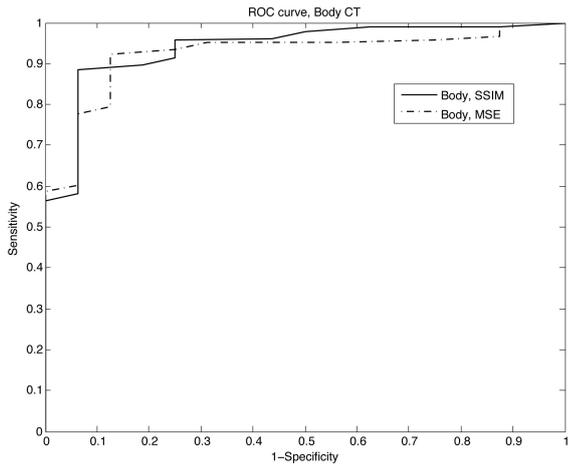


Figure 9. ROC curves corresponding to Body CT. The area under the curve for each of the types is: Body SSIM: AUC = 0.9389 Body MSE: AUC = 0.9226.

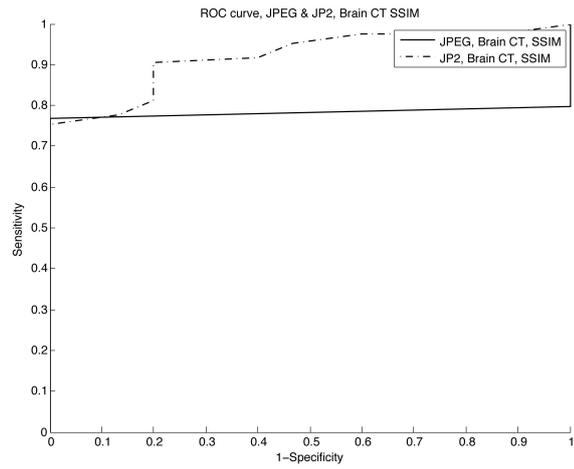


Figure 10. ROC curves corresponding to JPEG and JPEG2000 compressed Brain CT images and SSIM. JPEG Body SSIM: AUC = 0.7828 JP2 Body SSIM: AUC = 0.9204.

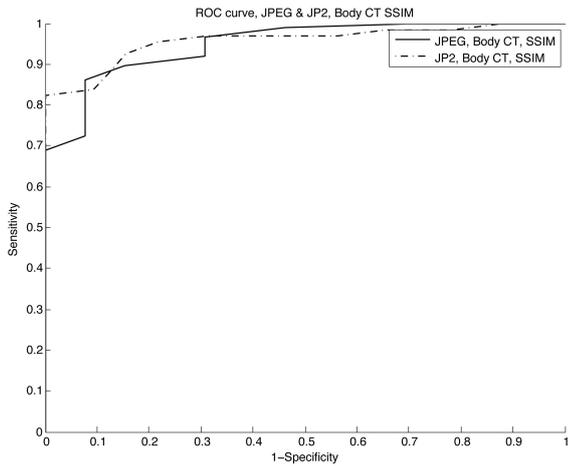


Figure 11. ROC curves corresponding to JPEG and JPEG2000 compressed Body CT images and SSIM. JPEG Brain SSIM: AUC = 0.9492 JP2 Brain SSIM: AUC = 0.9577.

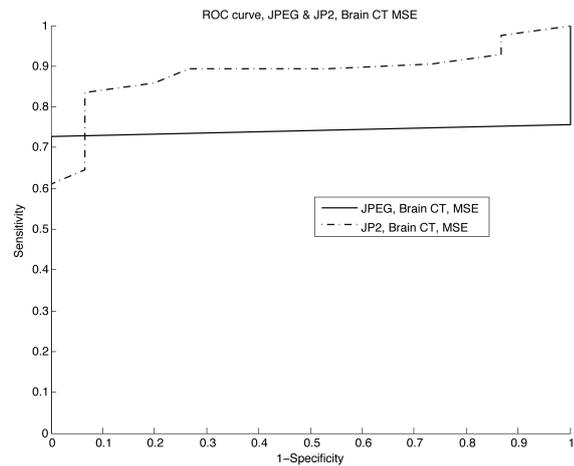


Figure 12. ROC curves corresponding to JPEG and JPEG2000 compressed Brain CT images and MSE. JPEG Body MSE: AUC = 0.7424 JP2 Body MSE: AUC = 0.8859.

