2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Ultrasound Image Quality Evaluation using a Structural Similarity Based Autoencoder

Karlo Nesovic, Ryan G.L. Koh, Azadeh Aghamohammadi Sereshki, Fatemeh Shomal Zadeh, Milos R. Popovic, and Dinesh Kumbhare

Abstract— **Ultrasound (US) imaging is a widely used clinical technique that requires extensive training to use correctly. Good quality US images are essential for effective interpretation of the results, however numerous sources of error can impair quality. Currently, image quality assessment is performed by an experienced sonographer through visual inspection, however this is usually unachievable by inexperienced users. An autoencoder (AE) is a machine learning technique that has been shown to be effective at anomaly detection and could be used for fast and effective image quality assessment. In this study, we explored the use of an AE to distinguish between good and poor-quality US images (caused by artifacts and noise) by using the reconstruction error to train and test a random forest classifier (RFC) for classification. Good and poor-quality ultrasound images were obtained from forty-nine healthy subjects and were used to train an AE using two different loss functions, with one based on the structural similarity index measure (SSIM) and the other on the mean squared error (MSE). The resulting reconstruction errors of each image were then used to classify the images into two groups based on quality by training and testing an RFC. Using the SSIM based AE, the classifier showed an average accuracy of 71%±4.0% when classifying images based on user errors and an accuracy of 91%±1.0% when sorting images based on noise. The respective accuracies obtained from the AE using the MSE function were 76%±2.0% and 83%±2.0%. The results of this study demonstrate that an AE has the potential to differentiate good quality US images from those with poor quality, which could be used to help less experienced researchers and clinicians obtain a more objective measure of image quality when using US.**

I. Introduction

Ultrasound imaging has long been used clinically to image tissues within the human body and recently it has been used in the diagnosis of musculoskeletal disorders [1]. Its popularity is largely due to the fact that it is able to obtain real-time recordings in the absence of ionizing radiation, and is low-cost compared to computed tomography (CT) and magnetic resonance imaging (MRI) [2]. Although ultrasound imaging has been used to diagnose different diseases, radiologists require a high level of experience and knowledge to analyze ultrasound images [3].

When using ultrasound, differences in tissue properties cause differential acoustic impedance that creates sound echoes and results in anatomical images. However, images from this modality are prone to problems such as signal dropout, attenuation, speckle noise, and shadows, all of which impair image quality. Different factors such as probe location and orientation, amount of ultrasound gel applied, as well as force induced by the probe affect image quality [4]. Due to the numerous variables that can affect image quality, being able to acquire high quality images with minimal training and expertise is highly desirable. Machine learning is a technique that can be used to quickly classify an image based on its quality and the level of noise it contains.

Machine learning can serve as a prospective clinical tool in medical imaging to improve diagnostic accuracy and the reliability of image interpretation [3,4]. These algorithms could allow for individuals with minimal training to be able to interpret an image's quality and could complement training for experienced users. It has been shown that machine learning techniques could improve image quality assessment and analytics [5]. In ultrasound, implementation of machine learning may help with classifying and segmenting images based on factors such as image quality and pathologies.

An autoencoder (AE) is a commonly used machine learning technique for feature extraction and classification that utilizes a neural network to first compress an input signal into a low-dimensional latent space and then reconstructs it based on the information from this space [6,7]. Studies [6,8] have shown that AEs are able to aid in denoising and image recovery of under-sampled ultrasound images. AEs have a loss function that represents the reconstruction error between the original and recreated image, with a common loss function being mean squared error (MSE), which looks at the differences in the square of the intensities between a target pixel and a reference pixel [9]. Although a convenient method, a per-pixel approach is ineffective as a metric for perceived visual quality and performs poorly when the AE creates an imperfect reconstruction due to factors such as defects or anomalies [9,10].