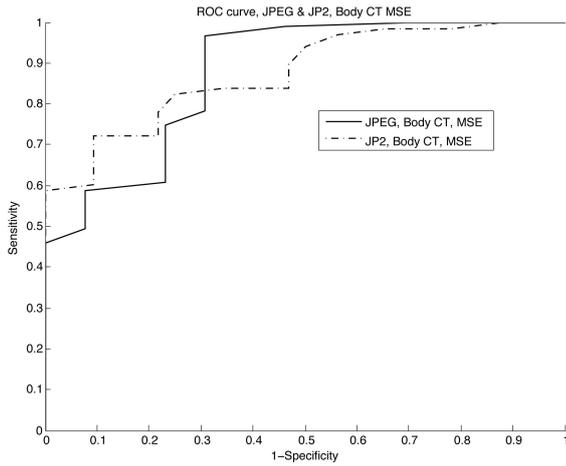


Figure 13. ROC curves corresponding to JPEG and JPEG2000 compressed Body JPEG Brain MSE: AUC = 0.8749 JP2 Brain MSE: AUC = 0.8750.

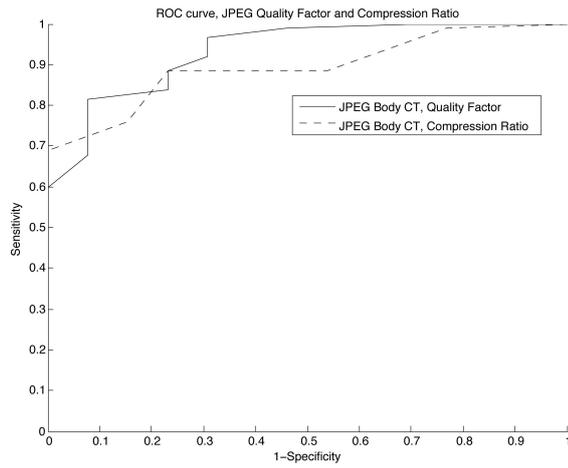


Figure 14. ROC curves corresponding to JPEG compressed Body images with respect to Quality Factor and Compression Ratio. JPEG Body, Quality Factor: AUC = 0.9332 JPEG Body, Compression Ratio AUC = 0.8926.

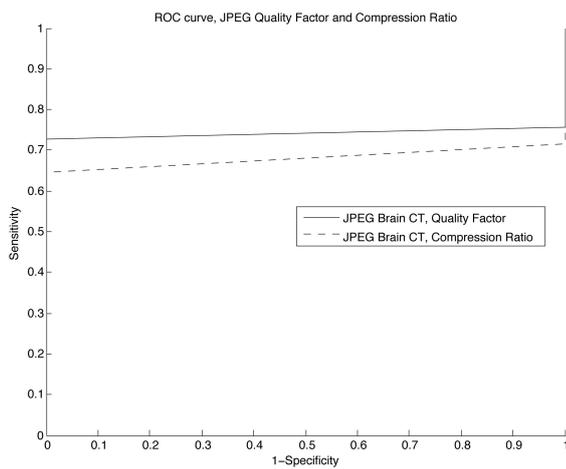


Figure 15. ROC curves corresponding to JPEG compressed Brain images with respect to Quality Factor and Compression Ratio. JPEG Body, Quality Factor: AUC = 0.7424 JPEG Body, Compression Ratio AUC = 0.6818.

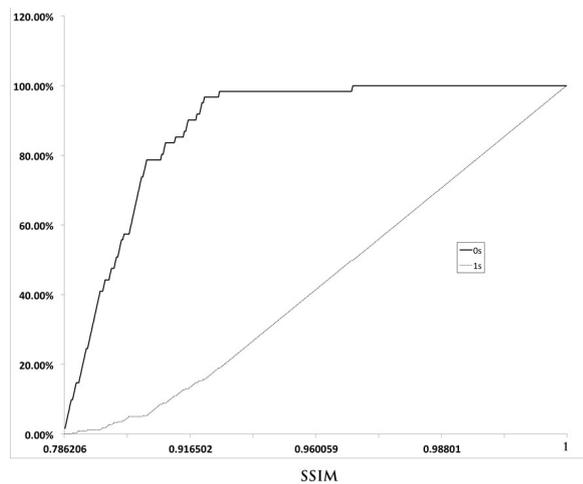


Figure 16. Cumulative distributions of subjective radiologist scores corresponding to SSIM. K-S = 81.09%.

We now summarize the results of applying the Kolmogorov-Smirnov (KS) test to the collected data. Figures 16 and 17 present the cumulative distributions of subjective radiologist scores (0's and 1's) corresponding to, respectively, SSIM and MSE quality measures for all data points. Figure 18 is a plot of the subjective radiologist scores for JPEG-compressed images using the JPEG quality factor Q as the quality measure. Finally, the distributions of 0's and 1's corresponding to compression ratio are shown in Figure 19. As expected, the KS statistic, i.e. the separation between the two cumulative distributions is largest for SSIM (81%), intermediate for JPEG quality factor Q (78%) and smallest for compression ratio CR (60%).

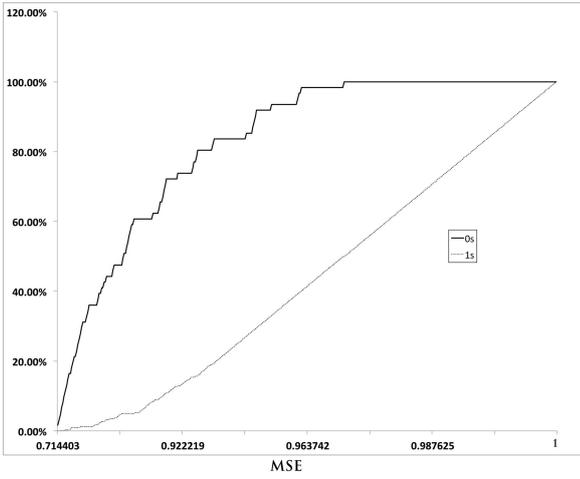


Figure 17. Cumulative distributions of subjective radiologist scores corresponding to MSE. K-S = 64.40%.

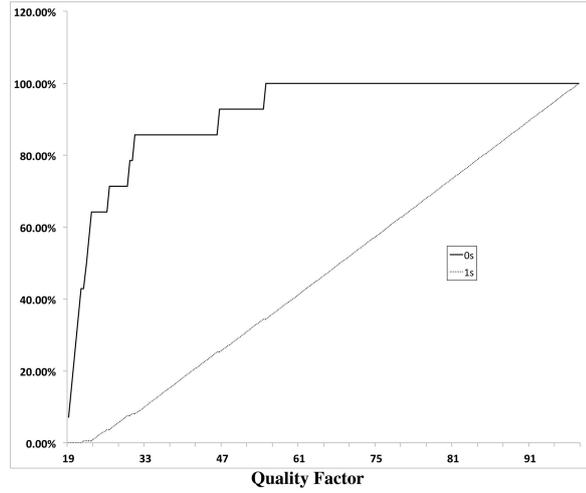


Figure 18. Cumulative distributions of subjective radiologist scores corresponding to JPEG quality factor. K-S = 77.65%.