An algorithm developed by the company *MvTec Software GmbH* successfully implemented a convolutional AE for anomaly detection utilizing a loss function based on the structural similarity index measure (SSIM) [9,11]. In the study, a database of fabric was used, and the algorithm was trained with normal images. During testing, both normal and anomalous images with defects were used to monitor the software's response. Overall, it was found that images with anomalies had higher reconstruction errors and that using SSIM as the AE loss function significantly outperformed other architectures tested (feature matching AE, MSE based AE, and variational AE) when it came to detecting anomalies,

K.N., R.G.L.K, A.A.S., F.S.Z., M.R.P., and D.K are with KITE – Toronto Rehabilitation Institute – University Health Network, Toronto, ON, M5G 2A2, Canada.

K.N., A.A.S., F.S.Z., M.R.P., and D.K are with the Institute of Biomedical Engineering, University of Toronto, Toronto, ON M5S 3G4, Canada. Corresponding author: K.N. (email: karlo.nesovic@mail.utoronto.ca)

978-1-7281-1178-0/21/$31.00 ©2021 IEEE

which was measured using area under the curve (AUC) values [9]. In this article, we investigated the effectiveness of an AE paired with a random forest classifier (RFC) for detecting low quality ultrasound images in order to determine its ability to find anomalies resulting from user errors when using the ultrasound probe, or from an increased level of noise.

II. Methods

A. Subject Recruitment

Subjects from our previous study [12] and from ongoing data collection were used in this study. Forty-nine healthy individuals (n = 49) demonstrating no symptoms or history related to neuromuscular disease participated in this study. All participants provided written consent prior to participating and the protocol was reviewed and approved by the Ethics Review Board of the University Heath Network (UHN) of Toronto. The procedure also adhered to the guidelines set out in the Declaration of the World Medical Association of Helsinki.

B. Image Acquisition Protocol

Images were acquired using a Sonosite X-Porte ultrasound system (Sonosite, Canada) at a depth of 2.3cm using a linear ultrasonic transducer of 15-6MHz (HFL50xp). Time gain compensation, depth, and window size settings were constant throughout each recording. During each measurement, the subject was seated and was instructed to rest their forearms on their thighs. To obtain good quality images, the transducer was placed over the center of the trapezius and enough gel was used to cover the area of skin where the transducer was placed. A ten second video of the trapezius muscle, recorded at a rate of 30 frames per second, was obtained by moving the transducer towards the acromioclavicular joint. Following each video, 300 images per subject were obtained for analysis. All ultrasound images were taken by an experienced sonographer trained in ultrasound imaging. To maintain consistency, the same sonographer took the measurements in each group and confirmed the image quality of the obtained images. Unique frames were extracted from each video using the complex wavelet structural similarity index measure (CW-SSIM) [13]. In order to remove redundant images, only frames with a CW-SSIM index of 0.5434 or less were used.

Images were taken under two conditions to provide inputs with varying quality. In the user error condition, images were taken with no gel (No Gel), a reduced amount of gel compared to what is regularly used in measurement (Less Gel), and with increased pressure applied on the skin with the probe (Max Pressure). In the noise condition, *Matlab* (2020a) was used to add multiplicative speckle noise to a set of good quality images to create a new set of noisy images. The noisy images were created by changing the variance (0.01, 0.025, 0.05, 0.1) of the noise added using the *imnoise* function in *Matlab*. Both conditions included good quality images (Healthy) taken following standard procedure.

C. Image Preprocessing

Following the extraction of the unique frames, each image was passed through an algorithm that automatically segmented the US image into regions of interest (ROIs): a muscle ROI and a fat ROI [14]. For this study, only the muscle ROIs were utilized in the training of the AE and RFC. To create a more homogenous input for the AE, each image was multiplied by a binary mask created from the original image that underwent global thresholding using Otsu's method [15], as this minimized variability in the images due to anatomical features unique to each participant (e.g., the distribution of fascia in the muscle). Following ROI extraction and thresholding, each muscle ROI was split into 64 x 720-pixel images in preparation for training of the AE, with the dimensions being chosen based on the most reasonable size output of the segmentation algorithm, while ensuring they remained consistent for the AE (**Figure 1)**. Any segmented images that had a dimension less than 64 pixels were excluded as the AE required equal sized inputs, and the remainders of images that could not be evenly divided into 64 x 720 images were truncated.