The result of the logistic curve-fitting model (described in the previous section) applied to our data is presented in Figure 20. Our application of cross validation followed by testing (on images not used in the cross validation procedure) in order to determine optimal thresholds will now be discussed. It is understood that the optimal thresholds for SSIM and MSE found in this work are dependent upon the collected data samples and may well be different for other data. In order to find optimal thresholds for a given quality measure, the experiment should employ a larger set of images and a larger number of radiologists as subjects. Moreover, the thresholds may vary according to image modality.

Cross validation was applied to two methods: the curve-fitting model (Figure 20) and the weighted Youden index $WY(s)$. The user specifies the parameter λ , which corresponds to the weighting of FPs versus FNs . Larger values of λ correspond to higher thresholds of SSIM and MSE and it means that having FPs is very costly. If λ equals 1, there are no TPs and too many FNs . Thus, λ equal or very close to 1 should not be considered. To demonstrate the results, we chose λ to be 0.95. Figure 21 shows the thresholds for SSIM and MSE.

The use of the weighted Youden index yields threshold values of 0.95 and 0.955 for the SSIM index and MSE, respectively. This results in almost a small number of FPs , i.e. there is no risk that an image with higher quality score will be marked as unacceptable.

6.2 Second experiment: Local quality

In this local analysis only JPEG images were considered. The ROC curves corresponding to the SSIM index, MSE and quality factor are shown in Figures 22, 23, 24 and 25.

Our examination of the data has shown that brain and body CT images should be analyzed separately. These two image types possess different characteristics (e.g. textures, distribution of intensities, variances of subblocks). We therefore expect, and observe, different types of degradations that become noticeable at different compression levels.

In the local analysis, the ROC curves for body CT images corresponding to SSIM, MSE and JPEG quality factor (Q) have very similar AUC. However, the ROC curves corresponding to brain CT images show poorer

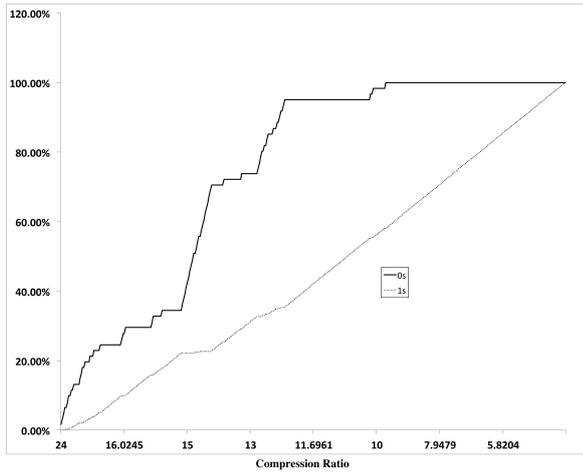


Figure 19. Cumulative distributions of subjective radiologist scores corresponding to Compression Ratio for JPEG and JPEG2000 compressed images. K-S = 59.68%.

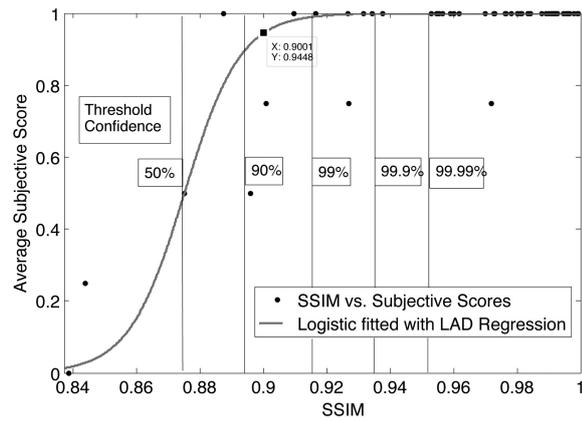


Figure 20. Logistic curve fitted with LAD Regression with threshold confidence.

SSIM Threshold with $\lambda = 0.95$	~ 0.95	TP = 57 FP = 0 TN = 16 FN = 27
SMSE Threshold with $\lambda = 0.95$	~ 0.955	TP = 55 FP = 1 TN = 15 FN = 29

Figure 21. SSIM index and SMSE thresholds obtained from cross validation and testing procedures using a fixed λ . The resulting totals of TP, FP, TN and FN are given for each case.

performance. We already observed a similar result for brain images in experiment 1. Again, for the brain, most radiologists' responses were positive, meaning that the images (the regions in Experiment 2) are acceptable. Due to too many "acceptables" for brain images in experiment 2, no proper local analysis is possible for the tested brain CT images.

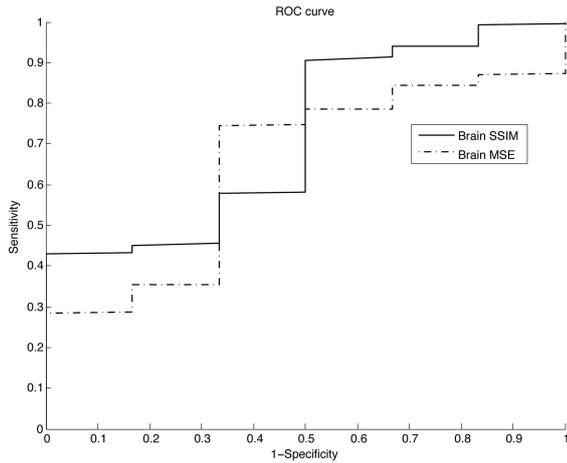


Figure 22. ROC curves corresponding to SSIM and MSE for Brain CT images. AUC SSIM = 0.7183 AUC MSE = 0.6481.

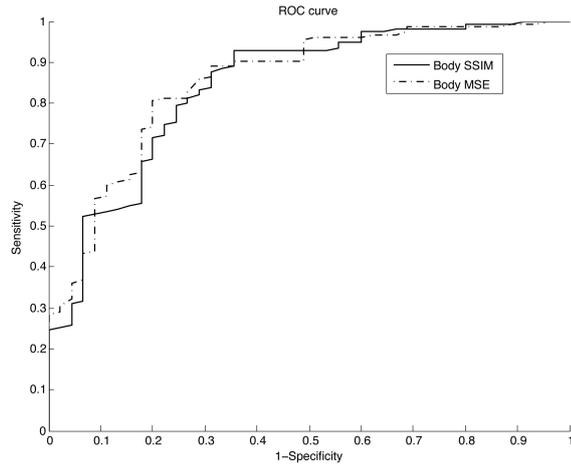


Figure 23. ROC curves corresponding to SSIM and MSE for Body CT images. AUC SSIM = 0.8422 AUC MSE = 0.8541.

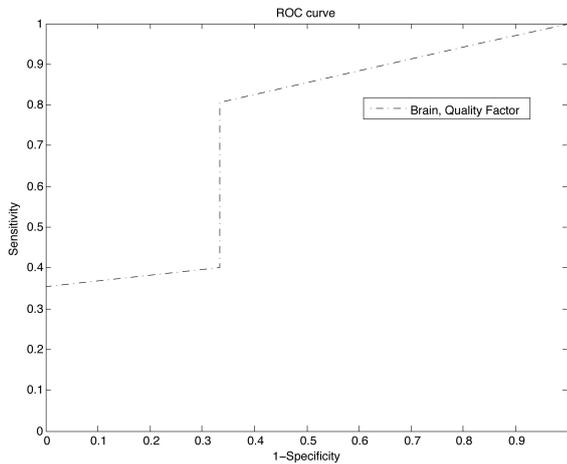


Figure 24. ROC curves corresponding to JPEG quality factor for Brain CT images AUC QF = 0.7284.

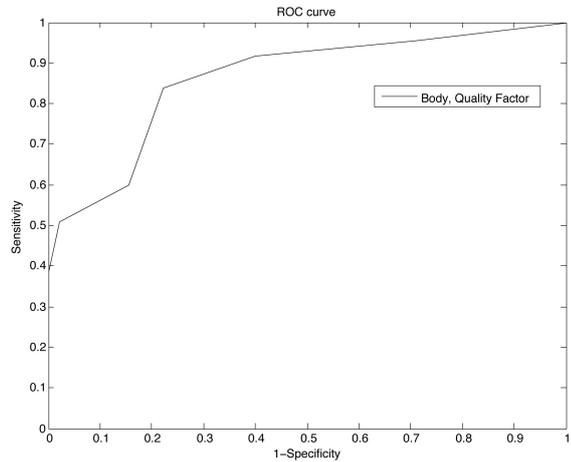


Figure 25. ROC curves corresponding to JPEG quality factor for Body CT images AUC QF = 0.8613.