D. Autoencoder Architecture

The images obtained were used to train a convolutional autoencoder for anomaly detection, which was programmed using Keras [16] in Python 3.8.7. The AE was trained with healthy, normal good quality images and the encoding layer architecture, which followed what was outlined in Bergman et al [9], can be seen in **Table I**. The same number of decoding layers are used to obtain a reconstruction of the input and they are built as a reversed version of the encoder in **Table I**.

Table I. Autoencoder architecture

Layer	Output Size	Kernel / Number of Nodes	Stride	Padding	Activation Function
Input	720,64,1				
Conv_1	360,32,32	4x4	2	same	*ReLu*
Conv_2	180,16,32	4x4	2	same	*ReLu*
Conv_3	180,16,32	3x3	1	same	*ReLu*
Conv_4	90, 8 ,64	4x4	2	same	*ReLu*
Conv_5	90, 8 ,64	4x4	1	same	*ReLu*
Conv_6	45, 4, 128	4x4	2	same	*ReLu*
Conv_7	45, 4, 64	3x3	1	same	*ReLu*
Conv_8	45, 4, 100	8x8	1	same	*ReLu*

"same" means that the output of the layer is the same size as the input

To train the algorithm to better represent and decode an image from the latent space and to ensure that the reconstruction error is minimized, a loss function based on the SSIM was used for the AE (SSIM-AE). The SSIM is a metric that can be used to mitigate the limitations of pixel-wise comparisons, as in addition to using pixel intensity (luminance), it also looks at the contrast and structure of each pixel within a specific odd sized square kernel of the image, thus giving a better measure of image quality than MSE [9,10]. The following equations (Eq. 1 and 2) show how SSIM can be calculated between two patches p and q in an image:

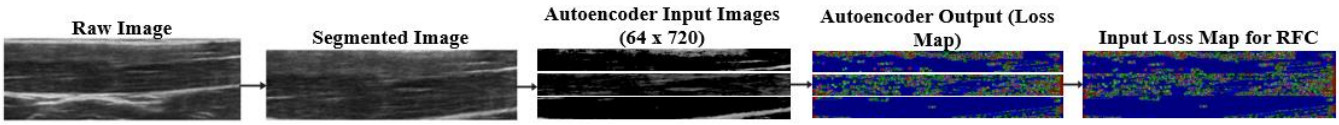


Figure 1: Example Image passing through the Processing Pipeline from Raw Image Acquisition to Random Forest Classifier Input

$$SSIM(p,q) = l(p,q)^{\alpha}\, c(p,q)^{\beta}\, s(p,q)^{\gamma} \quad (1)$$

$$SSIM(p,q) = \frac{(2\mu_p\mu_q + C_1)(2\sigma_p\sigma_q + C_2)}{(\mu_p^2 + \mu_q^2 + C_1)(\sigma_p^2 + \sigma_q^2 + C_2)} \quad (2)$$

where α, β, γ, C_1, and C_2 are user determined constants, *l*, *c*, and *s* are luminance, contrast, and structure respectively, μ_p, μ_q are the means of the pixels in patches *p* and *q* respectively, and σ_p, σ_q are the standard deviations of patches *p* and *q* respectively. The default values were used for the constants [10].

The SSIM loss function used was based on the structural dissimilarity (DSSIM) equation (Eq. 3), which was chosen to determine differences between the input and reconstruction images and used the following formula [17]:

$$DSSIM(X,Y) = 1 - SSIM(X,Y) \quad (3)$$

where X is the input image, Y is the reconstruction, and the SSIM function represents the SSIM between 0 and 1, where 0 indicates the images have no similarity and 1 indicates they are identical. The results of this were then compared against those from an AE using MSE as its loss function (MSE-AE), which uses the following equation (Eq. 4):

$$MSE\ (X,Y) = \frac{1}{N}(Y - X)^2 \quad (4)$$

where Y is the output image, X is the input image, and *N* is the number of pixels in the patch being used.