7. DISCUSSION

Using both the AUC (area under ROC curve) and the KS (Kolmogorov-Smirnov) analyses, our results for Experiment 1 (global analysis) indicate that compression ratio (CR) demonstrates the poorest performance of the four quality measures examined. Furthermore, MSE/PSNR performed inconsistently as an indicator of visual/diagnostic quality. Among the four image quality measures, SSIM shows the best performance, i.e. SSIM provides the closest match to the subjective assessments by the radiologists. Finally, quality factor performs nearly as well as SSIM but, of course, it applies only to JPEG compression.

Furthermore, we have utilized a weighted Youden index and a curve-fitting model to suggest thresholds for SSIM and MSE as indicators of acceptable compression levels for brain CT and body CT compressed images. The thresholds reported in this work correspond to the specific set of sample data collected in our experiments.

In order to obtain more reliable and statistically significant thresholds, the experiment should employ a larger set of images - separately for the various anatomical regions as well as modalities - and involve more radiologists as subjects.

The local analysis in Experiment 2 was performed separately on brain and body images. As expected, the performance of SSIM was quite satisfactory in terms of AUC (area under ROC curves). MSE and JPEG quality factor also performed well.

7.1 Training of the stability constant for the SSIM index

In Figure 26 are plotted ROC curves that correspond to the SSIM quality index (actually, the structure term of SSIM) for various values of the stability constant C . The AUC (area under ROC curve) values associated with these values of C are plotted in Figure 27. The monotonically decreasing behavior of the curve for values of C near 0 is quite interesting, and suggests that best ROC-performance (i.e., maximum AUC) is obtained. That being said, a nonzero stability constant of very low magnitude will be employed in order to avoid any possible division by zero in the computation of the SSIM index. Such “zero denominators” could arise in the case of “flat” image blocks, i.e., pixel blocks with constant greyscale value, hence zero variance, in both uncompressed and compressed images.

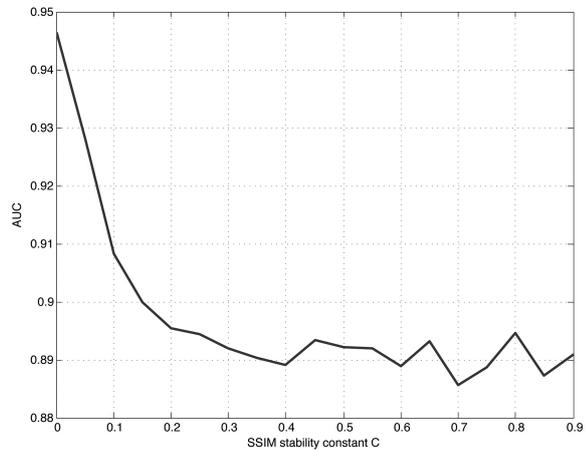
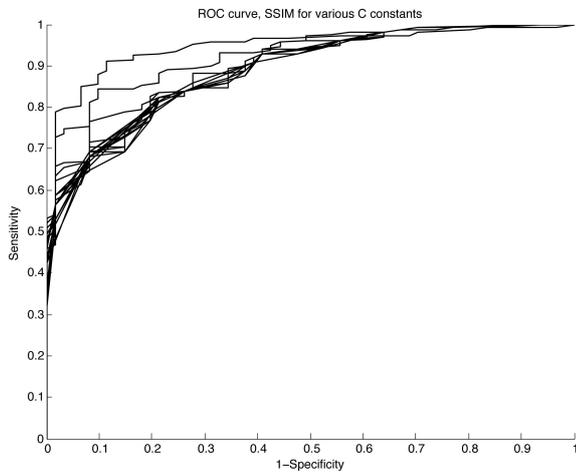


Figure 26. ROC curves corresponding to various values of the constant C in the structure term of SSIM.