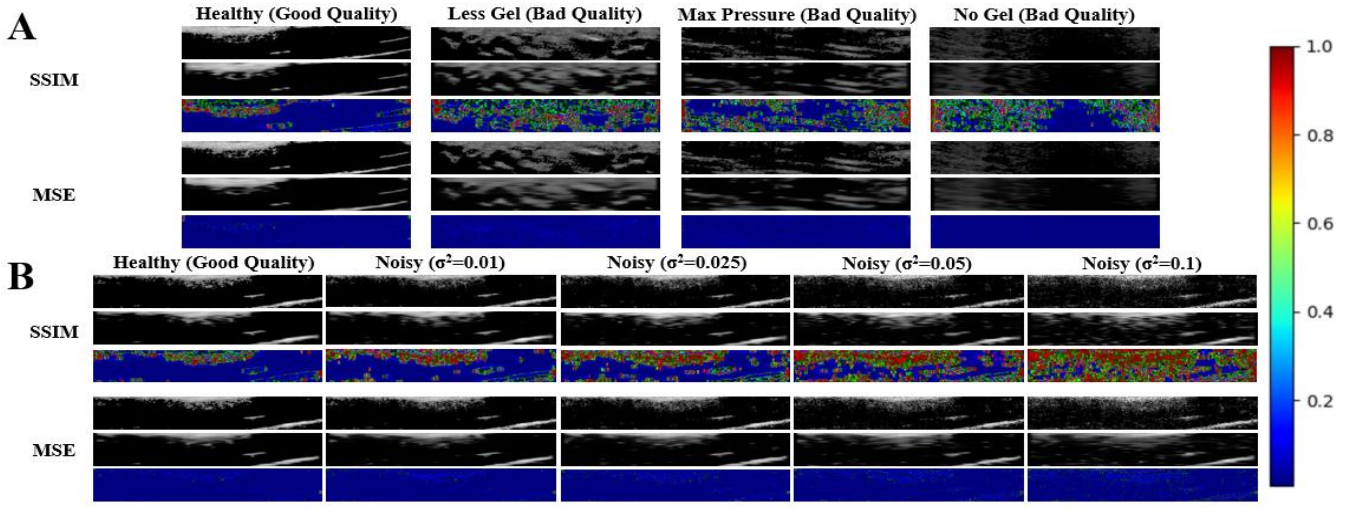
E. Autoencoder and Random Forest Classifier Testing

The images were divided into a training set for the AE and a set for training and testing the RFC. The AE training set included 874 images that were of healthy muscle and were of good quality as determined by the sonographer. The AE was trained for 50 epochs. The set used for the RFC consisted of the loss maps for reconstructed images in two conditions (user error or noise) from the AE. The user error condition used 199 new good quality images and 201 poor-quality images (67 Less Gel, 67 No Gel, 67 Max Pressure). In the noise condition set, four levels of noise were added to the 199 Healthy images used in the user error condition, and these were combined with the Healthy images to give a set with 995 images.

All images passed into the AE were 64 x 720. Following training of the AE, a reconstructed image and map of the reconstruction error for each image was generated, each with a size of 62 x 718 (the columns and rows on the edges were excluded to account for the edge effects when calculating SSIM). Since the original images were divided into 64 x 720 segments to be passed into the AE, following the training and testing of the AE the output images were grouped based on the original image they were obtained from (**Figure 1)**. Following image reconstruction, an RFC using 100 trees was used to classify images in the two different conditions. Adding more than 100 trees had minimal impact on the results, therefore this number was chosen for the study.

The RFC for the user error condition used the mean pixel intensities of the reconstruction error map as a feature to classify between good quality images and those of poor quality as a result of user error. To give a better estimate of the mean loss, the image was divided into 8 equally sized regions and the mean intensity of each was used as a feature for classification. For the noise condition, the mean intensity and the entropy [18] of the images were used as features.

The dataset from both conditions used a 5-fold cross-validation process. The mean classification accuracy and macro F_1-score from the five runs were used to evaluate RFC performance for both conditions. Accuracy was chosen as a metric due to it being a good measure of performance in balanced test sets (i.e., the user error condition), while the F_1-score is a good predictor of AE effectiveness when there is a much smaller sample of one group compared to the other.



C

User Error Condition

SSIM (n=80)	Predicted: Good	Predicted: Bad
Actual: Good	37	8
Actual: Bad	14	21

MSE (n=80)	Predicted: Good	Predicted: Bad
Actual: Good	33	10
Actual: Bad	10	27

D

Noise Condition

SSIM (n=199)	Predicted: Good	Predicted: Bad
Actual: Good	31	8
Actual: Bad	3	157

MSE (n=199)	Predicted: Good	Predicted: Bad
Actual: Good	13	26
Actual: Bad	6	154

Figure 2: A) Representative Sample of AE Outputs with User Error Conditions, B) Representative Sample of AE Outputs with Noise Conditions (in A and B, top panels are the input images to the AE, middle panels are the reconstructed outputs, bottom panels are the normalized maps of reconstruction error, with 1 representing a high reconstruction error and 0 representing no reconstruction error), C) Representative Confusion Matrices for User Error Conditions, D) Representative Confusion Matrices for Noise Conditions

III. Results

A. Number of Images

The resulting images remaining after the CW-SSIM and segmentation were 1722 images with a mean value of 37±36 images from each participant, with a maximum of 177 and a minimum of 3. 874 images were used to train the AE, and the remaining 848 were used in the training and testing of the RFC. In the user error condition, 199 images were Healthy, 244 were Max Pressure, 208 were Less Gel, and 197 were No Gel. To ensure a balanced training and test set for the RFC in the user error condition, only 67 images were used from the Less Gel, Max Pressure, and No Gel groups.

For the noise condition, the same 199 Healthy images were used, and four noisy versions of those 199 images were created.

B. Classification Accuracy and F_1-Score

In the user error condition, the RFC trained with the outputs from the SSIM-AE had a mean accuracy of 71%±4.0% for the SSIM-AE, while the MSE-AE yielded an accuracy of 76%±2.0%. The F_1-score of the SSIM-AE and MSE-AE were 0.72±0.03 and 0.77±0.02, respectively. **Figure 2A** shows representative samples of reconstructions from each case in the user error condition. **Figure 2C** displays confusion matrices from the user error conditions using both AE architectures.

For the noise condition, the mean accuracy was 91%±1.0% for the SSIM-AE and 83%±2.0% for the MSE-AE. The F_1-score was 0.84±0.03 for the SSIM-AE and 0.50±0.05 for the MSE-AE. **Figure 2B, 2D** show a representative sample of the input images and reconstructed images from the AEs, as well as representative confusion matrices from the noise condition.

IV. Discussion

This study has illustrated the potential for using the reconstruction error of an AE as a metric for assessing the image quality of US images. Both AEs were trained using only normal, good quality images. Based on the representative examples shown in **Figure 2A, 2B,** it is seen that on average, low quality images have higher loss when reconstructed, which allows us to discriminate them from good quality images based on the mean loss and entropy.

In the noisy condition, the SSIM-AE is more effective at differentiating between good quality and noisy images, as seen when comparing F_1-scores (SSIM-AE: 0.84, MSE-AE: 0.50) and from **Figure 2D**. The reconstruction error seen even when using noise with a variance of 0.01 is high, as seen in **Figure 2B,** suggesting that it would be possible to distinguish images with even higher levels of noise from good quality images due to the increased variability of the reconstruction.

When investigating the results of the user error condition, the accuracy of the classifier was lower (SSIM-AE: 71%, MSE-AE: 76%). Although these results indicate that evaluating image quality changes due to user error is possible, further investigation is necessary to reach clinical feasibility. The performance of the classifier in the user error condition may be explained by the images where the user error was that less gel was being used or that more pressure was being applied. In this study, the amount of pressure being applied, or gel being used, was not quantified but was determined by an experienced sonographer, which may result in images within these groups not corresponding to poor quality images.