Figure 27. Plot of AUC and the various SSIM stability constants.

8. CONCLUSION

In this work, we compared the performances of SSIM, MSE/PSNR, compression ratio CR and JPEG quality factor Q , based on experimental data collected in two experiments involving radiologists. The first experiment involved a global quality assessment of 100 brain and body CT images at various compression ratios. The radiologists evaluated compressed images as acceptable or unacceptable as compared to their uncompressed counterparts. An ROC and Kolmogorov-Smirnov analysis indicates that compression ratio is not always a good indicator of visual quality. Moreover, SSIM demonstrated the best performance, i.e., it provides the closest match to the radiologists’ assessments. We also show that a weighted Youden index and curve fitting method can provide SSIM and MSE thresholds for acceptable compression ratios. The second experiment involved a local/regional image quality analysis of these images by the radiologists. An ROC analysis once again shows that SSIM provides a closer match to subjective assessments.

Plans for future experiments include the determination of thresholds for SSIM and MSE on the basis of larger sample size data, for different modalities and a variety of anatomical regions. As well, we plan to involve

a much larger number of radiologists in future studies. We also plan to investigate the possible tuning of the SSIM index in order to correspond better with the subjective radiological scores. As well, we plan to conduct a subjective experiment in which the radiologist subjects will have three choices in the assessment of the quality of a compressed image: (1) not noticeable and acceptable, (2) noticeable and acceptable and (3) noticeable and not acceptable. Such a study should be helpful in determining a more reliable threshold for compression and tuning of the SSIM index.

REFERENCES

- [1] Yuden, J., “Index for rating diagnostic tests,” *Cancer* (5), 32–35 (1950).
- [2] Koff, D. and Shulman, H., “An overview of digital compression of medical images: can we use lossy image compression in radiology?,” *Canadian Association of Radiologists Journal* **57**(4), 211–217 (2006).
- [3] Koff, D., Bak, P., Brownrig, P., Hosseinzadeh, D., Khademi, A., Kiss, A., Lepanto, L., Michalak, T., Shulman, H., and Volkening, A., “Pan-Canadian evaluation of irreversible compression ratios (“Lossy Compression”) for development of national guidelines,” *Journal of Digital Imaging* **22**(6), 569–578 (2009).
- [4] Wang, Z. and Bovik, C., “Mean squared error: love it or leave it? - A new look at signal fidelity measures,” *IEEE Transactions on Signal Processing* **26**(1), 98 – 117 (2009).
- [5] Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E., “Image quality assessment: from error visibility to structural similarity,” *IEEE Transactions on Image Processing* **13**(4), 600–612 (2004).
- [6] “The Cancer Imaging Archive Sponsored by the Cancer Imaging Program, DCTD/NCI/NI.” <http://cancerimagingarchive.net>. Accessed: 2013-05.
- [7] “The SSIM Index for Image Quality Assessment, University of Waterloo.” <https://ece.uwaterloo.ca/~z70wang/research/ssim>. Accessed: 2013-05.
- [8] Hastie, Tibshirani, and Friedman, [*Elements of Statistical Learning, Data Mining, Inference, and Prediction, 2 Ed*], Springer-Verlag. (2009). Also available as <http://www.stat.stanford.edu/~tibs/ElemStatLearn>, Accessed 2013-05.
- [9] Kowalik-Urbaniak, I., Vrscay, E., Wang, Z., Cavaro-Menard, C., Koff, D., Wallace, B., and Obara, B., “The impact of skull bone intensity on the quality of compressed ct neuro images,” *Medical Imaging 2012: Advanced PACS-based Imaging Informatics and Therapeutic Applications, Proceedings of SPIE* **8319**(83190L) (2012).