Based on the results, the SSIM-AE appears to be most sensitive to random speckle noise, as it is accurate at classifying images based on if they are contaminated. Both AEs are sensitive to noise resulting from inadequate gel use or pressure, however further work must be done to understand how well the AE can distinguish between these user errors.

Although the classification accuracy of the two autoencoders was comparable, by observing the loss maps (**Figure 2A, 2B**), the SSIM-AE has the advantage of giving a more accurate visual depiction of where any anomaly or artifact is located. As outlined in previous research [9], this suggests that the SSIM-AE is more effective at highlighting disturbances in the images, compared to the MSE-AE which is more focused on pixel-wise differences between the input and output image.

The results suggest that an SSIM-AE can be used to classify images based on their quality, allowing it to detect quality loss due to noise and user error. Further improvements to the AE can enable researchers and clinicians to obtain high-quality ultrasound images without explicit training.

V. References

[1] E. Passmore, et al., "Application of ultrasound imaging to subject-specific modelling of the human musculoskeletal system," *Meccanica,* 52, 665-676. (2017).

[2] M. Mourtzakis & P. Wischmeyer. "Bedside ultrasound measurement of skeletal muscle," *Current Opinion in Clin. Nutrition and Metabolic Care*, 17(5), 23–30, 389-395. (2014).

[3] Q. Huang, et al., "Machine Learning in Ultrasound Computer-Aided Diagnostic Systems: A Survey," *Biomed Res. Int.*, (5137904). (2018).

[4] L. J. Brattain, et al., "Machine learning for medical ultrasound: status, methods, and future opportunities," *Abdom. Radiol. (New York)*, 43(4), 786–799. (2018).

[5] C. Charrier, et al., "Machine learning to design full-reference image quality assessment algorithm," *Signal Processing: Image Communication*, 27: 209-219. (2012).

[6] B. Li, et al., "Denoising Convolutional Autoencoder Based B-mode Ultrasound Tongue Image Feature Extraction," *ICASSP 2019 – 2019 IEEE Int. Conf. on Acoust., Speech and Signal Processing.* (2019).

[7] G. Litjens, et al., "A survey of deep learning in medical image analysis," *Medical Image Analysis*, 42, 60-88. (2017).

[8] D. Peridios, et al., "A Deep Learning Approach to Ultrasound Image Recovery," *2017 IEEE Int. Ultrasonics Symp. (IUS)*, 1-1. (2017).

[9] P. Bergmann, et al., "Improving Unsupervised Defect Segmentation by Applying Structural Similarity to Autoencoders," *Proceedings of the 14th Int. Joint Conf. on Comput. Vision, Imag. and Comput. Graphics Theory and Applications - Volume 5*, 372-380. (2019).

[10] Z. Wang, et al., "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, 13(4): 600-612. (2004).

[11] P. Bergmann, et al., "MVTec AD — A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection," *IEEE/CVF Conf. on Comput. Vision and Pattern Recognition*, 9584-9592. (2019).

[12] D. Kumbhare, et al., "Quantitative Ultrasound Using Texture Analysis of Myofascial Pain Syndrome in the Trapezius," *Critical Reviews in Biomedical Engineering*, 46(1):1-31. (2017).

[13] M.P. Sampat, et al., "Complex Wavelet Structural Similarity: A New Image Similarity Index," *IEEE Trans. on Image Process.*, 18(11), 2385-2401. (2009).

[14] M. Behr, et al., "Automatic ROI Placement in the Upper Trapezius Muscle in B-mode Ultrasound Images," *Ultrason. Imag.,* 41(4), 231-246. (2019).

[15] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Sys. Man. Cyber*. 9 (1): 62–66. (1979).

[16] F. Chollet, Keras (2015) https://github.com/fchollet/keras

[17] A. Loza, et al., "Structural Similarity-Based Object Tracking in Video Sequences," *2006 9th Int. Conf. on Information Fusion*. 1-6. (2006).

[18] Renyi, A., "On Measures of Entropy and Information," *Berkley Symp. on Math. Statist. and Probability*. Vol 4.1, 547-561. (1961).